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

Survey Methodology

Survey Methodology
43-1

Release date: June 22, 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

June 2017

•

Volume 43

•

Number 1

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 1, June 2017

Contents

Waksberg Invited Paper Series

Regular papers

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.
∞

Vol. 43, No. 1, p. 1
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

This issue of *Survey Methodology* opens with the fifteenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Louis-Paul Rivest (Chair), Tommy Wright, Kirk Wolter and J.N.K. Rao for having selected Don A. Dillman as the author of this year's Waksberg Award paper.

2016 Waksberg Invited Paper

Author: Don A. Dillman

Don A. Dillman, Ph.D. is Regents Professor, Department of Sociology and Deputy Director for Research in the Social and Economic Sciences Research Center at Washington State University, where he has been a faculty member since 1969. His 1978 book on mail and telephone survey methods, now in its fourth edition as *Internet, Phone, Mail and Mixed-Mode surveys: The Tailored Design Method*, has provided nearly 40 years of guidance for conducting surveys worldwide. From 1991-1995 he served as the Senior Survey Methodologist at the U.S. Census Bureau, providing leadership for the development of data collection procedures for the Decennial Census, for which he received the 2000 Roger Herriot Award for Innovation in Federal Statistics. At Washington State University he maintains an active research program in mixed-mode data collection. In 2017 his research team received the American Association for Public Opinion Research's Warren J. Mitofsky Innovators Award for development of the web-push methodology described in this issue of Survey Methodology.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 3-30
Statistics Canada, Catalogue No. 12-001-X

The promise and challenge of pushing respondents to the Web in mixed-mode surveys

Don A. Dillman[1]

Abstract

Web-push survey data collection that uses mail contact to request responses over the Internet, while withholding alternative answering modes until later in the implementation process, has developed rapidly over the past decade. This paper describes the reasons this innovative mixing of survey contact and response modes was needed, the primary ones being the declining effectiveness of voice telephone and slower than expected development of email/web only data collection methods. Historical and institutional barriers to mixing survey modes in this manner are also discussed. Essential research on the use of U.S. Postal address lists and the effects of aural and visual communication on survey measurement are then described followed by discussion of experimental efforts to create a viable web-push methodology as an alternative to voice telephone and mail response surveys. Multiple examples of current and anticipated web-push data collection uses are provided. This paper ends with a discussion of both the great promise and significant challenge presented by greater reliance on web-push survey methods.

Key Words: Surveys; Mixed-mode; Web-push; Web; Mail; Telephone; Address-based sampling; Visual communication; Response rates; Measurement differences.

1 Introduction

A surprising, but critical, development in survey design during the early 21st century is the extensive use of web-push data collection methods, i.e., the use of postal mail to obtain questionnaire responses from general public samples mostly over the Internet instead of paper questionnaires. Web-push methods are now being used as a replacement for paper mail-push procedures whereby an attempt is made to obtain responses by mail before using other modes of response such as telephone or personal interviews. Web-push methods are now being used in official government surveys and as a replacement for random digit dialing (RDD) voice telephone surveys.

For example, the American Community Survey, which serves as the major source of state and regional information on U.S. households, began using a web-push approach to data collection in 2013 that includes the possibilities of responding later in the implementation process by mail, telephone or in-person interview. Plans are now in place to use such a methodology for the 2020 U.S. Decennial Census. Web-push data collection with an initial mail request is also being used worldwide. Examples include the 2015 Japanese Census (Statistics Japan 2015), and the 2016 Censuses in Canada (Statistics Canada 2016) and Australia (Australian Bureau of Statistics 2016). Other examples of web-push procedures are household surveys in Switzerland (Roberts, Joye and Staehli 2016) and the United Kingdom's Community Life Survey (United Kingdom Cabinet Office 2016), which is being transitioned from a personal interview mode. In addition, the U.S. College Graduates Survey, conducted every 2-3 years by the National Science Foundation, has completed the shift from mail and telephone data collection to a web-push approach followed by the other two modes of data collection (Finamore and Dillman 2013). These examples are only a few of the major survey efforts around the world that are now using this methodology.

1. Don A. Dillman, Washington State University. E-mail: dillman@wsu.edu.

Use of web-push data collection methods has been encouraged by a number of considerations, ranging from seemingly unfixable problems of RDD telephone surveys to the fact that Postal Service residential address lists or country-wide registration lists now provide the most complete coverage of households. In addition, there are no acceptable ways of drawing probability samples of household email addresses as a means of household contact. Even if email addresses could be sampled, it is not likely that reasonable response rates could be obtained through email-only contact (Lozar, Bosnjak, Berzelak, Haas and Vehovar 2008).

The current heavy reliance on mail contact is surprising, despite the demonstrated potential of mail surveys for obtaining reasonable response rates in the late 20th century (Dillman 2000). Until recently, mailing address sample frames have been mostly unavailable and inadequate. In addition the general availability of a telephone alternative prior to the late 1990's meant that mail was infrequently used for government surveys, with the exception of official government censuses.

My purpose in this paper is first, in Section 2, to discuss the reasons that web-push survey methodologies have been developed and adopted worldwide. Secondly, in Section 3 and Section 4, I describe research efforts that have not only made web-push methodologies feasible, but are improving the effectiveness of such methods in producing reliable estimates of the opinions and behaviors of survey populations throughout the world.

This research has shown, see Section 5, that web-push methodologies are quite promising with regard to improving coverage and response rates, while reducing measurement differences across modes as well as total survey costs. It has also shown that there are many perils that threaten their use, ranging from respondent trust of the Internet to the plethora of devices now available for responding to such surveys, see Section 6. My focus in this paper is to present the substantial promise and many challenges associated with web-push methods for conducting sample surveys. Section 7 presents a summary and a conclusion.

2 Why Web-push data collection is needed

Fundamentally, making contact with households or individuals by one mode, such as mail or telephone, to request that they respond by another mode, is not an ideal data collection procedure. There is bound to be some friction between receiving a postal letter or phone call and then having to go to a different response mode. The switch by itself is likely to take a toll on response rates. Thus, it is not surprising that difficulties in conducting single mode telephone and e-mail/web surveys are the fundamental reason for seeking an alternative.

2.1 The declining effectiveness of telephone surveys

In the mid-20th century, most methodologists considered face-to-face interviews as the only acceptable means for conducting sample surveys (e.g., Parten 1950; Kerlinger 1965). In addition, sampling and surveying households was slow and costly, and therefore limited mostly to conducting national and other large area surveys.

Although telephone surveys had been used occasionally to support data collection (Nathan 2001), development of the telephone as a sole means of collecting survey responses did not occur until the early 1970's, a process described in detail by Nathan (2001). The first three books on methods for conducting telephone surveys appeared in rapid succession, developing marketing (Blankenship 1977), state and special population (Dillman 1978) and national (Groves and Kahn 1979) population survey perspectives. The use of telephone data collection methods advanced rapidly because of the expanding presence of telephone in households and development of the Mitofsky-Waxberg procedure for using random digit dialing methods of selecting households. In addition the declining costs for long distance calling resulted in RDD voice telephone surveys replacing most in-person interviewing (Dillman 2005).

Between 1997 and 2012 the Pew Research Center (2012), a major conductor of social surveys by telephone in the United States, reported declines in RDD response rates from 35 percent to about 9 percent. More recently, Dutwin and Lavrakas (2016) conducted an analysis of telephone response rates for nine organizations. They found that landline response rates declined from 15.7 percent in 2008 to 9.3 percent in 2015, while cell phone response rates declined during this period from 11.6 to 7.0 percent. They also reported that this roughly 40% decline in response is less the result of an increase in refusals than it is an increase in no answers and answering machines of 10 percentage points for landlines and 24 percentage points for cell phones.

However, these results present only the tip of the iceberg with regard to what is happening to telephone. The telephone has changed from being a household device, or landline, shared by all household members to a wireless individually possessed instrument, easily transportable from place-to-place. In the United States, half of all households and 60% of those with children are now wireless only (Blumberg and Luke 2017). At the same time, the presence of cell and/or landline phones in households has reached an all-time high of at least 95% in most European countries (Mohorko, de Leeuw and Hox 2013) and 97% in U.S. households (Blumberg and Luke 2017). One implication of the increased proportion of wireless phones is that household sampling has become much more difficult. It is possible to include mobile numbers in RDD sample frames. However, it has also become necessary to devote precious interview minutes to ascertaining a range of information including number and type of phones in a household in order to determine household selection probabilities.

In addition, one needs to learn whether the person who answers the phone is an adult, and select an appropriate respondent. Also, the landline "inconvenient time problem" of the respondent being interrupted while, for example, fixing dinner and not having time to talk has been expanded to needing to find out if a person who answers the phone is driving a car or engaged in another task where safety emerges as a serious issue. The inclusion of such items takes away from the ability to ask other questions in phone interviews for which considerable pressure exists to keep the length to only a few minutes. In sum, a major effect of the changes in how telephones are owned, regulated, and used has made its use for important data collection efforts, increasingly difficult.

Landline and cell phones jointly face a larger challenge. Fewer and fewer people engage in voice conversations by telephone. This is a huge change from the time when essential communication for business

discussions, maintaining social relationships, and coordinating daily activities in a timely way, were done mostly by voice telephone. Email and texting have largely replaced that use. Talking over the phone with a survey interviewer is increasingly out of sync with other aspects of people's daily lives.

Answering machines now take most incoming voice calls on both landline and cell phones. Not answering one's phone is no longer considered rude. Desired calls from children and other close relatives may be assigned a special ringtone to draw the call recipient's attention. Phone calls from specific numbers can also be blocked, or, on smartphones, swiped away. In addition, both landline and cell telephone numbers are now transportable across type of phone and area codes in the United States and different federal rules apply to automatic dialing of phones.

Yet, another emerging problem with telephones is that the repeated contacts necessary for achieving reasonable response rates for all types of phones are becoming less effective. Increasingly, telephone interviewers have only one chance, lasting just a few seconds, to persuade people to be interviewed. The appearance of the originating telephone number and/or source on caller identification screens makes it increasingly likely that follow-up phone calls can be avoided. Also, the plethora of marketing and fundraising calls has produced an environment in which fewer people are willing to answer the telephone, let-alone be interviewed. An additional challenge associated with cell phones is that their use has brought with them a greater likelihood that requests to be surveyed come when the call recipient is in the midst of business and work activities that are not conducive to people taking time to be interviewed.

The decrease in RDD telephone interviews was slowed down for awhile by research that has shown intensive callbacks to increase response rates does not improve the accuracy of results (Keeter, Miller, Kohut, Groves and Presser 2000) and other research that has suggested that occurrence of non-response error (the differences between respondents and non-respondents) is not closely related to response rates (Groves and Peytcheta 2008). The telephone's continued use was also encouraged by the great investment that organizations have in telephone hardware, software, and specialized personnel, many of whom had not done other types of survey data collection. However, the continued decline of telephone response rates in recent years noted by Dutwin and Lavrakas (2016) and measurement concerns have reduced the credibility of doing stand-alone telephone surveys.

2.2 The slower than anticipated emergence of email/web only surveys

In the mid 1990's when telephone response rates were starting their persistent decline, Internet surveying, the expected replacement, was beginning its rapid development (Dillman 2000, Chapter 11). Yet, two decades later, its use for general population surveys remains limited.

Household Internet penetration in the U.S. and many other developed nations now exceeds 85%, which is higher than for telephone when the rapid development of surveys by telephone occurred in the early 1970's (Nathan 2001). The lack of Internet in some households (e.g., 41% of U.S. adults 65+ and 26% of individuals with only a high school or less education) remains a concern (Anderson and Perrin 2016), but each year sees that becoming less of a problem. Internet use skills are now fundamental to the educational process, to organizational operations, and to accessing consumer services. Yet, the barriers to obtaining web responses for household surveys when using only email contacts remains huge.

There is no household or general population sampling algorithm for email addresses that will provide a known non-zero chance of being selected for survey participation, as calling random numbers has provided for telephone surveys. Email addresses do not exist in standard formats as is the case for our 10 digit telephone numbers that identify an area code, exchange, and the 10,000 number possibilities in each exchange. People within households are also likely to have multiple email addresses so that the probabilities of reaching specific households or other sample units cannot be calculated. In addition, some of the population that are most computer literate – young adults – have developed a reputation for minimizing their use of traditional email systems. Instead they focus heavily on Facebook, Snapchat, and other instant messaging applications for connecting with friends and acquaintances.

In addition, web response rates for random samples of existing email addresses are likely to be as low or lower than those achieved for today's telephone surveys (Lozar et al. 2008). And, they are likely to include disproportionately high numbers of individuals who are younger and better educated, despite the fact that many younger people rely on other ways of connecting electronically that make them only occasional users of traditional email. People's computer inboxes are typically a crowded space, with unsolicited and unwanted emails being more prevalent than the number of unwanted telephone calls once was. In addition, emails are often scanned and deleted based solely upon source or after reading only a few beginning words of the accompanying message.

Changing computer technologies are also contributing to web survey nonresponse. Smartphones that fit in one's purse, handbag, or pocket, now have far more computer power than desktop computers had when internet surveying began (e.g., Friedman 2016). Their constant presence with people has led to these devices being used as the first responders for scanning and discarding unwanted requests. Some users may defer answering survey requests until they get to a laptop or desktop with a full-scale keyboard. However, for some individuals, smartphones are now the dominant, or even only, device for responding to all emails.

When our dominant survey mode was telephone, interviewers could usually focus the respondent's attention on survey questions and guide that person through the interview. On desktops, laptops, and now tablets that are used in a person's office or home, considerable mental concentration by the respondent can often be achieved. In the smartphone era when people are as likely to be on the move from one place to another, concentration on survey content seems somewhat less likely to be achieved. It is evident that the proportion of surveys completed over smartphones is increasing as a proportion of all web completions (Couper, Antoun and Mavletova 2017). However, there appears to be no evidence that the smartphone delivery of web survey requests is increasing total survey response, and may in fact be lowering it. In addition, breakoff rates are much higher for smartphones than desktops and laptops.

Concern about the consequences of attempting to answer an electronic survey is another factor limiting the potential effectiveness of email/Internet surveying. The ease and low cost of sending out massive numbers of email survey requests, has increased the likelihood that people receive requests from organizations that they know nothing about. In addition, considerable fear exists that such requests may be originating from sources that are imitating legitimate sponsors, and attempting to deliver malware, ransomware, and/or collecting data that can be used for other nefarious purposes. Thus, people who are willing and able to respond to legitimate web surveys may be unwilling to take that risk. For many, the internet is a scary place where a "consumer beware" climate prevails.

For all of these reasons, it is hardly surprising that low cost email contact/web response surveys have not become the method of choice for conducting random surveys of the general public needed for public policy purposes. Even if the challenge of drawing probability samples could be solved, multiple issues including computer technologies, the circumstances in which potential respondents encounter survey requests, and mistrust about who is requesting a survey request and how data might be used, are limiting its ability to replace the telephone.

3 Overcoming barriers to the acceptance of mixed-mode designs

3.1 Historical barriers to mixing modes

Use of more than one survey mode, as a means of contacting and/or asking questions, was seldom done in the late 20th century. Gaining acceptance of mixed-mode designs for any purpose has been a slow process. The biggest barrier prior to the 1990's was simply the lack of perceived benefit. Response rates to in-person, telephone and mail surveys were considered high enough that use of a second or third mode was considered unnecessary. A significant exception was those surveys in which less expensive survey methods were used early in the data collection process, but in-person methods were necessary to achieve response rates over 90%. The U.S. Decennial Census from 1970-1990, which followed a mail questionnaire start with in-person and in some cases telephone follow-up is an example.

Another barrier to the mixing of survey modes was that the data collection technology of the times made it difficult to simultaneously implement multiple modes of data collection in a single survey. The lack of networked computers and software meant that using a second mode of data collection required finishing up data collection for one mode before switching the effort to a separate data collection unit for implementation of a second mode (Dillman, Smyth and Christian 2009, Chapter 8). An earlier review of the use of telephone in mixed-mode surveys in the late 1980's found that few mixed-mode surveys had been done, other than pre-letters to an anticipated telephone or in-person interview (Dillman and Tarnai 1988).

During the 1990's it became apparent that new methods of surveying needed to be developed. Response rates, especially for personal interviewing and telephone, had begun to decline (Brick and Williams 2013). Coverage problems were also increasing, as locked multi-unit buildings and gated-residential communities made it impossible to reach many households in-person. The landline coverage of households also began its long inexorable decline that now leaves about half of U.S. households without such connections.

Interest in coordinating the use of multiple modes in some way to improve response rates brought attention to interview concerns that had previously been ignored because of the practical barriers to mixing modes. For example, interviewed respondents often gave socially desirable responses so that estimates of desirable behaviors of e.g., "having voted in the last election" were higher than the actual behavior. In addition, estimates of undesirable behaviors, e.g., smoking marijuana or having sex outside of marriage were lower (de Leeuw 1992). Differences were also being observed between mail answers to survey questions on the one hand vs. telephone and in-person surveys where respondents gave more extreme positive answers on opinion questions (de Leeuw 1992; Tarnai and Dillman 1992). Research had also

suggested that respondents were more likely to choose the first response categories in mail surveys, a primacy effect, and the last categories in telephone surveys, described as a recency effect (Krosnick and Alwin 1987). As a consequence, it became increasingly difficult for survey sponsors to argue that telephone, or even in-person interviews, were superior survey modes.

Mixed-mode surveys were proposed as a potential, albeit imperfect, solution to the problems of individual survey modes. Five types of mixed-mode surveys were identified, ranging from collecting the same data from different members of a sample to using one mode only to prompt completion by another mode (Dillman 2000, page 219). The major advantage of combining modes appeared to be improvements that could be achieved in coverage and response rates. The major difficulty identified was the prospect of measurement differences when different response modes were used.

A critical article by de Leeuw (2005) influenced an important transition in thinking about mixed-mode surveys. She articulated a variety of accepted possibilities for combining survey modes and provided evidence that an increase in use of mixed mode surveys was occurring. She also noted the transition that occurred from debate on which survey mode was best for a particular study, to how modes could be used together and produce better results.

A contextual change was also underway as modern societies throughout the world began shifting from person-mediated activities (e.g., getting money from tellers in banks, making travel reservations through an agent, and purchasing goods in stores and from catalogues) to self-administration (Dillman 2000). But researchers had not yet answered the question of whether interviews by telephone could persist in the face of these self-administration trends.

3.2 Institutional barriers to the joint use of survey modes

Considerable reluctance, if not outright opposition, existed with regard to mixing survey modes and especially to giving up cherished ways of asking questions differently in different survey modes (Dillman 2000). One of the consequences of the emergence of new ways of collecting survey data in the last third of the 20th century is that data collection staff became quite specialized. Some organizations conducted surveys by only one mode. It was common for some data collection staff and their organizations to do only telephone surveys, and to a lesser extent mail surveys. A few large firms had in-person sampling and data collection units. A tendency existed to want to do surveys by the mode that a group knew best. This tendency was exacerbated in the late 90's as Internet-only survey organizations began to emerge.

In addition different styles of wording questions by individual modes had emerged. Interviews tended to withhold "don't know" categories, offering them only when respondents would object. Designers of paper and web surveys often used "mark all that apply" question formats to make responding easier, but the awkwardness of that format on telephone led to using only forced choice formats that obtained an answer after each individual item was presented on the telephone. The issue facing survey designers was whether to maximize question formats for the mode, or try to keep the same stimulus across all modes (Dillman and Christian 2005).

One of the factors supporting this tendency to stick with what surveyors knew best was the recognition that single mode surveys were best for many survey situations. That situation existed for each of the modes.

In-person interviews were the only way of obtaining adequate coverage for certain national surveys such as the Current Population Survey that produces employment rate estimates. RDD telephone surveys were the best means of conducting election and other cross-sectional household surveys. Mail was the most adequate means when one was conducting regional and local surveys for which only residential addresses were available. Interactive Voice Response surveys were most practical for many customer satisfaction surveys when people contacted calling centers to obtain a particular service. And Internet surveys quickly became the go-to methodology for client surveys and other situations for which email addresses had been previously collected.

This situation marked the development of interest in "tailored design," i.e., recognition that different modes of data collection fit better with surveys of particular populations, survey topics, and data collection situations. This trend in survey design now persists more powerfully than it did at the turn of the century. It is clear that picking the single best survey mode for a particular survey is increasingly inadequate because of negative effects on coverage, response rates and nonresponse error.

As the 20th century came to an end, there was much uncertainty with regard to where data collection methods might be headed. Prospects seemed dim for continuing sole reliance on either in-person or telephone surveys. The coverage challenges and costs were growing significantly, and it seemed unlikely that response rates were likely to improve for voice telephone surveys. Great interest existed in replacing interview methods with the Internet, but at the turn of the century only about half of the households in the U.S. as yet had computers, and even fewer had access to the internet (Dillman 2000).

4 Development and testing of web-push mixed-mode data collection

In the first decade of the 21st century, the idea of using multiple survey modes to contact individuals and obtain survey responses seemed to be an issue whose time for serious in-depth exploration had arrived (e.g., Tourangeau 2017; de Leeuw, Villar, Suzer-Gurtekin and Hox 2017). In addition, the same information technologies that brought on the internet carried with it the potential for managing the simultaneous use of multiple modes of data collection effectively as well as efficiently, thus removing the primary practical barrier to conducting mixed-mode surveys.

The effective development of web-push methods meant addressing multiple issues all at once in order to learn whether such an approach would be effective. These issues ranged from responding to household coverage problems and developing an understanding of how visual communication differed from aural communication, as well as expanding our theoretical thinking about what influences people to respond to survey requests.

A major, unanswered question was whether self-administration could replace human interviewing, and would the results be better or worse. A confounding challenge was that survey modes varied significantly in individual coverage, response rates, and bias in how people responded to the use of them, with each survey mode perhaps being better in some situations and worse in others.

4.1 U.S. Postal Service residential address lists now provide excellent household coverage

Because the U.S. Postal Service makes available, through vendors, complete residential address lists, it is possible to send mail requests to nearly all residences in the United States (Harter, Battaglia, Buskirk, Dillman, English, Mansour, Frankel, Kennel, McMichael, McPhee, Montaquila, Yancey and Zukerberg 2016). These computerized residential lists are provided without names, just as RDD telephone lists do not have names. The lack of names is not a barrier to obtaining response from households, as shown by a series of response rate studies using U.S. Decennial Census address lists (Dillman 2000, Chapter 9). In addition it does not direct a mailing to only one person in households whose occupants are now less connected to one another than when marriage rates were higher. It may also allow more accurate respondent selection by not having to overcome the limitations of mailings being associated with just one household member.

One of the first large-scale studies to evaluate the use of address-based sample (ABS) with postal data collection was by Link, Battaglia, Frankel, Osborn and Mokdad (2008). It found for a 2005 Behavior Risk Factor Surveillance System (BRFSS) questionnaire that a mail questionnaire sent to an ABS sample obtained significantly higher response rates than those obtained by RDD in five of the six states surveyed. The authors concluded with appropriate caution that the true potential of ABS might be in facilitating mixed-mode surveys that also involved telephone follow-up, and provided a strong recommendation for further study.

Other research at this time showed that ABS samples had very high coverage which was improving as city-style addresses were replacing less specific addresses, such as rural routes (O'Muircheartaigh, English and Eckman 2007; Battaglia, Link, Frankel, Osborn and Mokdad 2008). In addition, a series of studies showed that a two-step ABS mail survey (screening of households for the presence of school children, followed by a detailed questionnaire on a particular child) produced better results than a two-step RDD approach, with significantly higher response rates (Brick, Williams and Montaquila 2011; Williams, Brick, Montaquila and Han 2014).

These studies contributed significantly to establishing the high coverage qualities of address-based sampling as an alternative to RDD sampling. However, they stopped short of testing the possibility that the contacted households could be persuaded to respond by web to mailed requests.

4.2 Identifying and overcoming measurement differences between visual and aural surveys

A quite different concern that limited interest in address-based sampling with paper and/or Internet questionnaires was that people's answers to questions were likely to be different from telephone responses. There were two aspects to this concern. The first was that without an interviewer, respondents could not be given extra encouragement when they were unable or reluctant to answer a question, nor could misunderstandings of questions be corrected. The second was the long-standing evidence that social desirability and the tendency to agree (acquiescence) were greater for telephone than self-administered (mail) survey responses (de Leeuw 1992). Traditionally, the benefit of having an interviewer present was viewed as outweighing the potential bias from the latter.

A survey sponsored by the Gallup Organization in 1999 provided a new perspective on these differences. This test revealed that asking people to respond in an interview to aurally received stimuli (either by voice telephone or Interactive Voice Response) produced similarly more positive answers than those given to visually delivered stimuli, either by mail or Internet questionnaire (Dillman 2002; Dillman, Phelps, Tortora, Swift, Kohrell, Berck and Messer 2009).

Discoveries in how visual information is processed reported by Palmer (1999), Hoffman (2004) and Ware (2004), provided theoretical insights into the separate actions taking place as the eye takes in the information and the brain processes it to make sense of what is on the page or screen. Application of these concepts provided an understanding of the reasons that self-administered questionnaires often produced different answers than interview surveys, as revealed by the Gallup study. Respondents are guided through visual questionnaires by multiple languages that communicate meaning. They include symbols, numbers and their graphical composition (size, spacing, color, symmetry, regularity, etc.) that affect how information on paper and web pages is navigated, mentally grouped and interpreted (Dillman 2007, pages 462-497; Tourangeau, Couper and Conrad 2004). Additional research showed that compliance with branching instructions could be improved dramatically through changes in symbols, font size, font brightness (Redline and Dillman 2002; Christian and Dillman 2004), and the placement of those branching instructions in relation to answer choices (Redline, Dillman, Dajani and Scaggs 2003; Dillman, Gertseva and Mahon-Haft 2005).

Another major cause of measurement differences across modes became apparent: questions were often worded differently for each mode and presented using different structures (Dillman and Christian 2005). For example, researchers had a long tradition of asking forced choice questions individually on telephone surveys when surveying people's opinions on a list of items, but they often converted it to a check-all reply format for items presented as a group on mail questionnaires (Smyth, Dillman, Christian and Stern 2006). This practice was carried over to web surveys. New research showed that using forced-choice formats on both visual and aural modes would bring respondent answers much closer together (Smyth, Dillman, Christian and McBride 2009). Research also showed that open-ended questions to mail and web surveys would be comparable if similar visual construction was used for both mail and web (Smyth, Christian and Dillman 2008). In addition, it was learned that variations in scalar question formats (e.g., fully labeled vs. polar point labeled) produced dramatic differences in answers within visual modes (Christian, Parsons and Dillman 2009).

Unified mode construction – the use of the same wording and visual layout of survey questions – was proposed as a way of removing measurement differences across these modes (Dillman 2000). Unified construction could easily be accomplished for many types of questions (e.g., to present "don't know" categories to all respondents instead of only those who would not choose an offered answer choice), as typically done by telephone interviewers. However, in other instances construction that differed across modes was both practical and would reduce errors, e.g., automatic branching to the next appropriate question on web and telephone. This form of presentation cannot be accomplished with branching items for paper questionnaires where all options have to be printed because of not being able to anticipate how people will answer those items.

The major contribution of unified mode construction has been to reduce concern about measurement differences being encouraged by multiple modes of survey response. An exception is that strong evidence exists that telephone response to opinion scales using vague quantifiers are consistently more likely to produce extreme responses on the positive end of the scale and less use of intermediate categories, than are web and mail questionnaires (Christian, Dillman and Smyth 2008). The apparent reason for this difference is that the visual presentation of intermediate response categories is more visible, and therefore accessible to respondents than when those same categories are read over the phone, a process that makes the end categories more prominent in respondent minds, Dillman and Edwards (2016).

Another difference that unified mode construction does not resolve is how people answer socially desirable questions. However, self-administered (visual) questionnaires are generally thought to produce more honest answers.

The accumulation of research on visual vs. aural design issues has provided survey designers with crucial tools, the use of which partly eliminates measurement differences that might undermine coverage and response benefits of mixed-mode surveys. The practice of unified mode design was crucial for initial development and testing of the web-push methodology described below.

4.3 The sequential development of an effective web-push methodology

A sequence of ten tests of web-push data collection procedures was conducted by a team of researchers at Washington State University between 2007 and 2012 in five separate data collections. The plan that guided these experiments was to build upon what was learned from the initial tests to design and implement the later tests. All experimental comparisons used the equivalent of 12 page paper questionnaires, containing 50-70 numbered questions, requesting 90-140 potential answers. They were designed to be the equivalent of 20-30 minute interview questionnaires. The studies were on a variety of topics – community involvement and satisfaction, use of information technologies, economic and social effects of the 2008 recession, energy use attitudes, and understanding water quality and management. Researchers varied the topics in order to reduce concerns about the effect of topic on response rates and data quality.

The populations surveyed ranged from a rural region of Idaho and Washington and statewide surveys of Washington, Pennsylvania, and Alabama conducted from Washington State University, to surveys of Nebraska and Washington residents sent from the University of Nebraska and the same surveys sent to both states from Washington State University. Implementation procedures varied, but included from 4-5 mail contacts, with the mail-back questionnaire option provided in either the 3rd or 4th contact. A small token cash incentive was sent with the initial response request, and in some instances a smaller incentive was sent with the paper questionnaire when it was withheld until the 3rd or 4th contact. Detailed procedures for each of the studies are provided elsewhere (Smyth, Dillman, Christian and O'Neill 2010; Messer and Dillman 2011; Messer 2012; Edwards, Dillman and Smyth 2014; Dillman, Smyth and Christian 2014).

The initial test in a rural region of Idaho and Washington resulted in 55% of households responding to the web-push treatment, with 74% of those responses coming over the Internet. This test also revealed that enclosing a paper questionnaire and offering an immediate choice of modes produced a significantly higher

response rate of 63% (Smyth et al. 2010). Unfortunately, nearly 80% of those responses came by paper, too many to warrant the cost of setting up the web data collection. Because of that effect and the initial promise being shown of getting more than half the households to respond over the Internet in the web-push treatment, experimentation on the choice methodology was discontinued. We also found from this initial test that a paper-push treatment that withheld offer of a web option until the last contact produced only two percent of the responses over the Internet. Based upon this result the web follow-up was discontinued after two additional tests with similar results. In addition, results from this initial rural region study encouraged us to carry forward the web-push with paper follow-up for additional testing with statewide populations.

Across all ten experiments that involved five states, the web-push surveys produced a mean response rate of 43%, ranging from 31-55% (Figure 4.1). The mail-only comparisons produced a mean response rate of 53%, with a range of 38-71%. On average, 60% of the responses to the web-push treatments came over the Internet. Experimental treatments in one of the studies showed that the incentive enclosed with the web-push request improved the web response dramatically, from 13% to 31%, or about 18 percentage points (Messer and Dillman 2011). Although an RDD comparison was not included in any of the experiments, the results from the web-push procedures were undoubtedly much higher than would have been obtained by telephone for these long questionnaires, had such a comparison been included.

An item nonresponse comparison was made for three of the experiments to determine whether the mail follow-up questionnaires obtained higher item nonresponse rates than the web responses obtained in those treatment groups. For the regional study in 2007 and two statewide studies in 2009 the follow-up paper questionnaires produced item non response rates more than twice as high, 8.2% vs. 3.6% for those who responded by web. However, when the overall item nonresponses for the web-to-push treatment groups (web plus mail responses) were compared to mail only treatment group responses, there were virtually no differences between groups, being 5.3 and 5.7 respectively. The authors speculated that the initial web responses were being provided by “better” respondents, while the later responses by mail were generating responses from less able respondents, as indicated by being older and having less education. (Messer, Edwards and Dillman 2012).

Populations likely to be unfamiliar with Washington State University – the sponsor of these studies – had significantly lower response rates, especially among those responding on the web. For example, only 12% and 11% responded in Pennsylvania and Alabama respectively, compared to 28% in the Washington survey (Messer 2012). A water management study conducted by the University of Nebraska and Washington State University provided additional insight on this phenomenon by sending requests for responses to households in the other state. The web-push response was 6.1 percentage points lower among Washington residents and 14.7 percentage points lower among Nebraska residents when surveyed from the University located outside the state (Edwards et al. 2014). Virtually all of this decline occurred in the internet responses which decreased from 32 to 26 percent in Washington and from 38 to 23 percent in Nebraska, when the response requests came from the opposite state’s university. We speculated that responding over the Internet is more sensitive than mail responses to the lack of familiarity and trust of the survey sponsor.

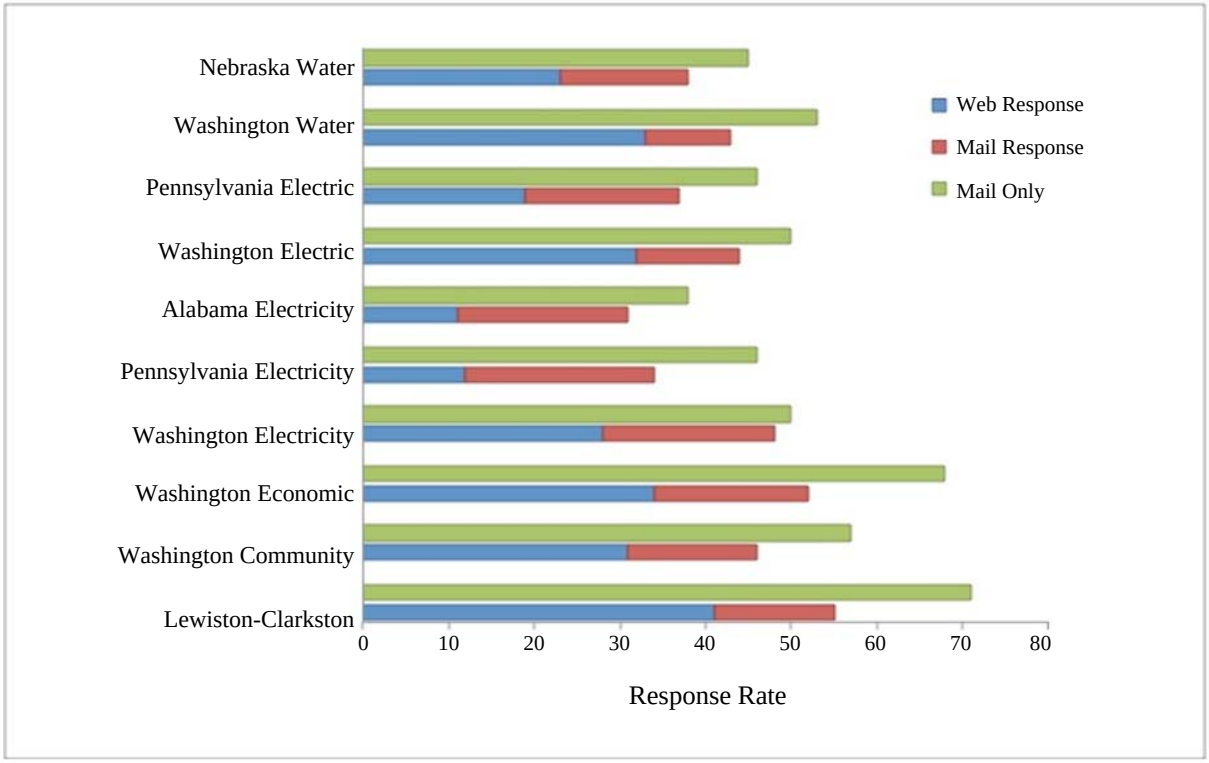

Figure 4.1 Mail-only treatment response rates vs. web-push response rates with proportion of that response received by each mode. (Dillman et al. 2014, Chapter 11).

The ten comparisons from these web-push studies revealed that those who responded over the Internet in the web-push treatment groups were significantly different than those who responded later to the mail questionnaire. For example, web respondents were younger, more educated, had higher incomes, and were less likely to live alone (Messer and Dillman 2011). However, the combined web and mail respondents in the web-push treatments were demographically quite similar to the mail-only treatment groups. The research concluded that individuals prone to respond by web could also be convinced to respond to the mail-only treatment. This finding was reinforced by the fact that a paper questionnaire follow-up to the web-only request produced significant improvement in response rates, whereas a web follow-up to a mail-only approach did not produce Internet responses that significantly improved overall response.

Although the web-push and mail-only treatment group responses were quite similar, the unweighted data exhibited nonresponse bias for certain demographics. Respondents had more education and children in the home than those who responded to the American Community Survey (discussed in more detail below) that now uses mail, web, phone and in-person interviews to obtain close to 97% response rates, and is relied on for producing official U.S. statistics for all U.S. states. Such comparisons were beyond the purpose and scope of these experiments and more investigation needs to be done to understand the nature of such differences. In addition, costs per respondent were not shown to be lower for Internet responses, because of contact costs being about the same for web-push and mail only methods, while producing fewer respondents (Messer and Dillman 2011). That seems likely to change as use of the Internet continues to expand to more people and areas of life.

Overall, the outcome of this coordinated set of studies made it clear that the web-push methodology offers considerable promise for obtaining web responses to household surveys. It was also clear that paper follow-up questionnaires would improve representation of people unable or unwilling to respond over the Internet.

4.4 Additional web-push tests on other populations and situations

In recent years the use of web-push data collection strategies has expanded and they are widely used in government, university, and private sector surveys in multiple countries. Uses have also spread beyond general public populations, and now involve survey situations where requests to respond via the Internet are not limited to mail contact. In addition, some surveys involve as many as three modes of contact and three modes of response, with the intent of getting very high response rates, while pushing as many respondents as possible to the web, in an effort to lower survey costs.

In 2013, the American Community Survey was converted from a sequence of mail-telephone-in person requests, to beginning with an Internet response, followed by the three remaining contact and response procedures (United States Census Bureau 2014, Chapter 7). The law requires U.S. citizens to respond to the American Community Survey (formerly the long form in the Decennial Census). Therefore, the overall response rate for occupied households was about 97%. Tests of web-push strategies began in 2011, when an initial experiment confirmed that web-push resulted in dramatically higher response rates (28% vs. 10%) over the Internet than did a "choice" strategy that also offered mail in the first contact (Tancreto 2012). In 2013, 28% of the responses from occupied households occurred over Internet, 22% by mail, 6% by telephone, and 43% by in-person interview. Thus, about 51% of the self-administered responses were over the Internet, a proportion that increased to 58% in 2015. Tests are now underway in support of plans to convert the 2020 Decennial Census to web-push methods with a similar follow-up.

The Japanese Census was converted to a web-push methodology in 2015 (City of Sapporo 2015). On-line response was about 37%, with one third of those responses coming from smartphones, which are used extensively in Japan. The remainder of the response was obtained by mail questionnaires and enumerator visits. The 2016 Australian Census and 2016 Canada Census were also conducted using web-push methodologies. Although final results are not yet available, it is known for Canada that 68% of households responded over the Internet, 20% by mail, with an additional 10% through enumerator visits, for an overall response rate of 98% (Statistics Canada 2016). The proportion of Internet responses in the Canadian Census is the highest I am aware of for a web-push household survey. In some areas of Canada a paper questionnaire was included with the request, providing a choice of response modes to respondents. The high Internet response (68%) and Internet plus mail response (88%) suggests great promise for use of a web-push methodology in that country and perhaps others with high Internet penetration.

A newly developed National Child Health Survey – developed as a replacement for a previous RDD household survey in the United States – plans to screen an address-based sample of children, then select a child for detailed reporting of health issues. However, instead of using two separate mail data collections, they tested in 2015 the possibility of reducing the process to do a one-step, in which the computer uses study criteria to immediately select and administer a topical health questionnaire for one child. This procedure

seeks to improve upon the two-step mail-only response process recently developed for the National Child Education Survey. Results from a pretest in 2015 were promising and it is now going through a second stage of testing.

The U.S. Residential Energy Consumption Survey, conducted for many years by the Energy Information Administration through in-person household interviews, is in the process of being changed to a web-push survey. This survey is noteworthy because it combines a cash incentive with the initial request to respond over the web, and also provides a post-incentive. The post incentive was deemed especially important because of the cost savings it provided by not having to send in-person interviewers to nonresponding households (Biemer, Murphy, Zimmer, Berry, Deng and Lewis 2015).

Not all web-push surveys use address-based sampling. The 2010 National Survey of College Graduates (NSCG) began sampling individuals who reported being a college-graduate in the previous year's American Community Survey and asked them to complete the NSCG, which is conducted every two years (Finamore and Dillman 2013). Postal addresses, as well as telephone numbers, were mostly available for the households where they had lived the previous year. Prior to 2010, households had been selected from the Decennial Census Long Form (last completed in 2000) with telephone, mail, and in some cases in-person interviews. In 2010, comparisons were made among pushing people to the telephone, pushing respondents to mail, and pushing respondents to the web, followed by use of the other two modes. All three of these treatments were followed by a final telephone effort in which responses could be made by that mode or either of the others. Two results were particularly important. First, all three response rates were within a few percentage points of each other, ranging from 74-77%, for this voluntary survey. However, the web-push strategy, in which 53% reported by web, proved to much less expensive, $48 per respondent vs. $66 for mail first and $75 for telephone first. It was concluded that the results from each procedure represented the original sample quite well.

A recent voluntary survey of spouses of U.S. military members compared a web-push strategy with a mail-push strategy. The web-push methodology produced a significantly higher response rate, 33% vs. 28%, with 87% of the web-push responses being received over the Internet (McMaster, LeardMann, Speigle and Dillman 2016). The web push strategy was also much less expensive, $61 per respondent vs. $89.

The success of the web-push strategies for the college graduate and military member studies may be for different reasons. All of the NSCG participants had at least a four-year college degree. Participants in the Family Study of Military Members were also relatively young. The authors of the latter study suggested that the fact that military members rely greatly on the Internet for communication with spouses during deployment might account for its greater effectiveness than mail-push methods.

Many other tests of a web-push methodology have emerged during the past decade. A Swiss study has shown that response rates of about 72 percent of households drawn from Swiss registration lists with 44% by web, 20% by mail and the remainder by telephone or in-person interviews (Roberts et al. 2016). In the United Kingdom, in-person interviews have been relied on far more extensively than telephone for conducting national statistical surveys. Recently, a decision was made to convert the Community Life Survey from an in-person interview to a web-push followed by mail strategy (United Kingdom Cabinet

Office 2016). This decision was made in order to lower costs, while also increasing the sample size. It remains to be seen what results will be obtained.

Private sector uses of web-push methods for specialized survey populations have also evolved. Nexant now conducts surveys of gas and electric utility customers with web-push methods. In the past telephone surveys were the dominant method. Companies whose customers are to be surveyed are able to provide postal addresses and telephone numbers for nearly all customers and email addresses from 20-40% of households (Sullivan, Leong, Churchwell and Dillman 2015). Following a procedure developed by Millar and Dillman (2011), emails are sent to those households to arrive shortly after the letter of request that contains as $2 incentive, followed by another email three days later, and if there is no response another paper survey is sent. Multiple tests have produced response rates of 40-80% with 80-90% of responses who received this email augmentation of the mail contact responding online, compared to about 35-70% of those without email addresses. Responses can be nudged upwards by 8-10 percentage points with a follow-up phone call to those without email addresses, compared to 1-2% for those with email addresses.

5 A promising future, but difficult challenges remain

5.1 Reasons for optimism

The development and deployment of web-push methodologies for survey data collection during the past decade provide reasons for optimism that higher quality survey data collection can be accomplished. That optimism arises less from the excitement over a specific approach of contacting people and convincing them to respond on the web than it does from a combination of considerations.

Address-based sampling now provides excellent household coverage and is conducive to the use of respondent selection procedures. Substantial proportions of survey populations can be contacted by one mode (mail) and encouraged to respond by another (web, or telephone). Survey sponsors not known to the recipient of the request can be legitimized through mail contact in ways that cannot be accomplished with email requests that go mostly unread or voice telephone requests that go mostly unanswered.

Postal contact also allows the sending of small token incentives with the request, thus providing motivation for making the transition from letter to computer and entering a URL (Uniform Resource Locator) and password. Multiple mail contacts provide the opportunity to offer more complete explanations of why a survey is being conducted and how the results will be used. Sending a paper questionnaire alternative in a later contact not only increases responses rates significantly, but brings in types of households not represented well among the initial internet responses. Several studies have also shown that the ability of web-push designs to bring in from half to three-fourths of all respondents quickly over the Internet, depending upon the sample frame and modes of contact, can reduce survey costs.

When email addresses are available for sample units, as is now the case for some survey populations, email augmentation (i.e., the sending of a quick email follow-up to the initial postal request to provide an electronic link that makes it easier for the recipient to respond over the internet) has been shown to improve web response considerably. Similarly, when telephone numbers are available, a telephone augmentation can be effective for improving response. The concept of using such contacts to augment previous mail contacts

encourages surveyors to think not just about stand-alone contacts, but how each contact becomes part of an overall response strategy.

As shown by the American Community Survey, The Canadian Census, National Science Foundation, and Nexant studies, multiple modes of contact and response provides the potential for achieving response rates that many survey sponsors thought were no longer possible. The ability to approach people with repeated requests to respond – and to do so in different modes – improves survey response more than any single mode of contact and/or response.

In addition, relying to a great extent on self-administration (internet and mail) achieves a better cultural fit with people than does a voice telephone conversation, which is increasingly out of sync with routine communication behavior that places great emphasis on texting and email. Also, changes in questionnaire construction methods from using different question structures and wording in each mode in the spirit of creating what's best for each mode through unified mode constructions assists in avoiding measurement differences across survey modes.

Over time, it seems likely that an increasing proportion of adults will be willing and able to respond to surveys over the Internet. Thus, web-push data collection procedures seem consistent with other societal trends that favor the internet over other forms of communications.

The promise of web-push survey methods stems from its ability to reduce survey error from coverage and survey nonresponse. In addition, our greater understanding of how visual vs. aural construction of questionnaires affects answers and the use of unified mode construction methods makes it possible to reduce measurement differences and error. It seems likely that the number of surveys using web-push methods is just beginning.

6 Challenges facing web-push data collection

Despite the potential of web-push data collection methods, there are also uncertainties regarding whether the use of Internet surveying will continue to expand in use. These concerns are the focus of the final section of this paper.

6.1 Fear of responding over the internet

When the push-to-web Australian Census began in 2016, a series of denial of service (DOS) attacks on the site prompted the Bureau of Statistics to turn off the system for fear of hackers. Such attacks are designed to overload a server with traffic, thus making it inaccessible to the intended users. This is only one kind of attack that might be made on a particular survey or computer user. Others include sending malware (e.g., spyware or ransomware) designed to gain access to or damage a computer that users unknowingly access by opening attachments or clicking on links. In addition, phishing emails are sometimes sent. They are designed to trick people into opening them, and providing personal information, for example, by appearing to be sent from a user well known to the recipient. The result of these various possibilities is to cause many people to worry about the security, or lack thereof, of the website, and information they provide in response

to web survey requests. The lack of trust in web surveys and concerns that information could be kept and used for non-survey purposes are also potential barriers to response.

Large scale surveys, especially those that have a great deal of public visibility, such as a nationwide Census that involves widespread prior communication inviting a response, present an inviting target for those hoping to harm the response process. Thus, even though sponsorship is known the perception of risk may be substantial. In the case of the Australian Census, the marketing focus on inviting everyone to respond on a particular "census day," made the situation worse than might otherwise have been the case. Thus, in addition to having to combat the potential of a cyber attack, survey sponsors face the challenge of restoring confidence in the data collection system.

Intentional attacks on individual computers and devices or on specific surveys are probably the largest peril facing surveying that involves the Internet. They are also a major justification for developing multiple response mode opportunities, and not relying entirely on the Internet. The reliance of web-push methodologies on multiple modes of responding provides some measure of protection for attacks on a particular survey, just as it now provides an alternative for those who now consider an internet response unacceptable. In especially large surveys, such as countrywide censuses, shifting away from asking everyone to respond on the same day may also lessen exposure as well as the impact of some of the potential Internet problems.

It is difficult to anticipate whether technological and social control advancements will negate risks associated with computer use. For now, this is an issue that threatens successfully surveying over the Internet that cannot be ignored.

6.2 Smartphones and the purse/pocket problem

A second, but quite distinct issue now challenging Internet data collection is the use of multiple devices for responding. Increasingly, people carry a computer device – mainly smartphones – with them. In most respects this is a very positive development. Because people carry the capability for providing a survey response with them throughout the day, survey requests can be responded to almost anytime from anywhere. This constant availability also brings to the fore what can be described as the pocket/purse problem. There are size preferences and probably limitations on the devices most people are willing to carry with them for use in cars, on public transportation, while working and when recreating.

Recent research has shown that while increasing portions of the population will respond to web requests on their smartphone, the small screen sizes present significant problems. Considerable research summarized elsewhere (Dillman, Hao and Millar 2016) has revealed that the proportion of smartphone responses has increased. In addition, it is difficult to ask many types of questions that seemed to work well in other survey modes. For example, Sarraf, Brooks, Cole and Wang (2015) have shown that the common question format of the item-on-the left with answer categories horizontally displayed to the right and the four-point scale placed below it, resulted in early abandonment of the response process and a dramatic increase in missing responses. In a later set of experiments, Barlas and Thomas (2016) have demonstrated the benefit of shortening scalar questions. These works bring into question the advisability of asking seven point fully

labeled scales, often favored in the past as ideal for interview surveys. Work by Stern, Sterrett and Bilgen (2016) suggest that grids – in which a general question establishing a set of response categories is followed by lists of items requiring an answer to each, a staple of paper and web questionnaires, are not an acceptable visual layout for smartphones.

An excellent review of the available research by Couper et al. (2017) concludes that questionnaires completed on mobile phones have lower response rates, higher breakoff rates, and longer completion times than do web surveys on personal computers. The authors note that part of the reason for these persistent problems may be that surveyors have not yet succeeded in optimizing design for mobile phones. Another factor that contributes to these problems may be competing demands for attention from smartphones and other activities as people are going about the daily rounds of life.

One of the challenges associated with designing for smartphones is maintaining unified question construction across all survey modes. This problem could be particularly acute when respondents in well-established surveys find that previously used question structures, wordings, and visual layouts are unilaterally changed for smartphone use. This challenge has been pointed out by Mistichelli, Eanes and Horwitz of the U.S. Census Bureau (2015). It's not yet clear whether survey designers are willing to change long standing ways of asking questions, (e.g., attitude questions with fewer categories and asking items-in-a series as individual items rather than a list of items introduced with a question that applies to the entire group of items being rated). If unified mode construction is to be used on smartphones, the needs of such devices are likely to be the major determinant of how items are presented across all modes.

The challenge now facing surveyors with regard to smartphones and mobiles also goes much deeper than how to present questions effectively in less space without the need for horizontal and vertical scrolling. In the early days of surveying, in-person interviewers could by their presence to engage the respondent's full attention. With mail, desktop, laptop, and tablets, it might be expected that respondents would often, if not normally, complete surveys at times they were not likely to be interrupted. Smartphones, by their nature are interruption devices, with the possibility of receiving texts, voice phone calls, emails at any moment, often while moving physically through one's daily activities. Answering some surveys may require consultation of records one does not have access to when away from home, or consultation with another household member, which seems harder to achieve if one tries to complete a questionnaire while on the move. The competition for attention that occurs with such devices might lead a surveyor to encourage a respondent not to fill out the survey on a smartphone and instead ask them to do it on their laptop or from home. The problem with that approach is for significant numbers of people smartphones may be their only computer or the only one they attend to on a daily basis. Also, it seems likely that the more one introduces barriers to answering a questionnaire "now", the less likely people are to answer at all.

Working through these issues is one of the largest challenges facing survey methodologists today. But, on a positive note, when multiple modes of contact are used and multiple ways of responding are offered, it seems easier to guide respondents to the most effective way for them to respond as well as for the success of the survey.

6.3 Sponsor Reluctance to undertake mixed-mode surveys and modify single-mode procedures

An additional peril facing web-push surveys is associated with the stress many organizations face in using multiple modes of survey contact and/or response. Each mode of contact and response requires specialized skills, equipment, and software. In order to be effective, it must also be effectively coordinated to deal with many issues at once, as described elsewhere (Dillman et al. 2014, Chapter 11).

Survey sponsors that have specialized in only one form of data collection, or who want to keep data collection activities simple, may be tempted to avoid the use of second or third modes of data collection. This is not likely to happen when high response rates are required (e.g., a national census) or when a substantial economic incentive exists for pushing early respondents to the web (e.g., Biemer et al. 2015). However, the development of do-it-yourself software has encouraged many surveyors to find ways of using only web data collection. Previous research has suggested that significantly biased results toward greater education and income will be produced if data collection stops with only web responses (Rookey, Hanway, and Dillman 2008; Messer and Dillman 2011). Over time this bias may be reduced, but appears not to have happened yet for general populations. Another source of low response rates and potential bias occurs when surveyors obtain only email addresses for a proposed survey, thus eliminating the potential for prior mail contact that allows inclusion of an incentive for encouraging respondents to respond over the Internet.

Making appropriate changes, even when the need is substantial, takes time. In the 1990's the U.S. Census Bureau developed a pre-notice, paper questionnaire, follow-up postcard strategy for data collection (Dillman, Clark and Sinclair 1995), which was used in the 2000 and 2010 Censuses. After the ACS push-to-web strategy was introduced in 2013, it continued to use this approach. The problem it presented is that the follow-up postcard could not provide password information (visible to anyone who picked up the postcard), thus creating the expectation that they would need to return to the web-request letter for that information. Also, the impression of the sequence of sending a pre-notice informing people they would receive a request to respond (by internet), a second letter asking them to go to the web using the provided information, and then a postcard reminder to follow through seemed unnecessarily laborious. Thus, a new procedure of abandoning the pre-notice and using a letter follow-up was introduced. A test of this procedure by the Census Bureau led to its adoption in August 2015 (Clark and Roberts 2016) and a significant increase of 2.5 percentage points in internet responses and a slight reduction in overall costs.

There are many other issues involved in shifting from single-mode thinking to widespread adoption of web-push surveys that involve multiple modes of response. For example, how do researchers overcome the frustration of willing respondents, who are irritated by being told they will have to wait for that request to come in a few weeks? Also, when telephone numbers are available, a phone call could be used as a follow-up reminder with encouragement, rather than simply trying to interview people over the phone. Experimental testing of these alternatives needs to be done.

6.4 Impacts of new discoveries and innovations

Anticipating the future is difficult. When telephone interviewing was rising towards prominence in the 1970's, personal computers were not yet available. And, virtually no one thought, or even imagined, that

only two decades later the telephone that had been tethered to our homes and workplaces would be carried with us nearly everywhere we went using wireless connections. When Internet surveying began in the 1990's few anticipated that not only would the large and clunky desktops that began occupying people's homes would transition to laptops that people would carry with them from place to place. And, that device would later transition to tablets and smartphones with touch screens – both with far more computing power than their original desktops and laptops.

A recent analysis by Friedman (2016) details the monumental changes in the capabilities and power of personal devices that are increasingly taken for granted by large portions of the worldwide population. He traces these capabilities to the exponential growth of each of five different components of today's computers: 1) integrated circuits that do the computing, 2) memory units that store and retrieve information, 3) networking systems that provide communications within and across computers, 4) the software applications that enable different computers to perform various tasks individually and together, and 5) sensors that detect movement, language, light, sound and other features of the environment and turn it into digitized data. He traces the rapid acceleration in these aspects to the development of the iPhone and related innovations occurring since 2007, and their melding into what he describes as the supernova (or cloud).

These developments were only dimly anticipated, even by many of the innovators who created them. Trying to imagine the future is no easier now than it was in the past. For example, voice activation of computer searches is rapidly replacing the individual tapping and swiping of commands on smartphones. Twenty percent of Google searches on Android-powered handsets in the United States are now input by voice (The Economist 2017). In addition, people can also dictate emails and text messages with reasonable success. Will voice-activated answers be the next wave of development for survey designers? It is easy to imagine one being interviewed by his or her smartphone. And, is it possible that simultaneous translations from one language to another, which can now be done with reasonable success, become common on surveys? But, herein lies a fundamental challenge, described by Friedman – the speed with which human beings and societies can adapt to those changes.

Many potential respondents of interest to surveyors still rely on feature phones, while others are racing madly to adopt the most advanced computing and communication device they find practical. And still others are reluctant to use any computer at all. The differences in people's capabilities and preferences require surveyors to be neither too far ahead nor too far behind where most people are.

This raises the issue of whether web-push methods are simply another transitional phase of survey design that may fade out as quickly as it has risen in prominence. The mixed-mode and tailored design focus that now appears to dominate the thinking of survey designers is recognition of the heterogeneity that exists among populations, whose opinions and behaviors surveyors seek to describe.

For a time it appeared that some surveyors thought the value of mixed-mode surveying was in offering people a choice of which mode they would use to respond to a survey request. However, this is only partly true. The real response power of mixed-mode designs for improving response rates stems from making multiple contacts effectively. Each contact gives an opportunity to provide new information about one's survey request and, in some cases, to reach people who cannot be contacted by other survey modes. When survey response requests are offered by different modes there is often an opportunity to improve coverage

(reaching people who can't be reached by another mode) and get people to attend to persuasive arguments for being a respondent. Also, the sequencing of those contacts may help with motivating people to respond (e.g., use of email augmentation of postal letters that makes it easier to respond).

7 Summary and Conclusion

Web-push data collection that begins with a postal mail request to respond over the Internet is one of the major survey design developments of the early 21st century, now offering promise of faster, less-expensive surveys. Many surveyors have been surprised by today's reliance on an initial postal contact. Although mail surveys had often been used to collect survey data, it was expected by many to disappear with the rise of Internet.

The critical development that encouraged reconsideration of mail contact methods greater use was by Link et al. (2008) and Battaglia et al. (2008). This research showed that residential address lists available from the U.S. Postal Service provided the best sample coverage of U.S. residences and could be used to support effective mail surveys of the general public. This work was encouraged by the strong desire to find alternatives to RDD telephone surveys that faced continually declining response and other challenges.

A series of studies, beginning in 2007, looked for ways to use mail contacts to push householders to the web from these address-based lists. This work focused on combining both Internet and paper responses. It was supported by several years of earlier research on measurement differences across modes that showed responses to web and paper questionnaires were quite similar so long as similar question structures, wordings, and visual layouts were used for both data collection methods. Ten experimental comparisons made in these studies received a web-push response rate of 43% of households, with about 60% of the responses coming over the internet and the remainder being obtained by a mail follow-up (Dillman et al. 2014). Major surveys in several countries have researched and adopted the use of web-push methods that rely on not only web and mail, but now include telephone and/or in-person follow-up in their protocols. The goal is to achieve greater response rates and data quality, which a decade ago were thought to be no longer possible in household surveys.

We are now in an era of tailored design in which different survey designs are used for different survey topics, populations, and survey situations. However, it seems likely that web-push data collection methods will see increased use throughout the industrialized world, as survey sponsors seek to benefit from the low cost of internet data collection in order to lower the overall cost of current surveys.

However such methods face challenges that need attention. One is the risk to surveys and respondents from malware, phishing, and server attacks. Another is the increased reliance on smartphones that may require significant changes in how questions are structured and presented to respondents. In addition, the reluctance of organizations and individuals to accept and master the greater complexity associated with shifting from single mode to mixed-mode surveys is a significant challenge.

The history of surveying over the last 75 years has involved significant transitions from the dominance of in-person interviews, to heavy reliance on voice telephone methods, and now to online and mixed mode surveys. It remains to be seen whether web-push methods – now growing in use as a replacement – have a

lasting presence, or will eventually be replaced with web-only data collection, or with other procedures that remain to be innovated or have not yet been conceived.

References

Anderson, M., and Perrin, A. (2016). 13% of Americans don't use the Internet, Who are they? Available at http://www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/. Accessed May 2, 2017.

Australian Bureau of Statistics (2016). Making Sense of the Census. Available at http://www.abs.gov.au/websitedbs/censushome.nsf/home/2016. Accessed May 2, 2017.

Barlas, F.M., and Thomas, R.K. (2016). Good questionnaire design: Best practices in the mobile era. *American Association for Public Opinion Research*, January 19th.

Battaglia, M.P., Link, M.W., Frankel, M.R., Osborn, L. and Mokdad, A.H. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly*, 72(3), 459-469.

Biemer, P., Murphy, J., Zimmer, S., Berry, C., Deng, G. and Lewis, K. (2016). A test of Web/PAPI protocols and incentives for the residential energy consumption survey. Unpublished paper presented at Annual Conference of the American Association for Public Opinion Research. May 13th.

Blankenship, A.B. (1977). *Professional Telephone Surveys*. New York: McGraw-Hill Book Company.

Blumberg, S.J., and Luke, J.V. (2017). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January-June 2016.

Brick, J.M., and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *Annals of the American Academy of Political and Social Science*, 645(1), 36-59.

Brick, J.M., Williams, D. and Montaquila, J.M. (2011). Address-based sampling for subpopulation surveys. *Public Opinion Quarterly*, 75(3), 409-428.

Christian, L.M., and Dillman, D.A. (2004). The influence of symbolic and graphical language manipulations on answers to paper self-administered questionnaires. *Public Opinion Quarterly*, 68, 1, 57-80.

Christian, L.M., Dillman, D.A. and Smyth, J.D. (2008). The effects of mode and format on answers to scalar questions in telephone and Web surveys. In *Advances in Telephone Survey Methodology*, (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link and R.L. Sangster). New York: Wiley-Interscience, 250-275.

Christian, L.M., Parsons, N.L. and Dillman, D.A. (2009). Designing scalar questions for Web surveys. *Sociological Methods and Research*, 37(3), 393-425.

City of Sapporo (2015). The Japanese government is conducting a Population Census. Available at https://www.city.sapporo.jp/city/english/news/news201508_1e.html. Accessed October 1, 2016.

Clark, S., and Roberts, A. (2016). Evaluation of August 2015 ACS mail contact strategy modification. *2016 American Community Survey Research and Evaluation Report Memorandum Series ACS16-0RER-13.*

Couper, M.P., Antoun, C. and Mavletova, A. (2017). Mobile Web surveys. In *Total Survey Error in Practice*, (Eds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker and B.T. West). New Jersey: Hoboken.

de Leeuw, E.D. (1992). Data quality in mail, telephone and face-to-face surveys. *TT-Publications Amsterdam.*

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* 21(2), 233-255.

de Leeuw, E., Villar, A., Suzer-Gurtekin, T. and Hox, J. (2017). How to design and implement mixed-mode surveys in cross National Surveys: Overview and guideline. In *Total Survey Error in Practice*, (Eds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker and B.T. West). New Jersey: Hoboken.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method, 2nd Edition*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2002). Navigating the rapids of change: Some observations on survey methodology in the early 21st century. *Public Opinion Quarterly,* 66(3), 473-494.

Dillman, D.A. (2005). Telephone surveys. In *Encyclopedia of Social Measurement,* (Ed., K. Kempf-Leonard), Volume 3. London, UK: Elsevier Press, 757-762.

Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method.* 2007 Update with New internet. Visual and Mixed-mode Guide. New Jersey: Hoboken.

Dillman, D.A., and Christian, L.M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods,* 17(1), 30-52.

Dillman, D.A., and Tarnai, J. (1988). Administrative issues in mixed-mode surveys. In *Telephone Survey Methodology*, (Eds., R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II and J. Waksberg), New York: John Wiley & Sons, Inc., 509-528.

Dillman, D.A., and Edwards, M.L. (2016). Designing a mixed-mode survey. In *The SAGE Handbook of Survey Methodology,* (Eds., C. Wolfe, D. Joye, T.W. Smith and Y.-c. Fu), Sage Publications, Thousand Oaks, CA, 255-268.

Dillman, D.A., Clark, J.R. and Sinclair, M.D. (1995). How prenotice letters, stamped return envelopes, and reminder postcards affect mailback response rates for census questionnaires. *Survey Methodology,* 21, 2, 159-165. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1995002/article/14394-eng.pdf.

Dillman, D.A., Gertseva, A. and Mahon-Haft, T. (2005). Achieving usability in establishment surveys through the application of visual design principles. *Journal of Official Statistics,* 21(2), 183-214.

Dillman, D.A., Hao, F. and Millar, M.M. (2016). Improving the effectiveness of online data collection by mixing survey modes. In *The Sage handbook of Online Research Methods, 2nd Edition,* (Eds., N. Fielding, R.M. Lee and G. Blank). Sage Publications, London, 220-237.

Dillman, D.A., Smyth, J.D. and Christian, L.M. (2014). *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition*. New Jersey: Hoboken.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. and Messer, B.L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1), 1-18.

Dutwin, D., and Lavrakas, P. (2016). Trends in telephone outcomes, 2008-2015. *Survey Practice*, 9(3). Available at http://www.surveypractice.org/.

Edwards, M.L., Dillman, D.A. and Smyth, J.D. (2014). An experimental test of the effects of survey sponsorship on Internet and mail survey response. *Public Opinion Quarterly*, 78(3), 734-750.

Finamore, J., and Dillman, D.A. (2013). How mode sequence affects responses by internet, mail and telephone in the national survey of college graduates. Presentation to European Survey Research Association, Ljubljana, Slovenia, July 18.

Friedman, T.L. (2016). *Thank you for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations*. New York, Farrar: Straus and Giroux.

Groves, R.M., and Kahn, R.L. (1979). *Surveys by Telephone*. New York: John Wiley & Sons, Inc.

Groves, R.M., and Peytcheta, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Harter, R., Battaglia, M.P., Buskirk, T.D., Dillman, D.A., English, N., Mansour, F., Frankel, M.R., Kennel, T., McMichael, J.P., McPhee, C.B., Montaquila, J., Yancey, T. and Zukerberg, A.L. (2016). Address-base sampling. *American Association for Public Opinion Research* Task Force Report. Available at http://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf.aspx, 140 pages.

Hoffman, D.D. (2004). *Visual Intelligence: How we Create What we See*. New York: Norton.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64(2), 125-148.

Kerlinger, F.N. (1965). *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston.

Krosnick, J.A., and Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, 51(2), 201-219.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L. and Mokdad, A.H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, 72(1), 6-27.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I. and Vehovar. V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, 50(1), 79-104.

McMaster, H.S., LeardMann, C.A., Speigle, S. and Dillman, D.A. (2016). An experimental comparison of web-push vs. paper-only survey procedures for conducting an in-depth health survey of military spouses. BMC Medical Research Methodology.

Messer, B.L. (2012). Pushing households to the web: Results from Web+mail experiments using address based samples of the general public and mail contact procedures. Ph.D. Dissertation. Washington State University, Pullman.

Messer, B.L., and Dillman, D.A. (2011). Surveying the general public over the Internet using address-based sampling and mail contact procedures. *Public Opinion Quarterly*, 75(3), 429-457.

Messer, B.L., Edwards, M.L. and Dillman, D.A. (2012). Determinants of item nonresponse to Web and mail respondents in three address-based mixed-mode surveys of the general public. *Survey Practice*, 5(2), 1-9. Paper available at http://wwww.surveypractice.org/.

Millar, M.M., and Dillman, D.A. (2011). Improving response to Web and mixed-mode surveys. *Public Opinion Quarterly*, 75(2), 249-269.

Mistichelli, J., Eanes, G. and Horwitz, R. (2015). Centurion: Internet Data Collection and Responsive Design. Presentation to Federal Economic Statistics Advisory Committee, June 12.

Mohorko, A., de Leeuw, E. and Hox, J. (2013). Coverage bias in European telephone surveys: Developments of landline and mobile phone coverage across countries and over time. *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=828.

Nathan, G. (2001). Telesurvey methodologies for household surveys – A review and some thoughts for the future? *Survey Methodology*, 27, 1, 7-31. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2001001/article/5851-eng.pdf.

O'Muircheartaigh, C., English, N. and Eckman, S. (2007). Predicting the relative quality of alternative sampling frames. *2007 Proceedings of the Survey Research Methods Section*, American Statistical Association, [CD ROM], Alexandria, VA: American Statistical Association.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. London: Bradford Books.

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

Pew Research Center (2012). Assessing the representativeness of public opinion surveys. Available at http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/. Accessed October 24, 2016.

Redline, C.D., and Dillman, D.A. (2002). The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little), New York: John Wiley & Sons, Inc.

Redline, C.D., Dillman, D.A., Dajani, A. and Scaggs, M.A. (2003). Improving navigational performance in U.S. census 2000 by altering the visual languages of branching instructions. *Journal of Official Statistics*, 19(4), 403-420.

Roberts, C., Joye, D. and Staehli, M.E. (2016). Mixing modes of data collection in Swiss social survey: Methodological report of the LIVES-FORS mixed mode experiment. Lives working paper 2016.48. Swiss National Centre of Competence in Research, a research instrument of the Swiss National Science Foundation.

Rookey, B.D., Hanway, S. and Dillman, D.A. (2008). Does a probability-based household panel benefit from assignment to postal response as an alternative to Internet-only? *Public Opinion Quarterly*, 72(5), 962-984.

Sarraf, S., Brooks, J., Cole, J. and Wang, X. (2015). What is the impact of smartphone optimization on long surveys? Presentation to American Association for Public Opinion Research Annual Conference, Hollywood, FL, May 16.

Smyth, J., Christian, L.M. and Dillman, D.A. (2008). Does 'Yes or No' on the telephone mean the same as check-all-that-apply on the Web? *Public Opinion Quarterly*, 72(1), 103-111.

Smyth, J.D., Dillman, D.A., Christian, L.M. and McBride, M. (2009). Open-ended questions in Web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, 73(2), 325-337.

Smyth, J.D., Dillman, D.A., Christian, L.M. and O'Neill, A.C. (2010). Using the Internet to survey small towns and communities: Limitations and possibilities in the early 21st century. *American Behavioral Scientist*, 53(9), 1423-1448.

Smyth, J.D., Dillman, D.A., Christian, L.M. and Stern, M.J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly*, 70(1), 66-77.

Statistics Canada (2016). 2016 Census of Population collection response rates. Available at http://www12.statcan.gc.ca/census-recensement/2016/ref/response-rates-eng.cfm. Accessed October 24, 2016.

Statistics Japan (2015). Almost 20 million households responded online in the 2015 Population Census of Japan. Available at http://www.stat.go.jp/english/info/news/20151019.htm. Accessed October 1, 2016.

Stern, M., Sterrett, D. and Bilgen, I. (2016). The effects of grids on Web surveys completed with mobile devices. *Social Currents*, 3(3), 217-233.

Sullivan, M., Leong, C., Churchwell, C. and Dillman, D.A. (2015). Measurement and Cost Effects of Pushing Household Survey Respondents to the Web for Surveys of Electricity and Gas Customers in the United States. Unpublished paper presented to the European Survey Research Association, Reykjavik, Iceland, July 16th.

Tancreto, J. (2012). 2011 American Community Survey Internet Tests: Results from First Test in April 2011. #ACS12-RER-13-R2. 2012 American Community Survey Research and Evaluation Report Memorandum Series, June 25th.

Tarnai, J., and Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail and telephone surveys. In *Context Effects in Social and Psychological Research*, (Eds., N. Schwarz and S. Sudman), New York: Springer Verlag, Inc. 115-129.

The Economist (2017). Now we're talking: Voice technology is making computers less daunting and more accessible. January 7th – 13th, 422 (No. 9022), 9.

Thomas, R., and Barlas, F. (2016). It's a Small Screen After All: Improving Measurement in an Ever-changing Online Survey World. GFK Webinar, September 27th.

Tourangeau, R. (2017). Mixing modes: Tradeoffs among coverage, nonresponse and measurement error. In *Total Survey Error in Practice*, (Eds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N. Clyde Tucker and B.T. West), New Jersey, Hoboken: John Wiley & Sons, Inc.

Tourangeau, R., Couper, M.P. and Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly,* 68(3), 368-393.

United Kingdom Cabinet Office (2016). Consultation Response: Community Life Survey: Development and implementation of online survey methodology for future survey years. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/539111/community_life_survey_consultation_response_final.pdf.

United States Census Bureau (2014). American Community Survey Design and Methodology, Version 2.0. Available at http://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html. Accessed October 15, 2016.

Ware, C. (2004). *Information Visualization: Perception for Design, 2nd Edition*. Karlsruhe, West German: Morgan Kaufman.

Williams, D., Brick, J.M., Montaquila, J.M. and Han, D. (2014). Effects of screening questionnaires on response in a two-phase postal survey. *International Journal of Social Research Methodology*, 19(1), 51-67.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 31-40
Statistics Canada, Catalogue No. 12-001-X

A layered perturbation method for the protection of tabular outputs

Jean-Louis Tambay[1]

Abstract

The protection of data confidentiality in tables of magnitude can become extremely difficult when working in a custom tabulation environment. A relatively simple solution consists of perturbing the underlying microdata beforehand, but the negative impact on the accuracy of aggregates can be too high. A perturbative method is proposed that aims to better balance the needs of data protection and data accuracy in such an environment. The method works by processing the data in each cell in layers, applying higher levels of perturbation for the largest values and little or no perturbation for the smallest ones. The method is primarily aimed at protecting personal data, which tend to be less skewed than business data.

Key Words: Confidentiality; Data perturbation; Tabular outputs.

1 Introduction

Statistical agencies are under pressure to provide more information from their data holdings to external users. Many now enable the creation of custom tables through on-line query systems. But the risks of a disclosure of confidential information increase with the quantity of outputs released. To address this problem agencies can go from one extreme, which is to severely limit the amount of information being released, to another, which is to generate outputs from model-based synthetic microdata. Perturbative methods, which add noise to microdata or aggregate results, lie somewhere in between. This paper proposes a perturbative method for quantitative administrative data, such as personal taxation data, in a custom tabulation environment. Section 2 provides some background information, outlines desirable objectives and reviews standard approaches for the protection of tables of magnitude. Section 3 presents the proposed Layered Perturbation Method (LPM) and provides some of its properties. An empirical evaluation is given in Section 4 and outstanding issues are discussed in Section 5.

2 Background

The proposed strategy aims to protect the confidentiality of tables of magnitude in a semi-controlled custom tabulation environment. It was primarily developed for administrative (census-like) data, notably personal taxation data. At Statistics Canada, such outputs are subject to disclosure control rules including minimum population sizes for identifiable geographic areas, the use of minimum-cell-size and dominance rules to suppress sensitive (confidential) cells, and the application of complementary cell suppression (CCS) to prevent the recuperation of sensitive cell values.

While personal data are inherently safer than business data, they are more readily used in custom tabulations. And with wider access to custom tabulations it becomes increasingly difficult to carry out CCS

1. Jean-Louis Tambay, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: jean-louis.tambay@canada.ca.

effectively. Alternative methods need to be considered. The proposed method consists of applying a perturbative technique, independently, in every non sensitive cell of every table. Only sensitive cells are suppressed, although some may become releasable if perturbed. The method is meant to protect sensitive cells in tables as well as to guard against residual disclosure from multiple tables – especially disclosure by the differencing of nested totals. The focus is on protecting two totals that differ by one unit.

It is assumed that we are in a semi-controlled environment where access is somewhat restricted, or at least not anonymous, so that some monitoring and control of requests is applied. This precaution is needed because offering unrestricted tabulations to anonymous hackers trying to exploit every vulnerability (in particular, through multiple requests involving carefully chosen sets of units) could lead to the approximate disclosure of unit values under certain conditions. The method is developed for census-like data, which are riskier, but it could undoubtedly be adapted to sample data if needed. The strategy is better suited to personal data as they are less subject to dominance than business data, and near-dominant cells get perturbed the most. But with some adaptation users may see to what extent the strategy could meet their needs for other types of data.

If possible, we would like the strategy to address other disclosure issues, such as the protection of ratios and of other types of outputs. Other desirable features are the ability to treat zeroes and negative values, the maintenance of data quality, the preservation of additivity in tables, and operational aspects such as computational simplicity and the use of minimal manual intervention.

In this paper we use a $P-$percent rule to identify sensitive cell totals, meaning that a cell is sensitive if the aggregate contribution from the smallest units, starting with the third-largest, is less than $P\%$ of the value of the largest unit (i.e., if $X - x_1 - x_2 < P\%\ x_1$, where X is the cell total and x_i is the contribution of its i^{th} largest unit). We assume that cells failing a minimum-cell-size rule are also sensitive.

We are interested in preserving quality and confidentiality for magnitude data in a custom tabulation environment. Techniques for tables of magnitude such as CCS (Cox and Sande 1979) and Controlled Tabular Adjustment (Cox and Dandekar 2004) do not work very well in such an environment. They require solving optimization problems to find table-specific solutions. Problems start to occur when trying to protect huge, complex and/or related (i.e., linked) tables, such as the inability to reach a solution, or the use of heuristics that may yield inconsistencies in suppression or perturbation patterns that can be exploited by hackers. It is far easier to perturb cell totals directly, e.g., by the application of random noise, but one still needs to look at the microdata to ensure adequate protection while controlling the impact on quality. And without additional measures it can lead to inconsistencies within and between tables that can be exploited by hackers.

Microdata perturbation, where data are perturbed at the microdata level, is better suited for our multi-table environment. Tables are additive and usually without suppression; with consistent results between tables. If custom tables are allowed it may be possible to recover some individual perturbed values directly or by differencing, so the noise level for each unit would need to be high enough to meet target ambiguity levels. As a result, the cumulated noise for specific aggregates can be large. A microdata perturbation method developed and used at the U.S. Census Bureau is the EZS method (Evans, Zayatz and Slanta 1998). EZS multiplies individual values x_i by a weight $w_i = 1 + \varepsilon_i$, where ε_i are i.i.d. random variables with mean

0 and variance σ_ε^2. Two distributions for ε_i of interest are the split triangular distribution (shaped like Figure 2.1) and the split uniform distribution (shaped like Figure 2.2) whose corresponding values of σ_ε^2 are $(3a^2+2ab+b^2)/6$ and $(a^2+ab+b^2)/3$, respectively. The ε_i (or w_i) are permanently attached to their unit i. Applying the same noise to all variables will not affect ratios. If it is necessary to protect ratios different weights w_i should be used for different variables, or unit-specific weights can be used jointly with unit-variable specific weights.

Figure 2.1 Split triangular distribution.

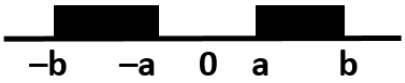

Figure 2.2 Split uniform distribution.

There are ways to attenuate the cumulative impact of microdata perturbation on quality. Massell and Funk (2007) suggest to balance the random noises within cells for a primary table to limit their impact there. Other methods perturb microdata, but not always the same way, allowing some inconsistencies in results. Giessing (2011) proposes to multiply unit values x_i by $w_i = 1 \pm |\varepsilon_i|$, for ε_i i.i.d. $\mathrm{N}(0, \sigma_0^2)$, except in sensitive cells, where the largest value gets multiplied by $w_i = 1 \pm (\mu_0 + |\varepsilon_i|)$. The value μ_0 is chosen to give an appropriate level of protection for sensitive cells, allowing a lower value of σ_0^2 to be used overall. But if σ_0^2 is too low the method may not sufficiently protect against disclosure by differencing. The Australian Bureau of Statistics' Top Contributors Method (TCM), developed for its TableBuilder remote access application, consists of perturbing the largest respondents in each cell in a semi-consistent way, i.e., where parts of their noise is applied consistently (Thompson, Broadfoot and Elazar 2013). The LPM uses some of the same concepts but, as will be explained, protects more against differencing.

Other commonly used strategies such as rounding, (sub-)sampling and swapping units, say between neighbouring areas, are better suited for the protection of frequency tables.

3 The Layered Perturbation Method (LPM)

3.1 Description

The LPM is a perturbative method for totals that focuses on disclosure from differencing. When used in tables of magnitude it allows cell suppression to be restricted to sensitive cells. Three basic ideas underlie the LPM. The first two are similar to the TCM approach.

The first basic idea is the attachment of pseudo-random hash numbers (PRNs) to units to produce consistent perturbation outcomes when needed. This discourages the use of repeated queries to improve the

estimation of unperturbed totals. The EZS method is used to multiply the value of a unit i by a weight $w_i = 1 + \varepsilon_i$, with $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ as above. To obtain consistent results ε_i are generated from a unit-specific PRN that is uniformly distributed over $[0,1)$. For example, use $h_i / 1000$, where h_i are generated from the Social Insurance Number (e.g., $h_i = Mod(SIN_i \bullet P, 1000)$ for P a large prime). Using h_i will always perturb unit i the same way. To perturb unit i the same way only when it appears in the same cell total, generate cell-unit level noise $w_i' = 1 + \varepsilon_i'$ from $h_i' = Mod(h_i + h_{tot}, 1000) / 1000$, where $h_{tot} = \sum_{i \in cell} h_i$. Primes are used to designate cell-unit specific noises and perturbations. All noise values are derived from h_i or h_i'.

The second idea is the application of perturbation to units in each cell by layers. The largest four units are perturbed in a random but *consistent* manner using perturbation weights w_i generated from h_i. The next largest units, say units 5 to 9, are perturbed in a semi-consistent manner. Their perturbation is a mixture of unit specific weights w_i and unit-cell specific weights w_i'. Smallest units are not perturbed. Their values are protected from differencing by the unit-cell perturbations of units 5 to 9 since adding or removing a unit in a cell, no matter how small, will affect the w_i' for those units. The number of units per layer is flexible, we have found that four and five, respectively gave satisfactory results.

A third set of measures mostly targets the issue of differencing. The direction of noise for even-ranked units is reversed (w_i are set from $(-1)^{i+1}\varepsilon_i$) to increase variances of differences when a top-ranked unit is changed. For units 5 to 9 a random mixture of w_i and w_i' is applied to lessen the risk when a small unit is added or removed. Finally, the noise for the top three units is amplified in nonsensitive cells with greater dominance. This allows lower levels of noise to be used generally, reducing the overall impact of the perturbation on data quality.

A suggested application of the LPM would consist of suppressing all sensitive and small cells (e.g., $n < 10$) and perturbing remaining cells. Because of the protection offered by perturbation, cells that are slightly sensitive may also be publishable. For other cells with cell total $X = \sum_{i \in cell} x_i$, set perturbed value Z as

$$Z = X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4 - \sum_{i=5}^{9} \left\{ (-1)^i \alpha_i \varepsilon_i - (1 - \alpha_i)\varepsilon_i' \right\} x_i .$$

K, L and M are set to increase the noise of Z, when needed (set K, L and $M \geq 1$). The α_i are random variables that are independent of ε_i, e.g., $\alpha_i \sim \text{Uniform}(0,1)$ or $\alpha_i = \text{Mod}(h_i, 8)/7$.

3.2 Some results

Let $\varepsilon_i, \varepsilon_i' \sim (0, \sigma_\varepsilon^2)$, $\alpha_i \sim \text{Uniform}(0,1)$, i.i.d. and let K, L and M be fixed (for now). It follows that:

$$E(Z) = X \text{ and } V(Z) = \left\{ K^2 x_1^2 + L^2 x_2^2 + M^2 x_3^2 + x_4^2 + \tfrac{2}{3} \sum_{i=5}^{9} x_i^2 \right\} \sigma_\varepsilon^2 .$$

Let X_{-1}, X_{-2}, X_{-3} and Z_{-1}, Z_{-2}, Z_{-3} equal X and Z for the cell after removing units 1, 2 and 3, respectively. Keeping subscripts from the original cell (i.e., subscript 2 refers to the unit that was second in X) we have:

$$Z_{-1} = X_{-1} + K\varepsilon_2 x_2 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10}\left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i,$$

$$Z_{-2} = X_{-2} + K\varepsilon_1 x_1 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10}\left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i, \text{ and}$$

$$Z_{-3} = X_{-3} + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10}\left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i.$$

We can obtain Z_{-i} for other units similarly. If we estimate the dropped units as $\hat{x}_i = Z - Z_{-i}$ it can be shown that, with $G = 2\frac{2}{3}x_5^2 + 2\sum_{i=6}^{9} x_i^2 + \frac{2}{3}x_{10}^2$,

$$E(\hat{x}_i) = x_i,$$

$$V(\hat{x}_1) = \left\{K^2 x_1^2 + (K+L)^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_2) = \left\{L^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_3) = \left\{M^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_4) = \left\{x_4^2 + G\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_5) = \left\{\tfrac{2}{3}x_5^2 + 2x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_6) = \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_7) = \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_8) = \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + \tfrac{2}{3}x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2,$$

$$V(\hat{x}_9) = \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2 + x_{10}^2\right\}\sigma_\varepsilon^2, \quad \text{and}$$

$$V(\hat{x}_i) = \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2\right\}\sigma_\varepsilon^2, \quad \text{for} \quad i > 9.$$

If we assume that K, L and M are fixed we can set them based on some requirement for $V(\hat{x}_i)$. For example, we may want to have $V(\hat{x}_i) = x_i^2/30$ since, for $z \sim N(0,1)$, $\Pr(|z| > 0.44) = 0.66$ which for $\hat{x}_i \sim N(x_i, x_i^2/30)$ gives $\Pr\{|\hat{x}_i - x_i| \geq 8\% x_i\} = 66\%$.

To obtain $V(\hat{x}_i) = x_i^2/NN$ we can solve (fixed) K, L and M in reverse order. This gives

$$M = \frac{\sqrt{(x_3^2 + x_4^2)(x_3^2/NN\sigma_\varepsilon^2 - G) - x_3^2 x_4^2} - x_4^2}{x_3^2 + x_4^2}$$

$$L = \frac{\sqrt{(x_2^2 + x_3^2)(x_2^2/NN\sigma_\varepsilon^2 - G - x_4^2(M+1)^2) - M^2 x_2^2 x_3^2} - Mx_3^2}{x_2^2 + x_3^2}$$

$$K = \frac{\sqrt{(x_1^2 + x_2^2)(x_1^2/NN\sigma_\varepsilon^2 - G - x_3^2(L+M)^2 - x_4^2(M+1)^2) - L^2 x_1^2 x_2^2} - Lx_2^2}{x_1^2 + x_2^2}$$

In practice, L and M are bounded below at 1 and above at some threshold value less than 2, and K is bounded below at 1 and can taper off above the threshold. Also, the target values of K, L and M depend on the situation in each cell. Here, for simplicity of illustration, they were assumed not to change when we removed observations from the cell.

Using the same noise and changing its direction for even-ranked units means that we take advantage of the correlation between the Z and Z_{-i} to increase the variance of $\hat{x}_i = Z - Z_{-i}$. For example, the contribution to $V(\hat{x}_1)$ from unit 2 is $(K+L)^2 x_2^2 \sigma_\varepsilon^2$. If we had used independent (or unit-cell specific) noises ε_i' instead of ε_i for units 1 to 4 the contribution from unit 2 would have been only $(K^2 + L^2) x_2^2 \sigma_\varepsilon^2$.

3.3 Comparison with the EZS and TCM approaches

With EZS the perturbed cell total is simply $Z = X + \sum_{i \in cell} \varepsilon_i x_i$, giving $V(Z) = \sum_{i \in cell} x_i^2 \sigma_\varepsilon^2$. For any unit i we have $E(\hat{x}_i) = x_i$ and $V(\hat{x}_i) = x_i^2 \sigma_\varepsilon^2$, which is smaller than the equivalent variance with the LPM for the same level of noise σ_ε^2 even when we set $K = L = M = 1$. A possible exception could be unit 5, if subsequent units are relatively quite small. This can be seen by examining $V(\hat{x}_5)$ above.

The TCM applies three multiplicative perturbation factors to the largest, say 4, units in each cell. A magnitude component M_i determines the relative size of the perturbation for the i^{th} ranked unit. The M_i are fixed; typically $M_1 > M_2 > M_3 > M_4$, e.g., $[0.6,\ 0.4,\ 0.3,\ 0.2]$. A permanent random factor $d_i = \pm 1$ fixes the direction of the noise for each unit i. A pseudo-random factor $s_i > 0$ determines unit-cell specific noises. This gives $Z = X + \sum_{i=1}^{4} M_i d_i s_i x_i$. The method can be represented in a form comparable to LPM, with $[M_1,\ M_2,\ M_3,\ M_4] = [K,\ L,\ M,\ 1]$, $d_i = \text{sign}(\varepsilon_i)$ and $s_i = |\varepsilon_i'|$. The way the d_i are fixed is a major difference with the LPM that greatly diminishes the protection offered to $\hat{x}_1$. To illustrate this, consider two adaptions of these methods that yield identical variances for Z:

$$Z_{LPM} = X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4, \quad \text{and}$$

$$Z_{TCM} = X + K\,sign(\varepsilon_1)\,|\varepsilon_1'|\,x_1 + L\,sign(\varepsilon_2)\,|\varepsilon_2'|\,x_2 + M\,sign(\varepsilon_3)\,|\varepsilon_3'|\,x_3 + sign(\varepsilon_4)\,|\varepsilon_4'|\,x_4,$$

where the same notational conventions as before are used, with fixed $K, L, M > 0$. This yields

$$V_{LPM}(\hat{x}_1) = \left\{K^2 x_1^2 + (K+L)^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + x_5^2\right\}\sigma_\varepsilon^2, \quad \text{and}$$

$$V_{TCM}(\hat{x}_1) = K^2 x_1^2 \sigma_\varepsilon^2 + \left\{(K^2 + L^2) x_2^2 + (L^2 + M^2) x_3^2 + (M^2 + 1) x_4^2\right\}\sigma_{|\varepsilon|}^2 + x_5^2 \sigma_\varepsilon^2.$$

Not only are factors such as $(K+L)^2$ larger than $(K^2 + L^2)$, but the variance for the noise, σ_ε^2, is often replaced with that of the absolute noise, $\sigma_{|\varepsilon|}^2$, which is much smaller. For the split triangular distribution it goes from $(3a^2 + 2ab + b^2)/6$ to $(b - a)^2/18$. When $b = 2a$ this means dropping from $11a^2/6$ to $a^2/18$.

This is not a legitimate comparison of the two methods. We are not using the actual LPM, and method parameters need not be identical. But it shows the impact of the different approaches taken for the d_i.

4 Empirical investigation

We applied the LPM and EZS methods to personal data from a taxation file. Two variables were used: $x = \text{income}\ (\text{if} > 0)$ and $y = x^2$ (to increase skewness). Cells of between 15 and 148 units were generated by combining age groups within postal code, sex and marital status. Different levels of noise (ε_i) from a split triangular distribution were tried. Results presented are those with $\sigma_\varepsilon^2 = 0.006$. Following the Risk-Utility Framework (Duncan, Keller-McNulty and Stokes 2001) the impacts of the methods on data accuracy and on risk were examined.

Table 4.1 shows the impact of the LPM on the quality of cell totals by cell size range. The LPM was applied 500 times in each cell. For each cell size range the table gives the number of cells, their average coefficient of variation (CV) after perturbation, and the percentage of times that the perturbed total was within 2%, 5%, 8% and 12% of the original cell total. For this study we assumed that cells that failed a P–percent sensitivity rule with $P = 15$ would be suppressed – so they were not included in the results. There were more such cells with variable y (which may resemble business data more). As expected, the impact of the perturbation was higher for smaller cells, and for variable y. All cells perturbed by more than 8% were near-sensitive and would have been suppressed with $P = 20$.

Table 4.1
Impact of layered perturbation method on cell totals

	Variable = Income (x)						Variable = Income² (y)					
Cell	Num.	Avg.	% times relative distance ≤				Num.	Avg.	% times relative distance ≤			
size	cells	CV	2%	5%	8%	12%	cells	CV	2%	5%	8%	12%
15 – 18	1,822	2.37	58.5	95.1	99.5	100.0	1,777	4.09	34.5	72.0	92.4	99.6
19 – 25	2,230	2.03	66.2	97.2	99.7	100.0	2,185	3.71	38.1	77.1	94.4	99.7
26 – 40	1,920	1.57	78.2	99.1	99.9	100.0	1,899	3.24	44.2	82.8	96.0	99.8
41 – 148	1,312	1.05	92.1	99.5	99.9	100.0	1,301	2.53	57.1	90.0	97.7	99.9
All	7,284	1.82	72.1	97.6	99.7	100.0	7,162	3.47	42.3	79.7	94.9	99.7

Note: values of 100.0 represent values above 99.95 that were rounded to 100.

Table 4.2 gives the impact of the EZS multiplicative noise, for the same σ_ε^2, on the cell totals. Results for income (x) are fairly similar, while results for y are noticeably better. Similar results were obtained when a value for σ_ε^2 near 0.014 was used (LPM was slightly better with x, EZS slightly better with y).

Table 4.2
Impact of EZS multiplicative noise on cell totals

	Variable = Income (x)						Variable = Income² (y)					
Cell	Num.	Avg.	% times relative distance ≤				Num.	Avg.	% times relative distance ≤			
size	cells	CV	2%	5%	8%	12%	cells	CV	2%	5%	8%	12%
15 – 18	1,822	2.33	58.7	97.1	100.0	100.0	1,777	3.19	41.2	86.4	99.8	100.0
19 – 25	2,230	2.08	64.5	98.5	100.0	100.0	2,185	2.93	45.2	90.0	99.9	100.0
26 – 40	1,920	1.74	73.9	99.6	100.0	100.0	1,899	2.59	51.4	93.8	99.9	100.0
41 – 148	1,312	1.30	86.9	99.9	99.9	100.0	1,301	2.09	63.4	97.1	100.0	100.0
All	7,824	1.91	69.6	98.7	99.9	100.0	7,162	2.76	49.2	91.4	99.9	100.0

Note: values of 100.0 in the 8% columns represent values above 99.95 that were rounded to 100.

We next examined the amount of protection offered to the largest units in each cell. For each cell, an estimate $\hat{x}_i$ for unit x_i was obtained by differencing perturbed cell totals with and without the unit. Relative differences $d_i = 100\left|\hat{x}_i - x_i\right| / x_i$ were calculated and incorporated in a score equal to $\sum_{cells} r_i$, where $r_i = 1$ if $d_i < 10$, $r_i = 0$ if $d_i > 15$ and $0 < r_i < 1$ otherwise. Table 4.3 shows the quartiles of d_i and the scores for variables x and y for the largest twelve units in each cell with LPM, and for the largest unit with EZS (EZS offers the same level of protection to all units).

With the LPM the largest three units tend to be protected the most, as expected. Patterns for variables x and y are different. If one looks at the quartiles of d_i for variable x, the level of protection gradually declines until unit 10 and increases afterwards. Since the $\mathrm{V}(\hat{x}_i)$ are the same for $i > 9$ results should keep improving after the 10th largest unit. The scores give a similar story. For variable y the descent is not as regular, with unit 5 being protected the least (unit 10 if one looks at Q1 only). The weaker protection around units 5 and 10 is predicted by the formulas for $\mathrm{V}(\hat{x}_i)$, whose basic form changes around those two units. Unit 10 is vulnerable to repeated targeted attacks the most, where an attack consists of obtaining an estimate $\hat{x}_{10}$ from totals for units 1 to 10, and for units 1 to 9, with some set of smaller units (e.g., obtain $\hat{x}_{10(i)}$ from totals excluding unit i and excluding units i and 10, for $i = 11, 12, 13\ldots$). Averaging the $\hat{x}_{10(i)}$, if there are enough of them, may give good estimates of x_{10}. Such attacks require carefully set up tabulation requests, which a semi-controlled custom tabulation environment could discourage.

Table 4.3
Protection of largest twelve units with LPM and of largest with EZS (quartiles for d_i)

	Variable = Income (x)					Variable = Income² (y)				
	Cells	**Q1**	**Med**	**Q3**	**Score (%)**	**Cells**	**Q1**	**Med**	**Q3**	**Score (%)**
Unit 1	7,962	7.9	15.7	26.6	3,196 (40)	7,823	7.6	14.4	23.2	3,365 (43)
Unit 2	7,962	8.6	17.5	29.3	2,895 (36)	7,782	7.2	15.0	25.2	3,311 (43)
Unit 3	7,962	8.1	16.9	28.7	3,021 (38)	7,782	6.6	14.1	24.2	3,522 (45)
Unit 4	7,962	7.2	15.5	26.2	3,314 (42)	7,799	6.1	13.3	22.5	3,726 (48)
Unit 5	7,962	6.4	13.9	23.8	3,647 (46)	7,808	5.5	11.9	20.5	4,052 (52)
Unit 6	7,962	6.4	13.9	23.3	3,614 (45)	7,811	6.0	12.6	21.6	3,885 (50)
Unit 7	7,962	6.2	13.3	22.4	3,765 (47)	7,814	6.0	12.6	22.2	3,868 (50)
Unit 8	7,962	6.3	13.4	22.3	3,731 (47)	7,818	6.5	13.8	23.7	3,581 (46)
Unit 9	7,962	5.1	11.5	19.9	4,267 (54)	7,818	5.7	13.0	24.2	3,750 (48)
Unit 10	7,962	3.3	10.7	20.9	4,373 (55)	7,818	4.4	13.5	27.4	3,704 (47)
Unit 11	7,962	3.8	11.8	22.4	4,121 (52)	7,818	4.8	15.7	32.1	3,422 (44)
Unit 12	7,962	3.8	12.2	24.7	4,031 (51)	7,820	5.8	17.9	37.9	3,110 (40)
U1/EZS	7,962	6.7	7.5	8.4	7,941 (100)	7,823	6.7	7.5	8.5	7,803 (100)

In contrast, results for EZS show that the level of protection offered to unit 1 (or for any unit for that matter) is fairly constant, and it is generally much poorer than that with the LPM. The score for EZS is almost 100%, a very poor outcome. But EZS was designed to offer protection for totals, not to protect from differencing. If protection from differencing is required then the level of noise would have to be set much higher to protect values at levels comparable to the LPM. But with EZS units around unit 10 would not be more vulnerable to repeated targeted attacks.

To investigate the roles of K, L and M we generated random values from a uniform distribution, but created an outlier in each cell by setting the value of x_1 as the highest value that would not make the cell sensitive, i.e., for $P = 15$, set $x_1 = \frac{100}{15}\sum_{i\geq 3} x_i$. The LPM was used with M set to 1, and K and L either calculated as suggested above or set to 1. For our generated data the calculated value of L never left 1. Table 4.4 shows that factor K is useful because when it is set to 1 the level of protection for the outlier is not high enough when $\sigma_\varepsilon^2 = 0.006$.

Table 4.4
Protection of outliers in artificial populations for 1,000 cells (quartiles for d_1)

Standard LPM (K ≥ 1)				LPM with $K = L = M = 1$			
Q1	**Med.**	**Q3**	**Score**	**Q1**	**Med.**	**Q3**	**Score**
11.1	12.6	14.2	472	6.7	7.5	8.6	996

5 Discussion and challenges

We presented a perturbative method for protecting tables of magnitude in a custom tabulation environment. The method is not resource intensive – it is only necessary to keep track of the largest units in each cell and their permanent random number. We have shown that the method is able to protect the largest units from a differencing attack.

Since perturbation is applied to the largest values, and sensitive cells are suppressed, there is less need to use variable-specific noise to protect ratios. Ratios can be calculated using perturbed values (Z). Likewise, means can be calculated using the Z values and perturbed (e.g., rounded) frequencies. Alternatively, if users prefer, means can be calculated by dividing Z by the true frequencies, and totals obtained by multiplying the perturbed means by perturbed frequencies.

Zeroes are not treated, but X (and Z) are suppressed for sensitive and small cells. If a non sensitive cell has less than 5 nonzero values then the addition of another zero-valued unit will not affect Z. So, in that particular situation, users may be able to tell if a unit added to the cell was zero-valued. If unit values x_i can be negative the largest absolute values $|x_i|$ in each cell could be treated (perturbed). Dominance rules would need to be adapted for negative values (e.g., see Tambay and Fillion 2013).

Residual disclosure issues with related outputs such as unperturbed totals and tables of distributions remain. If the Agency released some unperturbed totals, a hacker could try differencing attacks with the unperturbed total as the starting point. It would be preferable to keep unperturbed results to a minimum, e.g., only for official releases. Tables of distribution (e.g., total income by income range) may also present problems of residual disclosure because of the information conveyed by the ranges. One approach would be to severely restrict the ranges that can be used in such tables.

Table additivity is not maintained, and suppressed cells complicate the use of raking to restore additivity. One solution would consist of imputing those cells, raking, then suppressing the imputed cells. We could start by imputing lone suppressions in a row or column based on other cell values (bottom code at 0 if needed) and repeat this if it generated new lone suppressions in a row or column. Other methods can be used to impute values for remaining suppressed cells.

References

Cox, L.H., and Dandekar, R.A. (2004). A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. *Proceedings of the 2002 FCSM Statistical Policy Seminar*, Statistical Policy Working Paper 35, Federal Committee on Statistical Methodology, Washington, DC.

Cox, L.H., and Sande, G. (1979). Techniques for preserving statistical confidentiality. *Proceedings of the 42nd Session of the International Statistical Institute*, Manila, Philippines.

Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). *Disclosure Risk vs. Data Utility: The r-u Confidentiality Map*. Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences group, Los Alamos, New Mexico.

Evans, T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics*, 14, 537-551.

Giessing, S. (2011). Post-tabular stochastic noise to protect skewed business data. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Tarragona, Spain, October 26-28, 2011.

Massell, P., and Funk, J. (2007). Recent developments in the use of noise for protecting magnitude data tables: Balancing to improve data quality and rounding that preserves protection. *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia.

Tambay, J.-L., and Fillion, J.-M. (2013). Strategies for processing tabular data using the G-Confid cell suppression software. *Proceedings of the Survey Research Methods Section*, American Statistical Association Joint Statistical Meetings, Montreal, August 3-8, 2013.

Thompson, G., Broadfoot, S. and Elazar, D. (2013). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Ottawa, October 28-30, 2013.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 41-67
Statistics Canada, Catalogue No. 12-001-X

State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared errors estimation

Oksana Bollineni-Balabay, Jan van den Brakel and Franz Palm[1]

Abstract

Structural time series models are a powerful technique for variance reduction in the framework of small area estimation (SAE) based on repeatedly conducted surveys. Statistics Netherlands implemented a structural time series model to produce monthly figures about the labour force with the Dutch Labour Force Survey (DLFS). Such models, however, contain unknown hyperparameters that have to be estimated before the Kalman filter can be launched to estimate state variables of the model. This paper describes a simulation aimed at studying the properties of hyperparameter estimators in the model. Simulating distributions of the hyperparameter estimators under different model specifications complements standard model diagnostics for state space models. Uncertainty around the model hyperparameters is another major issue. To account for hyperparameter uncertainty in the mean squared errors (MSE) estimates of the DLFS, several estimation approaches known in the literature are considered in a simulation. Apart from the MSE bias comparison, this paper also provides insight into the variances and MSEs of the MSE estimators considered.

Key Words: Bootstrap; Hyperparameter; State space model; True MSE; Unemployment.

1 Introduction

Figures on the labour force produced by national statistical institutes (NSIs) are generally based on Labour Force Surveys (LFS). There is an increasing interest to producing these indicators at a monthly frequency (EUROSTAT 2015). Sample sizes are, however, hardly ever large enough even at the national level for producing sufficiently precise monthly labour force figures based on design-based estimators known from classical sampling theory (Särndal, Swensson and Wretman 1992; Cochran 1977). In such situations, small area estimation (SAE) techniques can be used to improve the effective sample size of domains by borrowing information from preceding periods or other domains, see Rao and Molina (2015) and Pfeffermann (2013). Repeated surveys in particular have a potential for improvement within the framework of structural time series (STS) or multilevel time series models.

STS models, as well as multilevel models, usually contain unknown hyperparameters that have to be estimated. If this uncertainty (here and further in this paper referred to as hyperparameter uncertainty) is not taken into account, estimated mean squared errors (MSEs) of the domain predictors become negatively biased. Within the framework of multilevel models, accounting for hyperparameter uncertainty is a necessary and common practice. It is routinely performed when those models are estimated with the empirical best linear unbiased predictor (EBLUP) or the hierarchical Bayesian (HB) approach, see Rao and Molina (2015), Chapter 6-7, 10. STS models, in turn, are not as widely used in SAE as multilevel models. The Kalman filter, usually applied to fit STS models, ignores the hyperparameter uncertainty, and therefore produces negatively-biased MSE estimates. Applications that give evidence for substantial advantages of STS models over the design-based approach treat the estimated model hyperparameters as known, see, e.g.,

1. Oksana Bollineni-Balabay, Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4481, 6401CZ Heerlen, the Netherlands. E-mail: oksana-bl@yandex.ru, obay@cbs.nl; Jan van den Brakel, Statistics Netherlands and Maastricht University School of Business and Economics, PO Box 616, 6200 MD Maastricht, the Netherlands; Franz Palm, Maastricht University, The Netherlands.

Bollineni-Balabay, van den Brakel and Palm (2016a), Krieg and van den Brakel (2012). Pfeffermann and Rubin-Bleuer (1993), Tiller (1992).

At Statistics Netherlands, a multivariate STS model, proposed by Pfeffermann (1991), is used to produce official monthly labour force figures for the DLFS. The DLFS is, as in many other countries, based on a rotating panel and features insufficiently large sample sizes for production of monthly figures. The STS model applied to the design-based estimates uses sample information from preceding time periods and accounts for the so-called rotation group bias (RGB) and for autocorrelation in the survey errors. In this way, sufficiently precise monthly estimates of the unemployed labour force are obtained (see van den Brakel and Krieg 2015). STS models are also applied in the production of official statistics at the US Bureau of Labor Statistics, Tiller (1992). Interest to this technique has been growing among several other NSIs spread around the world, for example at NSIs of Australia (Zhang and Honchar 2016), Israel and the UK (ONS 2015).

This paper presents an extended Monte-Carlo simulation study, where the DLFS model acts as the data generation process. Such a simulation is an insightful step into the process of model selection before implementation in the production of official statistics. First, evaluating distributions of the hyperparameter estimators under different model specifications provides additional insight into the importance of retaining certain hyperparameters in the model. Standard model diagnostics for state space models provide limited information on irrelevant hyperparameters. In case of model overspecification, not only may the distribution of redundant hyperparameter estimates largely deviate from normality, but estimation of other hyperparameters may also be disturbed. Therefore, even if the model diagnostics is satisfactory, it may still be wise to simulate the model and to examine the distribution of the maximum likelihood (ML) estimator of the model's hyperparameters.

Another aim of the simulation is evaluating to which extent uncertainty around the hyperparameter estimates affects estimation of the STS model-based MSEs. Ignoring the hyperparameter uncertainty in MSE estimation is only acceptable if the available time series are sufficiently long. Depending upon a particular application, the length of time required to be "sufficiently long" will vary. Most often, uninterrupted time series available at NSIs are relatively short, mainly due to survey redesigns. The literature offers several ways to account for the hyperparameter uncertainty in STS models: asymptotic approximation, bootstrapping and the full Bayesian approach (for the latter approach, see Durbin and Koopman (2012), Chapter 13). Among those approaches considered in this paper are the asymptotic approximation developed by Hamilton (1986), as well as parametric and non-parametric bootstrapping approaches developed by Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012). These methods are applied to the DLFS model to find the best MSE estimation method in this real life application. This paper also illustrates how the hyperparameter uncertainty problem decays as the DLFS time series increase from 48 to 200 months.

The contribution of the paper is four-fold. First of all, it shows how the Monte Carlo simulation can be used to check for model overspecification (i.e., for redundant hyperparameters). Secondly, it suggests the best of the proposed approaches to MSE estimation for the DLFS and offers a more realistic evaluation of the variance reduction obtained with the STS model compared to the design-based approach. Thirdly, this

Monte-Carlo study refutes the claim of Rodriguez and Ruiz (2012) about the superiority of their method over the bootstrap of Pfeffermann and Tiller (2005) in a more complex model. Finally, apart from MSE bias comparison, this paper also provides insight into the variance and MSEs of these MSE estimators. To the best of our knowledge, the variability of the above-mentioned bootstrap methods has not been studied yet.

The paper is structured as follows. Section 2 contains a description of the DLFS and the model currently used by Statistics Netherlands. Section 3 reviews the above-mentioned approaches to the MSE estimation. Details on the simulation setup specific to the DLFS are given in Section 4. Results are presented in Section 5. Section 6 contains concluding remarks.

2 The Dutch Labour Force Survey

2.1 The DLFS design

The DLFS has been based on a rotating panel design since October 1999. Every month, a sample of addresses is drawn according to a stratified two-stage sample design. Strata are formed by geographical regions; municipalities are the primary sampling units and addresses are the secondary sampling units. All households residing on one address are included in the sample. In this paper, the DLFS data observed from January 2001 until June 2010 are considered. During this period, data in the first wave were collected by means of computer assisted personal interviewing (CAPI) by interviewers that visit sampled households at home. After a maximum of six attempts, an interviewer leaves a letter with the request to contact the interviewer by telephone to make an appointment for an interview. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Respondents are re-interviewed four times at quarterly intervals. In these four subsequent waves, data are collected by means of computer assisted telephone interviewing (CATI). During these re-interviews, a condensed questionnaire is applied to establish any changes in the labour market position of the respondents. Proxy interviewing is also allowed during these re-interviews. Mobile phone numbers and secret land line numbers are collected in the first wave to avoid panel attrition. With the commencement of the rotating panel design for the DLFS, the gross sample size was about 6,200 addresses per month on average, with about 65% completely responding households. The response rates in the follow-up waves are about 90% compared to the preceding wave.

The general regression (GREG) estimator (Särndal et al. 1992) is applied to estimate the total unemployed labour force. This estimator accounts for the complexity of the sample design and uses auxiliary information available from registers to correct, at least partially, for selective non-response. Let Y_t^j denote the GREG estimate of the total number of unemployed in month t based on the j^{th} wave of respondents. Five such estimates are obtained per month, each of them being respectively based on the sample that entered the survey in month $t-l,\ \ l=\{0, 3, 6, 9, 12\}$. The GREG estimator for this population total is defined as:

$$Y_t^j = \sum_{k \in s} w_{k,t} \left(\sum_{i=1}^{n_{k,t}} y_{i,k,t} \right) \tag{2.1}$$

with $y_{i,k,t}$ representing the sample observations that are equal to 1 if the i^{th} person in the k^{th} household is unemployed, and zero otherwise; $n_{k,t}$ is the number of persons aged 15 or above in the k^{th} household; $w_{k,t}$ are the regression weights for household k at time t. The method of Lemaître and Dufour (1987) is used to obtain equal weights for all persons within the same household:

$$w_{k,t} = \frac{1}{\pi_{k,t}}\left[1+\left(\mathbf{X}_t - \sum_{k\in s}\frac{\mathbf{x}_{k,t}}{\pi_{k,t}}\right)\left(\sum_{k\in s}\frac{\mathbf{x}_{k,t}\mathbf{x}'_{k,t}}{\pi_{k,t}g_{k,t}}\right)^{-1}\frac{\mathbf{x}_{k,t}}{g_{k,t}}\right], \tag{2.2}$$

where $\pi_{k,t}$ is the inclusion probability of household k at time t, $g_{k,t}$ is the size of household k at time t; $\mathbf{x}_{k,t} = \sum_{i=1}^{n_{k,t}}\mathbf{x}_{i,k,t}$, with $\mathbf{x}_{i,k,t}$ being a J – dimensional vector with the weighting model auxiliary information on the i^{th} person in the k^{th} household at time t. Vector $\mathbf{X}_t$ contains population totals of auxiliary variables. The weighting model is defined by the following variables (with the number of categories in brackets): Age(5)Gender + Geographic Region(44) + Gender(2) × Age(21) + Age(5) × Marital Status(2) + Ethnicity(8), where × stands for interaction of variables, and Age(5)Gender is a variable classified into eight classes where Age has five categories, with the second, third and fourth age category being itemized into two genders.

The variance of the GREG estimator $Y_t^{\,j}$ is approximated by:

$$\widehat{\text{Var}}\left(Y_t^{\,j}\right) = \sum_{h=1}^{H}\frac{n_{h,t}}{n_{h,t}-1}\left(\sum_{k=1}^{n_{h,t}}\left(w_{k,t}\hat{e}_{k,t}\right)^2 - \frac{1}{n_{h,t}}\left(\sum_{k=1}^{n_{h,t}}w_{k,t}\hat{e}_{k,t}\right)^2\right), \quad j = \{1, 2, 3, 4, 5\}, \tag{2.3}$$

where the GREG residuals are $\hat{e}_{k,t} = \sum_{i=1}^{n_{k,t}}\left(y_{i,k,t} - \mathbf{x}'_{i,k,t}\hat{\boldsymbol{\beta}}_t\right)$; $n_{h,t}$ is the number of households in stratum h (with H being the total number of strata); vector $\hat{\boldsymbol{\beta}}_t$ is a Horvitz-Thompson type estimator for the regression coefficient that is obtained from regressing the target variable on the auxiliary variables from the sample.

2.2 The STS model for the DLFS

There are two reasons why Statistics Netherlands took a decision to switch to a time series model-based production approach in June 2010. One reason for that was inadequately small sample sizes for production of monthly estimates. With a net sample size of about 4,000 households in the first wave on average, the GREG estimates of the unemployed labour force had a coefficient of variation of about 4% at the national level, which was considered to be too volatile for official statistical publications. In addition to that, monthly unemployment figures must be published for six domains based on a classification of gender and age. The design-based estimates of these domains feature much higher coefficients of variation. Another problem with the DLFS is the so-called RGB, which refers to systematic differences between the estimates of different waves (see, e.g., Bailar 1975 or Pfeffermann 1991). Common reasons behind the RGB are panel attrition, panel-effects, and differences in questionnaires and modes used in the subsequent waves. In the case of the DLFS, the first wave estimates are assumed to be most reliable, with the subsequent waves systematically underestimating the unemployed labour force numbers. See van den Brakel and Krieg (2009) for a more detailed discussion.

Both problems are solved with an STS model, which uses five series of GREG estimates for the five different waves as input. With an STS model, an observed series is decomposed into several unobserved components, e.g., trend and seasonal. The Kalman filter, optionally in combination with a smoothing algorithm, can be applied to extract these components from the observed time series. By doing so, estimates of the components that define the signal for unemployment are separated from unexplained variance of the population parameter and from the sampling variance. This generally results in less volatile point estimates, with substantially smaller standard errors compared to those of the GREG estimates. By modelling the systematic differences between the five input series, the model also accounts for the RGB of the rotating panel.

In each month t, a five-dimensional vector $\mathbf{Y}_t = \left(Y_t^1\ Y_t^2\ Y_t^3\ Y_t^4\ Y_t^5\right)'$ is observed, containing GREG estimates of the total number of the unemployed labour force based on the five waves. Based on Pfeffermann (1991), van den Brakel and Krieg (2009) developed the following model for the GREG estimates $\mathbf{Y}_t$:

$$\mathbf{Y}_t = \mathbf{1}_5 \xi_t + \boldsymbol{\lambda}_t + \mathbf{e}_t, \tag{2.4}$$

here, $\mathbf{1}_5$ is a five-dimensional column vector of ones, ξ_t is the unknown (scalar) true population parameter, $\boldsymbol{\lambda}_t$ is a vector containing state variables for the RGB, and $\mathbf{e}_t$ is a vector of the survey errors that are correlated with their counterparts from previous waves (the structure will be presented later). For the true population parameter, it is assumed that: $\xi_t = L_t + \gamma_t + \varepsilon_t$, which is the sum of a stochastic trend L_t, a stochastic seasonal component γ_t, and an irregular component $\varepsilon_t \overset{\text{iid}}{\sim} N\left(0, \sigma_\varepsilon^2\right)$.

For the stochastic trend L_t, the so-called smooth-trend model is assumed:

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, \end{aligned}$$

where L_t and R_t represent the level and slope of the true population parameter, respectively, with the slope disturbance term being distributed as: $\eta_{R,t} \overset{\text{iid}}{\sim} N\left(0, \sigma_R^2\right)$.

For the seasonal component γ_t, the trigonometric model is assumed:

$$\gamma_t = \sum_{l=1}^{6} \gamma_{t,l},$$

where each of these six harmonics follows the process:

$$\begin{aligned} \gamma_{t,l} &= \cos\left(h_l\right)\gamma_{t-1,l} + \sin\left(h_l\right)\gamma_{t-1,l}^* + \omega_{t,l}, \\ \gamma_{t,l}^* &= -\sin\left(h_l\right)\gamma_{t-1,l} + \cos\left(h_l\right)\gamma_{t-1,l}^* + \omega_{t,l}^*, \end{aligned}$$

with $h_l = \frac{\pi l}{6}$ being the l^{th} seasonal frequency, $l = \{1, \ldots 6\}$. The zero-expectation stochastic terms $\omega_{t,l}$ and $\omega_{t,l}^*$ are assumed to be normally and independently distributed and to possess the same variance within and across all the harmonics, such that:

$$\begin{aligned} \text{Cov}\left(\omega_{t,l}, \omega_{t',l'}\right) &= \text{Cov}\left(\omega_{t,l}^*, \omega_{t',l'}^*\right) = \begin{cases} \sigma_\omega^2 & \text{if } l = l' \text{ and } t = t', \\ 0 & \text{if } l \neq l' \text{ or } t \neq t', \end{cases} \\ \text{Cov}\left(\omega_{t,l}, \omega_{t,l}^*\right) &= 0 \text{ for all } l \text{ and } t. \end{aligned}$$

The second component in (2.4) is the RGB. It is assumed that the first wave is unbiased, as motivated in van den Brakel and Krieg (2009). The RGBs for the follow-up waves are time-dependent and are modelled

as random walk processes. The rationale behind this is that field-work procedures are subject to frequent changes. Apart from that, response rates change gradually over time. This makes the RGB time-dependent, as illustrated by van den Brakel and Krieg (2015), Figure 4.3. The RGB vector for the five waves can be written in the following form: $\boldsymbol{\lambda}_t = \left(0\ \lambda_t^2\ \lambda_t^3\ \lambda_t^4\ \lambda_t^5\right)'$, with:

$$\lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,t}^j, \quad j = \{2, 3, 4, 5\}.$$

It is assumed that the RGB disturbances are not correlated across different waves and are normally distributed: $\eta_{\lambda,t}^j \overset{\text{iid}}{\sim} \left(0, \sigma_\lambda^2\right)$, with equal variances in all the four waves.

The last component in (2.4) contains the survey errors for the five GREG estimates, i.e., $\mathbf{e}_t = \left(e_t^1\ e_t^2\ e_t^3\ e_t^4\ e_t^5\right)'$. To account for sampling error heterogeneity caused by changes in the sample sizes over time, the sampling errors are modelled proportionally to the design-based standard errors according to the following measurement error model proposed by Binder and Dick (1990): $e_t^j = \tilde{e}_t^j z_t^j$, where $z_t^j = \sqrt{\widehat{\text{Var}}\left(Y_t^j\right)}$ and $\tilde{e}_t^j$ are standardised sampling errors that follow a stationary process defined later in the text. Here, $\widehat{\text{Var}}\left(Y_t^j\right)$ are the design-based variance estimates obtained from the micro data using (2.3). They are treated as a priori known sampling variances in the STS model.

Since the sample in the first wave has no overlap with samples observed in the past, $\tilde{e}_t^t$ can be modelled as a white noise with $E\left(\tilde{e}_t^1\right) = 0$ and $\text{Var}\left(\tilde{e}_t^1\right) = \sigma_{\nu_1}^2$. The variance of the survey errors e_t^t will be equal to the variance of the GREG estimates if the maximum likelihood estimate of $\sigma_{\nu_1}^2$ is approximately equal to unity.

The survey errors in the follow-up waves are correlated with the survey errors from the preceding waves. This autocorrelation coefficient is estimated from the survey data using the approach proposed by Pfeffermann, Feder and Signorelli (1998). The autocorrelation structure is modelled with an AR(1) model where the autocorrelation coefficient is obtained with the Yule-Walker equations (van den Brakel and Krieg 2009):

$$\tilde{e}_t^j = \rho \tilde{e}_{t-3}^{j-1} + \nu_t^j,\ \nu_t^j \overset{\text{iid}}{\sim} N\left(0, \sigma_{\nu_j}^2\right), \quad j = \{2, 3, 4, 5\}.$$

It is assumed that the first-order autocorrelation coefficient is common for all the four waves. Its estimate is used as a priori information in the model. Since $\tilde{e}_t^j$ is an AR(1) process, $\text{Var}\left(\tilde{e}_t^j\right) = \sigma_{\nu_j}^2 \big/ \left(1 - \rho^2\right)$. The variance of the sampling error e_t^j is approximately equal to $\widehat{\text{Var}}\left(Y_t^j\right)$ if the maximum likelihood estimate of $\sigma_{\nu_j}^2$ is approximately equal to $\left(1 - \rho^2\right)$. Five different hyperparameters $\sigma_{\nu_j}^2$, $j = \{1, 2, 3, 4, 5\}$, are assumed for the survey error components of the five waves.

The disturbance variances, together with the autocorrelation parameter ρ, are collected in a hyperparameter vector called $\boldsymbol{\theta} = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{\nu_1}^2\ \sigma_{\nu_2}^2\ \sigma_{\nu_3}^2\ \sigma_{\nu_4}^2\ \sigma_{\nu_5}^2\ \rho\right)'$, and the vector containing only the disturbance variances is called $\boldsymbol{\theta}_\sigma = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{\nu_1}^2\ \sigma_{\nu_2}^2\ \sigma_{\nu_3}^2\ \sigma_{\nu_4}^2\ \sigma_{\nu_5}^2\right)'$. To avoid negative estimates, the disturbance variance hyperparameters in $\boldsymbol{\theta}_\sigma$ are estimated on a log-scale. The quasi-maximum likelihood method is used (see e.g., Harvey 1989), where $\hat{\rho}-$estimates are treated as known. Numerical analysis of this paper is conducted with OxMetrics 5 (Doornik 2007) in combination with *SsfPack 3.0* package (Koopman, Shephard and Doornik 2008).

3 MSE estimation approaches

Linear structural time series models with unobserved components are usually fitted with the help of the Kalman filter after putting them into a state space form. See Bollineni-Balabay, van den Brakel and Palm (2016b) for the state space representation of the STS model for the DLFS. The state vector $\boldsymbol{\alpha}_t$ contains the state variables defined in the previous section, i.e., the trend, slope, seasonal harmonics, RGB, the population white noise and survey errors. All the non-stationary state variables are initialised with a diffuse prior (i.e., with a zero-mean and a very large variance). The five survey error components $\tilde{e}_t^j$, $j = \{1, 2, 3, 4, 5\}$ and the population white noise ε_t are stationary state variables that are initialised with zeros. The initial variance of the first-wave sampling errors is taken equal to unity, whereas the one of the other waves is taken equal to $(1-\rho^2)$. One could try a small value for the initial variance of ε_t.

Filtered estimates of the state vector $\boldsymbol{\alpha}_t$ and its covariance matrix $\mathbf{P}_{t|t}$ are usually extracted with the Kalman filter (see Harvey 1989). $\mathbf{P}_{t|t}$ thus contains MSEs extracted by the filter conditionally on the information up to and including time t:

$$\mathbf{P}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right], \tag{3.1}$$

where $\boldsymbol{\theta}$ is assumed to be the true hyperparameter value, and the expectation is taken with respect to the joint distribution of the state vector and the Y – values at time t. In practice, the true hyperparameter vector is replaced by its estimate $\hat{\boldsymbol{\theta}}$ in the Kalman filter recursions. Then, the MSE in (3.1) is no longer the true MSE and is called "naive" as it does not incorporate the uncertainty around the $\hat{\boldsymbol{\theta}}$ – estimates. The true MSE then becomes:

$$\mathbf{MSE}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)'\right],$$

which is larger than the MSE in (3.1) and can be decomposed as the sum of the filter uncertainty and parameter uncertainty, provided the error terms are normal:

$$\mathbf{MSE}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right] + \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)'\right]. \tag{3.2}$$

The first term, the filter uncertainty, is what is estimated by the naive $\mathrm{MSE}-$estimates $\mathbf{P}_{t|t}$ delivered by the Kalman filter. Estimation of the second term, the parameter uncertainty, requires some additional effort. The literature on MSE estimation proposes two main approaches: asymptotic approximation and bootstrapping. Bootstrapping can be performed in a parametric or non-parametric way. A few remarks have to be mentioned about these methods in the context of STS models and of the DLFS model specifically.

For the parametric bootstrap, the state disturbances, say, $\boldsymbol{\eta}_t$, are drawn from their estimated joint conditional multivariate normal density $\boldsymbol{\eta}_t \overset{\text{iid}}{\sim} \mathrm{MN}\left(\mathbf{0}, \hat{\boldsymbol{\Omega}}\right), \hat{\boldsymbol{\Omega}}$ being evaluated at the hyperparameter estimate $\hat{\boldsymbol{\theta}}$, and are used in the Kalman filter state recursions to generate the state variables. Non-parametric bootstrap, in turn, has an advantage of not depending on any particular assumption about this joint distribution. Unlike in the parametric case where state disturbances are drawn from their estimated distribution, in the non-parametric case, standardised innovations are resampled with replacement from the standardized innovations that are based on the original hyperparameter estimates. The resampled

standardized innovations are further used to generate bootstrap series $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ by running the so-called innovation form of the Kalman filter, see Harvey (1989) or Bollineni-Balabay et al. (2016b) for details. In the DLFS model, the first 13 time points of standardised innovations are not subject to resampling, as they constitute the so-called diffuse sample (this is the time needed to construct a proper distribution for the non-stationary state variables; see Koopman (1997) for initialisation of non-stationary state variables).

If an STS model contains non-stationary components, as is the case with the DLFS model, the generated series are likely to diverge from the original dataset they have been bootstrapped from, i.e., from $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$. Therefore, a special procedure is required for bootstrap samples to be brought in accordance with the pattern of the given dataset. This can be done with the help of the simulation smoother algorithm developed by Durbin and Koopman (2002). Technical details for implementation can be found in Koopman et al. (2008), Chapter 8.4.2. The survey errors, generated as described in either parametric or non-parametric unconditional state recursion, do not need any adjustments as they constitute (autocorrelated) noise.

The following sections contain a brief presentation of the asymptotic approach, as well as of the recent Rodriguez and Ruiz (2012) bootstrap approaches (hereafter referred to as the Rodriguez and Ruiz (RR) bootstrap) and of Pfeffermann and Tiller (2005) (hereafter the Pfeffermann and Tiller (PT) bootstrap) bootstrap approaches.

3.1 Rodriguez and Ruiz bootstrapping approach

Rodriguez and Ruiz (2012) developed their bootstrap method for MSE estimation conditional on the data, which means that bootstrap hyperparameters are further applied to the original data series for obtaining bootstrap estimates of the state variables. Bootstrapping can be done both parametrically and non-parametrically, following the steps below:

1. Estimate the model and obtain the hyperparameter estimates $\hat{\boldsymbol{\theta}}$.
2. Generate a bootstrap sample $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ using $\hat{\boldsymbol{\theta}}$, either parametrically or non-parametrically, as described in the introduction to this section. If the model is non-stationary, the bootstrap sample has to be corrected with the help of the simulation smoother.
3. The bootstrap dataset $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ is used to obtain both the survey error autocorrelation parameter estimates $\hat{\rho}^b$ and bootstrap ML estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^b$. Thereafter, the Kalman filter is launched using the original series $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ and the newly-estimated $\hat{\boldsymbol{\theta}}^b$, which produces $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ and $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.
4. Steps 2-3 are repeated B times. Then, the MSE are estimated in the following way:

$$\widehat{\mathbf{MSE}}_{t|t}^{\mathrm{RR}} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]', \tag{3.3}$$

where $\overline{\boldsymbol{\alpha}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.

Equation (3.3) is applied both for the parametric and non-parametric bootstrap $\mathrm{MSE}-$estimators (denoted hereafter as $\mathrm{MSE}^{\mathrm{RR1}}$ and $\mathrm{MSE}^{\mathrm{RR2}}$, respectively).

3.2 Pfeffermann and Tiller bootstrapping approach

The bootstrap developed by Pfeffermann and Tiller (2005) is an unconditional bootstrap. This implies that bootstrap state variables are derived from the bootstrap dataset $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$, i.e., not from the original data $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ as in Rodriguez and Ruiz (2012). Pfeffermann and Tiller (2005) prove that they approximate the true MSE up to the order of $O(1/T^2)$ (Pfeffermann and Tiller 2005, Appendix C):

$$\widehat{\mathbf{MSE}}_{t|t}^{\mathrm{PT}} = 2\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'. \quad (3.4)$$

Equation (3.4) is applied both for the parametric and non-parametric bootstrap MSE – estimators (denoted further as $\mathrm{MSE}^{\mathrm{PT1}}$ and $\mathrm{MSE}^{\mathrm{PT2}}$, respectively). MSE – calculation in (3.4) requires two Kalman filter runs for every bootstrap series. In the first run, $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ is estimated from the bootstrap data set $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ and the bootstrap parameters $\hat{\boldsymbol{\theta}}^b$. In this run, $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ can also be obtained based on $\hat{\boldsymbol{\theta}}^b$, since matrix $\mathbf{P}_{t|t}$ does not depend on the data. The second Kalman filter run is needed to produce the state estimates $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ based on $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ and $\hat{\boldsymbol{\theta}}$ – estimates that were obtained from the original dataset. The bootstrap procedure is summarized below:

1. Estimate the model using the original dataset and obtain the hyperparameter vector estimates $\hat{\boldsymbol{\theta}}$. Apart from that, save the "naive" MSE estimates $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ for future use in (3.4).
2. Use the parametric or non-parametric method to generate a bootstrap sample $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$. Apply the simulation smoother correction to it if the model is non-stationary.
3. Estimate bootstrap hyperparameter estimates $\hat{\boldsymbol{\theta}}^b$ from the newly generated bootstrap dataset. Run the Kalman filter once to get $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ and $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$, and another time to obtain $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$, as described under (3.4).
4. Repeat steps 2-3 B times. Then, estimate the MSE using (3.4).

Pfeffermann and Tiller (2005) note that, in the case of the parametric bootstrap, the second Kalman filter run can be avoided because the true state vector is generated (and thus known) for every bootstrap series. Thus, the state estimates $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ in (3.4) can be replaced by the true vector $\boldsymbol{\alpha}_t^b$ to obtain the following MSE estimator:

$$\widehat{\mathbf{MSE}}_{t|t}^{\mathrm{PT1}} = \mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'. \quad (3.5)$$

There is only one $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ in the right-hand side of (3.5). This is due to the fact that the new term $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'$, corresponding to the last term on the right-hand side of (3.5), can itself be decomposed, in the same fashion as in (3.2), into the measure of parameter uncertainty $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'$ and the filter uncertainty term $\mathbf{P}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) = \mathrm{E}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$, $\hat{\boldsymbol{\theta}}$ being the true parameter vector the bootstrap state variables $\boldsymbol{\alpha}_t^b$ are generated with. However, the bootstrap average term $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$ replacing $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ may need much more bootstrap iterations to converge. Further, this simplified method may result in an additional bias if the normality assumption about the model error terms is violated. Then,

the decomposition of the term $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}^b_t\right]'$ according to (3.2) will also contain a non-zero cross-term: $\mathrm{E}\left\{\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\right\}$. In this application, the non-zero cross-term bootstrap averages have turned out to be negligible, but the bootstrap average $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}^b_t\right]'$ exhibited large departures (in both directions) from the term it was meant to replace. This may be explained by the fact that the true Kalman filter MSE in (3.1) can be obtained from simulated series if the distribution of the state-vector is sufficiently dispersed. When bootstrapping non-stationary models, however, the bootstrap series are forced to follow the pattern of the underlying original series, as it has been mentioned in the description of the simulation smoother algorithm. Therefore, the term $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}^b_t\right]'$ that replaces $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ in (3.5) may not be sufficiently close to it. For this reason, both parametric (denoted as PT1) and non-parametric (PT2) bootstraps in this application rely on the estimator in (3.4).

A few words have to be said about the role of the simulation smoother of Durbin and Koopman (2002), mentioned at the end of the introduction to this section. We suggest that it should be used at the bootstrap series generation step. Without it, the bootstrap hyperparameter distribution obtained from uncorrected series for a non-stationary model could be very different from what it should be for a particular realisation of the data at hand. At least in the case of the DLFS, omitting the simulation smoother step resulted in bootstrap hyperparameter distributions having a much wider range than the range of distributions obtained with the simulation smoother. Moreover, such bootstrap hyperparameter distributions obtained from uncorrected series in the DLFS are centred around values that are much larger than the hyperparameter values the series have been generated with. This results in an excessively large bootstrap average $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ (relatively to $\mathbf{P}_{t|t}(\hat{\boldsymbol{\theta}})$) and, subsequently, in $\mathrm{MSE}-$estimates that are even lower than the naive ones. The term $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]'$ also becomes very unstable over time and excessively large compared to when the simulation smoother is used, but that does not compensate for the negative bias obtained from (3.4) without the simulation smoother.

3.3 Asymptotic approximation

An asymptotic approximation (AA) to the true MSE in equation (3.2) was developed by Hamilton (1986) and can be expressed as an expectation over the hyperparameter joint asymptotic distribution $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, conditional on the given dataset $\mathbf{Y} \equiv \{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$. In the present application, the part of the hyperparameter vector estimated by the $\mathrm{ML}-$method, $\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\right)$, depends on the estimated value of the autoregressive parameter $\hat{\rho}$. Therefore, the joint asymptotic distribution of the hyperparameter estimator has the following form: $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right) = \pi\left(\hat{\rho}\middle|\mathbf{Y}\right)\pi\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\middle|\hat{\rho},\mathbf{Y}\right)$. The MSE is approximated as follows:

$$\mathbf{MSE}_{t|t} = \mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)\right] + \mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right) - \hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right) - \hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)'\right], \qquad (3.6)$$

where $\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}$ is an expectation taken over the hyperparameter estimator joint asymptotic distribution $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, and $\hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)$ are the state vector estimates when the hyperparameters are not known (i.e., $\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}[\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}},\mathbf{Y})]$).

Distribution $N(\hat{\rho}, \mathrm{Var}(\hat{\rho}))$ is chosen as $\hat{\rho}$'s asymptotic distribution $\pi(\hat{\rho}|\mathbf{Y})$, from which random $\hat{\rho}-$realisations are drawn. Generally, the sampling distribution of the correlation coefficient has a complex form, but it may be well approximated by a normal distribution, which was the case in this application (the normal distribution fitted both the simulated and the bootstrap distribution of $\hat{\rho}$ very well). From equation (3) in Bartlett (1946), and using the fact that the autoregressive coefficient in an AR(1) process is equal to the correlation for lag 1, the variance estimator of $\hat{\rho}$ becomes: $\mathrm{Var}(\hat{\rho}) \approx (1-\hat{\rho}^2)/T$. In the case of the DLFS, where $\hat{\rho} = 0.208$, this means that $\widehat{\mathrm{Var}}(\hat{\rho}) \approx 0.96(1/T)$. Taking into account the fact that $\hat{\rho}$'s standard error is used for making draws from the asymptotic distribution, and that the square root is a concave function, the sample standard deviation would be an underestimate. Therefore, making $\hat{\rho}-$draws by means of $1/\sqrt{T}$ as the asymptotic distribution's standard deviation would be a reasonable choice.

A sample of B draws from the hyperparameter asymptotic distribution is obtained in the following way. After a value, say, $\hat{\rho}^a$, is drawn from $\pi(\hat{\rho}|\mathbf{Y})$, the other hyperparameters are re-estimated from the original data to obtain $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{ML}}\left|\hat{\rho}^a, \mathbf{Y}\right.$ and the information matrix $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{ML}}\left|\hat{\rho}^a, \mathbf{Y}\right.\right)$. Finally, a $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a-$ draw is made from distribution $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{ML}}, \hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{ML}}\left|\hat{\rho}^a, \mathbf{Y}\right.\right)\right)$. The Kalman filter is run again using $\hat{\rho}^a-$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a-$ realisations to obtain the state estimates $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a, \mathbf{Y}\right)$ and their MSEs $\hat{\mathbf{P}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right)$. The procedure is repeated until B $\hat{\boldsymbol{\theta}}^a-$ draws are obtained, whereafter (3.6) is obtained by averaging the necessary quantities over B iterations. If all the hyperparameters of the model are estimated within the $\mathrm{ML}-$ procedure, B draws can be made directly from $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}^{\mathrm{ML}}, \hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}^{\mathrm{ML}}\right)\right)$.

The first term in (3.6) can be approximated by the average value of the Kalman filter variance $\mathbf{P}_{t|t}$ across B realizations of the hyperparameter vector, and the second term by the variance of the state vector estimates across the same B realisations. An asymptotic approximation for the MSE could therefore be obtained in the following way:

$$\widehat{\mathbf{MSE}}_{t|t}^{\mathrm{AA}} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right) + \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a, \mathbf{Y}\right) - \bar{\hat{\boldsymbol{\alpha}}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a, \mathbf{Y}\right) - \bar{\hat{\boldsymbol{\alpha}}}_{t|t}\right]', \tag{3.7}$$

where $\hat{\boldsymbol{\theta}}^a$ is the a^{th} draw from the $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$ asymptotic distribution. As Hamilton (1986) suggests, the sample average $\bar{\hat{\boldsymbol{\alpha}}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a, \mathbf{Y}\right)$ can replace $\hat{\boldsymbol{\alpha}}_{t|t}(\mathbf{Y})$ in (3.6). Further, he states, such a decomposition of the total uncertainty into the filter and parameter uncertainty resembles the well-known decomposition: $\mathrm{var}(X) = E[\mathrm{var}(X|Y)] + \mathrm{var}[E(X|Y)]$. Obviously, this $\mathrm{MSE}-$ estimator is entirely based on the assumption of asymptotic normality of the hyperparameter vector estimator. Apart from that, this approach usually produces significant biases if the series is not of a sufficient length, in which case the assumed asymptotic normal distribution would fail to approximate the finite (usually skewed) distribution of maximum-likelihood estimates.

Another problem with the asymptotic approach can occur if some of the hyperparameters are estimated to be close to zero. This can happen to the initial model estimates or during the procedure itself, e.g., due to certain extreme $\hat{\rho}-$ draws. In these cases, the asymptotic variance of such hyperparameters will be very large, which will inflate the $\mathrm{MSE}-$ estimates of the signal and its unobserved components. It may as well lead to a failure in inverting the information matrix for the hyperparameter vector.

4 The DLFS-specific simulation setup

The performance of the five MSE estimation methods is examined on series of the original length from the DLFS survey (114 monthly time points from 2001(1) until 2010(6)), as well as on shorter series of lengths 48 and 80 months, and on longer ones of length 200. For each of these series lengths, a Monte-Carlo experiment is set up where multiple series (1,000) are simulated on the basis of the DLFS model for the number of unemployed. MSEs for each of these series are estimated based on $B = 300$ bootstrap series; for asymptotic approximation, however, at least $B = 500$ draws turned out to be needed. This number has been found sufficient for the approximated MSEs to converge. MSEs delivered by the five methods and averaged over the 1,000 simulations are compared to $\mathrm{MSE}-$averages produced by the "naive" Kalman filter. However, for the latter MSE estimates to converge to a certain average value, at least 10,000 simulations are needed.

The above-mentioned artificial series $\mathbf{Y}_t^s$ for simulations $s = 1,\ldots,1,000$ (or 10,000) are generated parametrically in the following way. First, the hyperparameter ML estimates $\hat{\boldsymbol{\theta}}_{\sigma}$ are obtained from fitting the STS model to the original series. Thereafter, state disturbances (recall that survey errors are also modelled as state variables) are randomly drawn from their joint normal distribution $N\left(\mathbf{0}, \boldsymbol{\Omega}\left(\hat{\boldsymbol{\theta}}_{\sigma}\right)\right)$, and series are generated using the Kalman filter recursion. Since the system is non-stationary, the generated series $\mathbf{Y}_t^s$ may take on negative or implausibly large numbers of the unemployed. In order to avoid an excessively large number of series with negative values, the state variables recursion is launched from the states' smoothed estimates at one of the highest points of the observed series. Further, the first 30 time points are discarded in order to prevent that the series start at the same time-point. With an assumption that unemployment in the Netherlands will not exceed 15 percent of the total labour force, the simulation data set is restricted to contain only series with values between 0 and 1 mln of unemployed (this value comprised about 15 percent of the Dutch labour force in 2010); other series are discarded. Keeping the artificial series below the upper bound is also done in order not to extrapolate outside of the original data range when simulating the design-based standard errors z_t^j.

Every series of simulated GREG point-estimates needs its own series of simulated design-based standard error estimates, z_t^j's. The original known design-based standard error estimates $\sqrt{\widehat{\mathrm{Var}}\left(Y_t^j\right)}$ would not be suitable for this simulation because the sampling error variance is proportional to the corresponding point-estimate. The following variance function is used to generate design-based variances for the simulated series of point-estimates (see Appendix B in Bollineni-Balabay et al. (2016b) for details):

$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^1\right)\right] = \ln\left[\left(z_t^1\right)^2\right] = c + \beta_1 \ln\left(l_t^1\right) + \varepsilon_t^1,\ \varepsilon_t^t \sim N\left(0, \left(\sigma_\varepsilon^1\right)^2\right);$$
$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^j\right)\right] = \ln\left[\left(z_t^j\right)^2\right] = \psi_j \ln\left[\left(z_{t-3}^{j-1}\right)^2\right] + \beta_j \ln\left(l_t^j\right) + \varepsilon_t^j,\ \varepsilon_t^j \sim N\left(0, \left(\sigma_\varepsilon^j\right)^2\right),\ j = \{2, 3, 4, 5\}, \quad (4.1)$$

where l_t^j, $j = \{1, 2, 3, 4, 5\}$ is the wave-signal being the sum of the trend, seasonal and RGB components. The regression coefficients in (4.1) are time-invariant and are obtained by regressing $\ln\left(z_t^j\right)^2$ on $\ln\left(l_t^j\right)$ and $\ln\left(\left(z_{t-3}^{j-1}\right)^2\right)$ from the original DLFS series. The superscripts are used to denote the wave these coefficients belong to. The coefficient estimates are presented in Table 4.1, together with the adjusted $\mathrm{R}-$square goodness of fit measure.

Table 4.1
Regression estimates for the design-based standard error process

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\hat{c}$	12.219	-	-	-	-
$\hat{\beta}_j$	0.630	0.468	0.354	0.414	0.413
$\hat{\psi}_j$	-	0.717	0.786	0.749	0.751
$\hat{\sigma}_\varepsilon^j$	0.202	0.204	0.228	0.225	0.267
R^2_{adj}	0.351	0.373	0.386	0.477	0.342

The simulation proceeds as follows. For each series length considered and in each simulation s, five simulated signals $l_{t,s}^j$, $j = \{1, 2, 3, 4, 5\}$ are used to generate five sets of the design-based standard errors $z_{t,s}^j$ according to the process defined by (4.1) and using the regression coefficients from Table 4.1. As soon as an artificial data set is generated, $\hat{\rho}_s$ estimate is obtained, whereafter the rest of the hyperparameters are estimated with the quasi-ML method. Note that the same set of design-based standard errors $\mathbf{z}_{t,s}$ is used to generate all bootstrap series within a particular simulation.

In order to obtain the true MSEs, the DLFS model is simulated a large number of times $(M = 50{,}000)$, with each of these replications being restricted to the same limits as before, i.e., between zero and 1 mln of the unemployed. The true MSE is calculated in the following way using the true state vector $\boldsymbol{\alpha}_{m,t}$ values known for every simulation m:

$$\mathbf{MSE}_t^{\text{true}} = \frac{1}{\mathrm{M}} \sum_{m=1}^{M} \left[\left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right) \left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right)' \right]. \tag{4.2}$$

The true MSE of the signal is calculated in the same way by using the wave-signal values $\mathbf{l}_{m,t}$.

5 Results

5.1 Alternative model specifications for the DLFS

STS models are usually selected and evaluated by means of formal diagnostic tests for normality, homoscedasticity and independence of the standardised innovations. Parsimonious parameterisation is based on log-likelihood ratio tests or on information criteria (e.g., AIC or BIC). The outcomes of such tests, however, depend on the particular point estimates of hyperparameters rather than on their entire distributions. Simulated distributions of the hyperparameter estimators, obtained with the Monte-Carlo simulation described in Section 4, give additional insight into the adequacy of the STS model. The simulated distributions give an indication as to whether or not the model tends to be overspecified in the sense that some state variables may be modelled as time invariant.

This study considers four models that differ in numbers of hyperparameters to be estimated with the ML method. The most complete model - Model 1 - is the one currently in use at Statistics Netherlands, but with the white noise component ε_t removed from the true population parameter ξ_t. This component has turned out to have an implausibly large variance and disturbed estimation of other marginally significant hyperparameters (the seasonal and RGB disturbance variances) in the case of the DLFS. Removing the

irregular component ε_t from the model has mitigated the instability in the two above-mentioned hyperparameters. This formulation implies that the population parameter ξ_t does not exhibit irregularities that cannot be picked up by the stochastic structure of the trend and seasonal components. This assumption can be advocated by a relative rigidity of labour markets. Alterations of unemployment levels are usually gradual and therefore must be largely incorporated into the stochastic trend movements. The other three models are special cases of Model 1, i.e., all with the irregular component ε_t removed (see Table 5.1).

Table 5.1
Hyperparameters estimated in the four versions of the DLFS model; the disturbance variances are estimated on a log-scale

Models	Description	Parameters estimated
M1	complete model	$\rho,\sigma^2_{\eta_R},\sigma^2_{\omega},\sigma^2_{\eta_\lambda},\sigma^2_{\nu_1},\sigma^2_{\nu_2},\sigma^2_{\nu_3},\sigma^2_{\nu_4},\sigma^2_{\nu_5}$
M2	seasonal time-independent	$\rho,\sigma^2_{\eta_R},\sigma^2_{\eta_\lambda},\sigma^2_{\nu_1},\sigma^2_{\nu_2},\sigma^2_{\nu_3},\sigma^2_{\nu_4},\sigma^2_{\nu_5}$
M3	RGB time-independent	$\rho,\sigma^2_{\eta_R},\sigma^2_{\omega},\sigma^2_{\nu_1},\sigma^2_{\nu_2},\sigma^2_{\nu_3},\sigma^2_{\nu_4},\sigma^2_{\nu_5}$
M4	seasonal, RGB time-independent	$\rho,\sigma^2_{\eta_R},\sigma^2_{\nu_1},\sigma^2_{\nu_2},\sigma^2_{\nu_3},\sigma^2_{\nu_4},\sigma^2_{\nu_5}$

The simulated distributions of the hyperparameter estimators under Model 1 indicate that variance hyperparameters of the seasonal and, in particular, of the RGB component are often estimated to be close to zero. This causes bi-modality in the distribution of these variance estimates with a significant mass concentrated close to zero. Apart from that, an attempt to estimate both $\ln\left(\hat{\sigma}^2_{\omega}\right)$ and $\ln\left(\hat{\sigma}^2_{\eta_\lambda}\right)$, as in Model 1, distorts the other hyperparameters' ML estimator distribution that is expected to be normal. For instance, normality in $\ln\left(\hat{\sigma}^2_{\nu_3}\right)$, $\ln\left(\hat{\sigma}^2_{\nu_4}\right)$ and $\ln\left(\hat{\sigma}^2_{\nu_5}\right)$ is severely violated with extreme outliers and/or a huge kurtosis (see Figure A.1 in Appendix, where the $x-$axis is extended due to the outliers), while the corresponding variances are less likely to exhibit extreme values as they are supposed to fluctuate around 1. Making the seasonal component time-invariant, as in Model 2, hardly changes the situation for the trend and RGB hyperparameters. Instead, it may even be seen as suboptimal due to more extreme outliers and excess kurtosis in the distribution of all the five survey error hyperparameters (Figure A.2). By contrast, under both models where the $RGB-$component is fixed over time (Models 3 and 4), all hyperparameter estimates corresponding to the survey error component have turned out to be normally distributed, see Figure A.3 and Figure A.4. Under Model 3, distributions are still skewed for the slope and seasonal components (skewness of -0.88 and -0.72, and kurtosis of 5.56 and 4.61, respectively). Fixing the seasonal hyperparameter to zero under Model 4 results in only a marginal improvement: the distribution of $\ln\left(\hat{\sigma}^2_{\eta_R}\right)$ is negatively skewed (-0.81) with an excess kurtosis of 1.76.

This simulation evidence suggests that the preference in modelling the DLFS series may be given to the more parsimonious Model 3, where only the RGB disturbance variance is set equal to zero. However, since the RGB itself depends on the numbers of unemployed, its variance hyperparameter is retained for production purposes at Statistics Netherlands to secure sufficient flexibility against gradual changes in the underlying process.

The likelihood ratio test can be used to test if the hyperparameters of the seasonal and RGB components are significantly different form zero, since Models 2 through 4 are nested in Model 1. The test-statistic has very low values for all the three alternative models with respect to Model 1 (0, 0.18 and 0.18 for Models 2,

3 and 4, respectively, where the absence of differences between Models 2 and 1, as well as between Models 3 and 4 is due to a very low hyperparameter value of the seasonal component). These tests, thus, do not indicate that the more parsimonious models perform worse compared to Model 1. Another way of evaluating the adequacy of the four models is to compare their predictive power using the Root Mean Squared Differences (RMSD) between the GREG estimates and the one-step-ahead predictions for the signals. This is done for each wave separately: $\text{RMSD}^j = 1/(T-d)\sum_{t=d}^{T}\left(\hat{l}_{t|t-1}^{j} - Y_t^j\right)^2$, with d taken equal to 20, 30 and 60 months. Results presented in the Appendix (Table B.1), however, show that there is hardly any difference in the performance of the four models when applied to the original series. The more parsimonious models show a slight increase in the RMSD.

The distribution of the estimator of the survey error autoregressive parameter ρ across the 1,000 simulated series does not seem to be affected by model reformulations: it approaches the normal distribution quite closely and ranges between 0 and 0.4 when $T = 114$, which is in line with the approximation of its asymptotic distribution mentioned in Subsection 3.3. The range is slightly wider for the shorter time series and narrower when $T = 200$. The simulation procedure described in the previous section and the analysis of bootstrap methods that follows is performed separately for all the four models.

5.2 MSE estimation

The focus of this simulation study is MSE estimation for the trend and for the population signal, the latter being the sum of the trend and seasonal components. The performance of the Kalman filter and of the five MSE estimation methods discussed in Section 3 is evaluated by use of the relative bias and MSE of the MSE estimators. First, the filtered MSE estimates from (3.3), (3.4) and (3.7) are averaged over 1,000 simulations (where the average is denoted with a bar: $\overline{\text{MSE}}_{t|t}$) whereas the Kalman filter MSE estimates are averaged over 10,000 simulations, as mentioned at the beginning of Section 4. These averaged filtered MSE estimates for Model 3 (except for the AA – method, see below why) are depicted in Figure 5.1 – 5.4 for $T = 48$, $T = 80$, $T = 114$, and $T = 200$, respectively, skipping the first $d = 30$ time points of the sample (d should exceed the number of time points at the beginning of the series required to eliminate the effect of the diffuse filter initialization). Note that the analysis is based on filtered, rather than smoothed estimates, because filtered estimates better mimic the process of official figures production. MSEs in the four figures exhibit declining patterns, as expected, since the accuracy of the filtered estimates increases if more information becomes available over time for estimating the state variables. An exception is the true MSEs in Figure 5.2. A possible explanation is that, in this application, the signal MSEs are proportional to the signals themselves through the design-based standard errors, with the true MSEs being based on another (much larger) set of simulated series (50,000 for true MSEs; 1,000 for MSE – estimators). Note that the lines in Figure 5.1 look much smoother because they are stretched over a smaller number of time points. Further, the patterns in Figure 5.2 – 5.3 look more erratic because the scale of the y – axis is finer, compared to Figure 5.1 and Figure 5.4.

The percentage relative bias is calculated as $\text{RB}_t^f = 100\%\left(\overline{\text{MSE}}_{t|t}^{f} \Big/ \text{MSE}_{t|t}^{\text{true}} - 1\right)$, where f defines a particular estimation method and $\text{MSE}_{t|t}^{\text{true}}$ is defined in (4.2). The percentage relative MSE biases averaged over time (skipping the first $d = 30$ time points) for the signal, the trend and seasonal components are presented in Tables 5.2, 5.3, 5.4, and 5.5.

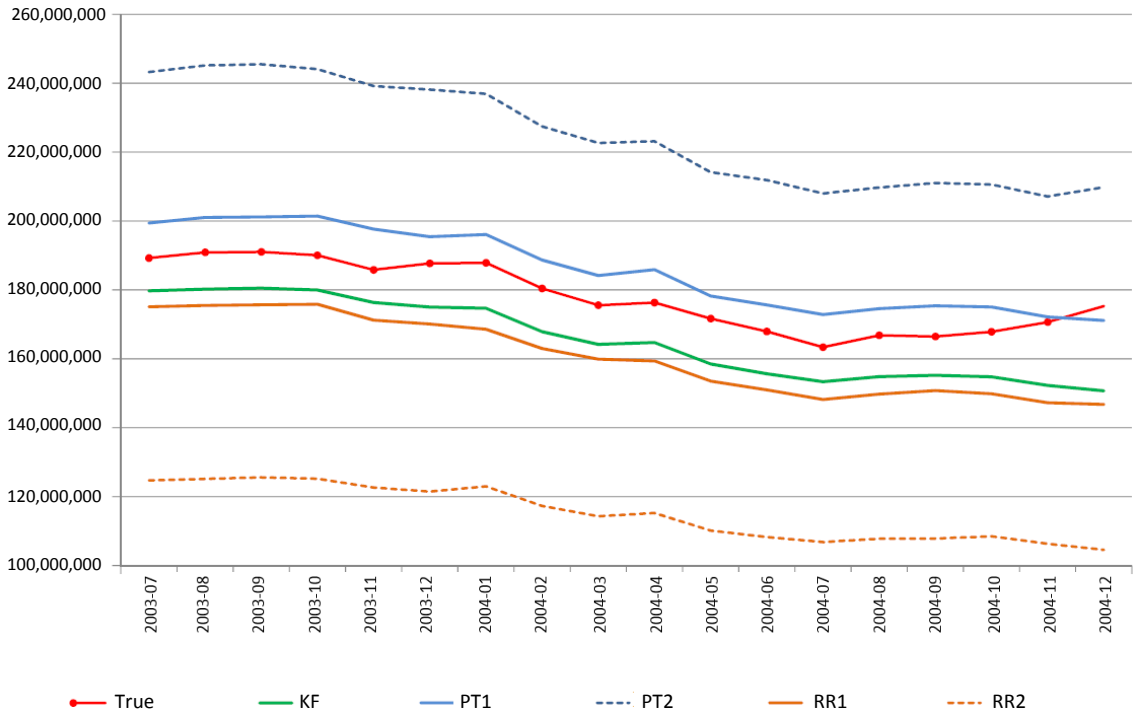

Figure 5.1 True MSEs and average MSE estimates for filtered true population parameter (trend plus seasonal) from Model 3, $T = 48$ months.

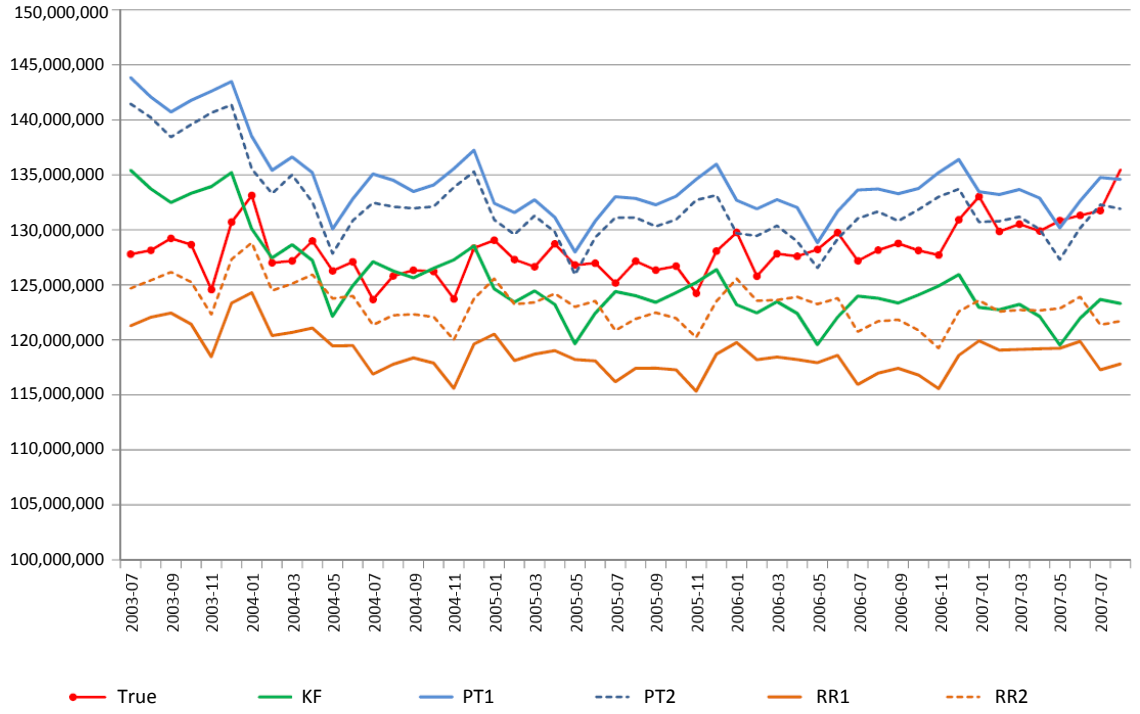

Figure 5.2 True MSEs and average MSE estimates for filtered true population parameter (trend plus seasonal) from Model 3, $T = 80$ months.

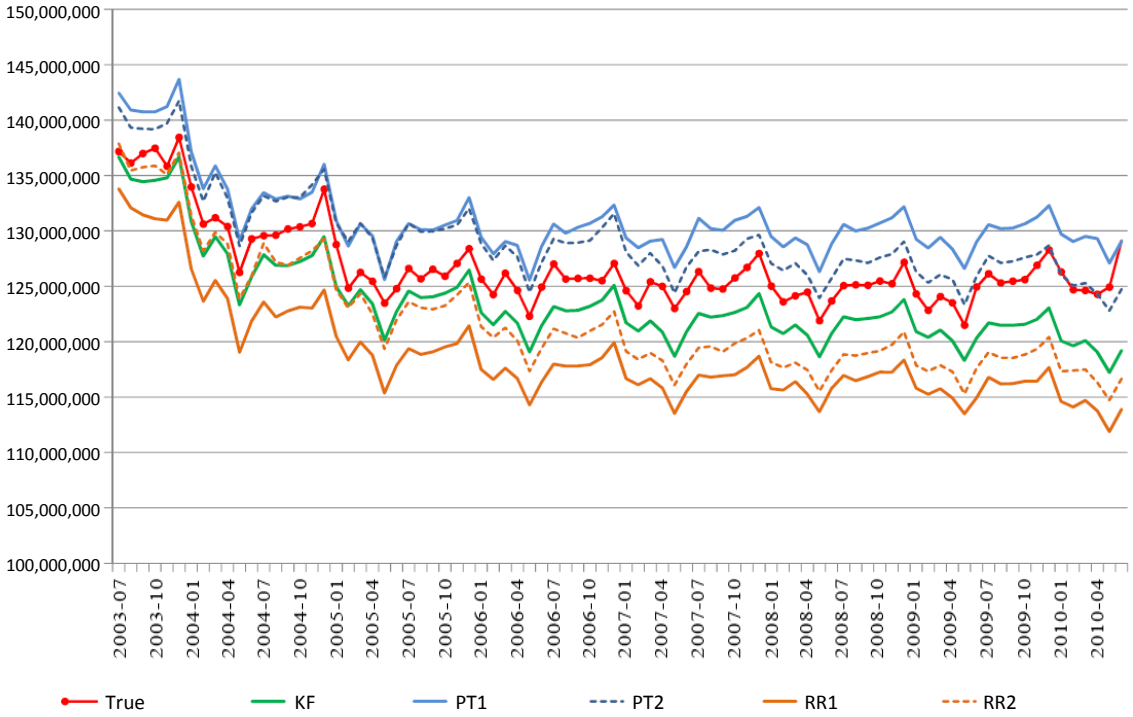

Figure 5.3 True MSEs and average MSE estimates for filtered true population parameter (trend plus seasonal) from Model 3, $T = 114$ months.

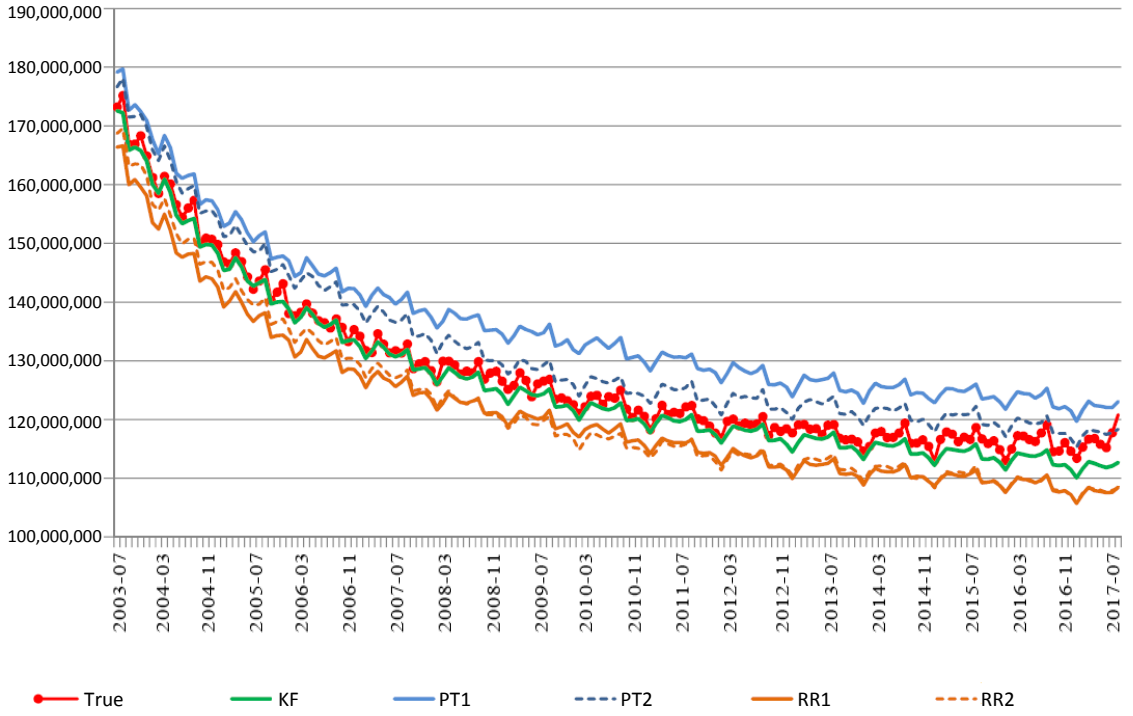

Figure 5.4 True MSEs and average MSE estimates for filtered true population parameter (trend plus seasonal) from Model 3, $T = 200$ months.

Table 5.2
Average percent bias of the MSE estimators under the DLFS model, $t = \{31,\ldots,T\}, T = 48$

Models	Signal*				Trend				Seasonal			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	N/A	N/A	-7.1	-7.6	N/A	N/A	-6.5	-6.6	N/A	N/A	-6.7	-7.0
PT1	N/A	N/A	*4.4*	*1.4*	N/A	N/A	8.7	6.4	N/A	N/A	4.9	2.4
PT2	N/A	N/A	26.2	-4.4	N/A	N/A	22.4	-3.1	N/A	N/A	25.6	-4.6
RR1	N/A	N/A	-9.8	-10.8	N/A	N/A	-13.9	-13.8	N/A	N/A	-9.5	-10.1
RR2	N/A	N/A	-35.3	-5.6	N/A	N/A	-29.9	-3.2	N/A	N/A	-29.7	-5.1

* Signal is the sum of the trend and seasonal component.

Table 5.3
Average percent bias of the MSE estimators under the DLFS model, $t = \{31,\ldots,T\}, T = 80$

Models	Signal*				Trend				Seasonal			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-3.0	-3.2	-2.1	-2.2	-3.5	-3.8	-2.5	-2.5	8.8	2.5	2.9	2.4
AA	N/A	N/A	N/A	14.9	N/A	N/A	N/A	15.0	N/A	N/A	N/A	14.9
PT1	8.6	6.7	4.9	6.2	10.6	8.9	7.1	8.4	20.8	10.7	10.3	11.1
PT2	*4.8*	*3.7*	*1.4*	*2.1*	4.8	4.9	2.1	2.3	17.3	8.2	6.9	7.1
RR1	-7.2	-9.0	-7.3	-7.2	-9.6	-11.2	-9.6	-9.5	-3.8	-9.0	-6.7	-6.6
RR2	6.7	-3.5	-3.9	-4.2	5.3	-4.1	-4.6	-5.4	18.6	-4.7	-4.1	-4.3

* Signal is the sum of the trend and seasonal component.

Table 5.4
Average percent bias of the MSE estimators under the DLFS model, $t = \{31,\ldots,T\}, T = 114$

Models	Signal*				Trend				Seasonal			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-2.1	-2.6	-2.4	-2.2	-2.3	-2.7	-2.4	-2.3	2.5	-3.2	-3.1	-2.6
AA	N/A	N/A	N/A	5.2	N/A	N/A	N/A	4.1	N/A	N/A	N/A	12.5
PT1	8.1	5.7	3.3	5.5	10.0	7.9	5.2	7.6	4.9	1.4	1.4	0.3
PT2	*2.2*	*3.2*	*1.9*	*1.5*	3.3	4.3	3.1	2.8	1.2	-2.0	1.0	0.6
RR1	-8.3	-7.8	-6.4	-6.5	-10.7	-9.9	-8.7	-8.9	-3.1	-7.2	-5.5	-5.6
RR2	-1.1	-6.0	-3.9	-3.5	-3.0	-7.6	-5.5	-5.0	7.3	-5.9	-3.2	-3.0

* Signal is the sum of the trend and seasonal component.

Table 5.5
Average percent bias of the MSE estimators under the DLFS model, $t = \{31,\ldots,T\}, T = 200$

Models	Signal*				Trend				Seasonal			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-1.3	-1.6	-1.3	-1.3	-1.7	-1.8	-1.6	-1.6	3.8	-1.7	-1.6	-1.6
AA	N/A	N/A	N/A	5.9	N/A	N/A	N/A	5.6	N/A	N/A	N/A	5.6
PT1	*6.3*	6.2	6.3	5.5	7.5	7.7	7.8	7.1	10.8	2.6	3.0	3.0
PT2	6.8	*4.0*	*3.0*	*2.3*	7.6	4.9	4.2	3.6	12.5	2.1	1.3	0.6
RR1	-8.0	-8.0	-4.9	-5.9	-10.0	-9.9	-6.8	-7.1	-1.1	-5.3	-3.8	-3.9
RR2	-5.1	-5.6	-4.5	-5.0	-7.0	-7.4	-6.0	-6.4	3.6	-3.1	-3.3	-3.9

* Signal is the sum of the trend and seasonal component.

Table 5.6
Average estimated variance and MSE of the MSE estimators for the numbers of unemployed under the DLFS model (divided by 10^{15}), $t = \{31,\ldots,T\}, T = 48$

Models	Signal*				Trend				Seasonal			
	M3		M4		M3		M4		M3		M4	
	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}
PT1	3.39	3.46	3.64	3.66	3.61	3.83	3.67	3.81	0.59	0.61	0.64	0.65
PT2	5.03	7.26	3.03	3.10	4.02	5.27	2.56	2.61	1.00	1.50	0.52	0.54
RR1	2.51	2.83	2.68	3.06	2.03	2.51	2.13	2.62	0.44	0.51	0.48	0.55
RR2	1.59	5.93	2.74	2.85	1.52	3.97	2.50	2.56	0.55	1.28	0.50	0.52

* Signal is the sum of the trend and seasonal component.

Table 5.7
Average estimated variance and MSE of the MSE estimators for the numbers of unemployed under the DLFS model (divided by 10^{15}), $t = \{31,\ldots,T\}$, $T = 80$

Models	Signal*				Trend				Seasonal			
	M3		M4		M3		M4		M3		M4	
	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}	Var_{MSE}	MSE_{MSE}
PT1	2.24	2.29	2.43	2.52	1.82	1.91	1.97	2.09	0.27	0.30	0.27	0.31
PT2	2.20	2.23	2.14	2.18	1.71	1.74	1.66	1.69	0.27	0.28	0.27	0.29
RR1	1.86	1.95	1.74	1.82	1.42	1.56	1.33	1.46	0.22	0.23	0.22	0.23
RR2	1.98	2.01	1.94	1.97	1.57	1.60	1.49	1.54	0.23	0.23	0.23	0.23

* Signal is the sum of the trend and seasonal component.

The main conclusions from the simulation study are as follows:

1. For $T = 48$, and when averaged over time (starting from $t = 31$), the relative bias of the signal MSE obtained with the use of the Kalman filter is around -7 percent. This bias tends to decrease as the series length increases. The Kalman filter (KF) bias is quite small for the case of $T = 200$, such that none of the estimation methods offers an improvement over the $\text{KF}-$based $\text{MSE}-$estimates. One could still apply the best estimation method with positive biases in order to get a range of values containing the true MSE.

2. The $\text{AA}-$method turned out to be inapplicable to the models with marginally significant hyperparameters. When some of the hyperparameters are estimated close to zero, the matrix $\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\text{ML}} \middle| \rho^{a}\right)$ is numerically either singular, leading to a failure in the procedure, or nearly singular. In the latter case, the asymptotic variance becomes excessively large and thus not reliable. Taking this into account, the $\text{AA}-$method could only be considered for Model 4. As expected, the method performs poorly in short series, with positive biases of about 15 percent. The performance for $T = 114$ and $T = 200$ is comparable to that of the $\text{PT1}-$bootstrap method, but significantly worse than the PT2 method's performance.

3. As can be immediately observed, the use of the $\text{RR}-$bootstrap results in a negative bias, whereas the use of the $\text{PT}-$method produces a positive bias. Contrary to the claim of Rodriguez and Ruiz (2012) that their approach has better finite sample properties compared to the approach of Pfeffermann and Tiller (2005), the case of the DLFS suggests that the $\text{RR}-$based MSE estimates, both the parametric and non-parametric ones, have larger negative biases than the $\text{KF}-$based MSE estimates across all the models and series lengths (except for RR2 in Model 4 when $T = 48$, and in Model 1 when $T = 80$ and $T = 114$). While the $\text{PT}-$bootstrap method is shown to have satisfactory asymptotic properties in Pfeffermann and Tiller (2005), Rodriguez and Ruiz (2012) illustrate the superiority of their method in small samples based on a simple model (a random walk plus noise). The present simulation study reveals that the $\text{RR}-$method may not behave well in more complex applications. The $\text{PT}-$methods have never produced negative biases for the DLFS, which makes these methods conservative (except for PT2 in Model 4 when $T = 48$, with the negative bias still being smaller than that of the Kalman filter). Another striking outcome for $T = 48$ is that the PT2 positive bias and the RR negative bias take on very large values in Model 3. However, with such a short series length and with so many non-stationary components like in the DLFS model, it is difficult to obtain reliable estimates from non-parametric bootstrap methods, since the burn-in period (or the diffuse

sample) necessary for the non-parametric generation of the series takes more than a quarter of the series length (13 months out of 48).

4. For the series of lengths $T = 114$ and $T = 80$, the positive biases produced by the $PT2-$method slightly exceed the $KF-$biases in absolute value in models with insignificant hyperparameters (Models 1 and 2). In the more stable models (Models 3 and 4), the positive biases are smaller than the KF negative biases in absolute value. For $T = 48$, bootstrap results are presented only for Models 3 and 4 (Models 1 and 2 that tend to be overspecified are not considered due to numerical problems). As could be expected, the biases are larger for this series length: the negative KF and RR biases become larger in absolute value, and so do the PT positive biases, with an exception of the above-mentioned result for PT2 in Model 4.

The signal MSE of Model 3, which could be considered as the better option for the production of official DLFS figures, is best estimated by the PT2 approach, with the relative bias of 1.4 and 1.9 percent for $T = 80$ and $T = 114$, respectively. The $PT2-$bootstrap method also seems to be the best method for $T = 200$, but, as already noted, the negative KF biases are already quite small for series of this length. For very short series like $T = 48$, the parametric PT1 bootstrap seems to be the best option.

5. For both the $PT-$ and $RR-$methods (except for RR2 in Model 4, $T = 48$), the absolute values of the relative biases are smaller in the case of the non-parametric approaches, compared to their parametric counterparts. The superiority of the non-parametric approach over the parametric one can be explained by the distorted normality of the error distribution in the models. Therefore, non-parametric bootstraps should be preferred unless time series are very short.

6. Apart from the bias of the MSE estimators, their variability may also give important insights into their reliability. To our knowledge, this has not been yet presented in the statistical literature. Tables 5.6 and 5.7 contain variances and MSEs of the four bootstrap MSE estimators for the signal, trend, and seasonal components for the two most interesting series lengths: $T = 48$ and $T = 80$ months (Models 1 and 2, as well as the asymptotic approximation, are not considered due to the aforementioned numerical problems). For both Model 3 and Model 4, the MSEs of the two PT MSE estimators are larger than the MSEs of the two RR MSE estimators. The RR MSE estimators' seemingly superior performance, reflected by their smaller MSEs, is due to their smaller variances. The biases, however, are sometimes large enough to bring MSEs of these MSE estimators almost to the level of MSEs of the PT estimators. More importantly, the biases of the RR MSE estimators are mostly negative, often exceeding those of the Kalman filter. This phenomenon makes $RR-$bootstraps hardly applicable in this application.

Apart from the above-mentioned simulation results, it is also interesting to see if the STS model-based approach still offers more precise predictors than the design-based variance estimates even after correcting for the hyperparameter uncertainty. For this purpose, $STS-$model-based Root MSEs (RMSEs) obtained with the different MSE estimation procedures for the original series $(T = 114)$ are compared to the standard errors (SEs) of the GREG estimator. Such Mean Differences in the Standard Errors (MDSE) under the time series model m $(m = \{1, 2, 3, 4\})$ are defined as: $\mathrm{MDSE}_m^f = 100\%/(T-d)\sum_{t=d}^{T}\left[\mathrm{RMSE}^f\left(\hat{l}_{t|t}^m\right) - \mathrm{SE}\left(Y_t\right)\right]\Big/\mathrm{SE}\left(Y_t\right)$ and are presented in Table 5.8, with $\hat{l}_{t|t}^m$ being the filtered estimate for the true population parameter, defined

as trend plus seasonal, under model m. Results are shown for the Kalman filter (labelled as "KF" in the table), i.e., when the hyperparameter uncertainty is neglected, as well as for cases when the five MSE estimation methods are applied to take the hyperparameter uncertainty into account. The true RMSEs from (4.2) are also compared to the GREG standard errors (see row "True" in Table 5.8). Note that the RGB and, particularly, the seasonal hyperparameter estimates obtained from the original DLFS data set are quite small. Therefore, there are no noticeable differences between the signal point-estimates of the four models. The AA, being the most unreliable approach, produces overestimated SEs (compare the 18- to 20-percent reduction based on the true RMSEs) due to nearly singular information matrices of the hyperparameter ML estimates. Keeping that in mind, one should feel more confident with the use of the PT – estimators. Although the simulation study presented in this paper shows that PT2 usually outperforms the PT1 parametric approach, for this particular series, the PT1 – based SEs are closest to the true RMSEs, offering about a 20 percent reduction in the estimated GREG standard errors. This means that the model-based approach offers a significant variance reduction compared to the traditional design-based approach, even after accounting for the hyperparameter uncertainty.

Table 5.8
Percentage mean differences in the SEs (MDSEs) between the GREG- and model-based estimators for the original DLFS series, $d = 30$; percentage increase in the Kalman filter-based SEs after applying the MSE correction (in parentheses)

	Model 1	Model 2	Model 3	Model 4
KF	-24.1	-24.1	-24.5	-24.5
True	-20.0 (5.56)	-20.1 (5.5)	-20.6 (5.4)	-20.7 (5.3)
AA	-18.8 (6.9)	-19.0 (6.7)	-19.1 (7.1)	-19.5 (6.6)
PT1	-20.1 (5.2)	-20.1 (5.2)	-21.1 (4.6)	-21.2 (4.4)
PT2	-22.9 (1.6)	-21.2 (3.8)	-22.2 (3.1)	-22.5 (2.6)
RR1	-26.5 (-3.2)	-26.6 (-3.4)	-26.5 (-2.7)	-26.5 (-2.7)
RR2	-24.0 (-0.1)	-25.4 (-1.8)	-25.6 (-1.4)	-25.7 (-1.6)

6 Concluding remarks

There is a gradually increasing interest among NSIs in the use of STS models for the production of monthly figures on the labour force. In the Netherlands, such a model has been applied in the production of official LFS figures since 2010. STS models constitute a type of small area estimation (SAE), where sample information from preceding periods is used to obtain more precise estimates, as well as to account for the rotating panel design, often used in Labour Force Surveys.

Ignoring the hyperparameter uncertainty in the MSEs of STS model-based estimates results in underestimation of the MSEs of domain estimates. Particularly when series are short, which is often the case at NSIs, the bias due to ignoring hyperparameter uncertainty can be substantial. Most applications of SAE procedures in the literature are based on multilevel models, where it is common practice to account for hyperparameter uncertainty. The literature on STS models applied in the context of SAE is rather limited, with most applications ignoring hyperparameter uncertainty in the MSE estimates. Whether the bias in the

obtained MSEs becomes substantial, depends on the structure of the model and on the length of the series. The present work describes a Monte-Carlo simulation applied to the STS model used by Statistics Netherlands for estimating monthly unemployment. The simulation serves two purposes. Firstly, it establishes the amount of bias in the DLFS MSEs when hyperparameter uncertainty is ignored. In addition to that, several MSE estimation methods available in the literature for the STS framework are compared in this simulation, and the best approach for the Dutch LFS is established. Secondly, simulating the distributions of the hyperparameter estimators is useful for obtaining better insights into the dynamics of the unobserved components in the STS model, and thus, ascertain the necessity to model the components as time-variant. In the case of the DLFS, the simulation shows that it might be worth considering a more restricted version of the model, where the rotation group bias is time-invariant and the population white noise is ignored. For both reasons, it is advisable to conduct a simulation as described in this paper as part of the model implementation process into official statistical production.

The comparison of the MSE estimation procedures also sheds new light on their properties. The asymptotic approximation is not applicable to cases where hyperparameters are close to zero because the information matrix of the hyperparameter estimates becomes (almost) singular. The non-parametric bootstraps, being less dependent on normality assumptions, perform better than their parametric counterparts under both Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012) approaches, except in very short series. The most important finding is that the PT bootstraps have positive biases and consistently outperform the RR bootstraps, where the biases are generally negative and larger (in absolute terms) than those produced by the Kalman filter. This is contrary to the claim of Rodriguez and Ruiz (2012) about the superiority of their method in short time series. Apparently, their findings are purely heuristic and are based on a simple model (random walk plus noise), while Pfeffermann and Tiller (2005) prove that their bootstrap approach produces MSE estimates with a bias of correct order.

The variances of the PT MSE estimators are larger than the variances of the RR MSE estimators. Differences between MSEs of the PT and RR MSE estimators are modest to moderate (MSEs of the RR MSE estimators are 28 to 8 percent lower than those of the PT estimators, depending on the model and the time series length). More importantly, the tendency of the RR MSE estimators to have negative biases, sometimes exceeding those of the Kalman filter, renders these bootstrap methods inapplicable. Hence, the $PT-$methods should be generally considered for other survey data too, despite the fact that these methods may occasionally be outperformed by the $RR-$methods.

For very short time series, the non-parametric bootstraps do not seem to be an option for a model of the presented complexity. The PT parametric bootstrap, however, corrects the negatively biased MSE up to a small positive bias (1.4 to 4.4 percent, depending on the model). For the present series length of 114 months, the negative MSE bias can be reduced from about -2.4 to 1.9 percent with the non-parametric method of Pfeffermann and Tiller (2005) in the model with a time-invariant RGB. The true Kalman filter root MSEs are about 20 smaller than the standard errors of the GREG estimates in all the four models applied to the DLFS data. In general, the biases in the Kalman filter MSE estimates are relatively small in the DLFS application. Therefore, it may be deemed sufficient to rely on these naive MSE estimates for publication purposes.

Acknowledgements

We thank the National Statistical Office of the Netherlands, Statistics Netherlands, for funding this research, as well as the Associate Editor and the two anonymous reviewers for careful reading of this manuscript and valuable comments. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

Appendices

A. Simulated densities of the hyperparameters under the four versions of the DLFS model

This appendix presents the hyperparameter density functions obtained from simulations where the four versions of the DLFS model (see Table 5.1) act as the data generating process. The x-axes depict variance hyperparameters on a log-scale, while the y-axes stand for frequencies. The x-axis may be extended due to outliers.

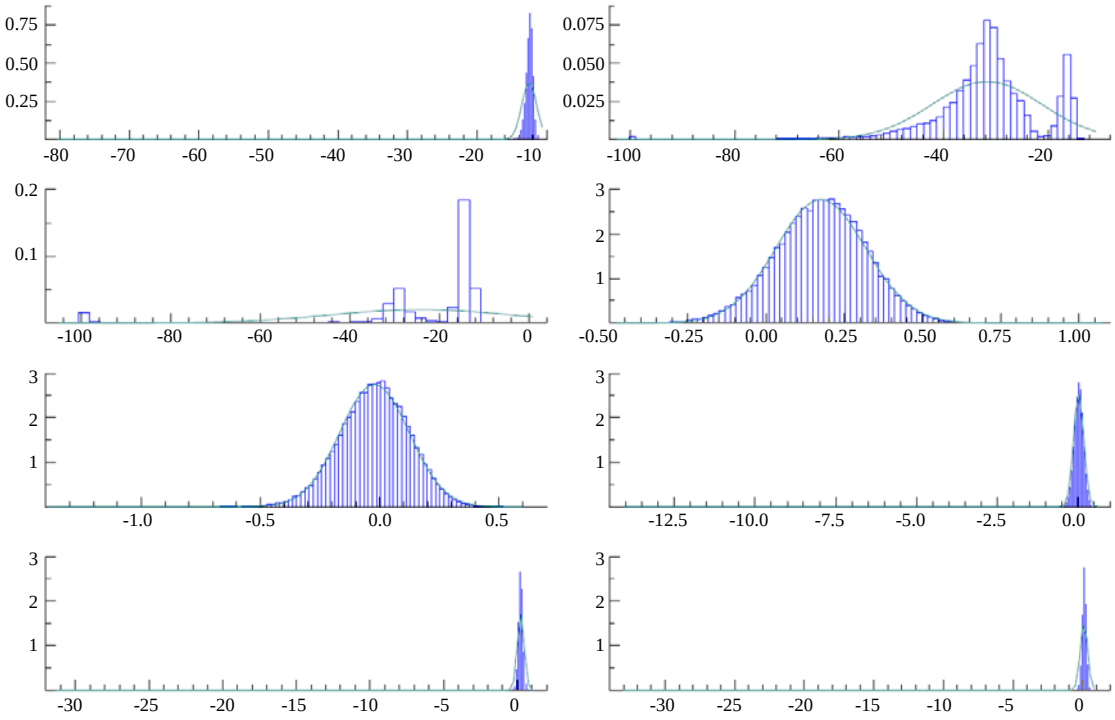

Figure A.1 Hyperparameter distribution under the complete DLFS model (Model 1), left to right on the x-axes: $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^l})$, $\ln(\hat{\sigma}^2_{v_t^{l-3}})$, $\ln(\hat{\sigma}^2_{v_t^{l-6}})$, $\ln(\hat{\sigma}^2_{v_t^{l-9}})$, $\ln(\hat{\sigma}^2_{v_t^{l-12}})$; the normal density with the same mean and variance superimposed; 50,000 simulations, $T = 114$.

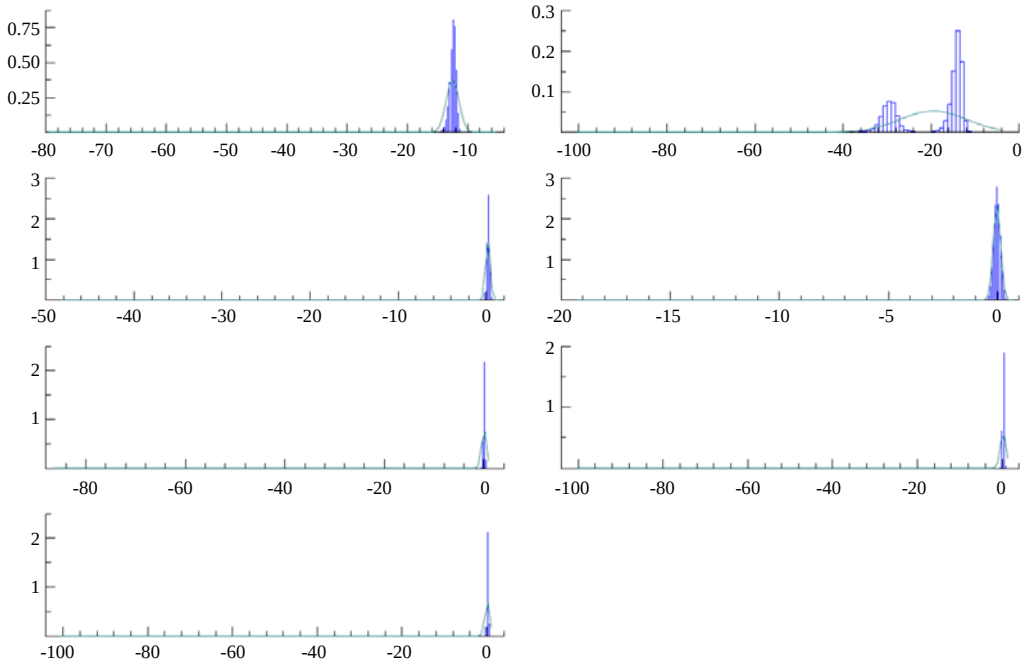

Figure A.2 Hyperparameter distribution under Model 2, left to right on the x-axes: $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; the normal density with the same mean and variance superimposed; 50,000 simulations, $T = 114$.

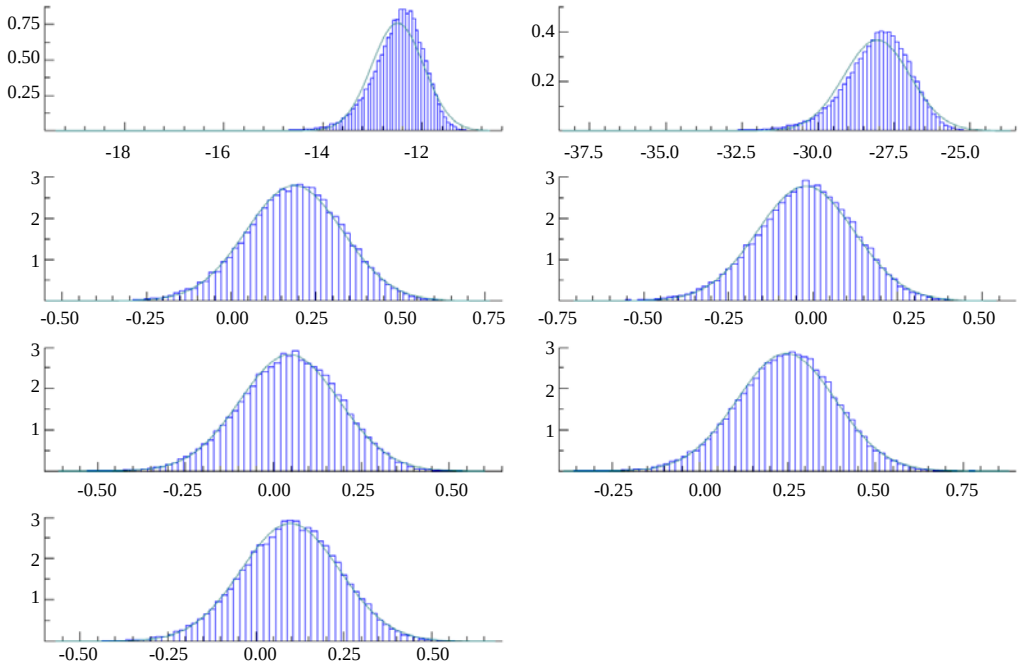

Figure A.3 Hyperparameter distribution under Model 3, left to right on the x-axes: $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; the normal density with the same mean and variance superimposed; 50,000 simulations, $T = 114$.

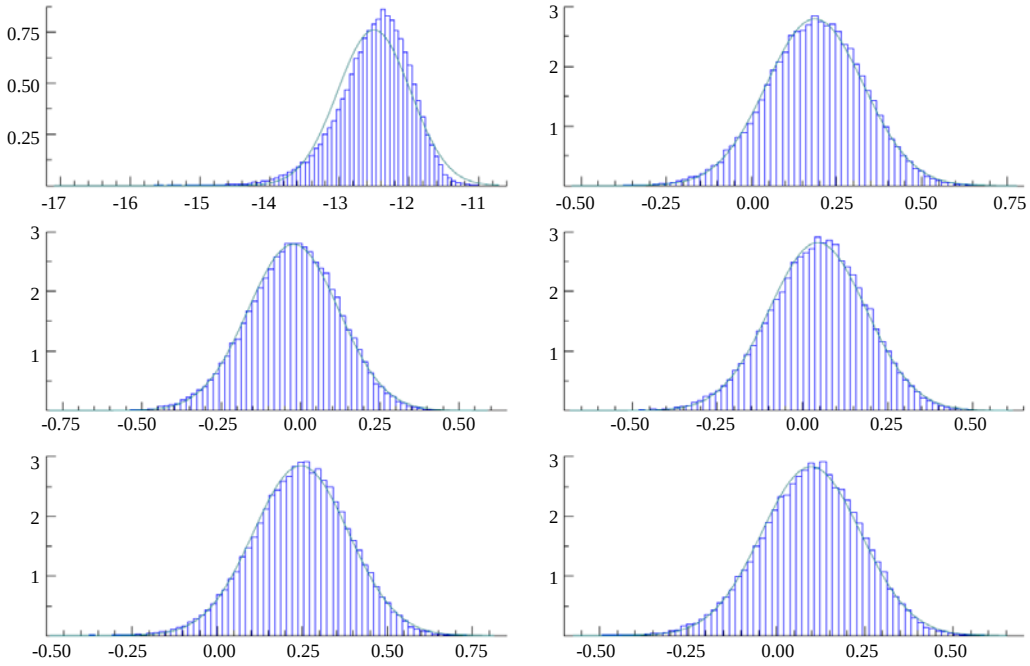

Figure A.4 Hyperparameter distribution under Model 4, left to right on the x-axes: $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; the normal density with the same mean and variance superimposed; 50,000 simulations, $T = 114$.

B. Predictive performance of the four DLFS models

Table B.1
Root mean square deviations of GREG estimates of the numbers of unemployed from their one-step-ahead predictions, per wave

W	Model 1			Model 2			Model 3			Model 4		
	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$
1	34,370	33,582	34,641	34,370	33,582	34,641	34,518	33,754	34,881	34,525	33,757	34,885
2	30,130	29,770	29,410	30,130	29,770	29,410	30,138	29,780	29,418	30,144	29,779	29,409
3	35,792	32,631	34,654	35,792	32,631	34,654	35,714	32,535	34,499	35,716	32,532	34,499
4	39,647	38,556	36,797	39,647	38,556	36,797	39,753	38,640	36,891	39,743	38,633	36,889
5	38,271	37,622	36,341	38,271	37,622	36,341	38,183	37,528	36,225	38,177	37,523	36,226

References

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bartlett, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8, 27-41.

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.

Bollineni-Balabay, O., van den Brakel, J. and Palm, F. (2016a). Multivariate state space approach to variance reduction in series with level and variance breaks due to survey redesigns. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179, 377-402.

Bollineni-Balabay, O., van den Brakel, J. and Palm, F. (2016b). State space time series modelling of the Dutch Labour Force Survey: Model selection and MSE estimation, - Extended version. Discussion paper, Statistics Netherlands, Heerlen. https://www.cbs.nl/en-gb/background/2016/41/state-space-time-series.

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

Doornik, J. (2007). *An Object-Oriented Matrix Programming Language Ox 5*. Timberlake Consultants Press, London.

Durbin, J., and Koopman, S.J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89, 603-615.

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Number 38. Oxford University Press.

EUROSTAT (2015). Task force on monthly unemployment - revised report. Working group labour market statistics.

Hamilton, J. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33, 387-397.

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

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

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

Krieg, S., and van den Brakel, J. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis*, 56, 2918-2933.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 2, 199-207. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1987002/article/14607-eng.pdf.

ONS (2015). A state space model for LFS estimates: Agreeing the target and dealing with wave specific bias. Report of the 29th meeting of the Government Statistical Service Methodology Advisory Committee. http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/previous-meeting-papers-and-minutes/mac-29-papers.pdf.

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

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40-68.

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 Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26, 893-916.

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 and Economic Statistics*, 16, 339-348.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rodriguez, A., and Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis*, 56, 62-74.

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

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

van den Brakel, J., 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., 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.

Zhang, M., and Honchar, O. (2016). Predicting survey estimates by state space models using multiple data sources. Paper for the Australian Bureau of Statistics' Methodology Advisory Committee.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 69-92
Statistics Canada, Catalogue No. 12-001-X

Bayesian predictive inference of a proportion under a two-fold small area model with heterogeneous correlations

Danhyang Lee, Balgobin Nandram and Dalho Kim[1]

Abstract

We use a Bayesian method to infer about a finite population proportion when binary data are collected using a two-fold sample design from small areas. The two-fold sample design has a two-stage cluster sample design within each area. A former hierarchical Bayesian model assumes that for each area the first stage binary responses are independent Bernoulli distributions, and the probabilities have beta distributions which are parameterized by a mean and a correlation coefficient. The means vary with areas but the correlation is the same over areas. However, to gain some flexibility we have now extended this model to accommodate different correlations. The means and the correlations have independent beta distributions. We call the former model a homogeneous model and the new model a heterogeneous model. All hyperparameters have proper noninformative priors. An additional complexity is that some of the parameters are weakly identified making it difficult to use a standard Gibbs sampler for computation. So we have used unimodal constraints for the beta prior distributions and a blocked Gibbs sampler to perform the computation. We have compared the heterogeneous and homogeneous models using an illustrative example and simulation study. As expected, the two-fold model with heterogeneous correlations is preferred.

Key Words: Blocked Gibbs sampler; Hierarchical Bayesian model; Intracluster and intercluster correlations; Goodness of fit; Unimodality; Weakly identifiable.

1 Introduction

We assume that there are several small areas, each area consists of several clusters and each cluster consists of a number of units (individuals). A random sample of clusters is taken from each area and within each sampled cluster a random sample of units is taken. This is a two-fold sampling design; see Rao and Molina (2015). When there is cluster sampling, the units within a cluster are generally positive and this correlation can have a significant impact on inference. We consider this situation for binary responses; see Nandram (2015) who defined an intracluster (between two units in the same cluster) correlation and an intercluster (between two units in two different clusters in the same area) correlation. We extend the model of Nandram (2015), who assumes that the correlation remains constant over areas, to accommodate the situation in which the correlations can be different. We are interested in the finite population proportion for each area, and like Nandram (2015), we use a hierarchical Bayesian model for this purpose.

Given such correlated data, a statistical problem arises from the intracluster correlation, leading to a smaller effective sample size and therefore larger variability in the estimates. Thus, when there is clustering effect, analyses that assume independence of the units will generally result in smaller $p-$values (i.e., rejection when it is otherwise). Rao and Scott (1981, 1984) have studied this problem and presented simple corrections to standard chi-squared statistic for the test of independence in two-way contingency tables under a complex sample design (e.g., two-stage cluster sampling).

Nandram and Sedransk (1993) presented a hierarchical Bayesian model under two-stage cluster sampling. This is the design we have within each area in a two-fold sample design with binary responses.

1. Danhyang Lee, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A. E-mail: danhyang@iastate.edu; Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609, U.S.A. E-mail: balnan@wpi.edu; Dalho Kim, Department of Statistics, Kyungbuk National University, 80 Daehakro, BukGu, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr.

As a discrete analogue of the model for two-stage cluster sampling with normal data (Scott and Smith 1969), this model makes inference about the overall finite population proportion. This model was also extended by Nandram (1998) to multinomial data which can be viewed as a Bayesian analogue of the multinomial-Dirichlet model for cluster sampling (Brier 1980).

For two-fold modeling, there are a limited number of studies for continuous response variables, and almost none for discrete (binary) data. Most of the analyses for two-fold modeling are based on the empirical Bayes framework. Fuller and Battese (1973) introduced one-fold and two-fold nested error regression models. Ghosh and Lahiri (1988) studied multistage sampling under posterior linearity using Bayes and empirical Bayes methods. Under two-stage and three-stage cluster sampling, estimation of regression models with nested error structure and unequal error variances were further studied by Stukel and Rao (1997). Small area models under two-fold nested error regression models were also studied by Stukel and Rao (1999); see Rao and Molina (2015) for a review. Nandram (2015) proposed a hierarchical Bayesian model for binary data arising from a two-fold sample design.

Nandram (2015) showed that it is important to consider the sample design within each area. Specifically, similar to Rao and Scott (1981, 1984), he showed that if a model does not capture the two-stage cluster sample design within each small area, the result will be too optimistic. That is, the variability will be too small. It is also true that the point estimates could be different when the two-stage cluster sample design is ignored. He also noted that there are other situations where the result could be the opposite. For example, if there is a stratified design, rather than a two-stage cluster sampling design, there will be increased precision within each area (i.e., the design effect for each area will be smaller than one). See Nandram, Bhatta, Sedransk and Bhadra (2013) for a Bayesian analysis on this problem.

To gain flexibility and generality over the two-fold hierarchical Bayesian model, Nandram (2015), we generalize it to incorporate unequal intracluster correlations. Our idea is to extend the model of Nandram (2015) by considering an additional layer for intracluser correlation to vary over areas in the two-fold sample design and to compare the two-fold model with homogeneous correlation (equal over areas) and heterogeneous correlations (vary over areas). Like the homogeneous model, the heterogeneous model has weakly identified parameters. When a Markov chain Monte Carlo sampler is used to fit such a model, there can be long-range dependence, and there will be difficulties in monitoring convergence of a Gibbs sampler. Nandram (2015) showed how to overcome the difficulty with these weakly identified parameters using random draws. Similar random draws are discussed in Molina, Nandram and Rao (2014) and Toto and Nandram (2010) who avoided Markov chain Monte Carlo modeling fitting completely. Unfortunately, it is not simple to use random draws to fit the heterogeneous model; we are forced to use the Gibbs sampler.

We use the blocked Gibbs sampler to fit our two-fold small area model. There are two difficulties we face. First, the conditional posterior densities of the correlation parameters can be multimodal. Second, some parameters can be related in a complex manner. Both of these issues can give difficulties in using a Markov chain Monte Carlo sampler leading to long-range dependence in the iterates. Thus, to help relieve these difficulties, we have restricted the prior densities of the area parameters to be unimodal and we have used the blocked Gibbs sampler to draw groups of parameters simultaneously. Both strategies lead to additional complexities but much better fitting samplers.

As a summary, we extend Nandram (2015) to accommodate heterogeneous correlations. The model with heterogeneous correlations is desirable because one may assume that the correlation does not vary with area when it actually does and this can lead to inaccurate results. Evidently, this is an important contribution beyond Nandram (2015). However, we encounter three difficulties.

1. The heterogeneous correlations introduce weakly identifiable parameters into the model.
2. Unlike Nandram (2015) Markov chain Monte Carlo methods are needed to fit the model.
3. A useful unimodal restriction is imposed on the hyperparameters to help proper mixing.

We have an innovative construction of a griddy blocked Gibbs sampler to fit the model with heterogeneous correlations. We have extensive testing of our model beyond Nandram (2015).

In this paper we consider Bayesian predictive inference of the finite population proportions of a number of small areas when there is a cluster sample design within each area. In our main contributions we use a hierarchical Bayesian model, which has unequal intracluster correlations, to make posterior inference about the finite population proportion of each area. In Section 2 we have a detailed description of the heterogeneous model. Specifically, first for motivation and updating, we give a brief review of the homogeneous model, Nandram (2015). We show that some parameters can be weakly identified. We also describe the computation to draw a random sample from the posterior distribution using the blocked Gibbs sampler. In Section 3, to compare the models with homogeneous correlation and heterogeneous correlations, we present an illustrative example on the Third International Mathematics and Science Study (TIMSS) and a small-scale simulation study. Finally, Section 4 contains concluding remarks and future research directions. Appendices A and B contain proofs and additional information.

2 Bayesian two-fold small area models and computations

We consider a finite population of ℓ areas and M_i clusters within the i^{th} area, and we assume there are N_{ij} individuals in j^{th} cluster within i^{th} area. The binary responses are y_{ijk} for $i = 1,\ldots,\ell$, $j = 1,\ldots,M_i$, $k = 1,\ldots,N_{ij}$. We assume that a simple random sample of m_i clusters is taken from the i^{th} small area and a simple random sample of n_{ij} individuals is taken from the m_i sampled clusters from the i^{th} area. Here, we assume the survey weights are the same within all clusters in each area. Let $n_i = \sum_{j=1}^{m_i} n_{ij}$, $s_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$ and $s_i = \sum_{j=1}^{m_i} s_{ij}$.

Our target is the finite population proportion of the i^{th} area which is given by

$$P_i = \frac{\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}} y_{ijk}}{N_i}, \quad i = 1,\ldots,\ell,$$

where $N_i = \sum_{j=1}^{M_i} N_{ij}$. Let $T_{ij}^{(1)} = \sum_{k=n_{ij}+1}^{N_{ij}} y_{ijk}$ denote the nonsampled totals of the sampled clusters $(j = 1,\ldots,m_i)$, and $T_{ij}^{(2)} = \sum_{k=1}^{N_{ij}} y_{ijk}$, the totals of the nonsampled clusters $(j = m_i+1,\ldots,M_i)$. Letting $n_i = \sum_{j=1}^{m_i} n_{ij}$, $\hat{p}_i = \sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} y_{ijk} \big/ n_i$, we can express our target, P_i as

$$P_i = \frac{n_i \hat{p}_i + \sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}}{N_i}, \quad i = 1, \ldots, \ell. \tag{2.1}$$

To make inference about the P_i, we fit hierarchical Bayesian models to the data. Using the beta-binomial representation, these models accommodate the two-fold design structure. We describe two models, one with homogeneous correlation and the other with heterogeneous correlations, our main contribution beyond Nandram (2015). In Section 2.1 we review the hierarchical Bayesian model with homogeneous correlation, Nandram (2015) and we show how to make it comparable to our hierarchical Bayesian model with heterogeneous correlations which we describe in Section 2.2. In Section 2.3 we describe the blocked Gibbs sampler to fit our model with heterogeneous correlations.

2.1 A review of two-fold model with homogeneous correlation

Nandram (2015) described the two-fold small area model with homogeneous correlation. Here we give a brief review of its main assumptions which are

$$y_{ijk} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(p_{ij}\right), \tag{2.2}$$

$$\mu_i \,|\, \theta, \gamma \overset{\text{iid}}{\sim} \text{Beta}\left[\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right], \tag{2.3}$$

$$\rho, \theta, \gamma \overset{\text{iid}}{\sim} \text{Uniform}(0,1), \tag{2.4}$$

where ρ and γ represent the intracluster and intercluster correlation, respectively. It is assumed that $0 < \theta, \rho, \gamma < 1$ strictly. Note that within the same area the intracluster correlation ρ, the correlation between two units in the same cluster, is $\text{cor}\left(y_{ijk}, y_{ijk'} \,|\, \mu_i, \gamma, \rho\right) = \rho, k \neq k'$. Similarly, within the same area the intercluster correlation γ, the correlation between two units in two different clusters, is $\text{cor}\left(y_{ijk}, y_{ij'k'} \,|\, \theta, \gamma, \rho\right) = \gamma, j \neq j', k \neq k'$. Here, it is ρ that makes a difference between the one-fold and two-fold models, and when ρ goes to zero, the two-fold model becomes the one-fold model, Nandram (2015).

To fit the model specified by (2.2) – (2.4), Nandram (2015) used random sampling and Gaussian quadrature to perform one-dimensional numerical integrations. He also used Gibbs sampling for comparison and found minor differences. However, our generalization to heterogeneous correlations (increased number of parameters) leads to additional weakly identified parameters and model fitting becomes more difficult. So we incorporate unimodality constraints on the prior distributions of the area parameters, thereby making it possible to analyze sparse data. To make fair comparisons between the two models, one with homogeneous correlations and the other with heterogeneous correlations, we also impose the unimodality constraints in the model specified by (2.2) – (2.4). Our results under this slightly modified homogeneous model are similar to those in Nandram (2015).

The methods introduced in this paper allow unimodality to be imposed on some distributions to assist in the estimation of weakly identified parameters. The unimodality restrictions are flexible enough to avoid over-restricting the models. For a full nonparametric Bayesian procedure, see Damien, Laud and Smith

(1997). Thus, throughout all our computations, we apply the unimodality restriction to hyperparameters of $\mu_i\,(i=1,\ldots,\ell)$,

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma},\ 0<\gamma<\frac{1}{3}. \tag{2.5}$$

We also use similar unimodality restrictions in Section 2.2 for the model with heterogeneous correlations. Henceforth, we call the model specified by (2.2) – (2.5) the HoC model.

To fit the model, Nandram (2015) use the multiplication rule by obtaining p_{ij} after drawing random samples of $(\boldsymbol{\mu}, \rho, \theta, \text{ and } \gamma)$ from their joint posterior density, where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_\ell)'$. The conditional posterior density of the p_{ij} is given by

$$p_{ij}\,|\,s_{ij},\mu_i,\rho \overset{\text{ind}}{\sim} \text{Beta}\left\{s_{ij}+\mu_i\frac{1-\rho}{\rho}, n_{ij}-s_{ij}+(1-\mu_i)\frac{1-\rho}{\rho}\right\},$$

and letting $s_{ij}=\sum_{k=1}^{n_{ij}} y_{ijk}$ and collapsing over the p_{ij}, we get

$$\pi(\boldsymbol{\mu},\rho,\theta,\gamma\,|\,\mathbf{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\frac{B\left(s_{ij}+\mu_i\frac{1-\rho}{\rho}, n_{ij}-s_{ij}+(1-\mu_i)\frac{1-\rho}{\rho}\right)}{B\left(\mu_i\frac{1-\rho}{\rho},(1-\mu_i)\frac{1-\rho}{\rho}\right)}$$

$$\times\frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma},(1-\theta)\frac{1-\gamma}{\gamma}\right)},\ 0<\mu_i,\rho<1,\quad i=1,\ldots,\ell,\ \frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma},\ 0<\gamma<\frac{1}{3}.$$

Because $T_{ij}^{(1)}\,|\,p_{ij} \overset{\text{ind}}{\sim} \text{Binomial}\left(N_{ij}-n_{ij},p_{ij}\right)$ and $T_{ij}^{(2)}\,|\,p_{ij} \overset{\text{ind}}{\sim} \text{Binomial}\left(N_{ij},p_{ij}\right)$ and, given p_{ij}, $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ are independent, once samples of the p_{ij} are obtained, it is easy to make Bayesian predictive inference. See Nandram (2015) for details.

2.2 A two-fold model with heterogeneous correlations

We extend the HoC model to accommodate the heterogeneous correlations. Our assumptions are

$$y_{ijk}\,|\,p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(p_{ij}\right), \tag{2.6}$$

$$p_{ij}\,|\,\mu_i,\rho_i \overset{\text{ind}}{\sim} \text{Beta}\left[\mu_i\frac{1-\rho_i}{\rho_i},(1-\mu_i)\frac{1-\rho_i}{\rho_i}\right], \tag{2.7}$$

$$\mu_i\,|\,\theta,\gamma \overset{\text{iid}}{\sim} \text{Beta}\left[\theta\frac{1-\gamma}{\gamma},(1-\theta)\frac{1-\gamma}{\gamma}\right], \tag{2.8}$$

$$\rho_i\,|\,\phi,\delta \overset{\text{iid}}{\sim} \text{Beta}\left[\phi\frac{1-\delta}{\delta},(1-\phi)\frac{1-\delta}{\delta}\right], \tag{2.9}$$

$$\theta,\gamma,\phi,\delta \overset{\text{iid}}{\sim} \text{Uniform}(0,1). \tag{2.10}$$

Note that the intracluster correlation coefficient ρ introduced in the HoC model is replaced by $\rho_i\,(i=1,\ldots,\ell)$ to provide the hierarchical Bayesian model with heterogeneous correlations.

Similar to the HoC model, a priori we also impose two sets of unimodality constraints,

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma},\ 0 < \gamma < \frac{1}{3} \text{ and } \frac{\delta}{1-\delta} < \phi < \frac{1-2\delta}{1-\delta},\ 0 < \delta < \frac{1}{3}. \tag{2.11}$$

Appendix B contains simple proofs of the above inequalities as unimodality criterion and how to incorporate these constraints into our computation. Henceforth, we call the hierarchical Bayesian model specified by (2.6) – (2.11) the HeC model.

Again, similar to Nandram (2015), under the HeC model, we show in Appendix A that

$$\text{cor}\left(y_{ijk}, y_{ijk'} \,\middle|\, \mu_i, \gamma, \rho_i\right) = \rho_i, \quad k \neq k', \tag{2.12}$$

$$\text{cor}\left(y_{ijk}, y_{ij'k'} \,\middle|\, \theta, \gamma, \rho_i\right) = \gamma, \quad j \neq j', \quad k \neq k'. \tag{2.13}$$

That is, within the i^{th} area, the intracluster correlation coefficient is ρ_i and the intercluster correlation coefficient is γ.

Using Bayes' theorem in the HeC model, the joint posterior density $\pi(\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y})$ is easy to write down. (This is the density without the normalization constant.) Henceforth, we would call this joint posterior density the HeC posterior.

In order to make inference about the finite population proportion, P_i, we draw samples from $\pi(\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y})$ using the multiplication rule and blocked Gibbs sampler. This procedure is described in Section 2.3.

2.3 Computations in the HeC posterior

First, note that we collapse HeC posterior over the p_{ij} and then use the Gibbs sampler to fit the joint marginal posterior density. After obtaining the samples, we can draw samples of the p_{ij} from the conditional posterior densities of the p_{ij} by applying the multiplication rule.

As in the HoC model, the conditional posterior density of p_{ij} is

$$p_{ij} \,\middle|\, \mu_i, \rho_i, \theta, \gamma, \phi, \delta, \mathbf{y} \overset{\text{ind}}{\sim} \text{Beta}\left\{s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right\}, \quad 0 < p_{ij} < 1.$$

Thus, it is easy to draw samples of the p_{ij} once samples are obtained from the joint posterior density of $(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta)$. After integrating out the p_{ij} from the HeC posterior, the marginal joint posterior density is given by

$$\pi(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)} \times \frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}, \ 0 < \mu_i, \rho_i < 1, \ i = 1, \ldots, \ell,$$

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma},\ 0 < \gamma < \frac{1}{3},\ \frac{\delta}{1-\delta} < \phi < \frac{1-2\delta}{1-\delta},\ 0 < \delta < \frac{1}{3}.$$

The conditional posterior densities are

$$\pi\left(\mu_i \,|\, \rho_i, \theta, \gamma, \phi, \delta, \mathbf{y}\right) \propto \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)} \times \mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1},$$

$$\pi\left(\rho_i \,|\, \mu_i, \theta, \gamma, \phi, \delta, \mathbf{y}\right) \propto \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)} \times \rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\delta)\frac{1-\delta}{\delta}-1},$$

and letting $G_1 = \left\{\prod_{i=1}^{\ell}\mu_i\right\}^{1/\ell}$ and $G_2 = \left\{\prod_{i=1}^{\ell}(1-\mu_i)\right\}^{1/\ell}$,

$$\pi(\theta \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \gamma, \phi, \delta, \mathbf{y}) \propto \left\{\frac{G_1^{\theta\frac{1-\gamma}{\gamma}-1} G_2^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}\right\}^{\ell},$$

and

$$\pi(\gamma \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \phi, \delta, \mathbf{y}) \propto \left\{\frac{G_1^{\theta\frac{1-\gamma}{\gamma}-1} G_2^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}\right\}^{\ell}.$$

Similarly, letting $H_1 = \left\{\prod_{i=1}^{\ell}\rho_i\right\}^{1/\ell}$ and $H_2 = \left\{\prod_{i=1}^{\ell}(1-\rho_i)\right\}^{1/\ell}$,

$$\pi(\phi \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \delta, \mathbf{y}) \propto \left\{\frac{H_1^{\phi\frac{1-\delta}{\delta}-1} H_2^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}\right\}^{\ell},$$

and

$$\pi(\delta \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \mathbf{y}) \propto \left\{\frac{H_1^{\phi\frac{1-\delta}{\delta}-1} H_2^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}\right\}^{\ell}.$$

The problem with this procedure is that θ and γ are correlated because intuitively they both depend on only $\{\mu_i\}$ through two numbers, G_1 and G_2, not the data, $\mathbf{y}$. This gives poor mixing in the Gibbs sampler. For instance, $E\left(\mu_i \,|\, \theta, \gamma\right) = \theta$, $\text{Std}\left(\mu_i \,|\, \theta, \gamma\right) = \theta\sqrt{\gamma(1-\theta)/\theta}$ and $\mu_i \approx \theta\left\{1 + z_i\sqrt{\gamma(1-\theta)/\theta}\right\}$, where $E\left(z_i\right) = 0$ and $\text{Var}\left(z_i\right) = 1$, Nandram (2015). That is, $\{\mu_i\}$ is correlated with θ and γ. Similar problems

occur in $(\boldsymbol{\rho}, \phi, \delta)$. Therefore, in order to solve these weak identifiability problems, we use the blocked Gibbs sampler to draw random samples of $(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta)$.

The blocked Gibbs sampler is obtained by drawing from the conditional posterior density $(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ and $(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$ each in turn until convergence as we describe below. The two joint conditional posterior densities are

$$\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}, \; 0<\mu_i<1, \quad i=1,\ldots,\ell, \; \frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3}$$

and

$$\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}, \; 0<\rho_i<1, \quad i=1,\ldots,\ell, \; \frac{\delta}{1-\delta}<\phi<\frac{1-2\delta}{1-\delta}, \; 0<\delta<\frac{1}{3}.$$

To run the blocked Gibbs sampler, we apply the multiplication rule in $\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ and $\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$; see, for example, Molina et al. (2014) and Toto and Nandram (2010).

First, we consider $\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$. We integrate out $\boldsymbol{\mu}$ and obtain the joint conditional posterior density of (θ, γ) given $\boldsymbol{\rho}, \phi, \delta$ and $\mathbf{y}$,

$$p(\theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell}\left\{\int_0^1\left[\prod_{j=1}^{m_i}\frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}\right]\right.$$

$$\left.\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)} d\mu_i\right\}, \; 0<\mu_i<1, \quad i=1,\ldots,\ell,$$

$$\frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3}.$$

Here, the middle Riemann sum method is used to integrate out all $\mu_i, i=1,\ldots,\ell$. We partition the interval (0, 1) into G subintervals $(a_0, a_1], (a_1, a_2], \ldots, [a_{G-1}, a_G]$, where $a_0 = 0$, $a_i = i/G$, $i = 1,\ldots G$. Then we can compute the joint conditional posterior distribution of (θ, γ), as follows.

$$p(\theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell}\left[\lim_{G\to\infty}\sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F_1(a_{v-1})-F_1(a_v)\}\right],$$

$$g_i(\mu_i) = \prod_{j=1}^{m_i}\frac{B\left(s_{ij}+\mu_i\frac{1-\rho_i}{\rho_i}, n_{ij}-s_{ij}+(1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

and $F_1(\cdot)$ is the cdf corresponding to $f_1(\cdot)$ which is a density function of $\text{Beta}\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)$. Next, we also integrate out θ by using Gaussian quadrature via Legendre orthogonal polynomials,

$$p(\gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \approx \sum_{g=1}^{G}\omega_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_1\left(\mu_i, x_g, \gamma \,\middle|\, \rho_i, \phi, \delta, \mathbf{y}\right)d\mu_i\right\},$$

where $\{\omega_g\}$ are the weights and $\{x_g\}$ are roots of the Legendre polynomial with the interval $\left[\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma}\right]$. We have taken $G = 20$ in our computations (larger values of G make little difference).

Now, we can use a univariate grid method (e.g., Molina, Nandram and Rao 2014 and Toto and Nandram 2010) in order to draw samples of the posterior density of γ conditional on $\boldsymbol{\rho}, \phi, \delta$ and $\mathbf{y}$; see Ritter and Tanner (1992) for a description of the griddy Gibbs sampler. Then, conditional on γ, we get the posterior density of θ, as follows,

$$p(\theta \,|\, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \approx \sum_{g=1}^{G}\omega_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_1\left(\mu_i, \theta \,|\, \gamma, \rho_i, \phi, \delta, \mathbf{y}\right)d\mu_i\right\}.$$

Samples are obtained from the conditional posterior density of θ by using the univariate grid sampler again. Subsequently, conditional on (θ, γ), $\boldsymbol{\mu}$ is drawn from $p(\boldsymbol{\mu} \,|\, \theta, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ using the univariate grid sampler.

For the grid method, we divide the unit interval into sub-intervals of 0.01 width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the mid-points of these sub-intervals. Note that a uniform jittering is done within each selected interval to allow different deviates with probability one (Nandram 2015). Even when we used finer sub-intervals (e.g., using 0.005 width), the inference results turned out to be almost same. Thus, we use the sub-intervals of 0.01 width; see Molina et al. (2014). When most of the distribution is near one of the boundaries (e.g., 0 or 1), we make intervals with much smaller widths to capture small or large values of the parameter.

Second, we consider $\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$. We integrate out $\boldsymbol{\rho}$ and obtain the joint conditional posterior density of (ϕ, δ) given $\boldsymbol{\mu}, \theta, \gamma$ and $\mathbf{y}$,

$$p(\phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell}\left\{\int_0^1\left[\prod_{j=1}^{m_i}\frac{B\left(s_{ij}+\mu_i\frac{1-\rho_i}{\rho_i}, n_{ij}-s_{ij}+(1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}\right]\right.$$

$$\left.\times\frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}\right\},\ 0<\rho_i<1,\quad i=1,\ldots,\ell,\ \frac{\delta}{1-\delta}<\phi<\frac{1-2\delta}{1-\delta},\ 0<\delta<\frac{1}{3}.$$

Again, we apply the middle Riemann sum method to integrate out all $\rho_i, i = 1,\ldots,\ell$ and compute the joint conditional posterior distribution of (ϕ, δ),

$$p(\phi, \delta|\boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell}\left[\lim_{G\to\infty}\sum_{v=1}^{G} h_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F_2(a_{v-1}) - F_2(a_v)\}\right],$$

where

$$h_i(\rho_i) = \prod_{j=1}^{m_i}\frac{B\left(s_{ij} + \mu_i\frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

and $F_2(\cdot)$ is the cdf corresponding to $f_2(\cdot)$ which is a density function of $\text{Beta}(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta})$. Using Gaussian quadrature via Legendre orthogonal polynomials, we can integrate out ϕ and obtain the conditional posterior density of δ,

$$p(\delta|\boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \approx \sum_{g=1}^{G}\omega'_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_2\left(\rho_i, x'_g, \delta\middle|\mu_i, \theta, \gamma, \mathbf{y}\right)d\rho_i\right\},$$

where $\{\omega'_g\}$ are the weights and $\{x'_g\}$ are roots of the Legendre polynomial with the interval $[\frac{\delta}{1-\delta}, \frac{1-2\delta}{1-\delta}]$.

Then, we use the univariate grid method in order to draw samples of the posterior density of δ conditional on $\boldsymbol{\mu}, \theta, \gamma$ and $\mathbf{y}$. Therefore, the conditional posterior density of ϕ can be represented as

$$p(\phi|\delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \approx \sum_{g=1}^{G}\omega'_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_2(\rho_i, \phi|\delta, \mu_i, \theta, \gamma, \mathbf{y})d\rho_i\right\},$$

and we can get samples of θ by using the univariate grid sampler again. Finally, conditional on (ϕ, δ), $\boldsymbol{\rho}$ can be drawn from $p(\boldsymbol{\rho}|\boldsymbol{\mu}, \theta, \gamma, \phi, \delta, \mathbf{y})$, where we also use the univariate grid method.

This algorithm samples $\pi_1(\boldsymbol{\mu}, \theta, \gamma|\boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ by first drawing an iterate from $\pi_1(\gamma|\boldsymbol{\rho}, \phi, \delta, \mathbf{y})$, an iterate from $\pi_1(\theta|\gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ and then an iterate from $\pi_1(\boldsymbol{\mu}|\theta, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$. Then, it samples $\pi_2(\boldsymbol{\rho}, \phi, \delta|\boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$ by first drawing an iterate from $\pi_2(\delta|\boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$, an iterate from $\pi_2(\phi|\delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$ and then an iterate from $\pi_2(\boldsymbol{\rho}|\phi, \delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$. The entire procedure continues until convergence. It is like using a Gibbs sampler with two conditional posterior densities which is, in fact, the blocked Gibbs sampler. The construction of the blocked Gibbs sampler is very efficient and it is one of our key contributions in this paper. In fact, we might call the blocked Gibbs sampler the blocked griddy Gibbs sampler (Ritter and Tanner 1992).

We have monitored the convergence of the blocked Gibbs sampler using trace plots, autocorrelation plots and Geweke test of stationarity. The trace plot, iterates versus time, gives information about how long a burn-in period is required to remove the effect of initial values. The autocorrelation plots display dependence in the chain, and thus, in the plots high correlations between long lags indicate a poor mixing chain. The Geweke test compares the means from the early and latter part of the Markov chain by using a $z-$score statistic, where the null hypothesis is that the chain is stationary; the $p-$values are all larger than 0.10. We have used the trace plots, autocorrelation plots, and Geweke test for each parameter to study convergence of each run of the blocked Gibbs sampler. For our data, we draw 2,000 samples and burn in

1,000 in order to obtain a sample of 1,000 iterates for inference. This burn-in period, which is based on the trace plots and Geweke test, is long enough to get random samples. The correlations are all nonsignificant, and interestingly, we do not have to thin the iterates. Also, Geweke test demonstrates stationarity of our sampler. Thus, we have a highly efficient blocked Gibbs sampler. The procedure takes a few minutes on R. We have applied the same procedure in our simulation study.

3 Numerical study and comparisons

In this section, we perform empirical studies to assess the performance of the HeC model that we compare with the HoC model. In Section 3.1, we discuss an illustrative example and, in Section 3.2, we present a simulation study.

3.1 An illustrative example

We use data from the Third Grade US population; see Nandram (2015) for a brief discussion of these data. The dataset, collected in 1999, consists of 2,477 students who participated in the Third International Mathematics and Science Study (TIMSS). Foy, Rust, and Schleicher (1996) described the probability proportional to size (PPS) systematic sampling design used in TIMSS data collection and Caslyn, Gonzales and Frase (1999) gave highlights from TIMSS. Areas are formed crossing four regions (Northeast, South, Central and West) and three communities of the US (village or rural area, outskirts of a town or city and close to the center of a town or city). Thus, there are twelve areas. The binary variable is whether a student's mathematics score is below average. Clusters are schools while units within the clusters are the students.

To assess the quality of the Bayesian predictive inference, as suggested by a referee, Nandram (2015) took a half sample of the original data, which he called a synthetic sample. The original sample was used as the population, and the half sample was used for analysis, thereby providing a method to assess the predictive power of the models in Nandram (2015). In the current paper, as suggested by a referee, we do not use a half sample and we use the original dataset available to us; see Table 3.1 for the entire dataset which we analyze in this paper. The predictive power of the HeC model is assessed mainly through the simulation study.

Unfortunately, as in many complex surveys, the sample fractions are unknown to secondary data analysts. However, typically for many of these complex surveys, the sample fractions are relatively small. For the TIMSS data we assume that the dataset is a 5% sample of the population. For example, if there are four sampled schools for an area, say i^{th} area $(i = 1,\ldots,\ell)$, the total number of clusters, M_i is assumed to be 80. If there are 17 observed students within a sampled school, say j^{th} school, the total number of students, $N_{ij}\,(j = 1,\ldots,m_i)$ is assumed to be 340. For the nonsampled schools, $N_{ij}\,(j = m_i + 1,\ldots,M_i)$ is assumed to be the average of the total number of students within the sampled schools for each area. Moreover, there are many schools in which all or many students were either below or above average. In other words, this dataset is far sparse, thereby making direct estimation difficult.

Table 3.1
Number of US students below average in mathematics within schools by area

Area	(*s*, *n*)	*m*	Schools															
NR	40	4	9	10	11	10												
	74		17	16	21	20												
NO	60	9	8	7	12	3	12	8	7	1	2							
	173		20	21	17	19	16	25	22	14	19							
NC	135	11	9	20	1	22	20	11	26	10	1	12	3					
	222		15	23	16	25	22	25	27	19	16	22	12					
SR	84	8	6	14	14	9	14	10	12	5								
	140		16	21	16	14	23	19	22	9								
SO	164	16	14	9	12	10	18	11	3	0	13	9	13	8	11	10	19	4
	298		19	14	13	18	22	18	21	16	18	15	26	9	19	22	25	23
SC	150	13	16	11	13	6	8	9	13	6	11	15	15	18	9			
	225		16	13	17	16	19	16	18	12	19	16	19	21	23			
CR	17	2	7	10														
	39		16	23														
CO	59	7	13	11	5	15	3	2	10									
	140		22	18	9	19	24	23	25									
CC	145	14	21	1	12	9	12	13	16	13	7	12	7	8	4	10		
	259		21	26	22	13	16	18	21	18	17	18	17	19	16	17		
WR	54	7	13	11	4	2	7	11	6									
	118		15	19	10	16	16	20	22									
WO	117	13	8	11	15	9	7	10	1	15	14	9	7	6	5			
	224		13	13	25	16	20	12	20	18	20	17	17	17	16			
WC	331	31	9	17	10	12	15	15	8	22	20	7	18	7	13	15	13	8
			6	8	17	13	9	6	12	7	11	4	9	8	2	3	7	
	515		18	22	10	14	15	15	8	23	22	7	18	10	26	29	13	17
			16	14	18	15	13	23	21	26	16	11	14	14	17	15	15	

Note: (*s*, *n*) represent *s* (top), the number of students scoring below average and *n* (bottom) the sample size [e.g., NR has 74 students sampled from $m = 4$ schools with a total of 40 students scoring below average]. The areas formed by crossing region (N: north, S: south, C: central, W: west) and community (R: rural, O: outskirt of a town or city, C: town or city).

We perform three goodness-of-fit procedures, the deviance information criterion (DIC), the Bayesian posterior predictive $p-$value (BPP) and the log pseudo marginal likelihood (LPML), which is a measure based on the same cross-validation (leave-one-out) procedure. We can assess the overall fit of the models with these procedures.

In the HeC model, $s_{ij} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Binomial}(n_{ij}, p_{ij})$, $p_{ij} \overset{\text{ind}}{\sim} \text{Beta}(\mu_i(1-\rho_i)/\rho_i, (1-\mu_i)(1-\rho_i)/\rho_i)$. Thus, by integrating out the p_{ij}, we can obtain the following beta-binomial probability mass function,

$$f(\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\rho}) = \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\binom{n_{ij}}{s_{ij}}\frac{B\left(s_{ij}+\mu_i(1-\rho_i)/\rho_i, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho_i)/\rho_i\right)}{B(\mu_i(1-\rho_i)/\rho_i, (1-\mu_i)(1-\rho_i)/\rho_i)}.$$

It is also true that $E\left(s_{ij} \,|\, \mu_i, \rho_i\right) = n_{ij}\mu_i$ and $\text{Var}\left(s_{ij} \,|\, \mu_i, \rho_i\right) = n_{ij}\left\{1+\left(n_{ij}-1\right)\rho_i\right\}\mu_i(1-\mu_i)$.

Let $\mu_i^{(h)}$ and $\rho_i^{(h)}$ $(i = 1,\ldots,\ell, h = 1,\ldots,H)$ denote the iterates from the blocked Gibbs sampler. Let $\bar{\mu}_i = \sum_{h=1}^{H} \mu_i^{(h)} / H$ $(i = 1,\ldots,\ell)$ and $\bar{\rho}_i = \sum_{h=1}^{H} \rho_i^{(h)} / H$. Letting $D(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\rho}}) = -2\log\{p(\mathbf{s}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\rho}})\}$ and $\bar{D} = -2\sum_{h=1}^{H} \log\{p(\mathbf{s}|\boldsymbol{\mu}^{(h)}, \boldsymbol{\rho}^{(h)})\}/H$, deviance information criterion is given by

$$\text{DIC} = 2\bar{D} - D(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\rho}}).$$

Models with smaller DIC are more preferred over those with larger DIC. However, since DIC tends to select over-fitted models, Nandram (2015) described the Bayesian predictive $p-$values as a backup. For the HeC model, the discrepancy function is

$$T(\mathbf{s}; \boldsymbol{\mu}, \boldsymbol{\rho}) = \sum_{i=1}^{\ell}\sum_{j=1}^{m_i} \frac{\{s_{ij} - E(s_{ij}|\mu_i, \rho_i)\}^2}{\text{Var}(s_{ij}|\mu_i, \rho_i)}.$$

Let $\mathbf{s}^{(\text{rep})}$ denote repeated (rep) samples from the posterior predictive distribution of $\mathbf{s}$. Then the BPP is $p\{T(\mathbf{s}^{(\text{rep})}; \boldsymbol{\mu}, \boldsymbol{\rho}) \geq T(\mathbf{s}^{(\text{obs})}; \boldsymbol{\mu}, \boldsymbol{\rho})|\mathbf{s}\}$, which is calculated over its corresponding iterates $(\boldsymbol{\mu}^{(h)}, \boldsymbol{\rho}^{(h)})$, $h = 1,\ldots, H$. If the value of this probability is close to 0 or 1, it indicates poor fit of the model. In fact, models with BPPs in (0.05, 0.95) are considered reasonable.

In addition to these quantities, we can evaluate the goodness-of-fit of models with another measure, the LPML which is a summary statistic of the conditional predictive ordinate (CPO) values, and it is based on a cross validation. Unlike the DIC, larger values of LPML indicate better fitting models (e.g., Geisser and Eddy 1979).

For the HeC model, the CPO can be estimated by

$$\widehat{\text{CPO}}_{ij} = \left[\frac{1}{H}\sum_{h=1}^{H} \frac{1}{f(s_{ij}|p_{ij}^{(h)})}\right]^{-1}, \quad j = 1,\ldots,m_i, \quad i = 1,\ldots,\ell,$$

where $p_{ij}^{(h)}$ is the samples from $p_{ij}|s_{ij}, \mu_i, \rho_i$ and $s_{ij}|p_{ij} \overset{\text{iid}}{\sim} \text{Binomial}(n_{ij}, p_{ij})$. Note that for each (i, j), $\widehat{\text{CPO}}_{ij}$ is the harmonic mean of the likelihoods $f(s_{ij}|p_{ij}^{(h)}), h = 1,\ldots,H$. Then, the LPML is

$$\text{LPML} = \sum_{i=1}^{\ell}\sum_{j=1}^{m_i} \log\left(\widehat{\text{CPO}}_{ij}\right).$$

These three model evaluation measures have similar forms under the HoC model. For the HoC (HeC) model, DIC = 774.421 (773.173), BPP = 0.349 (0.408), LPML = -352.064 (-346.171), thereby indicating that the HeC model gives a better fit. At a finer level, we also looked at the individual CPO values from the two models for each school. In Figure 3.1 we compare the CPOs from the HeC and the HoC models, and we found that generally CPO values for the HeC model are higher that those of HoC model. In fact, under the HoC (HeC) model we found that the percent of the CPOs less than 0.025 is 3.70% (2.96%) and percent of the CPOs less than 0.014 is 0.74% (0.00%). These results do not show any indication of serious departure from model assumptions; see Ntzoufras (2009). Therefore, these measures give prima facie evidence that the HeC model fits the TIMSS data somewhat better than the HoC model.

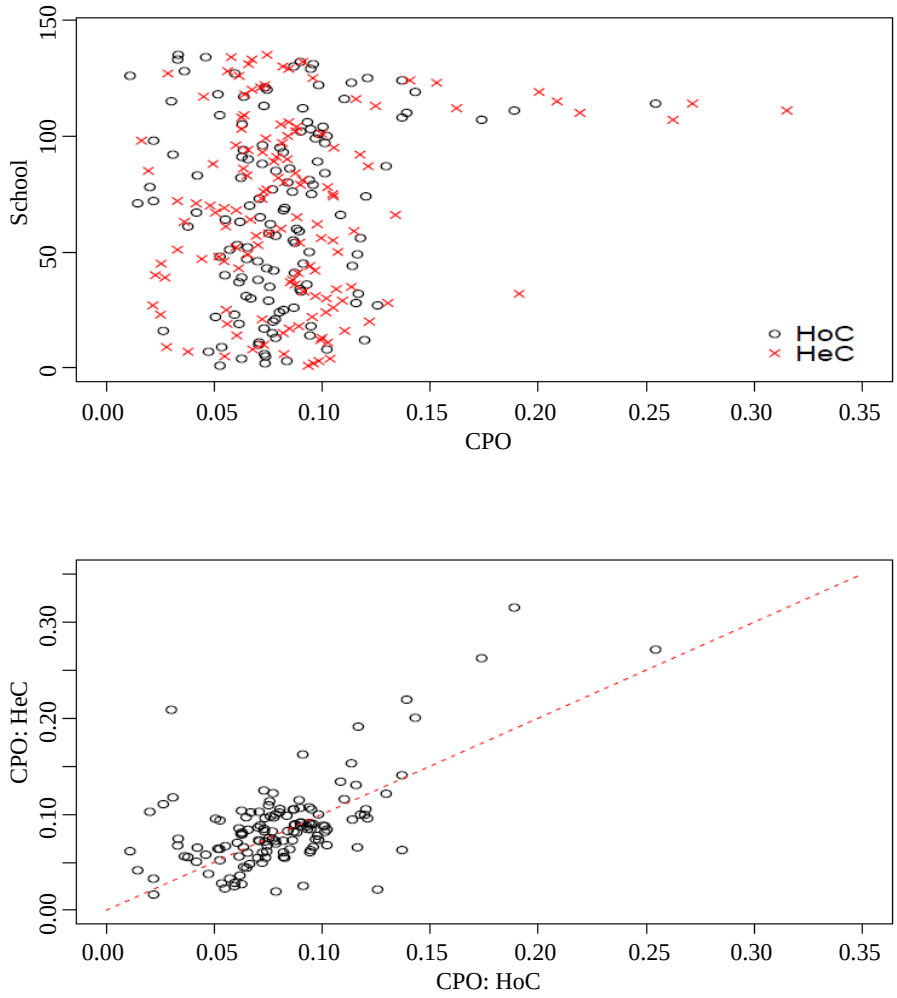

Figure 3.1 Scatter plots of the CPOs from the HoC and HeC models (Top panel: School vs. CPO) and (Bottom panel: HeC vs. HoC).

Now, consider inference about θ and γ. First, consider θ. Under the HoC model, the posterior mean (PM) is 0.519, posterior standard deviation (PSD) is 0.068 and 95% credible interval (Cre) is (0.390, 0.639). Under the HeC model PM = 0.515, PSD = 0.065 and 95% Cre is (0.383, 0.639). Second, consider γ. Under the HoC model PM = 0.207, PSD = 0.011, and 95% Cre is (0.190, 0.224). Under the HeC model γ PM = 0.208, PSD = 0.011 and 95% Cre is (0.190, 0.225). Thus, it is good that inference about θ and γ are very close for the two competitors (HoC and HeC models).

In Table 3.2, we present posterior inference about the finite population proportions for mathematics scores by areas. There are differences between the posterior means under the HoC and HeC models. Most of them are small but there are a few large differences. For NC, SR and CR, we have 0.560 (0.543), 0.568 (0.584) and 0.465 (0.445) under the HoC (HeC) model, respectively. The posterior standard deviations are also close but there are a few moderately large differences (e.g., for NR we have 0.113 under the HoC model and 0.077 under the HeC model). These differences are reflected in the credible and highest posterior density (HPD) intervals.

Table 3.2
Comparison of posterior inference from the two-fold models with homogeneous correlation (HoC) and heterogeneous correlations (HeC) for the finite population proportions for US students below average in mathematics by area

Area	HoC Model				HeC Model			
	PM	PSD	95% Cre	95% HPD	PM	PSD	95% Cre	95% HPD
NR	0.522	0.113	(0.299, 0.735)	(0.310, 0.741)	0.525	0.077	(0.363, 0.662)	(0.361, 0.658)
NO	0.365	0.075	(0.227, 0.524)	(0.227, 0.520)	0.359	0.072	(0.228, 0.511)	(0.236, 0.516)
NC	0.560	0.070	(0.420, 0.701)	(0.408, 0.680)	0.543	0.082	(0.370, 0.695)	(0.396, 0.710)
SR	0.568	0.080	(0.405, 0.725)	(0.424, 0.731)	0.584	0.062	(0.454, 0.699)	(0.456, 0.699)
SO	0.537	0.058	(0.423, 0.648)	(0.417, 0.639)	0.537	0.063	(0.409, 0.655)	(0.408, 0.653)
SC	0.646	0.064	(0.552, 0.766)	(0.522, 0.766)	0.654	0.059	(0.521, 0.763)	(0.544, 0.774)
CR	0.465	0.137	(0.195, 0.719)	(0.185, 0.709)	0.445	0.125	(0.212, 0.716)	(0.199, 0.700)
CO	0.437	0.085	(0.279, 0.603)	(0.276, 0.596)	0.439	0.091	(0.257, 0.620)	(0.265, 0.620)
CC	0.549	0.064	(0.415, 0.671)	(0.423, 0.672)	0.550	0.066	(0.414, 0.681)	(0.422, 0.685)
WR	0.461	0.086	(0.297, 0.629)	(0.295, 0.626)	0.460	0.085	(0.289, 0.626)	(0.276, 0.611)
WO	0.516	0.066	(0.384, 0.643)	(0.387, 0.644)	0.516	0.058	(0.401, 0.626)	(0.409, 0.633)
WC	0.670	0.042	(0.581, 0.748)	(0.586, 0.749)	0.662	0.047	(0.569, 0.748)	(0.568, 0.746)

Note: PM is the posterior mean, PSD is the posterior standard deviation, Cre is the equal-tail credible interval, and HPD is highest posterior density interval.

Table 3.3 shows summaries of the PM, PSD and 95% HPD for intracluster correlations under the HeC model. We can see that the intracluster correlations vary over the areas. The largest estimate is 0.337 for NC and the smallest one is 0.073 for SR. Both areas have a few large difference between the posterior means under the HoC and HeC models. The 95% HPD interval for the common correlation in the HoC model is (0.160, 0.260) and this interval is contained by all the intervals except for NR, NC, SR and WC. Thus, it is reasonable to study the HeC model.

Table 3.3
Posterior summaries for the intracluster correlations of the two-fold models with heterogeneous correlations for US students below average in mathematics by area

Area	PM	PSD	95% Cre	95% HPD
NR	0.076	0.084	(0.002, 0.301)	(0.001, 0.251)
NO	0.184	0.087	(0.053, 0.380)	(0.042, 0.358)
NC	0.337	0.087	(0.190, 0.520)	(0.184, 0.513)
SR	0.073	0.067	(0.003, 0.252)	(0.001, 0.216)
SO	0.237	0.075	(0.113, 0.393)	(0.110, 0.387)
SC	0.176	0.079	(0.055, 0.356)	(0.048, 0.329)
CR	0.149	0.147	(0.003, 0.523)	(0.001, 0.445)
CO	0.233	0.103	(0.079, 0.486)	(0.050, 0.434)
CC	0.235	0.077	(0.105, 0.388)	(0.099, 0.381)
WR	0.181	0.099	(0.033, 0.413)	(0.021, 0.378)
WO	0.181	0.075	(0.059, 0.362)	(0.048, 0.327)
WC	0.301	0.063	(0.191, 0.437)	(0.188, 0.434)

Note: Using the two-fold model with homogeneous correlation, PM = 0.211, PSD = 0.026, 95% Cre = (0.162, 0.266), and 95% HPD = (0.160, 0.260). PM is the posterior mean, PSD is the posterior standard deviation, Cre is the equal-tail credible interval, and HPD is highest posterior density interval.

In Figure 3.2, we compare the posterior densities of the intracluster correlations (twelve correlations) from the HeC model and the HoC model (one correlation). The distributions under the HeC model are more variable and are mostly to the left or right of those under the HoC model with not much overlap for some areas (e.g., NR, NC and SR).

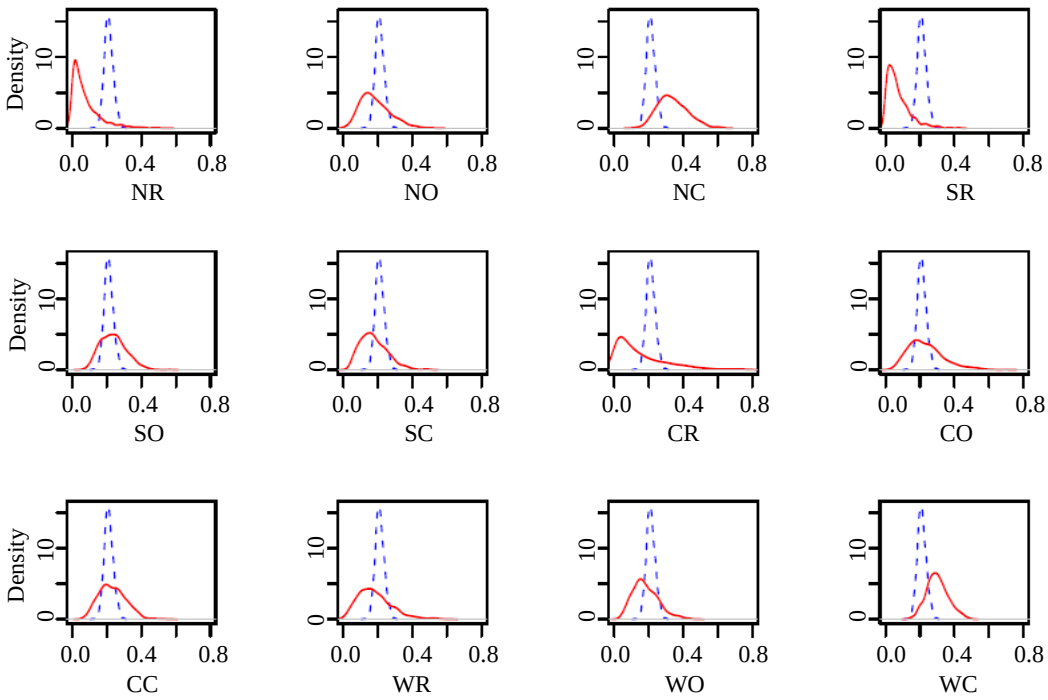

Figure 3.2 Plots of the posterior densities of intracluster correlations for mathematics scores by areas (solid: HeC model, dotted: HoC model).

In Figures 3.3, 3.4 and 3.5, we compare the posterior density plots of the finite population proportions for the mathematics score and all areas for the two models. There are noticeable differences between the HoC and HeC models (e.g., areas NR, NC, SR, CR and WC).

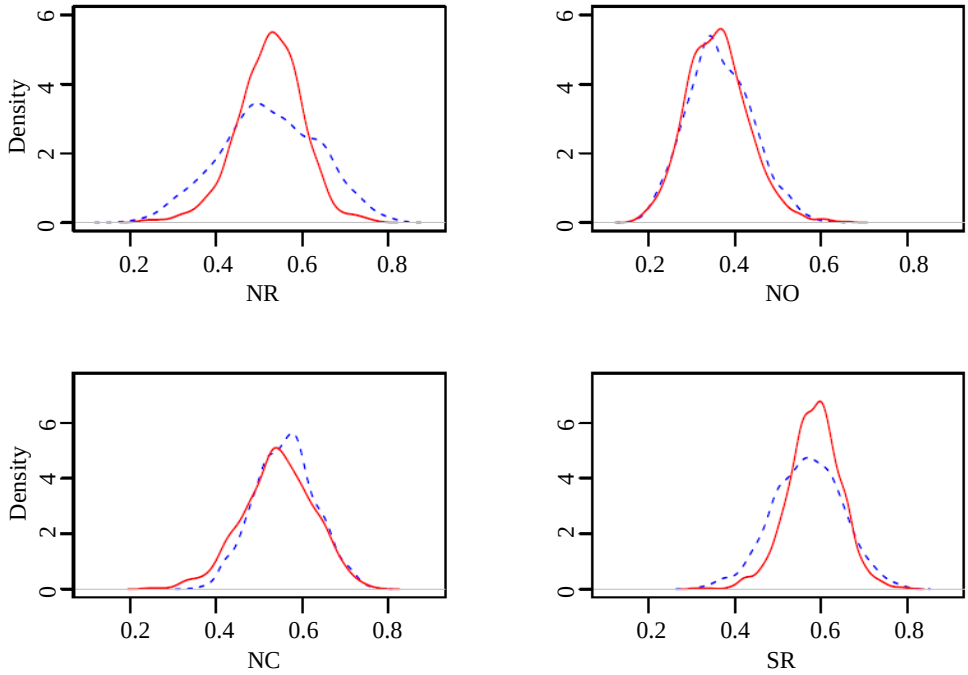

Figure 3.3 Plots of the posterior densities of finite population proportions for mathematics scores by areas (NR, NO, NC, SR) (solid: HeC model, dotted: HoC model).

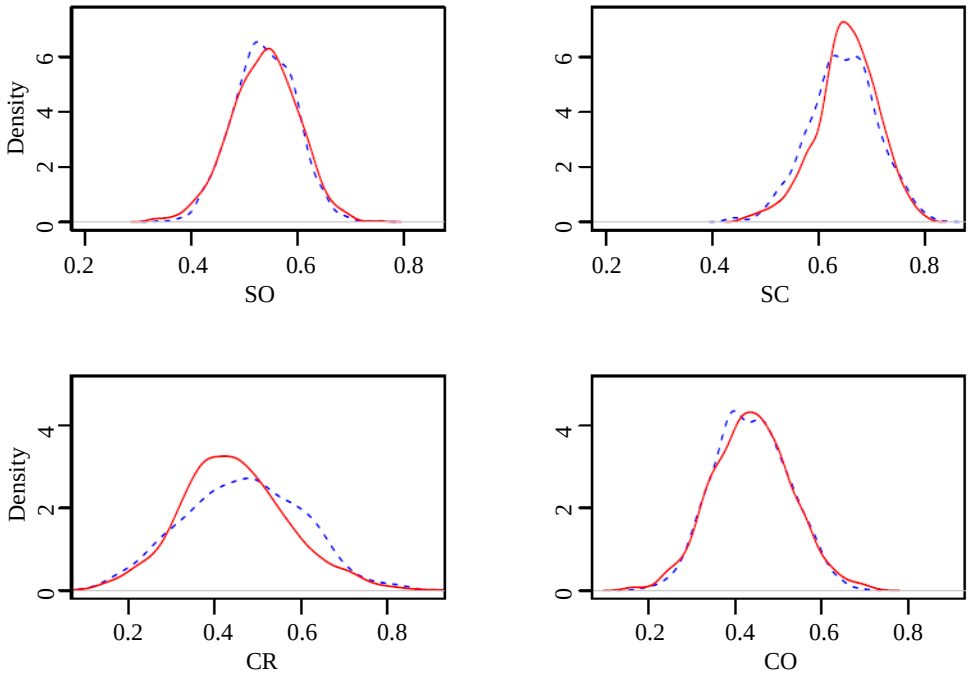

Figure 3.4 Plots of the posterior densities of finite population proportions for mathematics scores by areas (SO, SC, CR, CO) (solid: HeC model, dotted: HoC model).

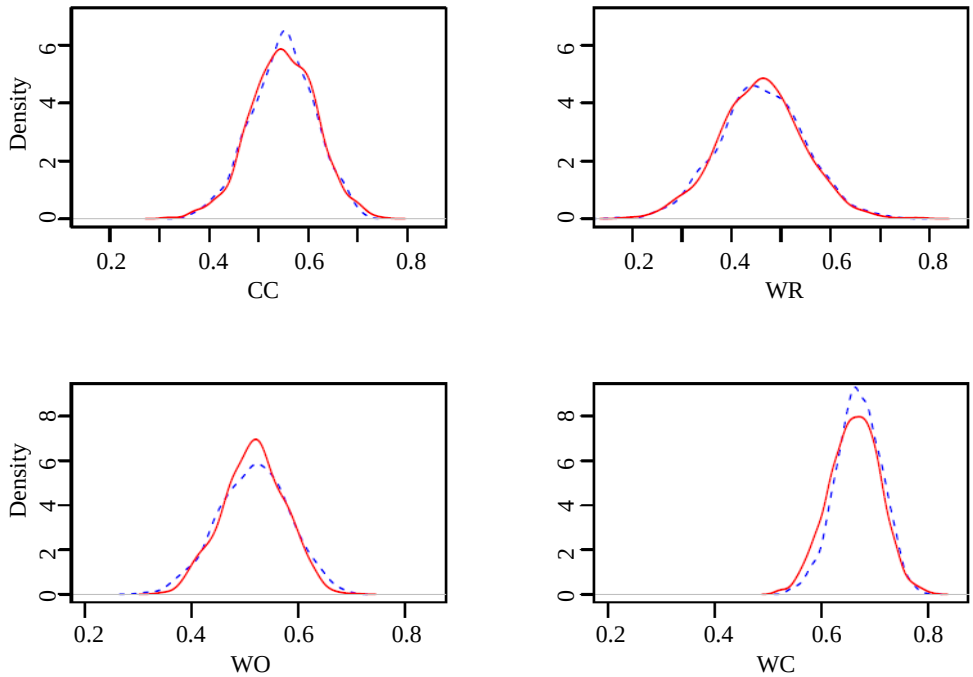

Figure 3.5 Plots of the posterior densities of finite population proportions for mathematics scores by areas (CC, WR, WO, WC) (solid: HeC model, dotted: HoC model).

3.2 Simulation study

In order to further assess the performance of the HeC model and to compare it to the HoC model, we perform a simulation study. Here we use two factors, each at three levels, to get nine design points.

We have set 100 as the number of clusters (schools) in each area and 15 as the number of individuals (students) within each cluster. In other words, we take $N_{ij} = 15$, $j = 1,\ldots,M_i$, $M_i = 100$, $i = 1,\ldots,\ell$ where $\ell = 12$. Let $\mathbf{a}$ denote a vector of posterior means and $\mathbf{b}$ denote the vector of posterior standard deviations corresponding to the μ_i or the ρ_i. Specifically, for the ρ_i we use $\mathbf{a}_1$ and $\mathbf{b}_1$ and for the μ_i we use $\mathbf{a}_2$ and $\mathbf{b}_2$. When we simulate data from the HeC model, the levels of the ρ_i are $(1{:}\,\mathbf{a}_1 - 0.5\mathbf{b}_1; 2{:}\ \mathbf{a}_1; 3{:}\,\mathbf{a}_1 + 0.5\mathbf{b}_1)$ and the levels of the μ_i are $(1{:}\,\mathbf{a}_2 - 0.5\mathbf{b}_2; 2{:}\,\mathbf{a}_2; 3{:}\,\mathbf{a}_2 + 0.5\mathbf{b}_2)$. For the twelve areas $\mathbf{a}_1$ takes values 0.09, 0.19, 0.32, 0.08, 0.22, 0.18, 0.15, 0.22, 0.23, 0.17, 0.18, 0.30; $\mathbf{b}_1$ 0.08, 0.09, 0.08, 0.06, 0.07, 0.08, 0.13, 0.09, 0.07, 0.09, 0.07, 0.06; $\mathbf{a}_2$ 0.53, 0.37, 0.54, 0.58, 0.54, 0.65, 0.46, 0.44, 0.55, 0.46, 0.52, 0.66; and $\mathbf{b}_2$ 0.08, 0.08, 0.08, 0.06, 0.06, 0.06, 0.12, 0.09, 0.07, 0.08, 0.06, 0.05.

We also take a simple random sample of five clusters among the 100 population clusters, and a simple random sample of ten individuals from each sampled cluster (i.e., $m_i = 5$ and $n_{ij} = 10$). These numbers are much smaller than those of data used in Section 3.1, which makes inference a little more challenging (Nandram 2015). Note that the dataset has about 7% of the sampled clusters where all students were either below or above average. We call this quantity the percent of sparseness. The setting of this simulation study also leads to even sparser data. For nine design points, all the average percents of sparseness are greater than 7% and most are around 10%. Figure 3.6 shows the histograms of sparseness percents for each design point.

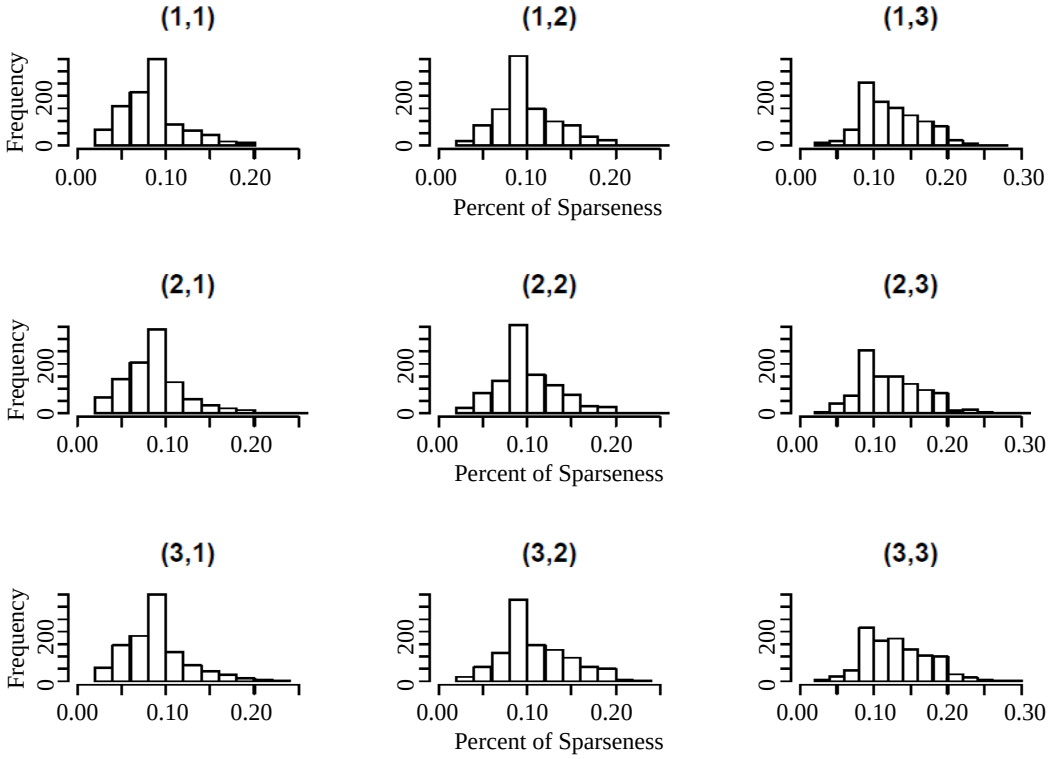

Figure 3.6 Histograms of the percent of sparseness when data are drawn from the HeC model by design point [(*i*, *j*): *i*, *j* = 1, 2, 3] in which the first factor corresponds to ρ and the second factor to μ.

We consider two scenarios. In the first scenario, we generate data from the HeC model and fit both models, and in the second scenario, we generate data from the HoC model and fit both models. When data are simulated from the HeC model, we have nine design points $[(1,1),(1,2),(1,3),\ldots,(3,1),(3,2),(3,3)]$,

the first factor corresponds to the ρ_i. When we simulate data from the HoC model, we have three design points $(1: \mathbf{a}_2 - 0.5\mathbf{b}_2; 2: \mathbf{a}_2; 3: \mathbf{a}_2 + 0.5\mathbf{b}_2)$ for the three levels for the μ_i; ρ is kept fixed at its posterior mean.

In the first scenario, at each design point we simulate binary data from the HeC model,

$$p_{ij} \mid \mu_i, \rho_i \overset{\text{ind}}{\sim} \text{Beta}\left[\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right],$$

$$y_{ijk} \mid p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}), \quad k = 1, \ldots, N_{ij}.$$

So we have the true values of $P_i = \sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}} y_{ijk} \Big/ \sum_{j=1}^{M_i} N_{ij}$ for $i = 1, \ldots, \ell$. We take 1,000 samples at each of the nine design points. For each sample we perform the blocked griddy Gibbs sampler in the same manner as for the data.

Like Nandram (2015), we calculate $\text{AB}_{ih} = \left|\text{PM}_{ih} - P_{ih}\right|$, $\text{RAB}_{ih} = \text{AB}_{ih} / P_{ih}$ and $\text{RPMSE}_{ih} = \sqrt{\text{PSD}_{ih}^2 + \text{AB}_{ih}^2}$ to study the frequentist properties of our procedure $(i = 1, \ldots, \ell,\ h = 1, \ldots, 1{,}000)$. We also obtain the 95% credible interval and HPD interval for each of the 1,000 simulated runs, and we study the width W_{ih} and the credible incidence I_{ih}. If the 95% credible (or HPD) interval of h^{th} run contains the true value P_i, I_{ih} is equal to one, otherwise it is equal to zero. Thus, the estimated probability content of the 95% credible interval for the i^{th} area is $C_i = \sum_{h=1}^{1{,}000} I_{ih} / 1{,}000$.

Table 3.4 shows comparison of the HoC and HeC models. Under the HeC model the coverages are much higher than those under the HoC model. Note that the coverages of HPD intervals for the HeC model are much closer to the nominal value of 95% and they are conservative. However, the 95% credible and HPD intervals are wider than those from the HoC model. These effects are much larger as $\boldsymbol{\rho}$ becomes larger. All measures AB, RAB and RPMSE under the HeC model are smaller than those under the HoC model. Thus, based on these measures, the HeC model is preferred over the HoC model.

Table 3.4
Simulation under the HeC model: Comparison of the HeC and HoC models using mean coverage and widths of 95% credible intervals and absolute bias, relative absolute bias and root posterior mean squared error for finite poulation proportions by design point

Design Point	Model	C-Cre	W-Cre	C-HPD	W-HPD	AB	RAB	RPMSE
(1,1)	HeC	0.989	0.620	0.961	0.603	0.112	0.227	0.206
	HoC	0.930	0.555	0.893	0.541	0.130	0.266	0.207
(1,2)	HeC	0.984	0.622	0.960	0.603	0.112	0.227	0.206
	HoC	0.926	0.558	0.889	0.545	0.132	0.249	0.209
(1,3)	HeC	0.980	0.623	0.955	0.608	0 .120	0.211	0.210
	HoC	0.923	0.558	0.892	0.546	0.134	0.236	0.212
(2,1)	HeC	0.982	0.621	0.953	0.603	0.119	0.242	0.212
	HoC	0.922	0.564	0.879	0.549	0.137	0.281	0.215
(2,2)	HeC	0.980	0.625	0.952	0.609	0.122	0.228	0.214
	HoC	0.918	0.566	0.879	0.552	0.139	0.264	0.217
(2,3)	HeC	0.981	0.628	0.956	0.611	0.121	0.211	0.214
	HoC	0.930	0.570	0.895	0.556	0.135	0.239	0.214
(3,1)	HeC	0.982	0.627	0.949	0.608	0.121	0.245	0.215
	HoC	0.934	0.583	0.892	0.566	0.136	0.278	0.218
(3,2)	HeC	0.980	0.628	0.947	0.610	0.123	0.242	0.217
	HoC	0.928	0.583	0.885	0.566	0.138	0.274	0.220
(3,3)	HeC	0.976	0.632	0.951	0.614	0.124	0.218	0.218
	HoC	0.928	0.581	0.889	0.565	0.139	0.246	0.220

Note: In the design point [(*i*, *j*): *i*, *j* = 1, 2, 3], the first factor corresponds to ρ and the second factor to μ. C-Cre and C-HPD are the probability contents of a credible interval and a HPD interval; W-Cre and W-HPD are the widths of a credible interval and a HPD interval. AB, RAB and RPMSE are the absolute bias, relative absolute bias and root posterior mean squared error.

In Table 3.5 we compare summaries of DIC, BPP and LPML. All the DICs under the HeC model are smaller than the corresponding ones under the HoC model and all the LPMLs under the HeC model are larger than those under the HoC model. Under the HoC model, all the BPPs vary in (0.06, 0.09) but under the HeC model they vary in (0.2, 0.4). Again, these measures show that the HeC model is superior to the HoC model.

In a similar manner, for the second scenario we generate binary data from

$$p_{ij} \mid \mu_i, \rho \overset{\text{ind}}{\sim} \text{Beta}\left[\mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho}\right],$$

$$y_{ijk} \mid p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}), \quad k = 1, \ldots, N_{ij}.$$

In Table 3.6 we present comparison of the HoC and HeC models. Here AB, RAB and RPMSE are only slightly smaller under the HoC model. The coverages of the credible and HPD intervals under the HeC model are closer to the nominal value of 95%, while those under the HoC model are smaller. Table 3.7 shows summaries of DIC, BPP and LPML. All the DICs under the HeC model are smaller than those under the HeC model, while the BPPs and LPMLs are similar for the two models, with those under the HoC model being slightly better.

Table 3.5
Simulation under the HeC model: Comparison of the HeC and HoC models using the deviance information criterion (DIC), the Bayesian predictive *p*-value (BPP) and the log pseudo marginal likelihood (LPML) by design point

Design Point	HoC Model			HeC Model		
	DIC	BPP	LPML	DIC	BPP	LPML
(1,1)	419.275	0.090	-285.452	402.044	0.429	-267.990
(1,2)	418.351	0.091	-286.250	400.647	0.439	-266.377
(1,3)	416.784	0.088	-286.290	400.414	0.446	-267.203
(2,1)	436.980	0.067	-307.028	416.264	0.300	-292.756
(2,2)	437.306	0.062	-308.816	414.955	0.318	-292.404
(2,3)	430.531	0.080	-302.258	410.436	0.351	-285.206
(3,1)	441.204	0.090	-316.126	424.010	0.227	-308.825
(3,2)	442.165	0.083	-318.223	424.363	0.235	-309.815
(3,3)	438.305	0.071	-315.159	418.827	0.260	-306.619

Note: In the design point [(*i*, *j*): *i*, *j* = 1, 2, 3], the first factor corresponds to ρ and the second factor to μ. DIC, BPP and LPML are summarized over the 1,000 simulation runs.

Table 3.6
Simulation under the HoC model: Comparison of the HeC and HoC models using mean coverage and widths of 95% credible intervals and absolute bias, relative absolute bias and root posterior mean squared error for finite poulation proportions by design point

Design Point	Model	C-Cre	W-Cre	C-HPD	W-HPD	AB	RAB	RPMSE
1	HeC	0.985	0.627	0.969	0.608	0.117	0.242	0.212
	HoC	0.944	0.575	0.919	0.559	0.107	0.240	0.210
2	HeC	0.988	0.634	0.952	0.616	0.122	0.234	0.216
	HoC	0.938	0.585	0.917	0.568	0.115	0.214	0.211
3	HeC	0.977	0.628	0.940	0.611	0.126	0.222	0.218
	HoC	0.933	0.572	0.908	0.556	0.113	0.202	0.208

Note: The design point [*i*: *i* = 1, 2, 3], corresponds to μ with ρ held fixed at its posterior mean. C-Cre and C-HPD are the probability contents of a credible interval and a HPD interval; W-Cre and W-HPD are the widths of a credible interval and a HPD interval. AB, RAB and RPMSE are the absolute bias, relative absolute bias and root posterior mean squared error.

Table 3.7
Simulation under the HoC model: Comparison of the HeC and HoC models using the deviance information criterion (DIC), the Bayesian predictive *p*-value (BPP) and the log pseudo marginal likelihood (LPML) by design point

Design Point	HoC Model			HeC Model		
	DIC	BPP	LPML	DIC	BPP	LPML
1	428.647	0.308	-300.526	416.626	0.302	-303.001
2	430.113	0.371	-295.191	417.557	0.317	-296.531
3	429.598	0.379	-295.613	414.877	0.335	-297.250

Note: The design point [i: i = 1, 2, 3], corresponds to μ with ρ held fixed at its posterior mean. DIC, BPP and LPML are summarized over the 1,000 simulation runs.

Thus, when data actually come from the HeC model, there are some important differences among the two models, with the HeC model being preferred. However, when data actually come from the HoC model, there are minor differences between the two models. Of course, the HeC model (unequal correlations) has more parameters than the HoC model (one correlation).

4 Concluding remarks

We have extended a homogeneous two-fold model, Nandram (2015) to a heterogeneous two-fold model which adds a degree of flexibility to our data analysis. Weakly identified parameters in these models posed serious computational problems. Therefore, we have done two additional things. First, we have introduced unimodal constraints on the parameters of the beta prior distributions. Second, we have used a blocked Gibbs sampler to perform the computations. To compare these models, we have performed a Bayesian predictive inference. As an illustrative example, we have used data from TIMSS, a study of the performance of US students at the third grade in mathematics. Also, we have performed a simulation study to compare these two two-fold models even further.

It is important to model the two-fold sample design using the heterogeneous model because for many applications the intracluster correlations may vary from area to area, making the heterogeneous two-fold model more appropriate than the homogeneous two-fold model. Indeed, using an illustrative example and the simulation study with several diagnostics, we have demonstrated that the heterogeneous two-fold model is to be preferred over the homogeneous two-fold model when the correlations vary significantly.

It is possible to extend our work to accommodate multivariate binary data. This can be viewed as a problem of pooling data from multinomial distributions in order to infer about the finite population proportions. For example, in TIMSS we can use both mathematics and science scores as bivariate binary responses (correlation). Then it is possible to develop a hierarchical Bayesian model for multinomial responses and a Dirichlet prior to the model of cell probabilities. In this study we can work with two issues. First, we can investigate how much the prediction will be improved when using the multivariate data. Second, we can also study how much the precision of inference will be improved when considering a model with heterogeneous intracluster correlations over one with homogeneous correlation with respect to the multivariate data.

Acknowledgements

The authors are grateful to the two reviewers for their careful reading of the manuscript and their suggestions. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2058954). Also this work was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

Appendix A

Proofs of formulas (2.12) and (2.13)

It is easy to show that

$$\begin{aligned}
\mathrm{Cov}\left(y_{ijk}, y_{ijk'} \middle| \mu_i, \gamma, \rho_i\right) &= \mathrm{Var}\left(p_{ij} \middle| \mu_i, \gamma, \rho_i\right) = \mu_i(1-\mu_i)\rho_i, \\
\mathrm{Var}\left(y_{ijk} \middle| \mu_i, \gamma, \rho_i\right) &= E\left[\mathrm{Var}\left(y_{ijk} \middle| p_{ij}, \mu_i, \gamma, \rho_i\right)\right] + \mathrm{Var}\left[E\left(y_{ijk} \middle| p_{ij}, \mu_i, \gamma, \rho_i\right)\right], \\
&= E\left(p_{ij} \middle| \mu_i, \gamma, \rho_i\right)\left[1 - E\left(p_{ij} \middle| \mu_i, \gamma, \rho_i\right)\right] = \mu_i(1-\mu_i).
\end{aligned}$$

Thus, $\mathrm{Cor}\left(y_{ijk}, y_{ijk'} \middle| \mu_i, \gamma, \rho_i\right) = \rho_i \, (k \neq k')$, thereby proving (2.12).

Similarly, it is easy to show that

$$\begin{aligned}
\mathrm{Cov}\left(y_{ijk}, y_{ij'k'} \middle| \theta, \gamma, \rho_i\right) &= E\left[\mathrm{Cov}\left(p_{ij}, p_{ij'} \middle| \mu_i, \theta, \gamma, \rho_i\right)\right] \\
&\quad + \mathrm{Cov}\left[E\left(p_{ij} \middle| \mu_i, \theta, \gamma, \rho_i\right), E\left(p_{ij'} \middle| \mu_i, \theta, \gamma, \rho_i\right)\right] \\
&= \mathrm{Var}\left(\mu_i \middle| \theta, \gamma\right) = \theta(1-\theta)\gamma, \\
\mathrm{Var}\left(y_{ijk} \middle| \theta, \gamma, \rho_i\right) &= E\left(\mu_i \middle| \theta, \gamma\right) - \left[E\left(\mu_i \middle| \theta, \gamma\right)\right]^2 = \theta(1-\theta).
\end{aligned}$$

Thus, $\mathrm{Cor}\left(y_{ijk}, y_{ij'k'} \middle| \theta, \gamma, \rho_i\right) = \gamma \, (j \neq j', k \neq k')$, thereby proving (2.13).

Appendix B

Computation with unimodality constraints

It is well known that a beta pdf with parameters α and β is unimodal if $\alpha > 1$ and $\beta > 1$. This can be established easily using calculus. In our case, $\mu | \theta, \gamma \sim \mathrm{Beta}\left\{\theta\frac{(1-\gamma)}{\gamma}, (1-\theta)\frac{(1-\gamma)}{\gamma}\right\}$. Therefore, we have two inequalities,

$$\theta\frac{(1-\gamma)}{\gamma} > 1 \text{ and } (1-\theta)\frac{(1-\gamma)}{\gamma} > 1,$$

and simple algebra gives

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}, \quad 0 < \gamma < \frac{1}{3}.$$

Next, we describe briefly how to apply these constraints to the computation in the two-fold model with heterogeneous correlations. Recall the conditional marginal posterior distribution of γ,

$$p(\gamma|\boldsymbol{\rho},\phi,\delta,\mathbf{y}) \approx \sum_{g=1}^{G} \omega_g\, p\left(x_g,\gamma|\boldsymbol{\rho},\phi,\delta,\mathbf{y}\right),$$

where $\{\omega_g\}$ are the weights and $\{x_g\}$ are roots of the Legendre polynomial. Here, we use the univariate grid method to sample γ. So, we divide the interval $(0,\frac{1}{3})$, the first constraint, into G1 subintervals $[\gamma_0,\gamma_1),[\gamma_1,\gamma_2),\ldots,[\gamma_{G1-1},\gamma_{G1}]$. For a uniform random number, u^*, from any grid, say, $[\gamma_{\nu-1},\gamma_\nu)$, we compute the height, i.e., the value of the conditional marginal posterior density function of γ as

$$\frac{1-3u^*}{1-u^*}\sum_{g=1}^{G^*}\omega_g^*\, p\left(x_g^*,u^*|\boldsymbol{\rho},\phi,\delta,\mathbf{y}\right),$$

where $\{\omega_g^*\}$ are the weights and $\{x_g^*\}$ are roots of the Legendre polynomial with the interval $\left[\frac{u^*}{1-u^*},\frac{1-2u^*}{1-u^*}\right]$, the second constraint. Similarly, we can apply unimodality criterion to sample (ϕ,δ).

References

Brier, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67, 591-595.

Caslyn, C., Gonzales, P. and Frase, M. (1999). Highlights from TIMSS. *National Center for Education Statistics*, Washington, DC.

Damien, P., Laud, P.W. and Smith, A.F.M. (1997). Bayesian estimation of unimodal distributions. *Communications in Statistics*, 26(2), 429-440.

Foy, P., Rust, K. and Schleicher, A. (1996). Sample design. In *TIMSS Technical Report, Volume I: Design and Development*, (Eds., M.O. Martin and D.L. Kelly), Chestnut Hill, MA: Boston College.

Fuller, W.A., and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

Ghosh, M., and Lahiri, P. (1988). Bayes and empirical Bayes analysis in multistage sampling. In *Statistical Decision Theory and Related Topics IV*, (Eds., S.S. Gupta and J.O. Berger), New York: Springer, Vol. 1, 195-212.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics*, 8(2), 852-885.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.

Nandram, B. (2015). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics* (accepted).

Nandram, B., and Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B,* 55, 399-408.

Nandram, B., Bhatta, D., Sedransk, J. and Bhadra, B. (2013). A Bayesian test of independence in a two-way contingency table using surrogate sampling. *Journal of Statistical Planning and Inference,* 143, 1392-1408.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New Jersey: Wiley, Hoboken.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

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 multi-way tables with cell proportions estimated from survey data. *Annals of Statistics,* 12, 46-60.

Ritter, C., and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association,* 87, 861-868.

Scott, A., and Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association,* 101, 1387-1397.

Stukel, D.M., and Rao, J.N.K. (1997). Estimation of regression models with nested error regression structure and unequal variances under two and three stage cluster sampling. *Statistics & Probability Letters,* 35, 401-407.

Stukel, D.M., and Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference,* 78, 131-147.

Toto, M.C.S., and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference,* 140, 2963-2979.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 93-106
Statistics Canada, Catalogue No. 12-001-X

Sample allocation for efficient model-based small area estimation

Mauno Keto and Erkki Pahkinen[1]

Abstract

We present research results on sample allocations for efficient model-based small area estimation in cases where the areas of interest coincide with the strata. Although model-assisted and model-based estimation methods are common in the production of small area statistics, utilization of the underlying model and estimation method are rarely included in the sample area allocation scheme. Therefore, we have developed a new model-based allocation named *g*1-allocation. For comparison, one recently developed model-assisted allocation is presented. These two allocations are based on an adjusted measure of homogeneity which is computed using an auxiliary variable and is an approximation of the intra-class correlation within areas. Five model-free area allocation solutions presented in the past are selected from the literature as reference allocations. Equal and proportional allocations need the number of areas and area-specific numbers of basic statistical units. The Neyman, Bankier and NLP (Non-Linear Programming) allocation need values for the study variable concerning area level parameters such as standard deviation, coefficient of variation or totals. In general, allocation methods can be classified according to the optimization criteria and use of auxiliary data. Statistical properties of the various methods are assessed through sample simulation experiments using real population register data. It can be concluded from simulation results that inclusion of the model and estimation method into the allocation method improves estimation results.

Key Words: Optimal area sample size; Criteria; Auxiliary information; Measure of homogeneity.

1 Introduction

In this paper we present a new model-based allocation method in stratified sampling where the areas of interest coincide with the strata. Our study is focused on the components of an efficient area allocation. A clear starting point for the allocation process is reached if the areas of interest are defined as early as in the design phase of the research and if it is also known how large a sample is allowed in consideration of the disposable resources (time, budget etc.). The choice of the allocation method depends on various factors such as the selected model, estimation method, available pre-information of the population and the optimization criteria set only on area or population level, or on both levels simultaneously.

We have selected six existing allocation methods and developed a new one which we call a model-based allocation. The general properties of these methods are examined in Section 2 and Section 3. Five of these allocations can be regarded as model-free. Two of them use only number-based information, such as the number of areas and the number of basic units in each area. Three other allocations need, in addition to number-based information, area level parameter information, such as area totals, standard deviation or coefficient of variation (CV). Because this information about the study variable is not available, a common solution is to replace it with a proper proxy variable. The last of the reference allocations, introduced by Molefe and Clark (MC) (2015), is a model-assisted allocation which is based on a composite estimator and a two-level model. We have named it MC-allocation.

The optimization criteria of the five model-free allocations differ from one another. Allocations based only on area-specific numbers can be computed easily, but their choice is reasonable under limited

1. Mauno Keto, University of Jyväskylä. E-mail: mauno.j.keto@student.jyu.fi; Erkki Pahkinen, Department of Mathematics and Statistics of University of Jyväskylä. E-mail: pahkinen@maths.jyu.fi.

circumstances. In each of the parameter-based allocations the optimization criterion is different. It can be set on the level of the population parameter estimate (Neyman allocation) or on area level estimates in average (Bankier allocation). The third allocation solution, which deviates from the two former ones, is the NLP allocation, in which the tolerances of estimates are set on both population and area level.

This article starts from the assumption that if model-assisted or model-based estimation is used in a survey the model and estimation method must be taken into account when the allocation of the sample into areas is designed. This was used as a starting point when the new model-based $g1-$allocation, presented in Section 2, was derived. Also, one of the reference allocations, model-assisted allocation, is based on a given model.

The comparison of performances of different allocation methods in real situations has been implemented by using simulation experiments and is presented in Section 4. An official Finnish register of block apartments for sale serves as the population. The structure of the register is introduced in Section 4.1. An auxiliary variable has been used in place of the study variable when computing the area sample sizes for each allocation except equal and proportional allocation. The comparison demonstrates clearly that these allocations lead to different sample distributions. The same kind of variety also concerns their performances. We have applied model-based EBLUP (Empirical Best Linear Unbiased Predictor) estimation on the allocations when estimating the area totals of the study variable. For measuring and comparing the performances of allocations, a relative root mean square error RRMSE% and absolute relative bias ARB% were used.

In Section 5 empirical simulation results are discussed as concluding remarks. They support the allocation solution in which not only auxiliary information, but also the model and estimation method should be determined as early as in the design phase of a survey. A good example is the $g1-$allocation presented in Section 2.2. The most accurate area estimates of area totals were obtained by using this method.

2 Allocations which utilize the model

2.1 Choosing the model

Pfeffermann (2013) presents a wide variety of models and methods for small area estimation. Our model is one of this assortment, a unit-level mixed model

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ \ k = 1,\ldots,N_d;\ \ d = 1,\ldots,D, \tag{2.1}$$

where v_d's are random area effects with mean zero and variance σ_v^2 and e_{dk}'s are random effects with mean zero and variance σ_e^2. Furthermore, $E\left(y_{dk}\right) = \mathbf{x}'_{dk}\boldsymbol{\beta}$ and $V\left(y_{dk}\right) = \sigma_v^2 + \sigma_e^2$ (total variance). Matrix $\mathbf{V}$ is the variance-covariance matrix of the study variable y. This model can be used when unit-level values are available for the auxiliary variables $\mathbf{x}$. We use one auxiliary variable in our study.

Two important measures are needed in developing one of these types of allocations. The first one is a common intra-area correlation ρ and the second one is the ratio δ between variance components. They are defined as follows:

$$\rho = \sigma_v^2 \big/ \left(\sigma_v^2 + \sigma_e^2\right) \text{ and } \delta = \sigma_e^2 \big/ \sigma_v^2 = 1/\rho - 1. \tag{2.2}$$

Before estimating area parameters, the variance components, regression coefficients and area effects must be estimated from the sample data. The BLUE estimator (Best Linear Unbiased Estimator) of $\boldsymbol{\beta}$, noted $\tilde{\boldsymbol{\beta}}$, is obtained according to the theory of the general linear model, and it is replaced with its EBLUP estimate $\hat{\boldsymbol{\beta}}$.

The EBLUP estimate (predicted value) for the area total Y_d of the study variable is the sum of the observed y – values and predicted y – values for units outside the sample:

$$\hat{Y}_{d,\text{Eblup}} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \mathbf{x}'_{dk} \hat{\boldsymbol{\beta}} + \left(N_d - n_d\right) \hat{v}_d. \tag{2.3}$$

We use the Prasad-Rao approximation (See Rao 2003) of MSE (Mean Squared Error) for finite populations:

$$\text{mse}\left(\hat{Y}_{d,\text{Eblup}}\right) = g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + 2g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right), \tag{2.4}$$

where the four components g_{1d}, g_{2d}, g_{3d} and g_{4d} are defined as follows:

$$\begin{aligned}
g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(1 - \hat{\gamma}_d\right) \hat{\sigma}_v^2, \\
g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(\overline{\mathbf{x}}_d^* - \hat{\gamma}_d \overline{\mathbf{x}}_d\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \left(\overline{\mathbf{x}}_d^* - \hat{\gamma}_d \overline{\mathbf{x}}_d\right), \\
g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(n_d^*\right)^{-2} \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \left(n_d^*\right)^{-1}\right)^{-3} \left[\hat{\sigma}_e^4 V\left(\hat{\sigma}_v^2\right)\right. \\
&\quad \left. + \hat{\sigma}_v^4 V\left(\hat{\sigma}_e^2\right) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \,\text{Cov}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)\right], \\
g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right) \hat{\sigma}_e^2.
\end{aligned} \tag{2.5}$$

The area sample sizes n_d^* depend on the sample and are not fixed. The component g_{1d} contains the area-specific ratio $\hat{\gamma}_d = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \big/ n_d^*\right)$. According to Nissinen (2009, page 53), the g_{1d} component (later simply $g1$) contributes generally over 90% of the estimated MSE. This component represents uncertainty as regards the variation between the areas. Of course this variation must be strong enough so that such a high proportion for $g1$ exists.

Unfortunately, the idea of an analytical solution, which means minimizing the sum of MSE's over areas subject to $n = \sum_{d=1}^{D} n_d$, is difficult and laborious to accomplish because components of the MSE approximation (2.5) include sample information which is unknown, and some components contain complex matrix and variance-covariance operations. We have examined this allocation problem for the first time in an experimental study (Keto and Pahkinen 2009). Now we have developed an allocation based only on the component $g1$ and auxiliary variable x. The reasoning for this solution is that because x and y are correlated, the between-area variation in x is transferred to y.

2.2 Model-based $g1-$allocation

The $g1-$allocation utilizes the auxiliary variable x and the adjusted homogeneity coefficient (Keto and Pahkinen 2014). This coefficient is an approximation of an intra-class correlation (ICC) known of cluster sampling. We regard one area as one cluster. First, simple ANOVA between areas is carried out, and then the adjusted homogeneity measure of variation between the areas can be computed:

$$R_{ax}^2 = 1 - R^2(x) = 1 - \mathrm{MSW}/S_x^2, \tag{2.6}$$

where $R^2(x)$ is the coefficient of determination from regression analysis, MSW (Mean Square within) is the mean SS (Sum of Squares) of areas and S_x^2 is the variance of the auxiliary variable x.

Because MSE of the area total is complex, we use only the component $g1$, which appears in (2.4) and (2.5), for the reason we have given in Section 2.1. We search for the minimum for the sum of $g1$'s over areas:

$$\sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)^2 (n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} \tag{2.7}$$

subject to $n = \sum_{d=1}^{D} n_d$.

We use Lagrange's multiplier method to find the solution. Therefore, we define the function F of sample sizes $\mathbf{n}' = (n_1, n_2, \ldots, n_D)$ and λ:

$$F(\mathbf{n}, \lambda) = \sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)^2 (n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} + \lambda\left(\sum_{d=1}^{D} n_d - n\right). \tag{2.8}$$

We set the derivative of F with respect to the area sample size n_d to zero and solve for n_d. The expression for area sample size n_d^{g1} is as follows:

$$n_d^{g1} = \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta = \frac{N_d n - (N - N_d D - n)(1/\rho - 1)}{N + D(1/\rho - 1)}, \tag{2.9}$$

where the ratio δ and the intra-area correlation ρ are defined in (2.2). The only unknown member in (2.9) is the intra-area correlation ρ. Therefore we substitute the known homogeneity measure (2.6) of the auxiliary variable x for ρ. Thus the final expression for computing area sample sizes is

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}. \tag{2.10}$$

It is easy to prove that $\sum_{d=1}^{D} n_d^{g1} = n$. The computed sample sizes are rounded to the nearest integer. Sometimes compromises must be made. It can be concluded by the examination of (2.10) that the sample size increases when the size of area N_d increases, but not proportionally. Under certain circumstances, such as low homogeneity coefficient, low overall sample size n or small size of area, N_d can lead to negative area sample size n_d^{g1}. In this situation the negative value is changed to zero. A special case occurs if the total variation is only between areas causing value one to the measure of homogeneity (2.6), and (2.10) is reduced to proportional allocation.

2.3 Model-assisted MC-allocation

Molefe and Clark (2015) have used the following composite estimator for estimating the mean of the study variable y for area d:

$$\tilde{y}_d^C = (1-\varphi_d)\,\bar{y}_{dr} + \varphi_d\,\hat{\boldsymbol{\beta}}'\bar{\mathbf{X}}_d. \tag{2.11}$$

This estimator is a combination of two estimators: the synthetic estimator $\hat{\bar{Y}}_{d(\text{syn})} = \hat{\boldsymbol{\beta}}'\bar{\mathbf{X}}_d$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient and $\bar{\mathbf{X}}_d$ is the area population means of auxiliary variables $\mathbf{x}$, and a direct estimator $\bar{y}_{dr} = \bar{y}_d + \hat{\boldsymbol{\beta}}'(\bar{\mathbf{x}}_d - \bar{\mathbf{X}}_d)$, where $\bar{y}_d$ and $\bar{\mathbf{x}}_d$ are the area d sample means of y and $\mathbf{x}$. We use one auxiliary variable in our study. The coefficients φ_d are set with the intent to minimize the MSE of the estimator (2.11). The approximated design-based MSE of the estimator under certain conditions and assumptions is given by the expression

$$\text{MSE}_p\left(\tilde{y}_d^C;\bar{Y}_d\right) \approx (1-\varphi_d)^2\, v_{d(\text{syn})} + \varphi_d^2 B_d^2, \tag{2.12}$$

where $v_{d(\text{syn})}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{d(\text{syn})}$ and $B_d = \boldsymbol{\beta}_U'\,\bar{\mathbf{X}}_d - \bar{Y}_d$ is the bias when $\hat{\bar{Y}}_{d(\text{syn})}$ is used to estimate $\bar{Y}_d$, with $\boldsymbol{\beta}_U$ denoting the approximate design-based expectation of $\hat{\boldsymbol{\beta}}$.

The population contains N units and D strata defined by areas, and stratified sampling is used. A random sample SRSWOR (Simple Random Sampling without Replacement) of n_d units is selected from stratum $d\,(d=1,\ldots,D)$ containing N_d units. The relative size of area d is $P_d = N_d/N$.

A two-level linear model ξ conditional on the values of $\mathbf{x}$ is assumed, with uncorrelated stratum random effects u_d and random effects ε_i:

$$\left.\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_d + \varepsilon_i \\ E_\xi(u_d) &= E_\xi(\varepsilon_i) = 0 \\ V_\xi(u_d) &= \sigma_{ud}^2 \\ V_\xi(\varepsilon_i) &= \sigma_{ed}^2 \end{aligned}\right\}, \tag{2.13}$$

where i refers to all units in stratum d. This model implies that $V_\xi(y_i) = \sigma_{ud}^2 + \sigma_{ed}^2$ for all population units and $\text{cov}_\xi(y_i, y_j)$ equals $\rho_d\,\sigma_d^2$ for units $i \neq j$ in the same stratum and zero for units from different strata, where $\rho_d = \sigma_{ud}^2/(\sigma_{ud}^2 + \sigma_{ed}^2)$. A simplifying assumption that $\rho_d = \rho$ are equal for all strata is defined.

After making some other simplifying assumptions and solving the optimal weight φ_d in (2.12), the final approximate optimum anticipated MSE or approximate model assisted mean squared error is obtained of (2.12):

$$\text{AMSE}_d = E_\xi \text{MSE}_p\left(\tilde{y}_d^C\left[\varphi_{d(\text{opt})}\right];\bar{Y}_d\right) \approx \sigma_d^2\rho(1-\rho)\left[1+(n_d-1)\rho\right]^{-1}. \tag{2.14}$$

Next the criterion F using anticipated MSE's of the small area mean and overall mean estimators for model-assisted allocation is defined and developed into the final approximative form:

$$\begin{aligned} F &= \sum_{d=1}^{D} N_d^q \text{AMSE}_d + GN_+^{(q)} E_\xi \text{var}_p\left(\hat{\bar{Y}}_r\right) \\ &\approx \sum_{d=1}^{D} N_d^q\,\sigma_d^2\,\rho(1-\rho)\left[1+(n_d-1)\rho\right]^{-1} + GN_+^{(q)}\sum_{d=1}^{D}\sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \end{aligned} \tag{2.15}$$

Optimal sample sizes for the areas are obtained by minimizing (2.15) subject to $\sum_d n_d = n$. Expression (2.15) follows the idea of Longford (2006). The weight N_d^q reflects the inferential priority (importance) for area d, with $0 \le q \le 2$, and $N_+^{(q)} = \sum_{d=1}^{D} N_d^q$. The quantity G is a relative priority coefficient on the population level. Ignoring the goal of estimating the population mean corresponds to $G = 0$, and the attention is then only focused on area level estimation. On the other hand, the larger the value of G, the more the second component in (2.15) dominates and the more the area level estimation is ignored.

We assume first that the population estimation has no priority $(G = 0)$ and the unit survey cost are fixed. In this case minimization of (2.15) with respect of n_d has a unique solution

$$n_{d,\text{opt}} = \frac{n\sqrt{\sigma_d^2 N_d^q}}{\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_d^2 N_d^q}}{D^{-1}\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} - 1\right). \tag{2.16}$$

The formula (2.16) contains two unknown parameters, the intra-class correlation ρ and the area-specific variance σ_d^2. We replace ρ with an adjusted homogeneity coefficient of the auxiliary variable x. This coefficient is an approximation of the ICC (Intra-Class Correlation) (Section 2.2). Parameter σ_d^2 is replaced with the variance of x in area d. The reason for both replacements is that y is correlated with x. If also the population estimation has a priority $(G > 0)$ then (2.16) does not apply and F must be minimized numerically by using, for example, the NLP method, as we have done (Excel Solver, NLP option).

Table 2.1
Summary of model-based and model-assisted allocations

Method	**Computing sample size n_d for area d**	**Optimality level**
Model-based $g1$	$n_d^{g1} = \dfrac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}$, where R_{ax}^2 is the adjusted homogeneity measure of auxiliary variable x.	Area
Model-assisted MCG0	$n_{d,\text{opt}} = \dfrac{n\sqrt{\sigma_d^2 N_d^q}}{\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} + \dfrac{1-\rho}{\rho}\left(\dfrac{\sqrt{\sigma_d^2 N_d^q}}{D^{-1}\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} - 1\right)$	Jointly area and population
MCG50	Minimization of $F = \sum_{d=1}^{D} N_d^q \sigma_d^2 \rho(1-\rho)\left[1 + (n_d - 1)\rho\right]^{-1} + GN_+^{(q)}\sum_{d=1}^{D}\sigma_d^2 P_d^2 n_d^{-1}(1-\rho)$ with respect of n_d. Parameter ρ is replaced with R_{ax}^2 and σ_d^2 with $S_d^2(x)$.	

3 Some model-free area allocations

The aim of this section is to list the five previously presented allocation methods in order to use them later as references. Depending on which kind of auxiliary information each one uses, they are divided into two groups: number-based and parameter-based allocations.

3.1 Number-based allocations

Two basic allocation solutions commonly used go under the names equal allocation and proportional allocation. Neither of these allocations contains any specific criterion on the area or population level. Their implementation requires only information on the number of strata D and the numbers of units N_d in each stratum.

In the equal area allocation the sample size n_d is simply a quotient

$$n_d^{\text{Equ}} = n/D. \tag{3.1}$$

It is recommended to choose the total sample size n so that the quotient is a whole number. This allocation method does not take differences between the areas into account in any way, which results in inaccurate area estimates. A natural lower limit of the sample size is min $n = 2D$.

Proportional allocation is a frequently used basic method. Area sample sizes are solved from

$$n_d^{\text{Pro}} = n\left(N_d/N\right). \tag{3.2}$$

If the sizes of the areas vary strongly, it can lead to situations where the allocated sample size $n_d^{\text{Pro}} < 2$ for one or more areas. This is an obstacle in calculating direct design-based estimates of standard errors. One solution is to apply the combined allocation proposed by Costa, Satorra and Ventura (2004). The idea is a weighted solution between the equal and proportional allocation depending on the situation. The combined area sample size is

$$n_d^{\text{Com}} = kn_d^{\text{Pro}} + (1-k)\,n_d^{\text{Equ}} \tag{3.3}$$

for a specified constant $k\,(0 \le k \le 1)$. A minor problem is present if for some areas $n/D > N_d$. A modified solution exists for this case.

3.2 Parameter-based allocations

These allocations use area-level information of the study variable y and in some cases of the auxiliary variable x correlated with y. The values of x are available for all population units. In practice the unknown y is replaced with a proper proxy variable y^* such as a study variable obtained from an earlier research of the same subject, or the values of y^* are generated with a suitable model developed of a small pre-sample. Also x can be substituted for y. Allocation criteria can be set on population level, only on area level or on combined population and area level.

The Neyman allocation aims at reaching an optimal accuracy concerning population parameters $\text{SD}(y)_d$ (Tschuprow 1923). The standard deviation of the study variable y or some proxy variable and the number of units in each area must be known. Allocation favors large areas with strong variation.

The Bankier or power allocation (1988) is based on a criterion set on the area level. Area CV values of y are weighted by area total transformations X_d^q which contain a tuning constant q. In practice y^* or x must be used in place of y. Allocation favors mainly large areas with high CV.

Choudhry, Rao and Hidiroglou (2012) present the NLP allocation method for direct estimation. This method uses non-linear programming to find a solution. Criteria for the allocation are defined by setting

upper limits for CV values of the study variable y in each area and in the population. In practice y^* or x replaces y. The program searches the minimum sample size $n = \sum_d n_d$ satisfying these conditions. The SAS (Statistical Analysis System) procedure NLP with Newton-Raphson option was used to find the solution. The allocation favors areas with high CV regardless of the area size N_d.

A summary of the model-free allocations and the formulas for calculating area sample sizes are presented in Table 3.1.

Table 3.1
Summary of number-based and parameter-based allocations

Allocation	Computing area sample size n_d	Optimality level
Equal	$n_d^{\text{Equ}} = n/D$	Area
Proportional	$n_d^{\text{Pro}} = n(N_d/N)$	Population
Neyman	$n_d^{\text{Ney}} = n\left(N_d S_d \Big/ \sum_{d=1}^{D} N_d S_d\right)$, where S_d is the standard deviation of y (in practise y^* or x) in area d.	Population
Bankier	$n_d^{\text{Ban}} = n\left(X_d^q \text{CV}(y)_d \Big/ \sum_{d=1}^{D} X_d^q \text{CV}_d(y)\right)$, where X_d is the area total of x, $\text{CV}_d(y) = S_d/\overline{Y}_d$ and q is a tuning constant. In practise y^* or x replace y.	Area
NLP	$n_{st}^{\text{NLP}} = \min\left(\sum_{d=1}^{D} n_d\right)$ satisfying tolerances $\text{CV}(\overline{y}_d) \le \text{CV}_{0d}$ and $\text{CV}(\overline{y}_{st}) \le \text{CV}_0$. In practise y^* or x replace y.	Jointly population and area

Some other parameter-based allocation methods are mentioned briefly. For example Longford (2006) introduced inferential priorities P_d for the strata d and G for the population and used those constraints for allocation. Another solution is presented by Falorsi and Righi (2008). This solution does not contain a direct imposition of quotas, but tries to solve the comprehensive collection of data by using a multi-stage sampling design, so that the area estimation can be implemented effectively.

4 Comparison of performances of allocations

In this section we study the performances of the allocation methods introduced in Sections 2 and 3. The estimated parameters are area and population totals of the study variable y. The overall sample size $n = 112$. Section 4.1 includes the description of the research data. Simulation experiments and comparisons of allocations are presented in Section 4.3.

4.1 Empirical data

Our research data is obtained from a national Finnish register of block apartments for sale. This register is maintained by a private company, Alma Mediapartners Ltd, whose customers are real estate agencies. They save all the necessary information of the apartments into this register as soon as they receive an assignment from the owners. The population we have used consists of 9,815 block apartments (these serve as sampling units) for sale selected from the register. They represent 14 Finnish districts, mainly towns, in spring 2011. The sizes of the smallest and largest area were 112 and 1,333, respectively. The study variable (y) measures the apartment price (1,000 €) and the auxiliary variable (x) measures the size (m^2). Area sizes (N_d), population summary statistics (totals, means, standard deviations and CVs) for y and x, as well as correlations between x and y, are given in Table 4.1. The characteristics of the areas have a wide range. The most diverging area is Helsinki.

Table 4.1
Population summary statistics

Area		Study variable y				Auxiliary variable x				Correlation
Label	N_d	Y_d	$\bar{Y}_d$	$S_d(y)$	$\mathrm{CV}_d(y)$	X_d	$\bar{X}_d$	$S_d(x)$	$\mathrm{CV}_d(x)$	r_{yx}
Porvoo town	112	25,409	226.86	207.82	0.916	8,940	79.82	50.67	0.635	0.877
Pirkkala district	148	30,323	204.88	87.82	0.429	11,149	75.33	23.78	0.316	0.823
South Savo county	493	64,863	131.57	72.90	0.554	32,644	66.22	20.25	0.306	0.437
Jyväskylä town	494	89,941	182.07	69.65	0.383	40,000	80.97	17.62	0.218	0.509
Lappi county	555	62,143	111.97	50.15	0.448	30,805	55.50	16.22	0.292	0.207
South-East Finland	585	98,504	168.38	106.78	0.634	47,750	81.62	21.68	0.266	0.601
Helsinki (capital)	621	437,902	705.16	562.38	0.798	76,931	123.88	57.98	0.468	0.753
West coast district	655	108,339	165.40	75.85	0.459	50,903	77.71	36.39	0.468	0.439
Trackside district	818	148,845	181.96	65.08	0.358	59,220	72.40	23.84	0.321	0.517
Kuopio district	871	126,867	145.66	75.79	0.520	64,103	73.60	23.27	0.324	0.580
Turku district	958	166,613	173.92	131.62	0.757	79,970	83.48	25.71	0.308	0.635
Oulu district	1,072	133,591	124.62	50.19	0.403	59,210	55.23	16.92	0.306	0.392
Metropol area	1,100	263,293	239.36	117.84	0.492	80,034	72.76	26.37	0.362	0.754
Lahti-Tampere distr.	1,333	262,400	196.85	110.76	0.563	105,804	79.37	25.54	0.322	0.602
Population	9,815	2,019,031	205.71	215.52	1.048	747,462	76.16	31.76	0.417	0.674

The adjusted measure of homogeneity of the auxiliary variable x is $R_{ax}^2 = 0.231$ indicating quite strong variability between the areas.

4.2 Allocations

In general, the overall sample size depends on the available time and financial resources in the research project. This aspect has not been taken into account now, because it is a question of an experimental study.

The value of the sampling ratio was determined as $f\,(\%) = 100 \times (112/9{,}815) = 1.14\%$. Method-specific allocations were produced according to the formulas presented in Table 2.1 and Table 3.1. Some details have been taken into account. In the Bankier allocation the value of a tuning constant q is 0.5. In the NLP allocation the selected CV limits 0.1258 (12.58%) for areas and the CV limit 0.0375 (3.75%) for the population lead to the overall sample size 112. We use the Excel Solver procedure with non-linear option for solving the NLP allocation problem. We use a modified proportional allocation to obtain an area sample size which is at least two. First we allocated one unit for every area and then allocated the rest 98 units by using proportionality. We have substituted x for y in every parameter-based allocation. In the model-assisted allocations the value of q was set to 1, and the quantity G was set to zero and 50. The final sample sizes in each allocation are presented in Table 4.2. The variation of sample sizes on area level is very strong between the allocations.

Table 4.2
Area sample sizes by allocation

Area		Model-based	Composite estim. Model-assisted		Number-based allocations		Parameter-based allocations		
Label	N_d	$g1^*$	**MCG0***	**MCG50***	**EQU**	**PRO**	**Ney_X**	**Ban_X**	**NLP_X**
Porvoo town	112	0	6	3	8	2	2	6	20
Pirkkala district	148	0	2	2	8	2	2	4	6
South Savo county	493	5	4	4	8	6	4	6	6
Jyväskylä town	494	5	3	4	8	6	4	5	3
Lappi county	555	6	3	4	8	6	4	5	5
South-East Finland	585	6	6	5	8	7	6	6	4
Helsinki (capital)	621	7	21	16	8	7	16	14	14
West coast district	655	7	12	11	8	8	10	11	14
Trackside district	818	10	8	8	8	9	9	8	7
Kuopio district	871	11	8	9	8	10	9	8	6
Turku district	958	12	10	11	8	11	11	9	6
Oulu district	1,072	13	6	8	8	12	8	8	6
Metropol area	1,100	13	11	12	8	12	13	11	8
Lahti-Tampere district	1,333	17	12	15	8	14	14	11	7
Total	9,815	112	112	112	112	112	112	112	112

* based on the adjusted coefficient of homogeneity (value 0.231) computed of x.

4.3 Comparison of performances of allocations

In this section we present the results based on design-based simulation experiments. For each allocation, 1,500 independent stratified SRSWOR samples were simulated with the SAS program and necessary calculations from the simulated samples were implemented with SPSS (Statistical Package for the Social Sciences) program. We have applied model-based EBLUP estimation on the samples for each allocation. For comparison of the allocations, we have computed two quality measures: $\mathrm{RRMSE}_d\%$ and $\mathrm{ARB}_d\%$ for each allocation.

Assume that r simulated samples are drawn in each allocation, and let $\hat{Y}_{di,\mathrm{EBLUP}}$ be the EBLUP estimate of the area total Y_d in the i^{th} sample $(i = 1,\ldots,r)$. Then $\mathrm{RRMSE}_d\%$ and $\mathrm{ARB}_d\%$ are defined as

$$\text{RRMSE}_d\,\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} - Y_d\right)^2} \Big/ Y_d\,,$$
$$\text{ARB}_d\,\% = 100 \times \left| 1/D \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} \Big/ Y_d - 1\right) \right|,$$

and their means over areas are computed as follows:

$$\text{MRRMSE}\% = 1/D \sum_{d=1}^{D} \text{RRMSE}_d\,\% \quad \text{and} \quad \text{MARB}\% = 1/D \sum_{d=1}^{D} \text{ARB}_d\,\%.$$

The estimate for the population total in the i^{th} simulated sample $(i = 1, \ldots, r)$ is the sum of the estimates of the area totals: $\hat{Y}_{i,\,\text{EBLUP}} = \sum_{d=1}^{D} \hat{Y}_{di,\,\text{EBLUP}}$. RRMSE% for the population total is computed as

$$\text{RRMSE}_{\text{pop}}\,\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} - Y\right)^2} \Big/ Y\,,$$

where Y is the true value of the population total, for which ARB% is computed as

$$\text{ARB}_{\text{pop}}\,\% = 100 \times \left| 1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} \Big/ Y - 1\right) \right|.$$

Tables 4.3 and 4.4 contain RRMSE% and ARB% values for areas, their means over areas and population RRMSE%s and ARB%s in each allocation. The evaluation of the results was based on two arguments. One was the mean value of the quality measure on the area level and the other was the value of the quality measure on the population level.

Table 4.3
Area and population RRMSE%s by allocation

Area	N_d	$g1$	MCG0	MCG50	EQU	PRO	Ney_X	Ban_X	NLP_X
Porvoo town	112	8.08	14.63	15.93	13.41	19.79	16.49	14.78	10.10
Pirkkala district	148	6.60	9.72	10.77	8.35	12.04	10.60	9.76	8.97
South Savo county	493	22.29	22.77	23.20	18.63	20.70	23.20	20.16	20.88
Jyväskylä town	494	15.36	24.55	20.70	13.61	14.43	20.83	18.33	21.98
Lappi county	555	21.72	28.19	26.19	19.91	21.34	25.45	23.97	22.59
South-East Finland	585	20.76	27.25	25.93	19.68	19.64	24.37	24.31	27.81
Helsinki (capital)	621	22.72	12.68	14.97	21.92	23.15	14.35	16.02	16.43
West coast district	655	21.15	22.43	21.57	20.35	19.92	21.75	20.67	18.91
Trackside district	818	11.93	12.86	13.63	12.31	11.38	13.73	12.76	13.47
Kuopio district	871	16.22	23.22	20.70	19.21	16.37	20.84	20.82	23.49
Turku district	958	17.56	24.75	21.66	20.94	17.74	21.57	22.70	26.44
Oulu district	1,072	14.39	25.40	21.14	16.96	14.34	21.22	19.00	19.81
Metropol area	1,100	9.59	11.31	10.86	12.14	9.78	10.16	10.78	11.55
Lahti-Tampere distr.	1,333	10.54	13.43	11.66	13.35	10.64	12.76	12.87	14.98
Mean over areas (%)		15.65	19.51	18.59	16.48	16.52	18.38	17.64	18.39
Population value (%)		6.15	6.53	5.88	6.13	5.97	6.07	5.89	6.62

The lowest RRMSE% mean over the areas (15.65%) was obtained in the $g1-$allocation developed in this study. Helsinki was an exception on area level because its RRMSE% value was clearly higher compared

with model-assisted and parameter-based allocations. Also equal and proportional allocations performed well on area level, with means 16.48% and 16.52%. The highest means were obtained in the model-assisted MC-allocations. On the population level, the lowest value for the quality measure was obtained in the model-assisted MCG50-allocation (5.88%) and the second lowest value in the Bankier allocation (5.89%), but in general, differences between the allocations on this level were small.

Table 4.4
Area and population ARB%s by allocation

Area	N_d	$g1$	MCG0	MCG50	EQU	PRO	Ney_X	Ban_X	NLP_X
Porvoo town	112	2.28	2.20	0.97	0.04	1.26	1.28	0.98	0.79
Pirkkala district	148	0.17	2.10	1.08	0.19	0.79	0.85	0.86	1.15
South Savo county	493	8.08	11.81	10.87	6.76	7.29	11.47	9.09	9.81
Jyväskylä town	494	6.09	19.78	15.36	6.10	5.82	14.33	12.16	16.31
Lappi county	555	2.08	5.27	3.14	1.45	2.70	2.44	1.22	1.44
South-East Finland	585	9.05	20.62	18.28	9.53	8.11	15.69	15.96	20.41
Helsinki (capital)	621	9.71	6.38	7.93	10.95	11.59	7.43	8.80	9.45
West coast district	655	7.83	12.34	11.60	9.07	8.16	12.69	10.52	10.87
Trackside district	818	1.21	3.11	1.78	1.76	0.96	2.61	2.10	2.94
Kuopio district	871	6.00	14.90	10.68	9.37	6.53	11.33	11.77	15.56
Turku district	958	5.26	16.46	12.59	8.48	5.78	11.54	13.27	16.91
Oulu district	1,072	0.81	10.17	6.08	1.88	1.84	6.47	4.71	4.00
Metropol area	1,100	3.06	5.84	5.11	5.29	3.37	4.39	5.12	5.76
Lahti-Tampere distr.	1,333	1.86	6.14	3.97	3.62	1.79	4.65	4.37	6.10
Mean over areas (%)		4.53	9.79	7.82	5.32	4.71	7.66	7.21	9.15
Population value (%)		0.01	3.33	2.05	0.18	0.50	2.26	1.83	3.01

The $g1-$allocation was the only allocation with absolute relative bias less than 10% on each area, and it had a practically zero bias on the population level. Also the equal and proportional allocations had low biases on both levels, but the model-assisted and parameter-based allocations had a clearly poorer performance. An interesting detail in the $g1-$allocation is that the accuracy of area estimates is fairly good and the relative bias is low also for the case of two areas with zero sample size. A common characteristic for these areas is that the means of variables y and x are close to corresponding population means. In any case, it is essential that the model-based estimation can produce reliable estimates for areas, which are not represented in the random sample.

5 Concluding remarks

This research was focused on seven different allocation solutions which were categorized into three groups according to the auxiliary data needed in their implementation. The least amount of auxiliary information is needed in equal and proportional allocation which are based on the number of areas and the number of statistical units in each area. The Neyman, Bankier and NLP allocations are based on pre-set optimization criteria, and application of these methods presumes area-specific parameter information such

as the standard deviation or CV of the study variable, and in the Bankier allocation the area totals of at least one auxiliary variable must be known. Because the study variable is unknown, it must be replaced with a suitable proxy or auxiliary variable to enable the use of these three methods. A common feature of the number-based and parameter-based allocations is that they are not based on any model, whereas the other three allocations utilize the underlying model, in addition to number-based information.

On the basis of the empirical results, the performance of the model-based $g1-$allocation can be regarded as the best compared with the other allocations tested in this research. Also equal and proportional allocations reached good results, but the model-assisted allocations and the parameter-based allocations had clearly weaker performances. The last three allocations are developed originally for direct design-based estimation, and their results can be understood from that point of view. Compared with $g1-$allocation, the MC-allocations are based on a different model and this fact seems to affect their results.

One of the characteristics of the $g1-$allocation is that when the sampling design is constructed, also the model and estimation method are fixed, meaning that they are regarded as given preliminary information. This allocation, which is based on a unit-level linear mixed model and EBLUP estimation method, needs only the homogeneity coefficient between areas which is computed by using the values of the auxiliary variable. In this respect, the $g1-$allocation differs from the other allocations used in the comparison. Also the starting point for choosing the final estimation method is different, because this allocation is focused on model-based estimation, not on direct design-based estimation using sampling weights. The choice of the model-based estimation is justified also for the reason that it is commonly used in small area estimation. On the other hand, the $g1-$allocation enables the use of small sample sizes, because information can be borrowed between areas when the model is applied. This can be significant in quick surveys or studies carried out by market research organizations, when a single measurement is expensive. However, it is important to examine the characteristics of the areas and especially the small areas, before the final sample sizes are determined.

As a recommendation, it would be justified to start a wider research to find out what advantages and disadvantages are encountered if the applicable computing technique for producing area statistics is decided as early as in the design of the research plan.

Acknowledgements

The authors thank the Editor, Associate Editor and two referees as well as Professor Risto Lehtonen for constructive comments and suggestions.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician,* 42, 174-177.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology,* 38, 1, 23-29. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2012001/article/11682-eng.pdf.

Costa, A., Satorra, A. and Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT,* 28(1), 69-86.

Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. *Survey Methodology,* 34, 2, 223-234. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2008002/article/10763-eng.pdf.

Keto, M., and Pahkinen, E. (2009). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". In *Survey Sampling Methods in Economic and Social Research,* (Eds., J. Wywial and W. Gamrot), 2010. Katowice: Katowice University of Economics.

Keto, M., and Pahkinen, E. (2014). On sample allocation for efficient small area estimation. *Book of Abstracts.* SAE 2014, Poland: Poznan University of Economics, page 50.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology,* 32, 1, 87-96. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2006001/article/9259-eng.pdf.

Molefe, W.B., and Clark, R.G. (2015). Model-assisted optimal allocation for planned domains using composite estimation. *Survey Methodology,* 41, 2, 377-387. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14230-eng.pdf.

Nissinen, K. (2009). *Small Area Estimation with Linear Mixed Models from Unit-Level Panel and Rotating Panel Data*. Ph.D. thesis, University of Jyväskylä, Department of Mathematics and Statistics, Report 117, https://jyx.jyu.fi/dspace/handle/123456789/21312.

Pfefferman, D. (2013). New important developments in small area estimation. *Statistical Science,* 28, 40-68.

Rao, J.N.K. (2003). *Small Area Estimation*. Hobogen, New Jersey: John Wiley & Sons, Inc.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron,* Vol. 2, 3, 461-493; 4, 646-683.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 107-124
Statistics Canada, Catalogue No. 12-001-X

A mixed latent class Markov approach for estimating labour market mobility with multiple indicators and retrospective interrogation

Francesca Bassi, Marcel Croon and Davide Vidotto[1]

Abstract

Measurement errors can induce bias in the estimation of transitions, leading to erroneous conclusions about labour market dynamics. Traditional literature on gross flows estimation is based on the assumption that measurement errors are uncorrelated over time. This assumption is not realistic in many contexts, because of survey design and data collection strategies. In this work, we use a model-based approach to correct observed gross flows from classification errors with latent class Markov models. We refer to data collected with the Italian Continuous Labour Force Survey, which is cross-sectional, quarterly, with a 2-2-2 rotating design. The questionnaire allows us to use multiple indicators of labour force conditions for each quarter: two collected in the first interview, and a third one collected one year later. Our approach provides a method to estimate labour market mobility, taking into account correlated errors and the rotating design of the survey. The best-fitting model is a mixed latent class Markov model with covariates affecting latent transitions and correlated errors among indicators; the mixture components are of mover-stayer type. The better fit of the mixture specification is due to more accurately estimated latent transitions.

Key Words: Gross flows; Labour market; Mixture models; Latent class models.

1 Introduction

Analysts can exploit panel data to estimate labour force gross flows - i.e., transitions in time between different states. Net flows measure variations in time in various market states, whereas gross flows provide information on the dynamics of the labour market.

A large body of literature on gross flows estimation is based on the assumption that errors are uncorrelated over time, i.e., they are Independent Classification Errors (ICE). The ICE assumption implies that: (i) classification errors referring to two different occasions are independent of each other conditionally on the true states, and (ii) errors only depend on the present true state. Thus, classification errors produce spurious transitions and consequently induce overestimation of changes.

However, in many contexts, the ICE assumption turns out not to be realistic, because of the survey design and data collection strategies. In these circumstances, classification errors may be correlated: observed states may also depend on true states at other times or on true transitions, or direct effects may exist between observed states (Bound, Brown and Mathiowetz 2001).

In this paper, we use a model-based approach to adjusting observed gross flows for classification errors. It combines a structural sub-model for unobserved true transition rates and a measurement sub-model relating true states to observed ones. A convenient framework for formulating our model is provided by latent class (LC) analysis.

1. Francesca Bassi, Department of Statistical Sciences, University of Padova, Italy, Via C. Battisti 241, 35121, Padova, Italy. E-mail: francesca.bassi@unipd.it; Marcel Croon and Davide Vidotto, Department of Methodology and Statistics, University of Tilburg, The Netherlands.

We apply our approach to observed gross flows among the three labour force states - Employed (E), Unemployed (U) and Not in the labour force (N) - taken from the Italian Continuous Labour Force Survey (CLFS), a quarterly survey with a 2-2-2 rotating design which yields two-wave panels one quarter, three quarters and one year apart. We consider data collected from 2005 to 2009.

The questionnaire allows us to use multiple indicators of labour force conditions for each quarter: (i) all respondents are classified as Employed, Unemployed, or Not in the labour force, according to the definition of the International Labour Office (ILO) on the basis of answers given to a group of questions; (ii) respondents are asked to classify themselves as employed, unemployed, or not in the labour force, the self-perceived condition; (iii) a retrospective question asks about respondents' state in the labour market one year before the interview. This approach provides a way of estimating labour market mobility by taking into account correlated measurement errors and the rotating design of the survey.

In detail, the best-fit model is a mixed latent class Markov (LCM) model with covariates affecting latent transitions and correlated errors among indicators. The mixture is obtained by assuming the existence of two unobservable sub-populations, movers, i.e., respondents who change their state in the labour market during the observation period, and stayers. A secondary result of our research is that the mover-stayer model and the LCM estimate the same amount of measurement error in the data. The better fit of the mixture specification is due to more accurately estimated latent transitions. Magidson, Vermunt and Tran (2007) also found that the mixed LC Markov model has a better fit to the data than the traditional one. However, in that case, the difference in fit was due to the fact that, as heterogeneity was not taken into account, the result was overestimation of measurement error.

Our paper follows recent contributions to the scientific literature on the topic of gross flows estimation with hidden Markov chain and multiple indicators. An accurate description of the model may be found in Langeheine (1994). The method was not only applied to estimation of labour market gross flows but also to many other contexts, longitudinal data being available. Paas, Vermunt and Bijmolt (2007), for example, estimated an LCM model to study acquisitional patterns in the financial product market; multiple indicators of ownership of financial products were used to identify not directly observable market segments among which customers could move on consecutive measurement occasions. Bartolucci, Lupparelli and Montanari (2009) estimated the same model in following changes in health status in a sample of patients over time. Manzoni, Vermunt, Luijkx and Muffels (2010) applied an LCM model to estimate gross flows in the Swedish labour market. In a more recent work, Pavlopoulos and Vermunt (2015) used a hidden Markov model to estimate the amount of measurement error in information from the Dutch Labour Force Survey and the Dutch Institute for Employee Insurance on the type of job (permanent or temporary).

The contribution of this paper to the scientific literature on the topic of gross flows estimation is that we have three indicators, one of them collected retrospectively, on labour force state and we can also take into account the rotating design of the survey. The paper also contributes to the literature on the quality of data from the CLFS (Bassi, Padoan and Trivellato 2012).

The paper is organised as follows. Section 2 introduces the traditional (or standard) and the mixed LCM model. Section 3 describes the survey and its data. Section 4 compares the performances of the traditional

versus mixed LCM models. Section 5 provides results, referring to the best fitting model to correct gross flows in the labour market from measurement errors. Section 6 concludes.

2 The latent class Markov model

Latent class analysis has been applied in a number of studies on panel data to separate true changes from observed ones affected by unreliable measurements. Relatively recent contributions include Bassi, Torelli and Trivellato (1998), Biemer and Bushery (2000), Bassi, Croon, Hagenaars and Vermunt (2000), Bassi and Trivellato (2009).

The true labour force state is treated as a latent variable and the observed one as its indicator. The model consists of two parts:

a) structural, describing true dynamics among latent variables;

b) measurement, linking each latent variable to its indicator(s).

Let us consider the simplest formulation of latent class Markov (LCM) models (Wiggins 1973), which assumes that true unobservable transitions follow a first-order Markov chain. As in all standard LCM specifications, local independence among indicators is assumed, i.e., indicators are independent conditionally on latent variables In the LCM model with one indicator per latent variable, the assumption of local independence coincides with the Independent Classification Errors condition.

Let X_{it} denote the true labour force condition at time t for a generic sample individual $i, i = 1,\ldots,n$; Y_{it} is the corresponding observed condition; $P\left(X_{i1} = l_1\right)$ is the probability of the initial state of the latent Markov chain, and $P\left(X_{it+1} = l_{t+1} \mid X_{it} = l_t\right)$ is the transition probability between state l_t and state l_{t+1} from time t to $t+1$, with $t = 1,\ldots,T-1$, where T represents the total number of consecutive, equally spaced time-points over which an individual is observed. In addition, $P\left(Y_{it} = j_t \mid X_{it} = l_t\right)$ is the probability of observing state j at time t, given that individual i at time t is in the true state l_t: this is also called the model measurement component.

It follows that $P(Y(1),\ldots,Y(T))$ is the proportion of units observed in a generic cell of the $T-$way contingency table. For a generic sample individual i, a LCM model is defined as:

$$P\left(\mathbf{Y}_i = \mathbf{y}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1\right) \prod_{t=2}^{T} P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}\right) \prod_{t=1}^{T} P\left(Y_{it} = j_t \mid X_{it} = l_t\right) \quad (2.1)$$

where $\mathbf{y}$ is the vector containing observed values for individual i, l_t and j_t vary over K classes (in our application, three labour force conditions). Equation (2.1) specifies the proportion of units in the generic cell of a $T-$way contingency table as a product of marginal and conditional probabilities.

In an LCM model with concomitant variables, latent class membership and latent transitions are expressed as functions of covariates with known distributions (Dayton and McReady 1988). $P\left(X_{i1} = l_1 \middle| \mathbf{Z}_{i1} = \mathbf{z}_1\right)$, where $\mathbf{z}_1$ is a vector containing the values of covariates for respondent i at time 1, estimates covariate effects on the initial state, and $P\left(X_{it} = l_t \middle| X_{it-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right)$, where $\mathbf{z}_t$ is a vector containing the values of covariates for respondent i at time t, estimates covariate effects on latent transitions.

On the basis of the above components, the complete model for individual i is given by:

$$P\left(\mathbf{Y}_i = \mathbf{y} \middle| \mathbf{Z}_i = \mathbf{z}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1 \middle| \mathbf{Z}_1 = \mathbf{z}_1\right) \prod_{t=2}^{T} P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right) \prod_{t=1}^{T} P\left(Y_{it} = j_t \middle| X_{it} = l_t\right) \tag{2.2}$$

When more than one (M) indicators per latent variable are observed, the model formulation becomes the following (Vermunt 2010):

$$P\left(\mathbf{Y}_i = \mathbf{y} \middle| \mathbf{Z}_i = \mathbf{z}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1 \middle| \mathbf{Z}_1 = \mathbf{z}_1\right) \prod_{t=2}^{T} P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right) \prod_{m=1}^{M} \prod_{t=1}^{T} P\left(Y_{mit} = j_t \middle| X_{it} = l_t\right) \tag{2.3}$$

In our application, the M indicators are given by the three pieces of information collected for all respondents on their labour market condition.

Typically, conditional probabilities are parameterised and restricted by logistic regression models. The parameters are estimated via maximum likelihood (Vermunt and Magidson 2013). Identification is a well-known problem in models with latent variables and, although the number of independent parameters must not exceed the number of observed frequencies, this is not a sufficient condition. According to Goodman (1974), a sufficient condition for local identifiability is that the information matrix is positive definite. Latent Gold software (Vermunt and Magidson 2008), provides information on parameter identification. Another problem linked to estimation is that of local maxima, to deal with which we estimated our models several times with different sets of starting values.

A mixed LCM model assumes the existence in the population of not directly observable groups moving across time, following latent chains with different initial state probabilities and different transition probabilities; the groups may also be assumed to have different response probabilities (van de Pol and Langeheine 1990). Such a model can be extended to include time-varying and time-constant covariates

(Vermunt, Tran and Magidson 2008). A special case of a two-class mixed LCM model is the mover-stayer model: the group of movers has positive probabilities of transferring from one state to another over time, and the group of stayers do not change. For the latter, transition probabilities between different states are imposed as zero. A two-class mixed LCM model with concomitant variables has the following form:

$$P\left(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{Z}_i = \mathbf{z}\right) = \sum_{w=1}^{2} \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P(W = w) P\left(X_{i1} = l_1 \mid \mathbf{Z}_1 = \mathbf{z}_1, W = w\right)$$
$$\prod_{t=2}^{T} P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t, W = w\right)$$
$$\prod_{j_t=1}^{K} \prod_{t=1}^{T} P\left(Y_{it} = j_t \mid X_{it} = l_t, W = w\right) \tag{2.4}$$

where W is a binary latent variable. The mover-stayer model is obtained assuming, for $l_t \neq l_{t-1}$, $P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}, W = 2\right) = 0$ and, consequently, for $l_t = l_{t-1}$ $P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}, W = 2\right) = 1$.

The likelihood function of an LC model can also be estimated if information is missing in the response variables. We exploit this opportunity to take into account the response patterns generated by the survey rotation design. Sampled households are interviewed for two consecutive quarters, do not participate in the survey for the subsequent two quarters, and are then re-interviewed on two other occasions (see Table 3.1). We assumed that missing information due to survey design is missing at random. In this case, each unit only contributes to the likelihood function with the information available (Vermunt 1997).

3 The data

The Continuous Labour Force Survey (CLFS), conducted by ISTAT (Italian Institute of Statistics), is the main and official source of statistical documentation on the Italian labour market. The CLFS has been conducted since 1969 and has been modified many times. In 2004, major updating was carried out, mainly dictated by the requirement to adapt the survey to new EU (European union) standards. The principal changes involved interviews distributed throughout the years of the study, new criteria to classify respondents' status in the labour market, computer-assisted data collection techniques, and dependent interviewing. Every year the survey collects information on about 280,000 households, for a total of about 700,000 individuals. The reference population consists of all household members officially resident in Italy.

The Italian CLFS sampling design has two stages: 1) municipalities were denominated as primary sampling units (PSUs) with stratification, and households as final sampling units (FSUs) with rotation. PSUs were stratified according to demographic size. Large municipalities, with population over a given threshold (also called self-representative municipalities), were always included in the sample; smaller municipalities (not self-representative) were grouped in strata, so that one municipality in each stratum was selected with probability proportional to its population; 2) households were randomly selected from the population registers in all municipalities drawn at stage 1.

The survey was quarterly with a 2-2-2 rotating design. Householders were interviewed in two consecutive quarters. After a two-quarter break, they were interviewed again, twice in the corresponding two quarters of the following year. As a result, each household was included in four waves of the survey over a period of 15 months. This rotation system meant that half of the sample remained unchanged in two consecutive quarters and in quarters one year apart, and 25% of the sample remained unchanged over three quarters.

All the following statistical analyses are made on the so-called longitudinal population. The CLFS is not designed as a proper panel: the initial population changes during the observation period due to demographic events and migrations. Although ISTAT has proposed a procedure to calculate longitudinal weights (Boschetto, Discenza, Lucarelli, Rosati and Fiori 2009), they are not available to researchers, so that we could not take into account the complex sample design. However, we consider that it was reasonable to assume that respondents belonging to the same households were independent.

Information on labour force condition in one reference quarter was collected three times: (i) each respondent was classified as employed, unemployed or not in the labour force according to the definition of the ILO on the basis of answers given to a selected group of questions; (ii) in a subsequent section of the questionnaire, all respondents were asked to classify themselves in the labour market, in order to collect the "self-perceived" condition; (iii) after one year, a retrospective question asked about respondents' state in the labour market one year before the first interview.

According to the ILO definition, respondents were classified as employed in the reference quarter if, aged 15 years or over, during the reference week they performed some kind of work, for at least one hour, for pay, profit or family gain, or were not at work but had a job or business from which they were temporarily absent because of illness, holidays, industrial dispute, or education and training. Respondents were classified as unemployed if, aged from 15 to 74, they were: (a) without work during the reference week; (b) currently available for work in the two weeks following the reference week; (c) actively seeking work, i.e., had taken specific steps, in the four-week period ending with the reference week, to seek work or who did not seek work but who had found a job to be started later, within a period of up to three months (International Labour Organization (ILO) 2008).

Current self-perception and the retrospective question classified respondents in eight categories: employed; unemployed looking for new employment; unemployed looking for first employment; fulfilling domestic tasks; student; retired; disabled for work; other.

Table 3.1 shows the rotating design of the survey for two consecutive calendar years. Letters identify rotation groups: four rotation groups were interviewed in each quarter. With reference to one calendar year, information on labour market condition came from nine rotation groups. However, the rotation design generates a specific pattern of missing data. For example, for units of rotation group A who are interviewed for the fourth time in the first quarter of year 1, only the ILO (I) indicator and self-perception (S) of labour market condition in the first quarter of year 1 are available. For units in rotation group F, who were first interviewed in the first quarter of year 1, we only have information on labour force state based on the ILO definition, self-perception and the retrospective question (R) for the first and second quarters of year 1.

Table 3.1
CLFS rotation design

Rotation Group	Year 1				Year 2			
	I quarter	II quarter	III quarter	IV quarter	I quarter	II quarter	III quarter	IV quarter
A	I-S							
B	I-S	I-S						
C		I-S	I-S					
D			I-S	I-S				
E	I-S-R			I-S	I-S			
F	I-S-R	I-S-R			I-S	I-S		
G		I-S-R	I-S-R			I-S	I-S	
H			I-S-R	I-S-R			I-S	I-S
I				I-S-R	I-S-R			I-S
L					I-S-R	I-S-R		
M						I-S-R	I-S-R	
N							I-S-R	I-S-R
O								I-S-R

I = ILO indicator, S = self-perception of labour market condition, R = retrospective indicator.

We examined data collected from 2005 to 2010. (Excluded from these analyses are data collected in 2004, the first year of implementation of the new labour force survey, because the data may not be totally reliable; with reference to 2010, here we use only information collected with the retrospective question and referring to labour condition in 2009.) Table 3.2 lists labour market composition in the first quarter from pooled data over the five-year period. The ILO indicator clearly counts a lower percentage of unemployed and a higher percentage of persons not in the labour force than the other two indicators. The two measures based on self-perception give a higher unemployment rate because ILO applies a very strict definition of unemployment. To be classified as unemployed, respondents between the ages of 15 and 74 must not be in employment at the moment of the interview but would accept suitable jobs in the next two weeks if the opportunity arose, and had actively looked for ways of obtaining jobs in the preceding two weeks. ILO provides these guidelines in order to facilitate comparisons of labour market performance over time and across countries (ILO 2008). However, this framework was set up when the prevailing type of employment was full-time and under permanent contract; since then, the employment situation has changed to one of more flexibility, with more part-time and fixed-term types of work, especially for those about to enter the labour market.

Table 3.2
Labour market composition 2005 – 2009 I quarter, % - pooled data

	E	U	N
ILO	43.07	3.60	53.33
S	41.73	6.73	51.54
R	41.55	6.49	51.96

E = Employed, U = Unemployed, N = Not in the Labour Force.

Other studies in the literature show that the distinction between labour market states is not always clear-cut: people may not know official definitions or perceive their labour condition as different from that arising from standard criteria (see, for example, Clark and Summer 1979; Flinn and Heckman 1983; Gonul 1992). In most cases, it is difficult to distinguish between unemployment and not in the labour force: the most

critical condition seems to be that of actively seeking a job, since respondents may perceive themselves as unemployed even when they are not actively looking for a job. Inconsistencies may consequently arise between information collected in surveys and effective behaviour. Another explanation of the differences between the ILO and the self-perceived classifications is that respondents with temporary jobs in terms of hours of work per week may not classify themselves as employed.

Table 3.3 lists inconsistencies, i.e., different labour conditions observed for the same respondent with two indicators, among the three indicators for the period in question. Data over quarters and years were pooled for reasons of space. The number of inconsistencies is clearly higher for the state of unemployment than for the other two states, and most of the misclassifications tend to refer to people out of the labour force rather than in employment, as many previous studies show (see, for example, Poterba and Summers 1986). Comparing the labour condition according to the ILO definition with that reported according to answers to the retrospective question generated the highest number of inconsistencies. Examining consistencies over quarters and years for couples of the three indicators (not reported here for reasons of space) we note that consistency tends to increase slightly over time, perhaps because all the actors involved in the survey process - interviewers, respondents, etc. - learn how to collect and supply good-quality information while participating in the survey. Although we did not observe seasonal effects in the number of inconsistencies, the number of inconsistencies indicated non-negligible measurement error in the data, which means that one of the two indicators, or both, were reported incorrectly.

Table 3.3
Inconsistencies 2005 – 2009, % - pooled data

	EU	EN	UE	UN	NE	NU
ILO – Self-perception	0.97	1.72	0.44	13.02	0.17	5.80
ILO – Retrospective	1.14	2.06	5.22	16.76	1.00	5.76
Self-perception – Retrospective	0.92	1.62	6.03	8.73	1.00	0.89

EU = Classified as Employed with first indicator but Unemployed with second indicator.
EN = Classified as Employed with first indicator but Not in the Labour Force with second indicator.
UE = Classified as Unemployed with first indicator but as Employed with second indicator.
UN = Classified as Unemployed with first indicator but Not in the Labour Force with second indicator.
NE = Classified as Not in the Labour Force with first indicator but Employed with second indicator.
NU = Classified as Not in the Labour Force with first indicator but Unemployed with second indicator.

However, the inconsistencies emerging from Tables 3.2 and 3.3 may also occur because all three indicators are exposed to measurement error. Previous studies have investigated the causes of labour condition misperception, finding that it is influenced by social, demographic, economic and institutional factors (e.g., Richiardi 2002). Inconsistencies between the two self-perceptions (actual and retrospective) may mainly be due to memory decay (Bound, Brown and Mathiowetz 2001). Lastly, the higher consistency between the self-perception indicators suggests the possibility of correlated measurement errors.

Table 3.4 lists observed quarterly transition probabilities among the three labour force conditions from the first to the second quarter of the years from 2005 to 2009 with the three indicators. The ILO indicator describes a much more dynamic labour market, especially for unemployed respondents, than that described by the self-perceived and retrospective indicators. This difference is another piece of evidence revealing measurement error in the data. From the existing literature, we know that even small degrees of classification

error may lead to severe bias in the estimation of transition probabilities (Hagenaars 1994; Pavlopoulos, Muffles and Vermunt 2012). If errors are uncorrelated over time, we can expect to observe a more dynamic labour market than the true one, and the opposite if error correlation over time also exists.

Table 3.5 compares observed gross flows, as an example from the first to the second quarter of 2005, by gender and age. The three age intervals were obtained by dividing the samples into three groups, with equal dimensions (i.e., 33rd and 66th percentiles). In detail, for the year 2005, in age 1 we find respondents aged between 16 and 36; in age 2 they are between 36 and 55, and in age 3 between 56 and 75. The evidence is that women are more dynamic, especially with regard to unemployment, than men. When leaving unemployment, women tend to leave the labour market more often than to become employed. There are also some important differences in observed gross flows across ages. The older respondents were more stable when out of the labour market and had higher probabilities of moving out of the labour market than of becoming unemployed after being employed. Younger respondents have lower probabilities than those in the second age-group of leaving unemployment and the condition of not being in the labour market by finding jobs. This evidence suggests that gender and age should be included as covariates in our model, to estimate corrected gross flows in the labour market.

Table 3.4
Observed gross flows I quarter to II quarter 2005 - 2009, %, International Labour Office (ILO), Self-perceived (S) and Retrospective (R) indicators

		EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	ILO	96.49	0.87	2.63	18.97	50.50	30.53	1.49	1.99	96.52
	S	96.99	1.33	1.69	15.32	69.85	14.83	1.29	1.50	97.21
	R	95.32	2.10	2.58	20.96	59.56	19.48	1.96	2.22	95.81
2006	ILO	96.13	0.78	3.09	20.40	45.21	34.39	2.42	1.74	95.84
	S	96.11	1.74	2.16	19.84	63.66	16.50	1.88	1.75	96.37
	R	95.55	1.72	2.73	17.93	66.57	15.50	2.00	1.75	96.25
2007	ILO	96.22	0.68	3.10	21.45	40.41	38.14	2.21	1.78	96.02
	S	96.08	1.74	2.16	19.84	63.66	16.50	1.88	1.75	96.37
	R	95.66	1.78	2.56	19.95	60.67	19.38	2.26	1.93	95.80
2008	ILO	97.05	0.80	2.16	19.82	48.50	31.68	1.87	1.87	96.26
	S	96.92	1.54	1.53	15.25	70.84	13.92	1.56	1.69	96.75
	R	95.76	2.13	2.11	19.04	62.60	18.36	2.02	2.26	95.72
2009	ILO	96.58	0.88	2.54	18.41	48.10	33.49	2.08	1.83	96.09
	S	96.14	1.76	2.10	15.17	70.09	14.75	1.59	1.61	96.80
	R	95.45	1.88	2.66	16.88	67.15	15.97	1.78	1.89	96.33

EE = Employed in both quarters.
EU = Employed in first quarter and Unemployed in second one.
EN = Employed in first quarter and Not in the Labour Force in second one.
UE = Unemployed in first quarter and Employed in second one.
UU = Unemployed in both quarters.
UN = Unemployed in first quarter and Not in the Labour Force in second one.
NE = Not in the Labour Force in first quarter and Employed in second one.
NU = Not in the Labour Force in first quarter and Unemployed in second one.
NN = Not in the Labour Force in both quarters.

Table 3.5
Observed gross flows I quarter to II quarter 2005, by gender and age, %, International Labour Office (ILO), Self-perceived (S) and Retrospective (R) indicators

		EE	EU	EN	UE	UU	UN	NE	NU	NN
Males	ILO	97.20	0.78	2.02	22.73	51.60	25.68	1.93	2.07	96.00
	S	97.63	1.08	1.29	18.97	73.80	7.23	1.36	1.10	97.53
	R	96.13	1.84	2.03	26.14	65.27	8.60	2.13	1.50	96.37
Females	ILO	95.43	1.01	3.57	15.70	49.31	34.99	1.23	1.98	96.79
	S	96.00	1.69	2.31	11.93	65.73	22.34	1.26	1.81	96.93
	R	94.14	2.46	3.40	16.24	53.56	30.19	1.86	2.71	95.43
Age 1	ILO	88.27	0.46	11.27	21.16	27.50	51.35	0.26	0.06	99.67
	S	89.66	0.56	9.78	10.20	60.09	29.71	0.31	0.10	99.60
	R	83.36	0.45	16.19	20.78	42.54	36.68	0.51	0.13	99.36
Age 2	ILO	97.65	0.55	1.80	21.62	43.01	35.37	2.72	2.95	94.33
	S	97.87	0.92	1.20	16.83	64.65	18.52	2.52	2.61	94.87
	R	97.04	1.23	1.74	24.60	53.42	21.98	4.05	4.24	91.70
Age 3	ILO	96.18	1.32	2.50	17.54	51.14	31.32	3.81	6.75	89.44
	S	96.83	1.89	1.28	14.77	71.97	13.27	3.17	4.82	92.01
	R	94.82	3.29	1.89	19.62	63.60	16.78	4.52	6.68	88.80

EE = Employed in both quarters.
EU = Employed in first quarter and Unemployed in second one.
EN = Employed in first quarter and Not in the Labour Force in second one.
UE = Unemployed in first quarter and Employed in second one.
UU = Unemployed in both quarters.
UN = Unemployed in first quarter and Not in the Labour Force in second one.
NE = Not in the Labour Force in first quarter and Employed in second one.
NU = Not in the Labour Force in first quarter and Unemployed in second one.
NN = Not in the Labour Force in both quarters.

4 Results: Comparisons of mixed and standard LCM models

We estimate various specifications of the standard and mixed LCM models. The standard model consists of two parts: structural, describing true dynamics among latent variables (true states) by a first-order Markov chain; and measurement, which links each latent variable to its indicators (observed conditions in the labour market). Some restrictions incorporating a priori information and/or assumptions are imposed on the parameters of the measurement part, based on evidence from observed data (inconsistencies and transitions) and on findings from the survey methodology and cognitive psychology literature on the error - generating mechanism. Only four of the nine rotation groups supplying information referring to one calendar year were interviewed in every quarter, and only for two of these groups do we have all three indicators of labour market conditions (see Table 3.1). For the other two groups, we do not have the information collected with the retrospective question. The pattern of missing information due to the rotation design of the survey is included in the estimated LCM models as data missing at random.

All estimated models share the following characteristics: true transitions follow a first-order Markov chain; (Due to the survey design, there were no individuals observed for three consecutive waves, i.e., a second-order Markov chain cannot be estimated, since the relative sufficient statistics are missing. However, although the labour market condition in one quarter may very plausibly affect the condition in the subsequent quarter, that it may do so in a significant manner after two quarters is far less plausible.) classification errors are assumed constant over time for each indicator; the ICE assumption is included.

Model fit is evaluated by the BIC (Bayesian Information Criterion) index because of the large sample size (average 250,000 units per year; see Table 4.1).

The specification of a mixed LCM model is also recommended by the fact that the sample may contain various groups of respondents with different behaviour in the labour market. As already noted, the recent literature shows that not taking unobserved heterogeneity in transitions into account when estimating LCM models may result in biased estimates of measurement error (Magidson et al. 2007). In addition, a mixed LCM model may give the data a better fit.

We estimate a mover-stayer LCM model with the assumption of constant measurement errors across the two latent groups. It should be noted that all estimated models were identified and that, in order to reduce the risk of detecting local maxima, estimation was performed several times with different sets of starting values. Latent Gold 4.5 software was implemented (Vermunt and Magidson 2008).

Table 4.1 compares the mixed and standard LCM models fitted to our five data samples, referring to the years from 2005 and 2009 and using the BIC index. The mixed model shows a better fit for all samples. Table 4.2 lists the percentages of movers and stayers in the first quarter of 2005, and the distribution of the two unobserved groups in the first quarter of each year. Clearly, unobserved heterogeneity is highly correlated with the initial state and, as expected, stayers are either employed or not in the labour market, i.e., only a very small percentage is unemployed.

Table 4.1
Comparison of standard and mixed LCM models: BIC index

Year	n	Standard	Mixture
2005	220,051	650,241	649,401
2006	206,037	587,794	587,058
2007	274,484	748,788	748,654
2008	277,363	667,399	666,335
2009	274,723	747,997	746,991

Table 4.2
Mixed LCM model: proportion of movers and stayers and distribution in initial state 2005, I quarter, %

	Proportion	E	U	N
Movers	10.23	39.85	39.09	21.06
Stayers	81.79	41.79	3.36	54.85

E = Employed, U = Unemployed, N = Not in the Labour Force.

As the data in Tables 4.3-4.5 show, (Labour market composition, estimated transitions and estimated measurement errors show the same pattern in the other three quarters of each year.) the better fit to the data of the mixed model is all due to the different estimated transition rates; labour market composition and estimated measurement errors are the same in both models. This result is the opposite of that obtained by Magidson et al. (2007), who compared the mover-stayer and standard LCM models applied to labour market transitions from the Current Population Survey. The above authors found that the mixed LCM model provides a better fit to the data than the standard LCM model and that the latter, not taking unobserved heterogeneity into account, overestimates the degree of measurement error with respect to the mover-stayer model. In detail, the above authors used simulated results to estimate a violation of homogeneous transition

probabilities, so that heterogeneity correlated with the initial state produces inflated estimates of measurement errors in a standard LCM model.

Table 4.3
Comparison of standard and mixed LCM models: labour market composition I quarter 2005, %

		E	U	N
2005	Standard	41.67	7.00	51.33
	Mixture	41.59	7.02	51.39

E = Employed, U = Unemployed, N = Not in the Labour Force.

Table 4.4
Comparison of standard and mixed LCM models: estimated transitions I quarter to II quarter 2005 – 2009, %

		EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	Standard	97.36	1.32	1.32	15.59	76.18	8.23	0.57	0.74	98.69
	Mixture	96.46	1.68	1.86	19.61	69.65	10.74	0.91	1.09	98.00
2006	Standard	96.75	1.68	1.56	19.52	71.27	9.21	1.01	0.99	90.00
	Mixture	96.22	1.92	1.87	22.11	66.96	10.93	1.25	1.22	97.54
2007	Standard	96.69	1.67	1.64	18.84	70.56	10.60	1.01	0.99	98.00
	Mixture	96.42	1.80	1.78	20.22	67.80	11.98	1.10	1.45	95.45
2008	Standard	97.56	1.41	1.03	15.86	79.73	4.42	0.53	0.62	98.85
	Mixture	96.45	1.89	1.66	19.56	73.25	7.19	0.83	0.89	98.28
2009	Standard	96.85	1.71	1.44	14.04	75.33	9.63	1.04	1.01	97.95
	Mixture	96.27	1.95	1.78	17.09	71.16	11.75	1.30	1.22	97.48

EE = Employed in both quarters.
EU = Employed in first quarter and Unemployed in second one.
EN = Employed in first quarter and Not in the Labour Force in second one.
UE = Unemployed in first quarter and Employed in second one.
UU = Unemployed in both quarters.
UN = Unemployed in first quarter and Not in the Labour Force in second one.
NE = Not in the Labour Force in first quarter and Employed in second one.
NU = Not in the Labour Force in first quarter and Unemployed in second one.
NN = Not in the Labour Force in both quarters.

Table 4.5
Comparison of Standard and mixed LCM models: estimated measurement errors I quarter 2005 – 2009, %, ILO indicator

		EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	Standard	99.82	0.01	0.17	6.17	45.04	48.80	0.89	0.50	98.61
	Mixture	99.82	0.01	0.17	6.16	45.06	48.78	0.90	0.51	98.59
2006	Standard	99.83	0.01	0.16	6.50	41.92	51.58	0.75	0.45	98.80
	Mixture	99.87	0.01	0.13	5.17	37.28	57.55	0.68	0.40	98.92
2007	Standard	99.75	0.01	0.24	6.84	39.83	53.34	0.75	0.47	98.79
	Mixture	99.75	0.01	0.24	6.77	39.92	53.31	0.77	0.47	98.76
2008	Standard	99.83	0.01	0.17	3.81	42.45	53.74	0.61	0.38	99.02
	Mixture	99.83	0.01	0.17	3.82	42.41	53.76	0.62	0.38	99.00
2009	Standard	95.34	0.98	3.68	18.30	41.17	40.53	2.06	1.61	96.33
	Mixture	95.22	2.34	2.44	15.60	68.02	16.37	1.74	2.14	96.13

EE = Truly Employed and classified as Employed by ILO indicator.
EU = Truly Employed but classified as Unemployed by ILO indicator.
EN = Truly Employed but classified as Not in the Labour Force by ILO indicator.
UE = Truly Unemployed but classified as Employed by ILO indicator.
UU = Truly Unemployed and classified as Unemployed by ILO indicator.
UN = Truly Unemployed but classified as Not in the Labour Force by ILO indicator.
NE = Truly Not in the Labour Force but classified as Employed by ILO indicator.
NU = Truly Not in the Labour Force but classified as Unemployed by ILO indicator.
NN = Truly Not in the Labour Force and classified as Not in the Labour Force by ILO indicator.

The mover-stayer model describes a more dynamic labour market, especially for unemployed respondents: the probability of remaining unemployed over the quarter is lower than that estimated by the standard model.

5 Results: Mixed LCM model with covariates and correlated measurement errors

The results shown in the previous section showed that a mixed LCM model gives a better fit to our data. Like the standard LCM model, it takes into account misclassification and the pattern of missing data assuming the latter at random, and also includes unobserved heterogeneity. Assuming that data are missing at random is explained by the fact that each rotation group is observed in two quarters, but not in the two subsequent quarters, and also that these data are missing by design and do not depend on respondents' true or reported status or other unobserved variables. In estimating our models, we simultaneously used information from all rotation groups, i.e., a Full Information Maximum Likelihood approach. Evidence from the observed gross flows, especially the fact that observed mobility is quite different between men and women and across ages (Table 3.5) indicated estimating a mixed LCM model with these two covariates affecting latent transitions.

Various models were estimated with the common following characteristic: mover-stayer and latent transitions follow a first-order Markov chain. In order to specify the measurement model, the following considerations were made: (i) the answer to the question on self-perceived condition in the labour market is given in the same interview after respondents answer the questions on which the ILO indicator is based; (ii) however, the ILO indicator is determined by ISTAT according to answers given to a series of questions following ILO guidelines, whereas S represents respondents' self-perceptions: it is plausible that respondents are not aware of the ISTAT classification; (iii) indicator S and the indicator resulting from retrospective interrogation describe a more stable labour market than that of ILO and show the highest level of consistency: respondents may be influenced by the answers they gave the previous quarter; (iv) information for R is collected one year after answers to ILO and S; (v) for individuals who are in a steady state, reporting labour force condition correctly is an easier cognitive task than for those who experience at least one change, and may consequently show higher probabilities of giving incorrect answers.

Among the various possible specifications, the best-fitting model, for all analysed years, was to assume that stayers report their labour market condition correctly and that, for movers, measurement errors are constant over time and that the two indicators based on self-perception, S and R, are correlated, i.e., a direct effect between these two indicators is inserted in the model specification. (All estimated models were identified and, in order to avoid local maxima, estimation was performed several times with different sets of starting values; to estimate more parsimonious models, all three variable interactions were set at 0.) As an example, Tables 5.1 to 5.3 list some of the estimation results: labour market composition and estimated flows for the overall population, movers and stayers together, (The complete set of estimation results is available from the authors.) and estimated measurement errors. On average, over the five years, the percentage of movers was 17.69.

Table 5.1
Estimated labour market composition I quarter 2005 – 2009, %

	2005	2006	2007	2008	2009
E	42.01	42.36	40.72	40.92	40.00
U	5.93	5.64	5.75	5.27	6.46
N	52.07	52.00	53.53	53.81	53.53

E = Employed, U = Unemployed, N = Not in the Labour Force.

Table 5.2
Estimated gross flows I quarter to II quarter 2005 - 2009, %, standard errors in brackets

	EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	96.70	1.60	1.61	17.41	71.80	10.78	0.97	0.70	98.29
	(0.0017)	(0.0012)	(0.0012)	(0.0133)	(0.0142)	(0.0079)	(0.0013)	(0.0011)	(0.0017)
2006	96.10	1.93	1.93	19.16	67.04	13.80	1.71	0.89	97.41
	(0.0027)	(0.0020)	(0.0020)	(0.0112)	(0.0150)	(0.0136)	(0.0011)	(0.0015)	(0.0018)
2007	96.30	1.79	1.89	18.11	67.95	13.94	1.42	1.24	97.34
	(0.0023)	(0.0016)	(0.0017)	(0.0145)	(0.0158)	(0.0094)	(0.0018)	(0.0018)	(0.0025)
2008	96.88	1.77	1.35	18.00	74.57	7.43	1.61	1.03	97.37
	(0.0037)	(0.0027)	(0.0028)	(0.0118)	(0.0157)	(0.0138)	(0.0013)	(0.0017)	(0.0020)
2009	96.50	1.83	1.62	15.04	71.62	13.35	1.55	1.10	97.35
	(0.0024)	(0.0019)	(0.0016)	(0.0153)	(0.0168)	(0.0092)	(0.0019)	(0.0014)	(0.0024)

EE = Employed in both quarters.
EU = Employed in first quarter and Unemployed in second one.
EN = Employed in first quarter and Not in the Labour Force in second one.
UE = Unemployed in first quarter and Employed in second one.
UU = Unemployed in both quarters.
UN = Unemployed in first quarter and Not in the Labour Force in second one.
NE = Not in the Labour Force in first quarter and Employed in second one.
NU = Not in the Labour Force in first quarter and Unemployed in second one.
NN = Not in the Labour Force in both quarters.

Table 5.3a
Estimated measurement errors 2005 – 2009 ILO indicator, %, standard errors in brackets

	EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	99.75	0.02	0.23	0.93	89.72	9.36	0.97	1.04	98.00
	(0.0002)	(0.0001)	(0.0001)	(0.0028)	(0.0050)	(0.0051)	(0.0004)	(0.0003)	(0.0005)
2006	99.75	0.01	0.24	1.17	89.39	9.44	0.55	0.99	98.46
	(0.0007)	(0.0004)	(0.0005)	(0.0025)	(0.0042)	(0.0035)	(0.0003)	(0.0002)	(0.0004)
2007	99.82	0.01	0.24	0.84	88.28	10.88	0.58	0.87	98.55
	(0.0002)	(0.0001)	(0.0002)	(0.0028)	(0.0050)	(0.0051)	(0.0004)	(0.0003)	(0.0005)
2008	99.44	0.10	0.46	1.16	89.36	9.48	0.57	1.38	90.05
	(0.0007)	(0.0004)	(0.0005)	(0.0025)	(0.0042)	(0.0035)	(0.0003)	(0.0002)	(0.0004)
2009	99.77	0.01	0.22	0.43	88.98	10.57	0.33	0.86	98.79
	(0.0001)	(0.0000)	(0.0001)	(0.0025)	(0.0038)	(0.0039)	(0.0003)	(0.0002)	(0.0003)

EE = Truly Employed and classified as Employed by ILO indicator.
EU = Truly Employed but classified as Unemployed by ILO indicator.
EN = Truly Employed but classified as Not in the Labour Force by ILO indicator.
UE = Truly Unemployed but classified as Employed by ILO indicator.
UU = Truly Unemployed and classified as Unemployed by ILO indicator.
UN = Truly Unemployed but classified as Not in the Labour Force by ILO indicator.
NE = Truly Not in the Labour force but classified as Employed by ILO indicator.
NU = Truly Not in the Labour Force but classified as Unemployed by ILO indicator.
NN = Truly Not in the Labour Force and classified as Not in the Labour Force by ILO indicator.

Table 5.3b
Estimated measurement errors 2005 – 2009 S and R indicators, %, standard errors in brackets

	True state	SR								
		EE	EU	EN	UE	UU	UN	NE	NU	NN
2005	E	94.83	1.17	2.28	0.22	0.18	0.11	0.44	0.07	0.70
		(0.0008)	(0.0006)	(0.0005)	(0.0002)	(0.0001)	(0.0002)	(0.0003)	(0.0004)	(0.0003)
	U	0.01	0.00	0.00	0.97	97.16	1.11	0.09	0.31	0.35
		(0.0001)	(0.0001)		(0.0006)	(0.0008)	(0.0004)	(0.0009)	(0.0004)	(0.0003)
	N	0.00	0.00	0.01	0.12	0.70	0.70	0.78	0.98	96.72
				(0.0001)	(0.0005)	(0.0009)	(0.0008)	(0.0004)	(0.0006)	(0.0008)
2006	E	94.86	0.96	2.21	0.16	0.11	0.10	0.45	0.06	1.06
		(0.0052)	(0.0006)	(0.0005)	(0.0001)	(0.0002)	(0.0009)	(0.0001)	(0.0004)	(0.0003)
	U	0.00	0.01	0.00	0.86	97.98	0.50	0.11	0.32	0.22
			(0.0001)		(0.0001)	(0.0006)	(0.0001)	(0.0002)	(0.0003)	(0.0003)
	N	0.01	0.00	0.01	0.13	0.82	0.74	0.71	0.74	96.83
		(0.0001)		(0.0001)	(0.0006)	(0.0005)	(0.0004)	(0.0004)	(0.0001)	(0.0005)
2007	E	95.17	1.06	1.06	0.16	0.11	0.10	0.45	0.06	0.82
		(0.0009)	(0.0003)	(0.0005)	(0.0002)	(0.0004)	(0.0005)	(0.0006)	(0.0004)	(0.0004)
	U	0.00	0.01	0.00	0.90	97.74	0.73	0.09	0.31	0.21
			(0.0001)		(0.0005)	(0.0009)	(0.0003)	(0.0005)	(0.0004)	(0.0002)
	N	0.01	0.01	0.01	0.15	0.59	0.66	1.10	0.89	96.59
		(0.0001)	(0.0001)	(0.0001)	(0.0005)	(0.0006)	(0.0008)	(0.0004)	(0.0004)	(0.0020)
2008	E	94.65	1.48	1.83	0.16	0.02	0.14	0.72	0.04	0.96
		(0.0006)	(0.0009)	(0.0005)	(0.0003)	(0.0006)	(0.0004)	(0.0003)	(0.0004)	(0.0002)
	U	0.00	0.03	0.00	1.32	97.39	0.82	0.05	0.33	0.05
			(0.0001)		(0.0002)	(0.0010)	(0.0009)	(0.0005)	(0.0004)	(0.0004)
	N	0.01	0.02	0.01	0.17	0.45	1.34	1.05	1.50	95.45
		(0.0001)	(0.0001)	(0.0001)	(0.0009)	(0.0005)	(0.0003)	(0.0006)	(0.0004)	(0.0003)
2009	E	96.11	0.65	1.21	0.12	0.24	0.10	0.42	0.10	1.04
		(0.0004)	(0.0002)	(0.0001)	(0.0002)	(0.0003)	(0.0008)	(0.0009)	(0.0008)	(0.0009)
	U	0.01	0.01	0.00	0.59	98.23	0.55	0.08	0.26	0.25
		(0.0001)	(0.0001)		(0.0004)	(0.0004)	(0.0002)	(0.0005)	(0.0006)	(0.0006)
	N	0.01	0.00	0.01	0.08	0.76	0.52	0.74	0.78	97.08
		(0.0001)		(0.0001)	(0.0004)	(0.0002)	(0.0002)	(0.0004)	(0.0003)	(0.0008)

E = Employed, U = Unemployed, N = Not in the Labour Force.
EE = Classified as Employed by Self-perceived and Retrospective indicators.
EU = Classified as Employed by Self-perceived indicator and Unemployed by Retrospective indicator.
EN = Classified as Employed by Self-perceived indicator and Not in the Labour Force by Retrospective indicator.
UE = Classified as Unemployed by Self-perceived indicator and Employed by Retrospective indicator.
UU = Classified as Unemployed by Self-perceived and Retrospective indicators.
UN = Classified as Unemployed by Self-perceived indicator and Non in the Labour Force by Retrospective indicator.
NE = Classified as Not in the Labour Force by Self-perceived indicator and Employed by Retrospective indicator.
NU = Classified as Not in the Labour Force by Self-perceived indicator and Unemployed by Retrospective indicator.
NN = Classified as Not in the Labour Force by Self-perceived and Retrospective indicators.

The estimated labour market composition in the first quarter, compared with the observed one (Table 3.2), shows a percentage of unemployment slightly lower than that obtained with the two self-perception indicators and higher than that with the ILO indicator.

Estimated transitions describe a more stable labour market than that observed with all three indicators, with the only exception of two transitions (see Table 3.4). Estimated gross flows are much more similar to those observed with self-perception and retrospective questions than those observed with the ILO indicator. This evidence also appears from the estimated measurement error (Table 5.3). An immediate objection to this result would be that we used two very similar indicators (the two self-perceptions) and a third one which was quite different (ILO). In fact, a similar result - lower measurement errors for self-perception than for the ILO indicator - was obtained by estimating an LCM model with only two indicators per latent variable: ILO and self-perception.

6 Concluding remarks

This paper presents a latent class approach to correct gross flows from correlated errors. The emphasis is on the capacity to account for correlated classification errors across panel data, due to the rotating design of the survey which generates patterns of missing data and of unobserved heterogeneity.

The latent class approach was applied to transitions in the Italian labour market among the three usual conditions of employed, unemployed and not in the labour force. The data refer to the years from 2005 to 2009 and were collected by the Continuous Italian Labour Force Survey on a sample of Italian households with a 2-2-2 rotating design over quarters. Information on labour force condition in one reference quarter was collected three times: (i) respondents were classified as employed, unemployed or not in the labour force according to the definition of the International Labour Office on the basis of answers to a selected group of questions; (ii) respondents were asked to classify themselves as employed, unemployed or not in the labour force (i.e., the self-perceived condition); (iii) a retrospective question asked about state in the labour market one year previously. This means that three indicators of labour condition were available. The three indicators gave quite different descriptions of the Italian labour market, revealing a significant degree of inconsistency. This evidence indicates measurement error in the data.

The best-fitting model was a mover-stayer LCM, in which latent transitions in the labour market follow a first-order Markov chain, stayers always report their market condition correctly; for movers, measurement errors were constant over time and correlated to the two self-perception indicators; the gender and age of respondents were included as covariates; the rotating design of the survey was treated as information missing at random. The model corrects observed gross flows towards a more stable labour market and estimates that the indicator of labour market condition based on the ILO definition is affected by the greatest degree of measurement error.

A second result found here is that, when unobserved heterogeneity occurs, a mixed LCM model fits the data better than the standard LCM model. This finding is consistent with other reports (e.g., Magidson et al. 2007). However, in our case, the two models estimate the same quantity of measurement error, the difference in fit being due to estimated flows. Instead, the above authors found an overestimation of measurement error when unobserved heterogeneity was not taken into account.

A final consideration regards the sample design of the survey, which is two-stage, as described in Section 3. In our analyses, we did not take into account the complex sample design, but estimated gross flows on the longitudinal population provided by the Italian Institute of Statistics. In subsequent research, it will be of interest to compare how results may be affected by incorporating methods for surveys on complex samples with our estimation strategy, an interesting reference to which was made by Lu and Lohr (2010).

References

Bartolucci, F., Lupparelli, M. and Montanari, A. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistics,* 3, 611-636.

Bassi, F., and Trivellato, U. (2009). A latent class approach for estimating gross flows in the presence of correlated classification errors. In *Methodology of Longitudinal Surveys*, (Ed., P. Lynn), Chichester: Wiley, 367-380.

Bassi, F., Padoan, A. and Trivellato, U. (2012). Inconsistencies in reported characteristics among employed stayers. *Statistica*, 1, 93-109.

Bassi, F., Torelli, N. and Trivellato, U. (1998). Data and modelling strategies in estimating labour force gross flows affected by classification errors. *Survey Methodology*, 24, 2, 109-122. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1998002/article/4348-eng.pdf.

Bassi, F., Croon, M., Hagenaars, J.A. and Vermunt, J.K. (2000). Estimating true changes when categorical panel data are affected by correlated and uncorrelated classification errors. An application to unemployment data. *Sociological Methods and Research*, 29, 230-268.

Biemer, P.P., and Bushery, J.M. (2000). On the validity of Markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26, 2, 139-152. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2000002/article/5534-eng.pdf.

Boschetto, B., Discenza, A.R., Lucarelli, C., Rosati, S. and Fiori, F. (2009). Longitudinal data for the analysis of Italian labor market flows. *Italian Journal of Applied Statistics*, 22, 129-150.

Bound, M., Brown, C. and Mathiowetz, N.A. (2001). Measurement error in survey data. In *Handbook of Econometrics*, (Eds., J.J. Heckman and E. Leamer), Amsterdam: Elsevier, 3705-3843.

Clark, K., and Summers, L.H. (1979). Labour market dynamics and unemployment: A reconsideration. *Brooking Papers on Economic Activity*, 1, 13-69.

Dayton, C.M., and McReady, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83, 173-178.

Flinn, C.J., and Heckman, J.J. (1983). Are unemployment and out of the labour force behaviourally distinct market states? *Journal of Labour Economics*, 1, 28-42.

Gonul, F. (1992). New evidence on whether unemployment and out of the labour force are two distinct states. *Journal of Human Resources*, 27, 329-361.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

Hagenaars, J.A. (1994). Latent variables in log-linear models of repeated observations. In *Latent Variable Analysis. Applications for Developmental Research*, (Eds., A. von Eye and C. Clogg), Thousand Oaks (CA): Sage, 329-352.

International Labour Organization (ILO) (2008). Resolution concerning updating the International Standard Classification of Occupation, Geneva. Paper available at http://www.ilo.org/public/english/bureau/stat/isco/docs/resol08.pdf.

Langeheine, R. (1994). Latent variable Markov models. In *Latent Variable Analysis. Applications for Developmental Research*, (Eds., A. von Eye and C. Clogg), Thousand Oaks (CA): Sage, 373-395.

Lu, Y., and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36, 1, 13-22. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2010001/article/11248-eng.pdf.

Magidson, J., Vermunt, J.K. and Tran B. (2007). Using a mixture of latent Markov model to analyze longitudinal U.S. employment data involving measurement error. In *New Trends in Psychometrics,* (Eds., K. Shigemasu, A. Okada, T. Imaizumi and T. Hoshino), Tokyo: Universal Academy Press, 235-242.

Manzoni, A., Vermunt, J.K., Luijkx, R. and Muffels, R. (2010). Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement errors. *Sociological Methodology*, 40, 39-73.

Paas, L.J., Vermunt, J.K. and Bijmolt, T.H. (2007). Discrete-time discrete-state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A*, 170, 955-974.

Pavlopoulos, D., and Vermunt, J.K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology,* 41, 1, 197-214. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14151-eng.pdf.

Pavlopoulos, D., Muffles, R. and Vermunt, J.K. (2012). How real is mobility between low pay, high pay and non-employment? *Journal of the Royal Statistical Society, Series A*, 170, 749-773.

Poterba, J.M., and Summers, L.S. (1986). Reporting errors and labour market dynamics. *Econometrica,* 54, 1319-1338.

Richiardi, M. (2002). What does the ECHP tell us about labour status misperception? A journey in less known regions of labour discomfort. *LABORatorio Revelli,* Working paper No. 69.

van de Pol, F., and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 33, 231-247.

Vermunt, J.K. (1997). *Log-Linear Models for Event History*. Thousand Oaks (CA): Sage.

Vermunt, J.K. (2010). Longitudinal research using mixture models. In *Longitudinal Research with Latent Variables,* (Eds., K. van Montfort, J.H.L. Oud and A. Satorra), Heidelberg: Springer, 119-152.

Vermunt, J.K., and Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., and Magidson, J. (2013). *Technical Guide for Latent Gold 5.0. Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., Tran, B. and Magidson, J. (2008). Latent class models in longitudinal research. In *Handbook of Longitudinal Research: Design, Measurement, and Analysis,* (Ed., S. Menard), Burlington, MA: Elsevier, 375-385.

Wiggins, L.M. (1973). *Panel Analysis: Latent Probability for Attitude and Behavior Processes*. New York: Elsevier Scientific.

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 125-140
Statistics Canada, Catalogue No. 12-001-X

Variance estimation in multi-phase calibration

Noam Cohen, Dan Ben-Hur and Luisa Burck[1]

Abstract

The derivation of estimators in a multi-phase calibration process requires a sequential computation of estimators and calibrated weights of previous phases in order to obtain those of later ones. Already after two phases of calibration the estimators and their variances involve calibration factors from both phases and the formulae become cumbersome and uninformative. As a consequence the literature so far deals mainly with two phases while three phases or more are rarely being considered. The analysis in some cases is ad-hoc for a specific design and no comprehensive methodology for constructing calibrated estimators, and more challengingly, estimating their variances in three or more phases was formed. We provide a closed form formula for the variance of multi-phase calibrated estimators that holds for any number of phases. By specifying a new presentation of multi-phase calibrated weights it is possible to construct calibrated estimators that have the form of multi-variate regression estimators which enables a computation of a consistent estimator for their variance. This new variance estimator is not only general for any number of phases but also has some favorable characteristics. A comparison to other estimators in the special case of two-phase calibration and another independent study for three phases are presented.

Key Words: Calibration; Multi-phase sampling; Generalized regression.

1 Introduction

Survey statistics makes use of available auxiliary information on known population totals in order to improve survey estimates. A calibration estimator uses calibrated weights which are as close as possible, according to a given distance measure, to the initial sampling design weights, while also satisfying a set of constraints induced by the auxiliary information. Arbitrary sampling designs are allowed at all phases of sampling and the auxiliary information can be used at any phase and is incorporated in the estimation process.

Multi-phase sampling along with calibration to known auxiliary information is a powerful and cost effective technique. The process of calibration has been extensively studied and among the multi-phase designs the special case of two phases was an exception that was elaborately investigated. Rao (1973) and Cochran (1977, chapter 12) provided the basic results for stratification and non-response in two-phase sampling. A detailed framework of the linear weighting approach in two-phase sampling appears in Särndal, Swensson and Wretman (1992, chapter 9). Other estimation procedures were investigated for important sampling designs such as cases when the second-phase sample has been restratified using information gathered from the first-phase sample (Binder, Babyak, Brodeur, Hidiroglou and Jocelyn 2000). The variance estimation has been a main subject of active research using different approaches such as the linearization method as presented in Binder (1996), using jackknife (Kott and Stukel 1997) or other replication procedures (Rao and Shao 1992; Fuller 1998; Kim, Navarro and Fuller 2006). More related to our work, Breidt and Fuller (1993) gave efficient estimation procedures for three-phase sampling in the presence of auxiliary information and Hidiroglou and Särndal (1998) studied the use of auxiliary information for two-phase sampling while allowing a minor modification in the distance function that results with additive

1. Noam Cohen, Dan Ben-Hur and Luisa Burck, Statistical Methodology Department, The Central Bureau of Statistics, 95464 Jerusalem, Israel. E-mail: avinoam.cohen@mail.huji.ac.il.

calibrating factors (also known as $g-factors$) rather than multiplicative ones. A common characteristic of these results is the presentation of last phase calibrated weights via calibrated weights of previous phases. This is a major drawback, as it requires computation of weights of all former phases in order to obtain those of later ones and as a consequence makes it difficult to provide a well established methodology of how to estimate the variance of the calibrated estimators in designs with more than two phases.

To address this problem we use the modification of the generalized least squares (GLS) distance function, introduced by Hidiroglou and Särndal (1998), to provide a presentation of the vector of multi-phase calibrated weights which are presented solely through the initial weights based on the sampling design and does not include $g-$ factors. From this presentation we are able to construct multi-phase calibrated estimators that have the form of multi-variate regression estimators which in turn enable to derive a general formula for a consistent estimator for the variance of multi-phase calibrated estimators that holds for any number of phases of calibration. A comparison in the relatively simple case of two phases, where an alternative formula for an estimator for the variance exists in the literature, shows that the two estimators fundamentally differ in form and interpretation. It is important to note that in that specific case the new proposed variance estimator does not show superiority (nor inferiority) in terms of its bias or variance, though it demonstrates some other favorable characteristics which will be discussed in section 3.2. However, the main goal of this paper is not to prove superiority in the two-phase case but to introduce the alternative approach under which the new presentation of the calibrated weights can produce a closed form formula for an estimator for the variance of multi-phase calibrated estimators that holds for any number of phases.

The paper is organized as follows. Section 2 sets up the notation which will be very similar to the one used by Hidiroglou and Särndal (1998). Section 3 provides the methodology and presents the special cases of two-phase and three-phase calibration in subsection 3.2 more elaborately. In Section 4 we present a simulation study to demonstrate some characteristics of the new approach. Finally, in Section 5 we state our concluding remarks and offer some areas for future study.

2 Notation

We use a similar notation to the one used by Särndal et al. (1992) and Hidiroglou and Särndal (1998). Consider a finite population $U=\{1,\ldots,k,\ldots,N\}$. A first phase probability sample $s_1\left(s_1\subseteq U\right)$ is drawn from the population U using a sampling design that generates the selection probability π_{1k} for the k^{th} unit in the population. Given that s_{i-1} has been drawn, the i^{th} phase sample $s_i\left(s_i\subseteq s_{i-1}\right)$ is selected from s_{i-1} through a sampling design with the selection probabilities $\pi_{ik|s_{i-1}}\equiv \Pr\left(k\in s_i\,\middle|\,k\in s_{i-1}\right)$. Note the conditional nature of the consequent phase selection probabilities. From this point on we work only with weights in the estimation process. The conditioned i^{th} phase sampling weight of unit $k\in s_i$ and its overall sampling weight will be denoted by $w_{ik}=1/\pi_{ik|s_{i-1}}$ and $w_{ik}^{*}=\prod_{j=1}^{i}w_{jk}$ respectively.

Let y_k be the value of the target variable for the k^{th} population unit with which an auxiliary vector $\mathbf{x}_k=\left(x_{1k},\ldots,x_{jk},\ldots,x_{Jk}\right)$ is associated. Denote by y the vector of elements of the target variable obtained at the last phase of sampling, p. As outlined in Särndal et al. (1992, chapter 9), we partition the vector $\mathbf{x}$

as $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_p'\right)'$ with $p \leq J$ so that at certain phases maybe more than one auxiliary variable is obtained. The population total of $\mathbf{x}$, $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$ is assumed to be unknown. However, some demographic totals may be known from relatively accurate sources such as census data or other types of administrative files. Without loss of generality let $\mathbf{x}_1$ be the vector of variables known for all units in the population U. Let $\mathbf{x}_2$ be the vector of variables obtained in the first phase sample s_1, and so on. For elements in s_r, $r \leq p$ the complete information is then summarized in the vector $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_r'\right)'$. Denote also $t_i = t_{\mathbf{x}_i}$.

Let X_r be the design matrix with n_r rows representing n_r sampled units, and a number of columns as the number of auxiliary variables in the vector $\mathbf{x}_r$. Note that X_r is obtained in sample s_{r-1} at the $r-1^{\text{th}}$ phase of sampling so we may think of U as sample s_0. In the setting that appears for example in Särndal et al. (1992) and Hidiroglou and Särndal (1998), the design matrix X_r includes all auxiliary variables $\mathbf{x}_1, \ldots, \mathbf{x}_r$, and not just $\mathbf{x}_r$, and is referred to as the *full vector*. The analysis however is the same in both cases.

The auxiliary information available at each phase of sampling can be used to obtain improved weights through the process of calibration which produces calibration factors to be used in the estimation process. We use the superscript "*" to denote overall weights, i.e., weights taking all phases into account. The superimposed symbol "$\sim$" denotes calibrated weights. The i^{th} phase $g-$factors are denoted by g_{ik}, resulting with i^{th} phase calibrated weights $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ for $k \in s_i$, where $\tilde{w}_{i-1,k}$ are the calibrated weights of the $i-1^{\text{th}}$ phase and $\tilde{w}_{0k} = 1$. For $k \in s_i$ the calibration with respect to all phases produces overall calibration factors denoted as g_{ik}^*. As a result we will have overall calibrated weights $\tilde{w}_{ik} = w_{ik}^* g_{ik}^*$ where w_{ik}^* is the overall sampling weight. Denote by w_i the vector with components $w_{ik}; k = 1, \ldots, n_i$, and D_i a diagonal matrix of size n_i with w_i on its diagonal. The same notation will be used with the vectors w_i^*, $\tilde{w}_i$ and g_i.

3 Calibration with GLS distance

Calibration requires the specification of a distance function measuring the distance between the initial weights and the new calibrated weights. Several distance functions have been studied, see a selected summary in Deville and Särndal (1992). We concentrate on the generalized least squares (GLS) distance measure. The conventional form of multi-phase calibration under the GLS distance finds the values $\tilde{w}_{ik}$ for the set $k \in s_i$ that minimize the expression

$$\sum_{k \in s_i} \frac{c_{ik} \left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{\tilde{w}_{i-1,k} w_{ik}} \tag{3.1}$$

subject to

$$\sum_{k \in s_i} \tilde{w}_{ik} x_{ik} = \sum_{k \in s_{i-1}} \tilde{w}_{i-1,k} x_{ik} \tag{3.2}$$

(alternatively, one can write $\tilde{w}_{i-1,k} w_{ik} g_{ik}$ instead of $\tilde{w}_{ik}$) where $\{\tilde{w}_{i-1,k}: k \in s_i\}$ are the initial weights at the beginning of phase i, i.e., the calibrated weights obtained at phase $i-1$; $\{\tilde{w}_{ik}: k \in s_i\}$ are the calibrated weights of phase i that we want to obtain; and $\{c_{ik}: k \in s_i\}$ are specified positive factors used to control the relative importance that we are willing to assign to each of the elements of the sum on the basis of the auxiliary information available for $k \in s_{i-1}$. For simplicity of notation assume from now on that $c_{ik} = 1$ for all i, k. The weights resulting from this calibration scheme are $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ where $g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik}$ with $T_i = \sum_{l \in s_i} w_{il}^* g_{i-1,l}^* x_{il} x_{il}'$. Hence, the calibration factors in this process operate multiplicatively with an overall calibration factor $g_{ik}^* = \prod_{j=1}^{i} g_{jk}$ for $k \in s_i$ at the end of phase i.

Distance measure (3.1) may be criticized, because the factors $1/\tilde{w}_{i-1,k} w_{ik}$ for some i may not all necessarily be finite and positive, as the terms $g_{i-1,k}$ that appear in $\tilde{w}_{i-1,k}$ in the denominator can be zero or negative, contradicting the notion of distance. An alternative choice of distance function, and the one that we shall use in our analysis, is to replace (3.1) with

$$\sum_{k \in s_i} \frac{\left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{w_{i-1,k}^* w_{ik}} \tag{3.3}$$

i.e., with non-calibrated weights in the denominator. It is easy to verify that the overall calibrated weights resulting from minimizing (3.3) subject to (3.2) are (for $p = 2$ see Hidiroglou and Särndal 1998)

$$\tilde{w}_{pk} = w_{pk}^* \left(g_{1k} + \ldots + g_{ik} + \ldots + g_{pk} - (p-1)\right) \tag{3.4}$$

where

$$g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik} \tag{3.5}$$

for $k \in s_p$ with $T_i = \sum_{l \in s_i} w_{il}^* x_{il} x_{il}'$. The choice of a distance measure in the construction of calibrated estimators is not critical since the resulting estimators within a wide range of distance measures are asymptotically equivalent to the one that uses the GLS distance measure (3.1), Deville and Särndal (1992). This is the case with distance measure (3.3) as well. Since the Horvitz-Thompson estimator $X_1' w_1^*$ is unbiased for t_1 with standard deviation of magnitude $N \cdot O\left(n_1^{-1/2}\right)$ then $g_{1k} = 1 + O\left(n_1^{-1/2}\right)$ for all $k \in s_1$ and hence $\tilde{w}_{1k} = w_{1k}^* \left(1 + O\left(n_1^{-1/2}\right)\right)$. Inductively $g_{ik} = 1 + O\left(n_i^{-1/2}\right)$ for all i and from (3.4) $\tilde{w}_{pk} / w_{pk}^* \to 1$ in probability with n_p. New techniques to improve estimation were suggested by Farrell and Singh (2002) by proposing other types of penalized chi-square distance function.

3.1 Estimation

The motivation to our next analysis comes from the recursive nature of $\tilde{w}_{ik}$ in (3.4), where calibrated weights of previous phases $1, \ldots, i-1$ are nested in each g_{ik}, thus require the computation of the calibrated

weights sequentially, i.e., one has to compute all calibrated weights of previous phases in order to obtain those of later ones. Let $\hat{B}_{ij}^{+}=\left(\sum_{k\in s_i} w_{ik}^{*}x_{ik}x_{ik}'\right)^{-1}\sum_{k\in s_j} w_{jk}^{*}x_{ik}x_{jk}'$ and $\hat{B}_{ij}^{-}=\left(\sum_{k\in s_i} w_{ik}^{*}x_{ik}x_{ik}'\right)^{-1}\sum_{k\in s_{j-1}} w_{j-1,k}^{*}x_{ik}x_{jk}'$ be estimators for $B_{ij}=\left(\sum_{k\in U} x_{ik}x_{ik}'\right)^{-1}\sum_{k\in U} x_{ik}x_{jk}'$, the regression coefficient of $\mathbf{x}_j$ on $\mathbf{x}_i$. The difference between the two estimators is that while $\hat{B}_{ij}^{-}$ uses the entire set of units known for $\mathbf{x}_j$ which is obtained in s_{j-1}, $\hat{B}_{ij}^{+}$ uses only the subset $s_j\subseteq s_{j-1}$ and thus more variables than $\hat{B}_{ij}^{-}$. Let $\hat{Z}_{ij}=\hat{B}_{ij}^{+}-\hat{B}_{ij}^{-}$ the difference between the two coefficients estimates which is consistent to zero. Denote also $\hat{Z}_{i_1i_2\ldots i_k}=\prod_{j=2}^{k}\hat{Z}_{i_{j-1}i_j}$ for $k\geq 2$ and $\hat{Z}_{i_1}=1$ for $k=1$. Let $\hat{t}_i^{-}=\sum_{k\in s_{i-1}} w_{i-1k}^{*}x_{ik}$ and $\hat{t}_i^{+}=\sum_{k\in s_i} w_{ik}^{*}x_{ik}$ be two Horvitz-Thompson estimators for t_i, based on the units obtained in samples s_i and s_{i-1} respectively. Note that all the estimators defined in this paragraph use overall design weights w^{*} and not calibrated weights. In the following lemma we provide a presentation of $\tilde{w}_p$, the vector of calibrated weights after p phases of calibration, that relies solely on the pre-known sampling design weights $\{w_i^{*}\}_{i=1}^{p}$.

Lemma 3.1 *Consider a multi-phase sampling design with a calibration scheme that produces additive $g-$ factors as defined in (3.3). A presentation of the calibrated weights at phase p that is based entirely on the design weights is*

$$\begin{aligned}\tilde{w}_p &= D_p^{*\prime}1_{n_p}+\sum_{i_1=1}^{p}A_{i_1}-\sum_{i_1<i_2}^{p}A_{i_1i_2}\\ &\quad+\ldots+(-1)^{k+1}\sum_{i_1<\ldots<i_k}^{p}A_{i_1i_2\ldots i_k}+\ldots+(-1)^{p+1}A_{i_1i_2\ldots i_p}\end{aligned}\tag{3.6}$$

where $A_{i_1i_2\ldots i_k}'=\left(\hat{t}_{i_1}^{-}-\hat{t}_{i_1}^{+}\right)'\hat{Z}_{i_1i_2\ldots i_k}\left(X_{i_k}'D_{i_k}^{*}X_{i_k}\right)^{-1}X_{i_k}'D_p^{*}$.

Proof. See Appendix A.

Note the "Inclusion-Exclusion" form of $\tilde{w}_p$ in Lemma 3.1. The k'^{th} summation involves $\binom{p}{k}$ summands $A_{i_1i_2\ldots i_k}$, for which each $\hat{Z}_{i_1i_2\ldots i_k}=\prod_{j=2}^{k}\left(\hat{B}_{i_{j-1}i_j}^{+}-\hat{B}_{i_{j-1}i_j}^{-}\right)$ contains 2^k summands. Thus, a total of $\binom{p}{k}2^k$ summands. The overall number of terms in (3.6) is therefore 3^p as acknowledged in the proof of the lemma. Note also that the terms $A_{i_1i_2\ldots i_k}$ involve the product of the components $\hat{t}_{i_1}^{-}-\hat{t}_{i_1}^{+}$ and $\hat{Z}_{i_1i_2\ldots i_k}$, both having zero expectation, so the calibrated weight $\tilde{w}_p$ therefore equals to $D_p^{*\prime}1_{n_p}$, the overall design weight, plus correction terms of lower orders of magnitude, and maintains the familiar characteristic of calibrated weights. In our discussion so far we have merely provided a presentation to the vector of weights in a multi-phase calibration process which is constituted of design parameters only and does not include $g-$factors. However, from this presentation of $\tilde{w}_p$ it is possible to deduce an innovative estimator for the variance of multi-phase calibrated estimators. Let y be some variable of interest for which we want to estimate the population total Y. Let $\hat{\beta}_j=\left(\sum_{k\in s_j} w_{jk}^{*}x_{jk}x_{jk}'\right)^{-1}\sum_{k\in s_p} w_{pk}^{*}x_{jk}y_k$, the regression coefficient of y on $\mathbf{x}_j$, and $\hat{Y}_{\mathrm{HT}_p}=1_{n_p}'D_p^{*}y$ the non-calibrated Horvitz-Thompson estimator computed over the elements in s_p. Rearranging the terms in (3.6) produces a more conventional presentation of the multi-phased calibrated estimator $\tilde{w}_p'y$ as a multi-variate regression estimator

$$\tilde{w}_p' y = \hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p} \left(\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+}\right)' \hat{\gamma}_{i_1} \tag{3.7}$$

where

$$\hat{\gamma}_{i_1} = \hat{\beta}_{i_1} - \sum_{i_1<i_2}^{p} \hat{Z}_{i_1 i_2} \hat{\beta}_{i_2} + \ldots + (-1)^{k+1} \sum_{i_1<\ldots<i_k}^{p} \hat{Z}_{i_1 i_2 \ldots i_k} \hat{\beta}_{i_k} + \ldots + (-1)^{p-(i_1-1)+1} \hat{Z}_{i_1 \ldots p} \hat{\beta}_p .$$

A derivation of a consistent estimator for the variance of multi-phase calibrated estimators is now straightforward in the sense that it roughly follows the steps used in the derivation of the variance in a one-phase multi-variate calibration scheme.

Theorem 3.1 *Let* $\hat{e}_{rk} = x_{rk}' \hat{\gamma}_r - x_{r+1,k}' \hat{\gamma}_{r+1}$ *for* $r < p$ *and* $\hat{e}_{pk} = x_{pk}' \hat{\gamma}_p - y_k$. *A consistent estimator for the variance of* $\tilde{w}_p' y$ *is*

$$\sum_{1\le r_1, r_2 \le p} \sum_{k \in s_{r_m},\, l \in s_{r_M}} \frac{w^*_{r_M l}}{w^*_{r_m l}} \left(w^*_{r_m k} w^*_{r_m l} - w^*_{r_m k l}\right) \hat{e}_{r_m k} \hat{e}_{r_M l} \tag{3.8}$$

where $r_m = \min(r_1, r_2)$ *and* $r_M = \max(r_1, r_2)$.

The proof involves evaluation of the highest orders of magnitude and the estimation of their variance. Special attention is given to the evaluation of the joint probability of events $\{k \in s_i, l \in s_j\}$ and estimation of the covariance between units from different phases of sampling.

Proof. In the first step we will show that the substitution of the coefficient estimators $\hat{\gamma}_i; i = 1 \ldots p$ by their true values γ_i affects the estimation of the variance by a factor of $N^2 o\left(n_p^{-1}\right)$ and hence not affecting the consistency of the substituted estimator. To this end note that $\hat{B}_{ij}^{+}, \hat{B}_{ij}^{-}$ are both consistent to B_{ij}. Write $\hat{B}_{ij}^{+} = B_{ij} + \left(\hat{B}_{ij}^{+} - B_{ij}\right)$ so $\hat{B}_{ij}^{+} = B_{ij} + O_p\left(n_j^{-1/2}\right)$. Recall that $\hat{Z}_{ij} = \hat{B}_{ij}^{+} - \hat{B}_{ij}^{-}$ where $\hat{B}_{ij}^{-}$ is based on s_{j-1} while $\hat{B}_{ij}^{+}$ over its subsample s_j and thus $\hat{Z}_{ij} = O_p\left(n_j^{-1/2}\right)$ and therefore $\hat{Z}_{i_1 i_2 \ldots i_k}$ is bounded by $O_p\left(n_{i_k}^{-1/2}\right)$. Likewise $\hat{\beta}_j$ is $\beta_j + O_p\left(n_p^{-1/2}\right)$ because y is observed only at the last phase of sampling s_p. Hence $\hat{\gamma}_i$ is consistent for γ_i for all i, where $\hat{\beta}_i$ in $\hat{\gamma}_i$ are replaced with β_i in γ_i. Consistency does not necessarily imply the convergence of the moments and specifically not of the variance. However, for a finite population, i.e., a finite probability space, the concepts coincide. It follows that for n_p large enough $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p} \left(\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+}\right)' \hat{\gamma}_{i_1}\right)$ and $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p} \left(\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+}\right)' \gamma_{i_1}\right)$ are asymptotically equivalent and following the above discussion the difference can be quantified by

$$\mathrm{Var}\left(\tilde{w}_p' y\right) = \mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{r=1}^{p} \left(\hat{t}_r^{-} - \hat{t}_r^{+}\right)' \gamma_r\right) + N^2 o\left(n_p^{-1}\right).$$

The estimator $\hat{t}_r^{+}$ is a summation over units in s_r while $\hat{t}_r^{-}$ is over s_{r-1}. Rearranging the terms, the variance on the right hand side can be written as $\mathrm{Var}\left(\sum_{r=1}^{p} \sum_{i \in S_r} w_{ri}^* e_{ri}\right)$ which is equal to

$$\sum_{1\le r_1,\, r_2\le p}\ \sum_{k\in U}\ \sum_{l\in U} w^*_{r_1k} e_{r_1k} w^*_{r_2l} e_{r_2l} \mathrm{Cov}\left(I_{k\in s_{r_1}},\, I_{l\in s_{r_2}}\right)$$

so a sample based estimator would be

$$\sum_{1\le r_1,\, r_2\le p}\ \sum_{k\in s_{r_1},\, l\in s_{r_2}} w^*_{r_1k}\hat{e}_{r_1k} w^*_{r_2l}\hat{e}_{r_2l}\left[1-\frac{P\left(k\in s_{r_1}\right)P\left(l\in s_{r_2}\right)}{P\left(k\in s_{r_1},\, l\in s_{r_2}\right)}\right]. \tag{3.9}$$

To compute the covariance between the indicators $I_{k\in s_{r_1}}$ and $I_{l\in s_{r_2}}$ we need to know the joint probability of events $\{k\in s_i,\, l\in s_j\}$. If $s_j\subset s_i$, then $P\left(k\in s_i,\, l\in s_j\right)$ equals the joint probability that both units k, l are in sample $s_i = s_{\min(i,j)}$, multiplied by the conditional probability that unit l is in sample s_j given that it belongs to s_i. Formally, if $s_j\subset s_i$ then $P\left(k\in s_i,\, l\in s_j\right) = \frac{w^*_{il}}{w^*_{jl}} w^{*-1}_{i,lk}$, hence eliminating the dependence on s_{r_2} in the brackets in (3.9) and the result follows.

Another way to write (3.8) is

$$\sum_{1\le r\le p}\ \sum_{k,l\in s_r}\left(w^*_{rk}w^*_{rl}-w^*_{rkl}\right)\hat{e}_{rk}\hat{e}_{rl} + 2\sum_{1\le r_m<r_M\le p}\ \sum_{k\in s_{r_m}}\ \sum_{l\in s_{r_M}} w^*_{r_mk}\hat{e}_{r_mk} w^*_{r_Ml}\hat{e}_{r_Ml}\left(1-\frac{w^*_{r_mkl}}{w^*_{r_mk}w^*_{r_ml}}\right).$$

When $p=2$ the terms γ_i coincide with the deviation units obtained from the decomposition of the sampling error of the *two-step estimator* of Breidt and Fuller (1993). Consistent estimates for the standard deviations of calibrated subpopulation total estimates are derived in the ordinary way by multiplying the target variable by an indicator variable for the specific subpopulation.

In our discussion so far we have provided a presentation of the vector of calibrated weights from which we have derived a new consistent estimator for the variance of multi-phase calibrated estimators. However, under certain cases the estimators can be furthermore simplified without loss of accuracy. Two scenarios will be discussed here briefly and are dependent on whether n_j is significantly smaller than n_{j-1} or not, that is, whether for all j subsample s_j is significantly smaller than s_{j-1}. A typical case of the first scenario is when we have a set of nested administrative files of significantly diminishing sizes. The first set may be, for example, a population registry file that contains a limited number of variables about the whole population, like age, gender, etc. The second set can be a sample data from a wide national survey where comprehensive household data were collected on all sampled units, but with an additional questionnaire for a subgroup of those units (say, every tenth unit). This subgroup of units can now be calibrated to those two former sources of information. An example of the second scenario is when a few phases of calibration are undertaken over the same set of data. In other words, contrary to the customary multi-phase process, the element of sampling is present only in the first phase but not in later phases. Such a scenario may arise if we want to calibrate a survey to many variables for which we don't have their cross sectional totals but only their marginals. In such cases a sequence of calibrations over the same sample, but with a different set of auxiliary variables on each phase, while usually assigning the last phases for the most important variables, may be a satisfactory compromise. This scenario may better be referred to as *sequential*. Under these

scenarios $\tilde{w}_p$ and its variance can be vastly simplified. These scenarios can be stated as corollaries of our analysis but we choose not to consider them here in order to focus on our current results.

3.2 Examples: Two-phase and three-phase calibration

Two-phase calibration. We will use the special case of two-phase calibration $(p = 2)$ to demonstrate the new methodology and its distinction from the alternative estimator commonly used in literature. The calibrated estimator under matrix notation is given according to (3.7) by

$$\tilde{w}_2' y = \hat{Y}_{\mathrm{HT}_2} + \left(\hat{t}_1^{-} - \hat{t}_1^{+}\right)' \hat{\gamma}_1 + \left(\hat{t}_2^{-} - \hat{t}_2^{+}\right)' \hat{\gamma}_2$$

where $\hat{\gamma}_1 = \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2$ and $\hat{\gamma}_2 = \hat{\beta}_2$. Explicitly in non matrix form

$$\tilde{w}_2' y = \sum_{k\in s_2} w_{2k}^* y_k + \left(\sum_{k\in U} x_{1k} - \sum_{k\in s_1} w_{1k} x_{1k}\right)\hat{\gamma}_1 + \left(\sum_{k\in s_1} w_{1k} x_{2k} - \sum_{k\in s_2} w_{2k}^* x_{2k}\right)\hat{\gamma}_2$$

where

$$\hat{\gamma}_1 = \left(\sum_{k\in s_1} w_{1k} x_{1k} x_{1k}'\right)^{-1}\left[\sum_{k\in s_2} w_{2k}^* x_{1k} y_k - \left(\sum_{k\in s_2} w_{2k}^* x_{1k} x_{2k}' - \sum_{k\in s_1} w_{1k} x_{1k} x_{2k}'\right)\hat{\gamma}_2\right]$$

$$\hat{\gamma}_2 = \left(\sum_{k\in s_2} w_{2k}^* x_{2k} x_{2k}'\right)^{-1} \sum_{k\in s_2} w_{2k}^* x_{2k} y_k .$$

This estimator produces identical estimates to the two-phase calibrated estimator used in Hidiroglou and Särndal (1998) or in Särndal et al. (1992) section 9.7. But once one has computed the estimator of the parameters γ_1, γ_2, the presentation of $\tilde{w}_2' y$ becomes simple and informative, having the structure of a simple multi-variate regression estimator. This linear estimator is based on the coefficients γ which encompass the total effect of the variable $\mathbf{x}$ they multiply and hence slightly differ from the β coefficients. $\hat{\gamma}_i$ encompasses the overall effect of the calibration to variable $\mathbf{x}_i$ on the estimation of Y. In the general case it takes into account the projection of y on $\mathbf{x}_i$, the projection of y on $\mathbf{x}_{i+1}$ multiplied by the projection of $\mathbf{x}_{i+1}$ on $\mathbf{x}_i$ and so on. Moreover, as we will now show, the variance estimators differ significantly both in estimates and presentation. Because of the complication in evaluating the variance of estimators that involve $g-$factors, the common practice used up till now in literature for two phases involved first approximating the $g-$factors by 1, and then use the law of total variation to obtain two components, one for each phase, according to

$$\hat{V}_C\left(\tilde{w}_2' y\right) = \sum_{k,l\in s_2} w_{2kl}\left(w_{1k}w_{1l} - w_{1kl}\right)\left(g_{1k}\breve{e}_{1k}\right)\left(g_{1l}\breve{e}_{1l}\right)$$
$$+ \sum_{k,l\in s_2} w_{1k}w_{1l}\left(w_{2k}w_{2l} - w_{2kl}\right)\left(g_{2k}\breve{e}_{2k}\right)\left(g_{2l}\breve{e}_{2l}\right) \qquad (3.10)$$

where the error terms $\breve{e}_{1k} = y_k - x'_{1k}\hat{\gamma}_1$ and $\breve{e}_{2k} = y_k - x'_{2k}\hat{\gamma}_2$ are both defined for $k \in s_2$ because y is observed only at s_2 and note the simple presentation of the error terms under the notation that uses the γ coefficients. The $g-$factors are defined as in (3.5). The approximation of the $g-$factors by 1 in the derivation of (3.10) may undoubtedly lead to unpredictable estimates as those factors depart from unity exactly in those situations where calibration was essential. On the other hand, the variance estimator proposed in (3.8) for a two-phase calibrated estimator is given by

$$\hat{V}_P\left(\tilde{w}'_2 y\right) = \sum_{k,l\in s_1}\left(w_{1k}w_{1l} - w_{1kl}\right)\hat{e}_{1k}\hat{e}_{1l} + \sum_{k,l\in s_2}\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)\hat{e}_{2k}\hat{e}_{2l}$$
$$+ 2\sum_{k\in s_1, l\in s_2}\frac{w^*_{2l}}{w_{1l}}\left(w_{1k}w_{1l} - w_{1kl}\right)\hat{e}_{1k}\hat{e}_{2l}. \qquad (3.11)$$

The difference in the variance estimator between the two methods represented by equations (3.10) and (3.11) is fundamental. It is expressed in a couple of aspects. While the error term of the second phase in both methods is the same, i.e., $\hat{e}_{2k} = \breve{e}_{2k}$, the error term of the first phase differs. $\breve{e}_{1k}$ is based on the difference between y_k and the regression predictor $x'_{1k}\hat{\gamma}_1$ while $\hat{e}_{1k}$ is based on the difference between two predictors of Y from phases one and two $x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$. This modification causes the first summand in (3.11) to be computed over s_1 and not s_2 where the sample is larger. Noticeably, the estimator (3.11) has a third summand which involves the product of the two error terms from both phases that has no parallel in (3.10). Although this product will often be close to zero whenever the error terms are not strongly correlated, it may still not be negligible whenever y is strongly correlated with $\mathbf{x}_1$. An evident advantage is the absence of the $g-$factors which makes the estimator much simpler to compute, i.e., once we have computed the parameters estimates $\hat{\gamma}_i; i = 1\ldots p$, the estimator (3.11) can be computed using design parameters only without carrying the $g-$factors from all phases of calibration. Last, and maybe from an operational point of view more important, as will be also shown in the simulation study, (3.11) has the advantage that in a wide range of designs the second summand constitutes the absolute majority of the variance while the summands in (3.10) are usually of the same order of magnitude. This characteristic stems from the fact that the term $\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)$ which involves the total sampling weights is very large in comparison with $w_{2kl}\left(w_{1k}w_{1l} - w_{1kl}\right)$ or $w_{1k}w_{1l}\left(w_{2k}w_{2l} - w_{2kl}\right)$. In the variance estimator the function $f(w) = w_k w_l - w_{kl}$ attains its maximum on the diagonal $k = l$ where it is proportional to w_k^2 and then it is multiplied by the second power of its remainder $\hat{e}_k$ which is a non-negative term. So when the sampling rate of the second phase is high enough it drastically increases terms which are dependent on total weights of that phase w^*_2, in comparison with a parallel term from the previous phase. Hence the second summand may therefore be a good estimator of the variance of the calibrated estimator practically on its own.

Three-phase calibration. Multi-phase calibration can be implemented when in a series of samples of diminishing (non-increasing) sizes each pair of consequent phases share some common variables. It can be held whether the samples are nested, i.e., s_i is a subsample of s_{i-1}, or not. In practice, the simplest and most common case of course is of two phases when a smaller sample (nested or not) is being calibrated to

a much bigger sample such as a Labor Force Survey which in turn is frequently calibrated to an administrative file with demographic variables. However, due to computational feasibility and development of methodology, designs with more phases of calibration are still popular and three-phase designs are second in line in terms of their simplicity and implementation. It is therefore worthwhile to elaborate on the estimator for this case a bit further.

The approximation (3.8) involves six different terms, three for the three phases of sampling and another three for the covariance between phases. We denote these terms by V_1, V_2, V_3 and C_{12}, C_{13}, C_{23} respectively. Each is a multiplication of a term that involves sampling weights multiplied by remainders from the relevant phases. The formulae for three-phase calibration are presented in appendix B. As discussed for the two-phase case, when $w_i > 1$ the V_i's are likely to follow a clear order $V_1 < V_2 < V_3$ and V_3 will become more and more dominant the bigger the sampling rates of the third phase will be. This is marked as case 3 in Table 3.1, and in our simulation it is manifested in rows 2 and 6 of Table 4.2 where w_{3k} were 10 and 5 respectively. Clearly, in reality this is many times not the case as the approximation also depends on the sizes of the remainder terms which rely on the choice of the calibrating variables and their specific correlations which may be very strong. In which cases the remainders will be very small and it would be better to use all terms of (3.8). As for the covariance terms, although C_{13} involves overall weights $\{w^*_{3k}\}$, it is unlikely to add any substantial value to the total variance due to the generally weak correlation between the remainders of phases 1 and 3. On the other hand, the term C_{23}, although weighted by overall 2^{nd} phase weights only, may be significant due to the strong correlation between the remainders of phases 2 and 3 as they both include the term $x'_{3k}\hat{\gamma}_3$ for $k \in s_3$. The relative importance of the terms for some general designs is specified in Table 3.1. The γ coefficients which encompass the total effect of the variables $\mathbf{x}$ they multiply now take a more interesting and complicated form. $\hat{\gamma}_1$ for example takes into account the projections of $\mathbf{x}_1$ on $\mathbf{x}_2$ and of $\mathbf{x}_1$ on $\mathbf{x}_3$, but deducted of the projection of $\mathbf{x}_1$ over the projection of $\mathbf{x}_2$ on $\mathbf{x}_3$.

Table 3.1
A general presentation of the relative importance of each of the terms in (3.8) for some specific scenarios. Black bullets represent highly dominant terms, dark-gray moderate, and light-gray non-dominant terms

Case	Description	V_1	V_2	V_3	C_{12}	C_{13}	C_{23}
1	Hardly any additional sampling in the second and third phases: $w_2 \approx w_3 \approx 1$.	●	●	●	○	●	○
2	Weights w_1, w_2, w_3 are of moderate sizes.	○	◉	●	○	●	○
3	n_3 substantially smaller than n_2, regardless the sizes of w_1, w_2.	○	○	●	○	○	○

4 A simulation study

The main objective of our analysis in this paper was to provide a consistent estimator for the variance of multi-phase calibrated estimators that holds for any number of phases of calibration. A simulation study could thus be executed to compare the innovative estimator with others found in the literature. As generally

no alternative estimators are found for schemes with three or more phases $(p \geq 3)$ we conducted the comparison mainly for the most investigated case of two phases. Another study was preformed for $p = 3$ to evaluate the deviation of the proposed estimator from the true simulated value. The studies are described here in general terms. They meant basically to demonstrate the relevancy of the proposed estimator, its concurrence with the "boundary condition" of two phases and its potential for designs with more than two phases. An extensive study to characterize the efficiency of the proposed estimator as a function of the design parameters such as the sampling rates, the choice of calibrating variables and their correlation with y, etc., is left for future research.

An estimation process of two-phase calibration was applied to data from a recent survey on career and mobility of Doctorate Holders' (DHs). As there exists no frame of DHs, data about higher education was extracted from a recent population census. However, only a sample S_1 that constitutes one fifth of the households enumerated in the census were given an elaborated questionnaire that includes also questions about higher education. A subsample S_2 from S_1 was drawn for the DHs survey in which a further elaborated questionnaire was given to those who were in fact DHs. Thus, a two-phase calibration scheme to estimate characteristics of DHs was in order. The first phase calibrated the joint variables of S_1 and S_2 to estimated totals computed from S_1. In the second phase, demographic data of S_1 was calibrated to the known totals from the full population register U. We conducted a simulation study on that data where the survey data served in our study as the true population. One thousand samples (realizations) $\{u, s_1, s_2\}$ of sizes $N = 1{,}000$, $n_1 = 200$, $n_2 = 50$ were randomly drawn from the dataset S_2 of DHs. To each sample we applied the same process of two-phase calibration utilizing the estimator given by (3.7) with equation (3.6) as a presentation of the calibrated weights $\tilde{w}_2$, and its variance estimator given by (3.11) as a special case of (3.8). As already indicated, when $p = 2$ the estimates $\hat{Y} = \tilde{w}_2' y$ are identical either under the new presentation or under the conventional one used so far in the literature, Särndal et al. (1992) so our focus was on the variance estimators (3.10) and (3.11) computed according to the two different methods. A typical pattern of the comparison between the two variance estimators in this special case of two-phase calibration is presented in Figure 4.1. It can be seen that although the fundamental difference between the two variance estimators, in most realizations, the difference between their estimates is quite small. Though on a certain one it can reach up to 20%. For that particular variable shown in the figure, the mean value of both estimators for the variance was very similar, namely, 54.17^2 and 54.65^2, while the true value in the simulation data was 54.46^2. Even the variance of their standard deviation estimator, namely, 5.73^2 versus 5.93^2 were almost the same for that variable. These results are reported in Table 4.1. The favorable characteristic of the proposed estimator stands out in the 5^{th} column. Contrary to the conventional estimator where the two terms of the variance estimator are of the same order of magnitude, the 2^{nd} term of (3.11) constitutes over 99% of the variance, with a variation of less than two percent between all 1,000 realizations. We discussed the explanation to this phenomenon in 3.2. The outcomes reported above repeated themselves for all variables studied and we found it irrelevant at this point to present other variables or investigate this specific data or the special case of two-phase calibration any further.

Table 4.1
Proposed(P) Vs Conventional(C) estimator for the standard deviation of a two-phase calibrated estimator

Variable	Mean value	Std	CI coverage	2nd term as percent of $\widehat{\text{Std}}\left(\tilde{w}_2' y\right)$
$\tilde{w}_2' y$	200.43	54.46		
$\widehat{\text{Std}}_C\left(\tilde{w}_2' y\right)$	54.65	5.93	95.2%	77% ± 7%
$\widehat{\text{Std}}_P\left(\tilde{w}_2' y\right)$	54.17	5.73	95.1%	99% ± 2%

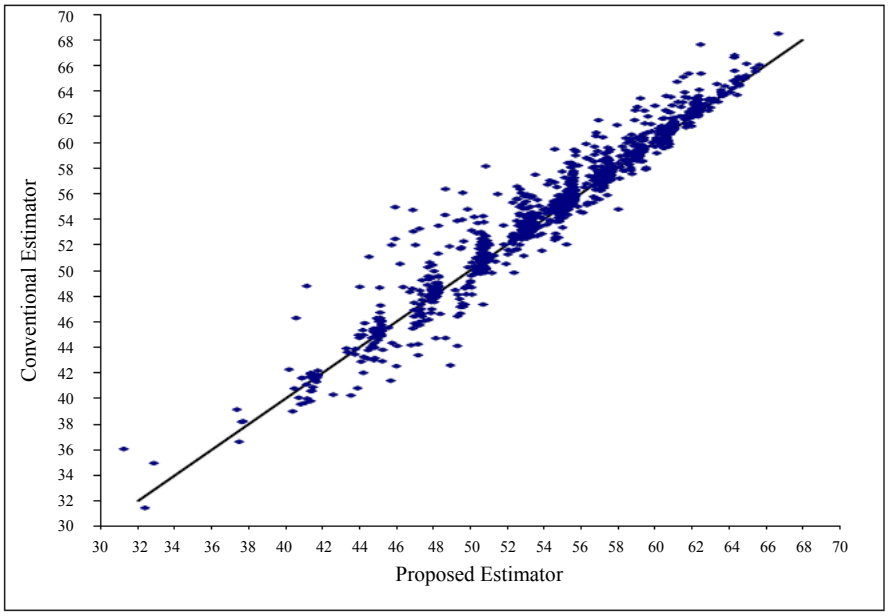

Figure 4.1 Variance estimates in two-phase calibration. A typical pattern of 1,000 realizations of the proposed estimator (equation 3.11) Vs the conventional estimator (equation 3.10) for the variance of a calibrated estimator of Y. The solid line is the main diagonal.

The similarity in estimates of the two variance estimators in the case of two phases is reassuring but a comparison in three or more phases could not be preformed because an alternative estimator to the variance does not exist. A replication method for two-phase stratified sampling was proposed by Kim et al. (2006) and a sketch for a generalization for a three-phase case is briefly outlined but with no explicit formulation or simulation results. In our simulation we added a third calibration phase using some variables, expertise field related, common with the second phase sample of DHs and conducted the study in the same manner as with the two-phase case. The simulation study has again demonstrated an excellent estimation for the variance of a three-phase calibrated estimator for all variables Y examined and all different sets of calibrating variables in all phases. Rapid convergence rates of the variance estimator are displayed even for very small sample sizes such as $n = 25$ or lower at the third phase. Some results for various design parameters are reported in Table 4.2. As portrayed earlier the simulation was performed over a population

size of 1,000 so the first three designs have overall weight of $w^* = 40$ and the next three of $w^* = 20$. So, as expected, the variance of the calibrated estimator for the first three designs is generally higher, although it also depends on the sample sizes of the 1st and 2nd phases as shown for example in the artificial case number 4 which depicts a generally impractical scenario. The relative biases $\frac{E\left(\widehat{\text{Std}}_P\right)}{\text{Std}} - 1$ are close to zero for all designs investigated and the 95% Confidence Interval (CI) coverages were also estimated and found to be mostly conservative and close to their nominal levels. The standard deviation of $\widehat{\text{Std}}_P$ are roughly about 5% - 10% of its value as presented in column 7.

Table 4.2
True and estimated values for the standard deviation of a three-phase calibrated estimator $\tilde{w}_3' y$ for various design parameters

Case	n1	n2	n3	True value	$\widehat{\text{Std}}_P$	Std of $\widehat{\text{Std}}_P$ in %	95% CI coverage
1	100	50	25	882.6	866.9	7.1%	94.9%
2	500	250	25	781.5	774.1	10.8%	95.2%
3	500	100	25	733.9	731.5	10.2%	96.0%
4	50	50	50	902.8	892.1	4.8%	95.6%
5	200	100	50	598.1	591.4	5.4%	94.4%
6	500	250	50	543.0	542.2	8.3%	96.3%
7	333	100	33	650.8	654.4	8.6%	95.3%

5 Concluding remarks

In this paper we have constructed a novel presentation of multi-phase calibrated weights that enables the presentation of a multi-phase calibrated estimator in the form of a one-phase multi-variate calibrated estimator. This presentation enables the derivation of a closed form approximation for the variance of multi-phase calibrated estimators for any number of phases. A comparison with another approximation known in literature for the two-phase case shows that although the two approximations are consistent yet they differ in their estimates, form and interpretation. We have discussed some advantages of the new approximation in the case of two phases and also demonstrated its consistency in a simulation study for three-phase calibration where it performed very well for all designs investigated. The efficiency of the proposed estimator as a function of the sampling rates and other design parameters is left for future research.

Appendix A

To shorten the notation we will conduct our analysis in matrix form. We shall use a convention that for $j > i$ the summation in the scalar products $X_i' w_j$ and $X_i' D_j$ (or with w_j^* or $\tilde{w}_j$) are over units $k \in s_j$ (and not s_i), i.e., over the sample indicated by the latest set of weights in the scalar product. Hence $\hat{Z}_{ij} = \left(X_i' D_i^* X_i\right)^{-1} X_i' \left(D_j^* - D_{j-1}^*\right) X_j$ under this notation.

Proof of Lemma 3.1. The weights that satisfy the calibration equation in the j^{th} phase with initial weights $\tilde{w}_{j-1}$ are given by equation (3.4). Under our matrix notation

$$\tilde{w}_j = D_j^* \left[g_1 + \ldots + g_j - (j-1) \right]$$

where $g_j = 1 + X_j T_j^{-1} \left(X_j' \tilde{w}_{j-1} - X_j' D_j \tilde{w}_{j-1} \right)$ (see equation (3.5)). So

$$\begin{aligned} \tilde{w}_j &= D_{j-1}^* D_j \left[g_1 + \ldots + g_{j-1} - (j-2) + g_j - 1 \right] \\ &= D_j \left[\tilde{w}_{j-1} + D_{j-1}^* \left(g_j - 1 \right) \right]. \end{aligned} \tag{A.1}$$

Plugging g_j gives $\tilde{w}_j = D_j \left[\tilde{w}_{j-1} + D_{j-1}^* X_j T_j^{-1} \left(X_j' \tilde{w}_{j-1} - X_j' D_j \tilde{w}_{j-1} \right) \right]$ which involves the weight $\tilde{w}_{j-1}$ from the previous phase of calibration and its scalar product with X_j' and $X_j' D_j$, while the rest of the multipliers are design parameters. The square brackets contain three summands and thus after j phases of calibration we would have 3^j summands that would involve design parameters only. Substituting $\tilde{w}_{j-1}$ of (A.1) into $X_j' D_j \tilde{w}_{j-1}$ yields

$$\begin{aligned} X_j' D_j \tilde{w}_{j-1} &= X_j' D_j \left\{ D_{j-1} \tilde{w}_{j-2} + D_{j-1}^* \left(g_{j-1} - 1 \right) \right\} \\ &= X_j' D_j D_{j-1} \tilde{w}_{j-2} + X_j' D_j D_{j-1}^* X_{j-1} T_{j-1}^{-1} \left(X_{j-1}' \tilde{w}_{j-2} - X_{j-1}' D_{j-1} \tilde{w}_{j-2} \right) \end{aligned} \tag{A.2}$$

and therefore also

$$X_j' \tilde{w}_{j-1} = X_j' D_{j-1} \tilde{w}_{j-2} + X_j' D_{j-1}^* X_{j-1} T_{j-1}^{-1} \left(X_{j-1}' \tilde{w}_{j-2} - X_{j-1}' D_{j-1} \tilde{w}_{j-2} \right). \tag{A.3}$$

Combining the terms results in an expression for $\tilde{w}_j$ that involves calibrated weights from phase $j-2$ only

$$\begin{aligned} \tilde{w}_j &= D_j D_{j-1} \tilde{w}_{j-2} \\ &\quad + D_j^* X_{j-1} T_{j-1}^{-1} \left(X_{j-1}' \tilde{w}_{j-2} - X_{j-1}' D_{j-1} \tilde{w}_{j-2} \right) \\ &\quad + D_j^* X_j T_j^{-1} \left(X_j' D_{j-1} \tilde{w}_{j-2} - X_j' D_j D_{j-1} \tilde{w}_{j-2} \right) \\ &\quad - D_j^* X_j T_j^{-1} \hat{Z}_{j-1,j}' \left(X_{j-1}' \tilde{w}_{j-2} - X_{j-1}' D_{j-1} \tilde{w}_{j-2} \right). \end{aligned} \tag{A.4}$$

Plugging (A.2) and (A.3) with $j = p$ into (A.1) and recursing $p-1$ times over the respective calibration groups, produces the desired result.

Appendix B

A consistent estimator for the population total in three-phase calibration can be presented by $\hat{w}_3' y = \hat{Y}_{\text{HT}_3} + \sum_{i=1}^{3} \left(\hat{t}_1^- - \hat{t}_1^+ \right)' \hat{\gamma}_i$, where

$$\hat{\gamma}_1 = \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2 - \hat{Z}_{13}\hat{\beta}_3 + \hat{Z}_{12}\hat{Z}_{23}\hat{\beta}_3$$
$$\hat{\gamma}_2 = \hat{\beta}_2 - \hat{Z}_{23}\hat{\beta}_3$$
$$\hat{\gamma}_3 = \hat{\beta}_3.$$

A consistent estimator for the variance is

$$\begin{aligned}\hat{V}_P\left(\tilde{w}_3' y\right) = & \sum_{k,l\in s_1}\left(w_{1k}^* w_{1l}^* - w_{1kl}^*\right)\hat{e}_{1k}\hat{e}_{1l} + \ldots + \sum_{k,l\in s_3}\left(w_{3k}^* w_{3l}^* - w_{3kl}^*\right)\hat{e}_{3k}\hat{e}_{3l} \\ & + 2\sum_{k\in s_1, l\in s_2} w_{2l}\left(w_{1k}w_{1l} - w_{1kl}\right)\hat{e}_{1k}\hat{e}_{2l} + 2\sum_{k\in s_2, l\in s_3} w_{3l}\left(w_{2k}^* w_{2l}^* - w_{2kl}^*\right)\hat{e}_{2k}\hat{e}_{3l} \\ & + 2\sum_{k\in s_1, l\in s_3} w_{2l}w_{3l}\left(w_{3k}^* w_{3l}^* - w_{3kl}^*\right)\hat{e}_{1k}\hat{e}_{3l}.\end{aligned}$$

where $\hat{e}_{1k} = x_{1k}'\hat{\gamma}_1 - x_{2k}'\hat{\gamma}_2$, $\hat{e}_{2k} = x_{2k}'\hat{\gamma}_2 - x_{3k}'\hat{\gamma}_3$ and $\hat{e}_{3k} = x_{3k}'\hat{\gamma}_3 - y_k$ as defined in Theorem 3.1.

References

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology,* 22, 1, 17-22. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1996001/article/14389-eng.pdf.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics,* 28, 751-764.

Breidt, J., and Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā,* 55, 297-309.

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

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

Farell, P.J., and Singh, S. (2002). Penalized chi-square distance function in survey sampling. *Proceedings of Joint Statistical Meeting,* NY, USA.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica,* 8, 1153-1164.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology,* 24, 1, 11-20. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1998001/article/3905-eng.pdf.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of American Statistical Association,* 101, 312-320.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology,* 23, 2, 81-89. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1997002/article/3621-eng.pdf.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika,* 6, 125-133.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

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

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 32, No. 4, December 2016

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. 1, March 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. 1, March/mars 2017

Issue Information

Original Articles

Acknowledgement

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 2, June/juin 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.