Journal of Official Statistics
vol. 30, i. 4 (2014)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 575–577, http://dx.doi.org/10.2478/JOS-2014-0038

Preface

Introduction to the Special Issue on Establishment Surveys

Welcome to this special issue of the Journal of Official Statistics containing articles emanating from the fourth International Conference on Establishment Surveys (ICES IV). We hope that it will present some interesting insights into the world of establishment surveys. If it's somewhere you don't normally tread, do come in and have a look around.

International Conference on Establishment Surveys

The first International Conference on Establishment Surveys (ICES) was held in 1993 in Buffalo, New York, filling a gap in the conference schedule for those working on surveys of businesses (or establishments), farms, institutions and other non-household populations. Many of these surveys are run in the public sector by National Statistical Institutes, although in North American countries such surveys are occasionally undertaken under contract. ICES II and III followed at seven-year intervals, in Buffalo in 2000 and Montreal in 2007.

By 2007, there was a general feeling that the pace of development in establishment surveys had quickened so that seven year conferences were too far apart, and ICES IV followed after five years. These conferences have been well attended: approximately 400, 450, 400 and 250 people respectively, and at least four people have managed to attend all four (the participation and registration lists on which this information is based have various quality issues, so there may be more). They have been likewise prolific, with more than 700 papers given over the conference series. Plans are in place for a four-yearly cycle in the future to fit around the World Statistics Congress. ICES V will take place in Geneva 20-23 June 2016, the first time that ICES has taken place outside North America (for more details of ICES V see www.ices-v.ch).

In her keynote address for ICES II, Susan Linacre (Linacre 2000) wrote that ICES I and ICES II had a striking amount in common, with incremental progress in many areas, but also further development in ICES II in some areas that were experimental or first put forward in ICES I, including additional countries applying ideas originating in other countries or agencies. She noted that this was a clear benefit of the ICES series. Looking through the range of papers presented at subsequent conferences, her comments are still relevant, and many people have had valuable experiences and insights to add to their own work and research through ICES. We hope that the papers presented in this issue will also spark some ideas and further developments, and look forward to seeing the fruits of that at future ICES conferences.

Organisation

The first conferences were put together by interested groups of people one conference at a time. But after ICES III, it became clear that more structure was needed. A Continuation Committee was formed, and the American Statistical Association, which had been strongly associated with ICES from the beginning, was selected as a permanent host organisation. One happy consequence is that the proceedings of all the ICES, previously somewhat difficult to find if you didn't actually attend one, are all available on the ASA's website at www.amstat.org/meetings/ices.cfm.

Trends in Topics

The original ICES highlighted topics that were specific to establishment surveys, including industrial classification, business register development and maintenance, dealing with outliers, sample coordination using permanent random numbers, disclosure avoidance practices in tabulations, and so on. Riviére (2002) would later summarise the characteristics of business statistics that make their methods rather different from those used by social surveys. On the data collection side, a lot of the techniques were similar to those used for social surveys, but the context was completely different, with challenges around reaching the right people to provide the information, evaluating the availability and quality of information in records systems, and developing collection modes that matched rapidly changing office technology. Many of the approaches specific to business surveys were not widely known and the book of invited papers (Cox et al. 1995) was an important reference for a long time.

Topics in ICES II and III reflected the main drivers of developments in survey-taking over the last 20 years, such as:

- electronic data collection and dissemination,
- generalized and integrated processing systems, and
- dealing with nonresponse.

The first two are often motivated by cost considerations. However, the general focus in the presented papers was on the development of quality instruments and processing systems, and the last driver is entirely about understanding and maintaining quality. None of these drivers is unique to establishment surveys, but the approaches needed to address them often are. ICES II featured sessions on improving response rates, including nonresponse management and priority follow-up of nonrespondents. Several countries shared their cutting-edge research on how to collect information with computer-assisted interviewing and through the web for business surveys. Two sessions presented ideas on the use of administrative data to supplement or replace data collection. ICES II was also notable for a number of papers dealing with data editing, a topic which was then a big focus for saving money by reducing editing resources.

ICES III continued several of these trends, with more sessions on electronic data collection, including Web and design interfaces, nonresponse and nonresponse bias, and unified statistical systems and architecture. It also saw a strong representation from the questionnaire testing community, with a wider range of countries using cognitive methods for developing business survey questionnaires and trying to get an understanding of the survey response process within establishments. On this topic, ICES III directly benefited from the first International Workshop on Business Data Collection Methodology in 2006, which brought together questionnaire design researchers and motivated ICES sessions and papers.

ICES IV

The fourth ICES saw the influence of greater use of administrative data to keep costs down and reduce response burden, along with a big push on model-based approaches to inference, traditionally regarded as challenging for establishment survey because of the nonignorability of the sampling, but finding a ready home in some applications in agriculture and retail where there are many smaller establishments. There was continued emphasis on generalized systems, and on alleviating nonresponse and assessing and mitigating non-response bias. More work on cognitive methods to understand survey responses and improve their quality was included, and there were several papers on respondent burden and motivation driven by the BLUE-ETS project in Europe (www.blue-ets.istat.it/).

This special issue highlights interesting developments and innovative research presented at ICES IV. The collection of articles covers the range of statistical processes across the Generic Statistical Business Process Model (GSBPM, www1.unece.org/stat/platform/display/GSBPM/Generic+Statistical+Business+Process+Model). The issue includes some approaches which are quite new for business statistics, such as the adaptive design methods presented in the articles by Beaumont et al. and by Earp et al. The article by Münnich et al. connects different parts of the GSBPM by examining the impact of sample design choices on small area estimation.

Other approaches have been implemented in different types of surveys or provide new "twists" on accepted practices, such as the application of R-indicators to business survey data in the Ouwehand and Schouten article, the ongoing research on mean square estimation with seasonally adjusted data in Sverchkov and Pfeffermann's article, and Cho et al. look at what can be used to predict the variability of surveys using generalized variance functions. Torres van Grinsven et al. examine what motivates people within establishments to respond to surveys, and how their participation can be encouraged, and Sigman et al. look at the influence of the timing of people's participation on the conclusions from a staff survey, a type of survey which has received little attention at ICES to date.

Several articles continue the development of topics which have been a long-running part of ICES. One is the assessment of sampling using coordinated permanent random numbers described in Lindstrom's article. Robbins continues the theme of compensating for non-response with an examination of the use of nonparametric transformations for imputation. Outliers are generally most important in establishment surveys and Mulry et al. compare M-estimation with Winsorization, continuing a line of ICES invited paper sessions on outliers. And Toth presents a new approach to disclosure limitation based on local averaging which has potential to make more establishment survey data available. All of the articles give an idea of the range of interesting topics in establishment surveys, and we hope that they will serve as an introduction and an incentive to learn more.

Acknowledgments

Thanks to all those who took the time to submit papers for this issue, and to the referees whose input has been so helpful. Special thanks to the volunteers from the ICES IV Organising and Programme committees who have served as Associate Editors for this special issue – Darcy Miller, Polly Phipps, Frank Potter, Paul Smith and Katherine (Jenny) Thompson.

References

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. 1995. *Business Survey Methods*. New York: John Wiley & Sons, Inc.

Linacre, S. 2000. "Establishment Surveys Since ICES I: What Has and Hasn't Changed, and What are the Issues for the Future?" ICES II: Proceedings of the Second International Conference on Establishment Surveys, Invited Sessions, 1–8. Alexandria, VA: American Statistical Association.

Riviére, P. 2002. "What makes business statistics special?" *International Statistical Review* 70: 145–159.

Paul Smith, Guest Editor for the Special Issue, and Program Committee Chair for ICES IV Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk

Polly Phipps, Associate Editor for the Special Issue, and Program Committee Chair for ICES V Office of Survey Methods Research, Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington DC 20212, U.S.A. Email: phipps.polly@bls.gov

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 579–606, http://dx.doi.org/10.2478/JOS-2014-0039

In Search of Motivation for the Business Survey Response Task

Vanessa Torres van Grinsven[1], *Irena Bolko*[2], *and Mojca Bavdaž*[3]

Increasing reluctance of businesses to participate in surveys often leads to declining or low response rates, poor data quality and burden complaints, and suggests that a driving force, that is, the motivation for participation and accurate and timely response, is insufficient or lacking. Inspiration for ways to remedy this situation has already been sought in the psychological theory of self-determination; previous research has favored enhancement of intrinsic motivation compared to extrinsic motivation. Traditionally however, enhancing extrinsic motivation has been pervasive in business surveys. We therefore review this theory in the context of business surveys using empirical data from the Netherlands and Slovenia, and suggest that extrinsic motivation calls for at least as much attention as intrinsic motivation, that other sources of motivation may be relevant besides those stemming from the three fundamental psychological needs (competence, autonomy and relatedness), and that other approaches may have the potential to better explain some aspects of motivation in business surveys (e.g., implicit motives). We conclude with suggestions that survey organizations can consider when attempting to improve business survey response behavior.

Key words: Data quality; incentive; organization; respondent; survey participation.

1. Introduction

It is a real challenge for today's survey organizations and researchers to collect information from their studied populations. This challenge is most visible in declining response rates (De Leeuw and De Heer 2002; Baruch 1999) that have stabilized at a low level in research on organizations (Baruch and Holtom 2008) but only because of response-enhancing techniques (Anseel et al. 2010). Less visible, though far from marginal, is the issue of the poor quality of reported data, which is the main (though not the only) reason for the high cost of data editing in governmental surveys of business organizations, which may represent as much as 30% (e.g., Adolfsson et al. 2010) of the total survey cost. Businesses describe statistical reporting as burdensome even if it only represents a tiny proportion of all administrative burdens (Haraldsen et al. 2013). The problems of declining or low response rates, poor data

[1] Faculty of Social Sciences, Utrecht University, Padualaan 14, 3584 CH, Utrecht, and Statistics Netherlands, CBS-weg 11, 6412 EX, Heerlen, Netherlands. Email: V.TorresvanGrinsven@uu.nl
[2] Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia. Email: irena.bolko@gmail.com
[3] Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia. Email: mojca.bavdaz@ef.uni-lj.si

Acknowledgments: The research reported herein was partly funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 244767. We acknowledge the valuable contributions of project colleagues that cooperated in the research at various stages. We would also like to thank the editors, referees, and our interviewees from businesses. The views expressed in this article are those of the authors and do not neccessarily reflect policies of their employers or the European Commission.

quality and burden complaints suggest that a driving force, namely the motivation for the business survey task, is insufficient or lacking.

Most commonly the term *motivation* is used to describe "why a person in a given situation selects one response over another or makes a given response with great energization or frequency" (Bargh et al. 2010, 268). Behavioral outcomes reflect, among other factors, the level of motivation to participate in a task and perform it well. The role of motivation has been acknowledged and indirectly tested in academic and commercial business surveys, for example, through research on incentives or survey design (for an overview of research on response enhancing techniques see, for example, Cycyota and Harrison 2006, or Jobber et al. 2004). More recently, a paradigm shift from a burden-centered to a motivation-centered approach seems to be occurring in governmental business surveys as increasing attention is given to improving the overall survey experience, especially through better communication and relationships with businesses and efforts to understand the business response environment (for a recent overview, see Snijkers et al. 2013). In the context of response burden, survey motivation is often associated with a respondent's perception of the usefulness of the statistics to the business and/or society; it has been considered an important factor for perception of response burden and through that for data quality and response rates (Dale and Haraldsen 2007). Very few studies, however, have established an empirical link between motivation, perceived burden and response behavior (e.g., Kennedy and Phipps 1995; Hedlin et al. 2005; Hedlin et al. 2008; Giesen 2012). These studies provide some evidence that higher motivation (i.e., higher perceived usefulness of the survey or greater interest in survey participation) may be related to lower perceived burden and/or better response behavior.

Several studies have given an account of factors that affect participation or data quality in business surveys (e.g., Davis and Pihama 2009; Giesen and Burger 2013; Janik and Kohaut 2009; Porter 2004; Seiler 2010). Theoretically, these accounts are largely based on one or a combination of the frameworks provided by Groves et al. (1992), Tomaskovic-Devey et al. (1994, 1995), and Willimack et al. (2002). Some of these studies explicitly suggest that the identified factors (e.g., survey design, time spent on a previous questionnaire's completion, business situation) affect participation through the motivation to respond; however, both the empirical accounts and the theoretical frameworks lack a detailed explanation about the precise role of motivation and how the factors affect response behavior or motivation for this behavior. These studies investigate neither the mechanisms about how motivation works nor the role of perceived response burden (for an exception addressing the latter, see Giesen and Burger 2013).

Recently, a psychological motivation theory, namely Self-Determination Theory (hereinafter SDT), has been applied to the field of household surveys (see Wenemark et al. 2010;

Fig. 1. A model of motivation according to Self-Determination Theory and its subtheory Organismic Integration Theory (based on Deci and Ryan 1985)

Wenemark et al. 2011). As illustrated in Figure 1, SDT posits motivation as a continuum between amotivation, that is lack of motivation, at one extreme, and intrinsic motivation, that is, completely self-determined, internally rewarding motivation, at the other extreme; extrinsic motivation, that is, originating from outside the individual, is in between (Deci and Ryan 1985; Gagné and Deci 2005). As Kruglanski (1975) puts it, with intrinsic motivation the task is an end in itself, whereas with extrinsic motivation the task is a means to an end. People may thus be completing business surveys because they find this kind of work interesting, or because some externally imposed reasons or incentives make them do it, for example, avoiding reminders or a superior's dissatisfaction, fulfilling duties to society, and so on. Given the importance of extrinsic motivation in the work environment where a business survey task usually takes place, a subtheory of SDT, Organismic Integration Theory, is used to detail the different variants of extrinsic motivation (Deci and Ryan 1985). As indicated in Figure 1, transitions from the least self-determined (i.e., external) to the most self-determined (i.e., integrated) extrinsic motivation are a matter of degree and may also change over time through processes of internalization and integration. Respondents can turn extrinsically motivating aspects of the business survey task into (more) internalized elements by making them more personal. Externally initiated motivation may become *introjected* if respondents accept an imposed regulation (though not as their own), or *identified* if respondents value certain behaviors for their congruence with their personal goals and identities, or even *integrated* if respondents completely identify specific behaviors with themselves.

Applying this theory allows survey participation theories to be broadened to also include task commitment and performance. Wenemark et al. (2011) use SDT as an inspiration to redesign data collection procedures and the questionnaire of a self-administered voluntary health-related survey of individuals. This redesign focused on promoting *competence*, *autonomy* and *relatedness*, which are regarded as innate psychological needs that facilitate intrinsic motivation according to Cognitive Evaluation Theory, another subtheory of SDT (Deci and Ryan 1980, 1985). Based on an experiment, they conclude that survey researchers should aim to enhance intrinsic motivation to achieve respondents' superior commitment to the task, as research suggests that the quality of experience and performance is higher when intrinsic motivation is stimulated (Ryan and Deci 2000), and that incentives undermine intrinsic motivation (see e.g., Deci et al. 1999). At the same time, they acknowledge that the topic studied may have been inherently interesting to respondents and that the possibilities of intrinsically motivating respondents may vary across different surveys and different populations.

Business surveys and businesses have many specific features (see e.g., Rivière 2002), which casts doubts on the applicability of Wenemark et al.'s (2011) conclusions for business surveys. The business participation decision and the survey response task occur in a business setting, where the response occupies participants' work time; respondents provide answers on behalf of their organization, and the task's rejection, inaccurate and late completion may have consequences for the participants and their organization (e.g., superiors' reprimands, or survey reminders, follow-up calls or even fines). To better understand survey response motivation in such a setting, we use a combination of primary and secondary data sources from qualitative research interviews conducted in businesses in two countries, the Netherlands and Slovenia. We use thematic analysis to identify sources of motivation according to theoretically defined types of motivation in the SDT

and its subtheories. We define *sources of motivation* as "those conditions that elicit, sustain, and enhance this special type of motivation" (Ryan and Deci 2000, 57). The data and methods are presented in the next section, followed by the presentation of results. In light of these exploratory data and specifics of the setting, we review and discuss the applicability of the SDT to business surveys. We propose that in the business setting: (a) extrinsic motivation calls for at least as much attention as intrinsic motivation, that (b) other sources of motivation may be relevant besides those stemming from the three fundamental needs in the Cognitive Evaluation Theory (competence, autonomy and relatedness), and that (c) other approaches may have the potential to better explain some aspects of motivation in business surveys than the SDT framework alone, for instance McClelland's (1985) dual system approach to motivation that treats implicit motives (for simplicity, these approaches are presented together with relevant quotes in Subsections 3.3 and 3.4). The article concludes with suggestions for improvement of motivation in business surveys and ideas for further research.

2. Data and Methods

The presented study is based on data collated from primary and secondary data sources, using the different sources of evidence to support validity of findings. The *primary data* were collected as part of the international research project BLUE-ETS (BLUE Enterprise and Trade Statistics; see www.blue-ets.eu) that sought, among other topics (e.g., use of internal and external data), to understand what motivates businesses to participate in and report accurately and on time to national statistical institutes' (NSI) surveys. Our study analyzed data collected from businesses in the Netherlands and Slovenia. The *secondary data* were collected as part of doctoral research that studied the actual response process to a specific business survey from start to finish (i.e., from the moment the survey instrument entered the business to the moment it left the business) in real business environments (see Bavdaž 2010). The survey studied, the Quarterly Survey on Trade, was a mandatory self-administered survey conducted in Slovenia by the Statistical Office of the Republic of Slovenia on a sample of units performing trade activities.

The two studies differed in many ways that might have an impact on the reported sources of motivation. One study addressed surveys in general so it collected general perceptions, while the other focused on a single survey when the experience of responding to that survey was still fresh and memories vivid (special attention was paid to minimizing the time that elapsed between the completion and the interview) so that it collected immediate perceptions about the situation as they arose. One study included units from different economic sectors that might have completely different attitudes towards data and surveys; the other included only units from the trade services that might be more homogeneous in this respect. One study addressed the motivation together with data use, thus extending the context to potential benefits of data reporting, while the other addressed motivation together with the response process, thus mainly focusing on the cost aspect of data reporting. One was conducted during the most recent economic downturn that might reduce motivation for survey response; the other was carried out in much better economic conditions. The secondary data source thus complements the primary data source well. More details about both data sources are given below.

2.1. Sample Selection

To ensure that the businesses had had some experience with business surveys, the sampling frame for primary data collection in the Netherlands was the sample of a large survey conducted by Statistics Netherlands. In Slovenia, the sampling frame for primary data collection consisted of all corporations as listed in the Slovenian commercial database GVIN. The sampling frame for secondary data collection was the sample of the studied trade survey.

In the case of both primary and secondary data sources, the selection of businesses aimed to gather as many different insights as possible in accordance with purposeful maximum variation sampling (Cutcliffe 2000). As suggested by Sandelowski et al. (1992) and Coyne (1997), this purposeful sampling was partially superseded by theoretical sampling: We targeted businesses of different size and economic activity because these two variables are generally hypothesized to influence survey response behavior the most and in multiple ways.

Businesses for the primary data collection were thus chosen from different size classes (small – fewer than 50 employees; medium – 50 to 250 employees; and large – 250 + employees) and diverse manufacturing, commercial and service activities. Three criteria guided our selection of the two-digit NACE activities from which we chose businesses: Activities had to have many businesses, because a high number of similar businesses increases the relevance of our findings; they had to be important for the national economy, because activities that have a significant contribution to national economic indicators typically receive considerable attention from survey organizations (they are surveyed more often and/or in more detail, which adds to a high response burden and increases the importance of motivation); or they had to have a large share of small businesses that deserve special attention, because they have a relatively high response burden given their modest resources (see Seens 2010) and are assumed to have problematic survey response behavior (such as nonresponse, item nonresponse or measurement errors). We selected businesses from activities that preferably satisfied more than one criterion. The national samples preferabely avoided more than one business sampled from any two-digit NACE codes. We also sought to ensure as much variability as possible with respect to other criteria that were not explicitly defined as inclusion criteria. We can say that we covered both services and industry, internationally oriented and locally oriented business, foreign and domestically owned business, and different locations. The secondary data source, on the other hand, was already limited to a single economic activity. Its sample was selected systematically across all business sizes, but businesses that were the largest in size in a particular trade activity and/or in trade as a whole were oversampled.

2.2. Sample Recruitment

In the case of the primary data source, initial contacts were established by phone. The recruiting strategy was to start with one interview per business agreed in advance (with either a business survey respondent or a data user; the latter sometimes being in the managing position), and then ask for another interview on the spot using the "foot in the door" technique. In some businesses, we first targeted business survey respondents, while in others we targeted data users (e.g., accounting, economic, analytical and (quality)

control departments). As can be seen from Table 1, the strategy was especially successful in Slovenia, where most on-site visits resulted in more than one interview. In the Netherlands, gift vouchers for use in many Dutch shops were given before or after the interview as a token of appreciation.

The recruiting approach was different in the case of the secondary data source. An advance letter was first sent to respondents of the Quarterly Survey on Trade. Then a telephone contact was established to obtain information about the timing of the questionnaire's completion. This information was later communicated to them in written form (mail and/or email). Subsequent telephone calls served as final confirmation of the date of the on-site visit, which had to coincide with or follow the completion of the questionnaire. As can be seen from Table 1, a group interview was conducted in three cases because respondents were working together very closely (e.g., a novice and the preceding respondent). After the interviews with respondents, an attempt was made to contact other mentioned key people involved in the survey response process besides the respondents (mainly providers of data to respondents, but also some authorities), but these contacts were sometimes short, structured telephone interviews. Altogether the study included 28 different-sized businesses covering various combinations of trade activities and various kinds of merchandise.

Table 1. Overview of interviewed people and businesses in achieved samples of primary and secondary data source

Country	Data source	Total number of interviewees by role	Total number of businesses included in the field study by size class
Netherlands	Primary (BLUE-ETS project)	13 interviewees: *7 data users* *5 business survey respondents* *1 interviewee in both roles*	11 businesses in different economic activities: *3 small* *4 medium* *4 large*
Slovenia	Primary (BLUE-ETS project)	16 interviewees: *8 data users* *7 business survey respondents* *1 interviewee in both roles*	9 businesses in different economic activities: *3 small* *3 medium* *3 large*
Slovenia	Secondary (research on the survey response process)	44 interviewees: *25 respondents* *6 respondents working in pairs* *13 other key people involved in the response process*	28 businesses mainly or partly involved in trade activities: *13 small* *5 medium* *10 large*

2.3. Data Collection

The primary data come from interviews conducted in the Netherlands and Slovenia between September 2010 and February 2011. Questions about the motivational aspects of business survey response behavior represented an important part of the interview guide, which also included questions on the use of data in businesses and the links within businesses between business survey respondents and those who use internal or external data as part of their job (labeled as data users). The interview guide was used in two waves of interviews, with a slight adaptation for the second wave. The semi-structured interviews had a fixed list of motivational topics and objectives (e.g., organizational decisions and norms on survey participation and responding; organizational aspects of the survey response process; beliefs about survey participation, organizational and interviewees' perceptions of NSI surveys; interviewees' perceptions of organizational norms, the survey task, the meaning of participation, etc.) but only a suggested list of questions within each topic (see appendix in Bavdaž 2011). The semi-structured interview guide acted as a frame of reference and as a reminder to ask about certain issues, while unstructured interviewing within this frame allowed interviewers to uncover previously unsuspected elements. All interviews were conducted on-site, except one conducted on the phone.

The secondary data come from on-site visits to businesses in Slovenia arranged around two consecutive deadlines for the completion of the Quarterly Survey on Trade in 2005. The qualitative research interview was the primary method of investigation in businesses. It largely relied on retrospective probing (Willis 2005) and ethnographic interviewing (Gerber 1999) of the principal respondent to the survey on-site. Other people with a minor role in the response process (e.g., a respondent that only answered a single survey question or a data provider that prepared some data input for the respondent) were sometimes reached over the phone. Although the focus of the interviews was on the survey response process, attention was also paid to contextual topics such as the role of authorities and other organizational issues as well as attitudes towards the NSI and (official) statistics. This often produced insights into the motivational aspects, which made the data source useful for the present analysis.

2.4. Data Analysis

Interviews from both primary and secondary data sources were recorded and transcribed so that a verbatim account of interviewees' verbal utterances would be available (an exception was made for some shorter interactions over the phone that were noted down immediately). From the primary data source, all interviews with respondents, and those interviews with data users that contributed any insight relevant to surveys (e.g., interviews with superiors deciding on survey participation) were included in the analysis. From the secondary data source, segments of transcripts and notes bearing information on motivation were identified and included in the analysis.

The purpose of the analysis presented in this article was to assess the fit of the data to psychological theories. It has to be noted, however, that the analysis meant a re-examination of the previously coded data from the primary data source, that is, the second round of analysis of these data. The first round of analysis mainly relied on an inductive, "bottom-up" approach with no specific framework in mind, even though some

preconceptions and background knowledge of potentially relevant or related theories may have contributed to topics and questions in the interview guide (see Coffey and Atkinson 1996; Dey 1993). This mainly data-driven process of coding resulted in the identification of several motivational themes that were then classified as either organizational motivation (corporate social responsibility, attention, prioritizing and statistical hub) or individual motivation (emotional aspects, habits and routines, worth attached to survey task and obligations) (see Torres van Grinsven et al. 2011).

The immersion in the data helped to achieve a deeper understanding of motivation in business surveys. As suggested by Stern (1980) and Strauss and Corbin (1994), we then systematically reviewed the literature, selected relevant psychological theories and brought in theoretically suggested themes. We also added the secondary data source. The second round of analysis that followed and is presented in this article relied on a deductive approach, in which the themes followed the SDT framework, namely the SDT and its subtheories Cognitive Evaluation Theory and Organismic Integration Theory (Deci and Ryan 1980; 1985). The sources of motivation that remained unassigned to the themes of the SDT framework were considered with respect to other relevant psychological theories.

Thematic analysis was applied in both rounds of data analysis. Thematic analysis can be defined as "a method for identifying, analyzing and reporting patterns (themes) within data" (Braun and Clarke 2006, 79). A theme is manifested through "expressions", that is, particular instances in data (Ryan and Bernard 2003) that are attributed to codes within that theme. We searched for expressions of motivation for business survey response behavior of the interviewees at the semantic or explicit level (as opposed to the latent level) within the realist/essentialist paradigm, which means that we reported the meaning, experiences and reality of interviewees without constructing or deriving other meanings from their words (see Braun and Clarke 2006). Codes sometimes applied to a longer passage of the interview, while at other times several themes applied to an interviewee's turn of speech. Codes were developed by the three authors using a standard iterative process (see MacQueen et al. 1998). Each coded passage was assessed individually in several rounds of discussions for agreement between authors on the codes and attribution of codes to themes.

3. Results

In this section we give an account of sources of motivation for business survey participation and accurate and timely response as expressed in our data. The sources of motivation were structured around the two main types of motivation they trigger or influence according to the SDT, that is, intrinsic and extrinsic motivation (see Table 2); amotivation is not included as it lacks a drive, an intention to act, while we were interested in the positive counterpart. The particular sources of extrinsic motivation found in our data were further attributed to the subthemes derived from the Organismic Integration Theory. The essential source of intrinsic motivation is "the fun or challenge entailed" that moves a person to act (Ryan and Deci 2000, 56). Three other sources of intrinsic motivation, that is, perceived competence, relatedness and autonomy, were derived from the Cognitive Evaluation Theory. Some sources of motivation remained unassigned to the themes derived from the SDT framework after the data analysis; these sources are presented at the end of the

Table 2. Themes, subthemes and codes for sources of motivation for the business survey task

Themes	Subthemes	Codes for sources of motivation
Extrinsic motivation	*Externally regulated motivation*	Legal mandate
	Introjected extrinsic motivation	Work tasks (explicitly assigned)
		Social responsibility: • Value for society in general • Value for specific purposes • Value for specific groups • Principle of reciprocity
	Identified extrinsic motivation	Value for own business or self
	Integrated extrinsic motivation	Work tasks (adopted)
Intrinsic motivation		*Enjoyment and challenge* *Perceived competence* *Autonomy* *Relatedness*
		Mood
		Human curiosity disposition
		Disposition for accuracy and precision
		Routines
		Task characteristics

Note: Terms in italics are taken from the SDT and its subtheories Cognitive Evaluation Theory and Organismic Integration Theory

results section together with a possible theoretical explanation and are debated further in the discussion section.

All quotes are accompanied by information about the interviewees. It is indicated in parentheses if the interviewee was a respondent to business surveys [Respondent], a superior to business survey respondent(s) [Superior], or if the role was more specific, for example self-employed, a user of official statistics, and so on.

3.1. Extrinsic Motivation

Sources of motivation in our data could be assigned to all four subtypes of extrinsic motivation as defined by the SDT and its subtheory, Organismic Integration Theory (see Table 2). Several verbal accounts expressing extrinsic motivation were identified in each interview.

3.1.1. Externally Regulated Motivation

Legal mandate

In the case of "pure" external regulation, the task is executed with the sole purpose of satisfying an external demand. In business surveys this demand often comes from legislation and represents a legal obligation for the business. External regulation seemed to be the most common source of motivation in governmental business surveys. While some interviewees stressed the importance of participation, others also expressed concern with accuracy and timeliness.

"The only reason to participate is the legal mandate." [Self-employed outsourcing reporting]

"We have to report, we are legally obliged to do it." [Respondent]

"It is something that has to be delivered in time. And it also concerns correctness. It has to be correct." [Respondent]

The obligation itself could be explicitly known or just assumed.

"I haven't checked, but I assume it's obligatory to report. If you are chosen and you agree on something, then you have to do it no matter what." [Respondent]

Response-enhancing practices based on legal mandates seemed to be highly effective in the minds of the interviewees. In the occasional event that a business was late with the response to the survey request, reminder phone calls and letters, and threats of fines would lead the business to respond. Reacting to letters threatening fines represented a form of externally regulated behavior while reminders represented a softer form of extrinsic regulation (i.e., introjected), mainly relying on feelings of guilt for not respecting the deadline.

"Preferably we want to prevent that we receive letters [with fines]." [Respondent]

"That one was also postponed for a while, and then there came serious letters with the possible fines. And that became rather nasty. [. . .] *So I caught up on that."* [Respondent]

Some other interviewees explained that the point at which they would finally respond was when the threats were communicated in a letter.

3.1.2. Introjected Extrinsic Motivation

Work tasks (explicitly assigned)

Obligations stemming from the organization and imposed on the respondent were an important source of motivation not only to participate in a survey, but also to respond in a timely and accurate manner. Introjected extrinsic motivation refers to behaviors performed under external pressure to avoid guilt and anxiety or to build self-esteem (Ryan and Deci 2000).

"The top management requests us to participate in as many surveys as possible in order to be more transparent." [User of official statistics]

"The agreement in this company is that we neatly comply with the request and send it [the questionnaire] *back in time."* [Respondent]

This obligation to comply was implicitly communicated by certain actions or explicitly part of one's work tasks and remuneration basis.

"When a survey comes in, he [the superior] *lays it down at my desk and just presupposes I will get it answered."* [Respondent]

"It's a part of my job tasks." [Respondent]

"It's in my work description." [Respondent]

These data showed that avoiding upsetting a superior was a reason to comply with the survey request, which would be an introjected type of motivation. From the point of view of the superior, though, this could be categorized as externally regulated extrinsic motivation, because from that perspective the avoidance of external punishments was salient.

"He instructs me to comply and to send those things back in time so that we don't get any reminders or anything. Because if we get a reminder by post he will come to my desk asking if I forgot or what's happening." [Respondent]

It has to be noted, however, that in some cases people exhibited a higher degree of internalization or self-determination of their work tasks. In such cases, motivation for these tasks could be part of the identified or even integrated extrinsic motivation.

Social responsibility: value for society in general, value for specific purposes, value for specific groups; principle of reciprocity

Verbal accounts of value for society as a source of introjected external motivation were also found. Businesses seemed to acknowledge the importance of their data for society and other businesses.

"The government needs data to function properly." [Self-employed]

"I think everybody has to just contribute their part, because the whole picture has to be right, because it will be used by politics, the national economic planning institution or any other institution." [Respondent]

"If I'm not selfish, then I have to say that as I need some specific data, others might need some other data that I might find useless, thus we should report them." [Superior]

"Data we are producing need to be accurate, that's the most important thing. We are informing the public, so we must provide accurate data." [Superior]

3.1.3. Identified Extrinsic Motivation

Value for own business or self

Identified extrinsic motivation refers to behaviors with which a person has identified so that he or she consciously values them (Ryan and Deci 2000). Our data showed that getting something back for the effort and time spent on responding to a business survey was an important source of this motivation. Value could be expressed with tangible benefits or rewards, or merely perceived as such.

[Referring to the gift voucher given for the interview] *"This is a good start. We, Dutch people, always want to have something. Get something."* [Self-employed outsourcing reporting]

"I think it is useful to send a thank-you note. Just to let them know you had the response and you appreciate it." [Self-employed outsourcing reporting]

"It gets a little on your nerves when you have to prepare it, and I say, 'oh, why', then you moan a little [about it] *but if you know, that if you want to find, get something, you have to do something for it, then you do it."* [Respondent]

[Referring to the value of (official) statistics] *"Having no statistical data is like driving a car by night without lights on – you have no idea where are you going."* [Superior]

"Look, everybody wants to receive data in return. And every company is very selfish in this." [Superior]

"One good deserves another." [Superior]

In fact, a commonly mentioned reason to participate in a voluntary survey was receiving results in return because they were relevant for the company's operations management.

"We pay to participate in surveys from which we get data back." [Superior]

"We participate in surveys if it's interesting for us to get data back." [Respondent]

When there were no perceived benefits, responding to the questionnaire was experienced merely as a cost.

"Replying to business surveys seems an extra job that doesn't give any benefit." [Respondent]

3.1.4. Integrated Extrinsic Motivation

WORK TASKS (ADOPTED)

Integrated extrinsic motivation refers to behaviors that are externally motivated but completely internalized (Ryan and Deci 2000). In the business context this can be interpreted as executing the tasks not because of an external requirement and control but because it is congruent with one's work-related values. So although some interviewees said that they took part in official surveys because they had to, this obligation was in some cases integrated to the extent that it was neither checked nor questioned but simply accepted as part of the job.

"Actually the CBS [the Dutch NSI] *surveys are all just answered."* [Superior]

"It's just something you have to do." [Respondent]

"This is not debated. It's just something that has to happen." [Respondent]

It is important to note that this integration affected not only participation, but also accuracy and timeliness.

"We just presuppose we will fill it in in good faith and accurately." [Respondent]

Business motivation to respond seemed also to be guided to some extent by the concern for the public image. An interviewee thus reported that their company carefully followed the media news on their company, and that the company would treat any survey request carefully so as to maintain its positive image and avoid any negative publicity.

"Sometimes qualitative information could ruin our image, reputation, although our quantitative data is showing a positive direction. We have to be aware of that." [User of official statistics]

3.2. Intrinsic Motivation

The sources of intrinsic motivation that are suggested by the SDT and its subtheory, Cognitive Evaluation Theory (see Table 2), were also all expressed in our data, as can be seen from the quotes below. Verbal accounts of intrinsic motivation were, however, fewer than those of extrinsic motivation. Still, as our data are qualitative, this does not necessarily mean that intrinsic motivation is less present or less important for the business survey response task than extrinsic motivation.

Enjoyment and challenge

Thematic analysis of the interview transcripts identified that some respondents enjoyed surveys and found them challenging, which is the essence of intrinsically motivating activities. They liked the survey task simply because they took pleasure in it, which showed their inherent motivation.

"I always found that the survey on finances and enterprises was a very enjoyable form. Yes, I like that. [. . .] *That's the kind of work I like to do."* [Respondent]

It also seems important that respondents enjoyed preparing data for surveys.

[Showing data in Excel files for reporting purposes] *"One has to be quite creative. If you enjoy it, then it's not a problem . . . "* [Respondent]

Perceived competence

Intrinsic motivation is triggered only if the person feels capable with respect to the task (Ryan and Deci 2000). Many respondents claimed that survey requests were intelligible and the questionnaires were clear and easy to them, which suggests that they perceived themselves as competent to perform the task.

"Questions seem to be clear enough, at least the majority of them." [Respondent]

"Well, I think that the surveys that arrive are clear and understandable." [Respondent]

Some respondents felt their capacity for successfully completing the task was low. In the first quote below, the survey task was outsourced and the respondent never completed the task alone. In the second quote, the respondent provided estimates because of inadequate information support, which made her feel frustrated.

"If I had to complete it [the questionnaire] *myself, well then I think you would not understand anything. Then it would be riddled with inaccuracies because I just don't know, you know. There is, there will be specific questions that are technical on accounting. Yes, for me it's counting up and deducting and the rest is up to the accountant."* [Self-employed outsourcing reporting]

"This [question about sales broken down by types of buyers] *is a problematic one, yes. It's done according to a feel, and percentages. Now, in the beginning, 15 years ago, we already had something similar.* [. . .] *It could be done at that time. Now we have 15 thousand buyers so it is very difficult to get data.* [The respondent explained that they contacted the NSI and got their permission to provide estimates but for her such a solution still represented a frustration:] *We are used to accurate numbers."* [Respondent]

Relatedness

Intrinsic motivation can also arise from connectedness to others in the business and the survey organization. In the data, there were several expressions of appreciation of a good personal relationship with the NSI. Respondents described how their personal relationship with the designated NSI staff had advantages and made them feel obliged to maintain a good relationship. If respondents received help from the NSI staff, then that could make them want to do something in return.

"The advantage is that you've seen each other a couple of times. When I'm talking to somebody on the phone now, then I think, I know his face." [Respondent]

"I think I have a good relationship. Yes, with X." [Respondent]

[About the interview] *"My colleagues asked why I should do this interview. Then I replied: I find this is important now, because I'm the one having the contact* [with the NSI], *therefore I want to do this. Because I want to maintain the contact in good shape, so I want to do this now."* [Respondent]

A friendly tone and language as an expression of a correct relationship seemed to be expected in communication that was addressed to businesses; they might have even been indispensable for survey requests to be considered.

[Discussing a polite tone and language] *"I think that's the way to cooperate. If you attack from one of both sides, then somebody might get blocked and that's worse."* [Respondent]

Some of the interviewed respondents also stated that they appreciated receiving a reaction when reported data seemed to be wrong or just an acknowledgment of receipt of the data. The awareness that the reported data were used made them feel that the time and effort they put into the questionnaire mattered, which enhanced a good relationship, contributed to positive feelings associated with the task, and influenced the perceived value of the performed task.

"But they do look at that, and yes, I like that, because if you get an answer then at least you know they do look at it. So that's pleasant." [Respondent]

Autonomy

Some respondents found it important that consultations and negotiations with the survey organization took place so that their working processes were considered and some autonomy about the deadline was granted.

"We don't have all the data available at the deadline and as we are a large company that represents a great share of aggregated data, we made an agreement with the NSI that we report with a few days of delay in order to assure accurate and reliable data." [Respondent]

3.3. Beyond the SDT Framework

Sources of motivation presented in Subsections 3.1 and 3.2 could be categorized under one of the (sub)themes derived from the SDT framework. However, some remaining sources of motivation that were identified in the first round of inductive coding did not seem to fit easily in the SDT framework. They are presented in this section along with possible theoretical explanations.

Mood

Some verbal accounts suggested that a person's mood affected motivation at least temporarily. Here we show an account pointing to the relevance of mood for the decision on survey participation.

"When I'm in a good mood then I usually participate in all those surveys, but if I'm in a bad mood then I probably reject." [Respondent]

"If they irritate me I just throw them away." [Self-employed]

Mood is shown to be an important factor for motivation in Seo et al. (2004), who provide a framework in which emotion is theorized to be the central construct affecting both the processes and outcomes of work motivation. Mood also affects information processing and task performance; people in a good mood are less easily distracted from the task than people in a bad mood (Bless et al. 1990).

Disposition

Our data showed the presence of two kinds of disposition that might be relevant for the business survey response task. One concerns a disposition for precision and accuracy that seems to be typically inherent in the accounting profession. This disposition stems from the accounting work methods that require precise balancing of accounts and from the accounting principles that require an accurate picture of business reality. Evidently, an accuracy-motivated disposition can also be present in other respondents regardless of profession. This disposition together with the explicit goal to respond to a survey promises to lead to an accurate response.

"If I do something, I do it well, that's in a bookkeeper." [Respondent]

"We do our best, we don't just put any random data thinking it's good enough for statistics." [Respondent]

The other type of disposition concerns human curiosity, which is visible in the attraction to performing new, demanding tasks (e.g., searching for new data and solutions, optimizing processes etc.) or learning new things. This disposition is likely to lead to enhanced intrinsic motivation when applied to the survey task.

"I'm a searcher in my soul. It's a challenge for me to search for new ways of obtaining and using data." [Superior]

"Next year we are facing an exciting event as two different branches of our company have to be merged. That's a challenge again, so I like to do that, yes." [Respondent]

These dispositions seem to be congruent with dispositions based on implicit motives. Implicit motives provide a general orientation toward certain types of goals such as a general trend to do things well (McClelland et al. 1989). They thus generally sustain behavioral trends over time, such as an accuracy-motivated disposition.

Routine

Our data also suggested that reporting ran smoothly on a routine basis after the first completion or introduction of changes. Once integrated into the usual work routine, the business survey response task became easier and less time consuming.

"So at least I personally have a structure that I fill in the questionnaires in a certain way and that I maintain this structure over the years." [Respondent]

"Well, yes, then you have a certain way of getting out of things. And if something changes, yes, then I have extra work but, ehem, you then finally figure out how to fill it in and then it runs again smoothly." [Respondent]

The consistency entailed by routine might even be considered more important than reporting accurately.

"We keep the method the same over the years to maintain comparable figures. Ehm, a mistake, if you make a mistake in 2008, then you have to do the same in 2009 because then at least the trend is visible." [Respondent]

Behaviors that were initially based on conscious (explicit) decisions may become habitualized and routinized through frequent execution and then be carried out independent of the implications of the original conscious judgment (Aarts and Dijksterhuis 2000; Ouellette and Wood 1998). According to recent research, due to characteristics of modern work life, a large proportion of daily cognitive and other processing is unconscious, occurring outside employees' awareness and control (Uhlmann et al. 2012; Johnson and Steinman 2009), which calls for attention to implicit processes within organizations (e.g., Bing et al. 2007; Haines and Sumner 2006).

Task characteristics

Interviewees mentioned some characteristics of the business survey task that seemed generally to support motivation for the successful completion of the task. One such characteristic was simplicity or simplification of the task. The most radical and preferred way, especially for mandatory surveys, was "*to make fewer surveys*". Other ideas mentioned were that "*the NSI should look for a junction with the tax declaration*", that the questionnaire should be adapted to the internal administration of companies, and other ways of simplifying response and "*automating things as much as possible*" in order "*to be as efficient as possible*".

Moreover, our data showed that it was important to maintain the same questionnaire items and item order over time in recurring surveys, and give prior notification of changes. Respondents often had a routinized approach to the questionnaire's completion in these surveys, especially if repeated frequently, which made it more difficult to adapt to changes. It thus seems important to make the survey task predictable.

> *"I mean, don't be too specific and once you have such a survey and you have figured things out* [. . .] *then let it be the same next year, and don't change too much."* [Respondent]

> *"When there are changes then I have to change my models and that costs extra time."* [Respondent]

> *"I find it important to get notifications on changes.* [. . .] *I do things automatically, thus it is important for me to know if there are any changes to the deadline or questionnaire. Then I check what is different."* [Respondent]

The task's simplicity, easiness, and predictability have previously been identified as factors affecting extrinsic motivation (Kruglanski 1975; Pittman et al. 1983; McClelland et al. 1989).

3.4. Dynamics Between Extrinsic and Intrinsic Motivation

Several sources of extrinsic motivation reflect the importance of authorities for the business survey response task: Authorities determine work tasks and expectations; they speak through companies' values, policies and routines. Although extrinsic motivation might become more integrated through the process of internalization, it does not mean that extrinsic motives are transformed into intrinsic ones (Ryan and Deci 2000). However, extrinsic motives could be replaced by intrinsic motives. In business surveys this could happen if a person started the business survey response task only to answer an external demand for obligatory reporting (and avoid fines), but experienced the activity as interesting, challenging and enjoyable, presumably also because the recurrent completion reduces burden. Such cases are consistent with Kehr (2004), who proposes that externally imposed goals fueling extrinsic motivation can become intrinsically motivating, provided they are congruent with the person's implicit motives.

> *"Well, I start enjoying it much more every time. (. . .) In the beginning because you're still looking for your way it's never pleasant."* [Respondent]

On the other hand, the negative side is that the same task might get boring over time, thus reducing the level of intrinsic motivation over time.

> *"If you do the same thing every year, then it gets boring."* [Respondent]

As mentioned above, sources of extrinsic motivation were expressed in all our interviews. At the same time some respondents showed that they experienced the business survey response task as intrinsically motivating, which suggests that both types of motivation coexisted and drove the respondents towards desirable outcomes. However, most respondents reported not liking or particularly enjoying the business survey response task,

which seems to indicate a lack of intrinsic motivation. In spite of this, the task was still carried out, indicating that another (extrinsic) motivation should be at play. Some of these respondents even reported executing the task on time and as accurately as possible. The first two of the following quotes explicitly express an extrinsic motive, namely the legal mandate, and at the same time an absence of intrinsic motivation. The same combination of an extrinsic motive and the absence of intrinsic motivation was also seen in the other two quotes.

> *"You take it as a necessary evil, you complete it. You complete everything that you have to complete. I don't think* [about it]." [Respondent]

> *"Well, it's not the greatest challenge to fill in those questionnaires. Yes, the obligation and, ehm, yes, of course, as accurate as possible. And on time."* [Respondent]

> *"Well we have nicer and less nice tasks, that everybody has in his job. And this is one of the standards. The tasks that are not always that enjoyable."* [Respondent]

> *"I think we fill in in good faith, but it's seen as a necessary evil."* [Superior]

4. Discussion and Suggestions for Improving Business Survey Response Behavior

In this article we have shown specific sources of motivation for the business survey response task based on interview data from the Netherlands and Slovenia from two sources. Our empirical data have provided support for the *existence of all different (sub)types of motivation suggested by SDT and its subtheories*, Cognitive Evaluation Theory and Organismic Integration Theory. Although the quantity of verbal accounts of extrinsic motivation compared to those of intrinsic motivation is by no means indicative of their prevalence, it is impossible to ignore their presence and relevance in the business setting. On the other hand, it seems that influencing intrinsic motivation also has some potential even if intrinsic motivation for the business survey response task seems relatively weak. Research findings suggest that intrinsic motivation can positively influence commitment (Ryan and Deci 2000), albeit in different kinds of settings. We therefore suggest not overlooking any of these types of motivation in the business setting.

Moreover, the results suggest that *SDT cannot explain all sources* of motivation expressed in our data. A large group of such sources might be considered implicit motives that are part of implicit, automatic processes, which seem to be pervasive in organizational life (Johnson and Steinman 2009). Some dispositions (e.g., for precision and accuracy) seem to have been built over the years through multiple repetitions and some routinized behaviors have lost connection with their original intent. It thus seems reasonable to expect that implicit motives and implicit processes also play a role in business survey response behavior. Kehr (2004) proposes that the presence (or absence) of an individual's implicit motives seems to determine if a task is experienced as intrinsically motivating (or not). Still, arousal of implicit motives does not necessarily lead to intrinsic motivation. It does not lead to intrinsic motivation if incompatible cognitive preferences exist at the same time (Kehr 2004). The influence of implicit motives may also disappear in the presence of powerful explicit motives, such as social constraints (Kehr 2004; McClelland et al. 1953). Some authors therefore propose a dual model of explicit and implicit

processes, in which the two types function in parallel and in interaction with one another (Fazion and Olson 2003; Strack and Deutsch 2004).

The concealed nature of implicit motives and processes, however, makes it difficult to recognize them and turn them to the benefit of the business survey response task. It might be easier to accomplish a mood change or simplify and predict the survey task, both of which also seem to influence the response behavior. The psychological literature also points to some other sources of motivation that have not been expressed in our data but might be relevant for the business survey response task. For instance, positive emotions are suggested to be essential elements of optimal functioning (Fredrickson 2001); accountability is shown to act as a source of motivation for more analytical cognitive processing (Tetlock 1985) and so forth.

Our study followed qualitative research in organizational psychology that studies how people "think, feel and behave in work and organizational contexts", including their motivation (Silvester 2008, 489). Some other sources of motivation may still be identified using other methodological approaches than qualitative research interviews, which mainly relied on reported sources of motivation. Especially for specific unconscious, concealed or otherwise latent sources of motivation, it might be necessary to use an experimental and/or laboratory setting to provide evidence of their existence and relevance in the business survey setting. Moreover, focusing on business surveys with specific design features (e.g., voluntary or interviewer-administered business surveys) or conducting research in other institutional environments might bring new insights, and using larger samples can further support the external validity of findings. According to respondents in several interviews, it was normal that the survey questionnaires "came" to them; any deviation from such a routine calls for more attention to be given to other people involved in the business survey response task (e.g., gatekeepers and authorities). Although our study involved some data users that were not simultaneously respondents to business surveys, involving people with other tasks could open up new perspectives. We only touched upon motivation of survey nonrespondents marginally, as some of our interviewees claimed not to respond to all survey requests. Still, people refusing to participate in our study might have some (other types of) motivation for the business survey response task which remain to be identified. Nevertheless, the presented study provided some indication of the drives underlying business survey response behavior.

Suggestions for Enhancing Motivation

Unfortunately, awareness of the sources of motivation is just a starting point for thinking of how to effectively and efficiently increase the motivation for the business survey response task. Still, the knowledge of sources of motivation identified thus far can be valuable in designing and testing actions and strategies for enhancing response and response quality in business surveys. We provide several suggestions below. Some of them are not new, but we iterate them here for the sake of completeness. The suggestions are focused both on intrinsic and on extrinsic motivation. They are presented as specific actions and strategies that can be applied to enhance motivation, which is in turn expected to positively affect the outcomes related to the survey response task. Although the identified sources of motivation are theoretically founded, further (preferably

experimental) research is necessary to determine how specific interventions targeting these sources affect motivation and the resulting response behavior. Nevertheless, it seems reasonable to expect that interventions triggering both intrinsic and extrinsic motivation produce a larger effect than just triggering a single motivation type, and that interventions triggering intrinsic motivation produce a larger effect because they directly involve the person unless there are negative consequences for the business. Respondents that already have some motivation might be more affected than those without or with a very low level of motivation. The effects on outcomes thus seem likely to be dependent on the initial level of motivation and respondents' perception of additional effort to improve behavior.

Before presenting suggestions for enhancing extrinsic motivation, it has to be noted that some previous research in behavioral psychology and the field of social surveys advises against using incentives as extrinsic rewards because they seem to undermine intrinsic motivation, which is considered better for performance than extrinsic motivation (see e.g., Deci et al. 1999; Wenemark et al. 2011). However, some studies suggest that such a conclusion is far from straightforward. A positive effect on intrinsic motivation is expected for praise (being delivered immediately, often and without clear stimuli) while a negative effect is limited to tangible rewards (Carton 1996). Rewards seem to enhance intrinsic motivation for low-interest tasks and also for high-interest tasks if they are linked to level of performance (Cameron et al. 2001). These findings have been taken into account when designing suggestions that focus mainly on enhancing respondents' extrinsic motivation, also taking note of limitations and cautions mentioned above:

- Current response-enhancing practices, that is, reminders and (threats of) fines in the case of nonresponse, seem to achieve their aim of assuring response though they typically represent negative, not positive *reinforcers*.
- *Value* of the survey, the survey organization and the survey outcomes should be improved and communicated. The value may be expressed with tangible benefits or rewards or merely perceived as such. Several 'stakeholders' should perceive this value, namely society, the economy, the business and the individual respondent. Influencing the *value in real terms* could be done by giving the businesses an appropriate incentive for the time and effort they have spent to fill in the questionnaire, though this might be costly or even have negative consequences if not tied to good performance. The *perceived value* can be increased by a communication strategy, for instance, by showing businesses in more concrete terms (e.g., with case studies and testimonials) what the data are used for and what the specific purpose of the requested data is.

Suggestions that focus mainly on enhancing respondents' intrinsic motivation are:

- Survey participation should be made as *enjoyable* as possible. Given that the task of answering survey questions is not attractive to most respondents, it might be necessary to think of other aspects of survey participation and make them enjoyable. The possibilities are greater or at least more convenient for electronic reporting and include, for instance, accessing an online questionnaire, printing the questionnaire, receiving a confirmation of receipt by email, and so on. These activities might become more enjoyable if accompanied by interesting figures, famous or wise thoughts, quiz-like questions, other

challenges and so forth. A respectful and friendly tone should be present in all communication.

- Respondents' *perceived competence* (or perceived abilities) should be enhanced, as it seems that perceived competence influences response behavior and, vice versa, having positive experiences with questionnaires influences perceived competence towards future questionnaires. This can be improved by using an appropriate communication strategy that stresses the easiness and the simplicity of the response task. This should of course be accompanied by a questionnaire that *is* both as easy and simplified as possible as well as user-friendly to make a good first impression.
- A good *relationship* with the business and the respondent should be built up to enhance relatedness. A good example of this is using dedicated staff (account managers) for businesses that are important for the aggregate statistics. However, such an approach is typically granted to only a handful of large businesses, so it is necessary to think of efficient ways of establishing and maintaining relationships with all businesses. Possible strategies are to target only new respondents, respondents involved in more surveys, respondents completing questionnaires for several businesses (e.g., in accounting firms), and so on. The relationship could be established through a live contact, with some tokens of appreciation, and so forth. It also seems important that nameless and faceless survey staff reveal their identity. Providing contact names is just the minimum; showing their picture and adding a few words about themselves could greatly deepen the feeling of relatedness. Given that a good relationship is typically based on reciprocity, giving different forms of feedback (from thank-you notes to statistical results) should also be useful in this regard, as explained below.
- Respondents should feel that they have some *autonomy* with regard to the business survey task. Two situations where more autonomy typically is or could be granted concern the deadline for reporting and the provision of estimates instead of precise figures, although some conditions could be set to avoid attributing less importance to deadlines and precise figures.

A different approach to enhancing motivation is to attempt to improve respondents' *mood*. As mood is a temporary state, it is important to focus on activities that immediately precede a questionnaire's completion. Given the impact of humor on people's mood, ideas for improving the mood could be sought, for instance, in humorous thoughts or instructive anecdotes of famous statisticians and so on. However, these ideas have to be applied with great delicacy to the business setting, which might exhibit certain expectations about professional behavior (Romero and Cruthirds 2006).

The business survey response task could also result in better outcomes if respondents were selected from those people that have *desirable dispositions*. As the disposition for precision and accuracy seems to be present in the accounting profession, we suggest treating them with special care, for example by targeting them through their professional organizations, events, publications, and so on. Better outcomes can also be expected if responding to a questionnaire is made as *easy, simple* and *predictable* as possible, which calls for the implicit processes in companies to be taken into account. This would include, for example, adapting the survey items as much as possible to the businesses'

administration; avoiding changes as much as possible, and if changes are made, communicating them early on and clearly to the business; using as much as possible a standardized format, for example in concordance with other data-requesting organizations such as the tax office, and so on.

It seems extremely important to reduce actual survey burden, defined as the time it takes to respond to the survey (Dale and Haraldsen 2007), because of its correlation with perceived response burden and data quality (see e.g., Berglund et al. 2013). The reductions of actual burden, however, may have a limit as some data have to be collected from businesses. The only way of improving survey outcomes thus seems to be through raising motivation. Whether raising motivation also affects perceived response burden or not might depend on the type of motivation invoked; intrinsic motivation and more internalized forms of extrinsic motivation presumably have more potential to further alleviate perceived response burden.

Some suggestions promise to affect intrinsic and extrinsic motivation, depending also on the exact form of implementation. This, for instance, holds for the provision of *feedback* that should be given to respondents because it can invoke motivation from several sources. For instance, a message confirming a questionnaire's receipt, thanking the respondent for a timely delivery and acknowledging their contribution to the timely release of official indicators should influence both intrinsic motivation (by enhancing a good relationship and perceived competence, thus making (the next) survey participation more pleasant) and extrinsic motivation (by influencing the perceived value of the task). Some kinds of feedback are already used on a regular or ad hoc basis, such as statistical results and thank-you notes. However, there is still a lot of potential for improving and diversifying even these two kinds of feedback. Immediate feedback should work better than delayed feedback. Statistical results can be customized and more tailored to the needs of a specific business, presented in a way to offer information (not only data) to the business or simply made more attractive, but survey organizations are not always aware of business data needs.

Acknowledging respondents' efforts can be supported with further-reaching marketing activities, such as rewarding the most deserved respondents once a year. Another strategy might be to send positive evaluations about good respondents to their superiors as well as requests for improvement of reporting, but a positive tone would have to be used to convey such messages so they can be clearly distinguished from reprimands and fines. Acknowledging organizational efforts, on the other hand, can be implemented with the cooperation of a reputable company that also excels in reporting and thus nourishes its social responsibility, or by publically naming companies that excel in reporting overall. These approaches, however, require NSIs to be able to determine the overall quality of reporting for every business (consisting of timeliness that is easy to measure and accuracy that is difficult to measure).

Many of these suggestions thus require good information about each business and each respondent. NSIs are typically in possession of such information, but not in a format that would allow further managing. It seems that without such support and a more customer-oriented focus, NSIs might have to relinquish more sophisticated and/or tailored forms of improving motivation.

5. Further Research

Research presented in this article is in line with the call for more research on motivation in business survey response behavior (Rogelberg and Stanton 2007; Rose et al. 2007). Application of motivation theories to business surveys promises to inspire new approaches to motivate business survey respondents to better respond to, and perform, the survey task. As our empirical data are limited, additional and somewhat different approaches might be recommended, especially if they are more focused on voluntary surveys, nonrespondents and reluctant respondents or other people involved in the response process. Nevertheless, our results should help with the development of actions and strategies enhancing motivation in the meantime.

Business surveys often embrace methodological advances from household surveys. A different perspective seems to be necessary to explain motivation for business survey response tasks which are done during work time, often require expert knowledge and rarely rely on monetary incentives. The theoretical framework should be expanded beyond the SDT framework, which seems to be insufficient to cover all specifics of the business setting. It seems necessary to bring in more research conducted in the work environment to understand the functioning of people involved in the response process.

More research is also necessary to conclude whether or not motivation to participate in business surveys and provide an accurate and timely response can be treated as a single and integral concept, what the relationship between motivation and perceived response burden is and whether it is appropriate to focus on respondents' motivation while acknowledging organizational motives as an important source of the individual motivation. An important step should be to implement and experimentally test interventions as a means to evaluate the proposed suggestions. Research is necessary to evaluate the specific impact of each of the different sources of motivation that appear to be of importance in the business survey response task, and also to evaluate in which cases each of the sources or a combination of the sources is most effective.

6. References

Aarts, H. and A. Dijksterhuis. 2000. "Habits as Knowledge Structures: Automaticity in Goal-Directed Behaviour." *Journal of Personality and Social Psychology* 78: 53–63. DOI: http://dx.doi.org/10.1037/0022-3514.78.1.53.

Adolfsson, C., G. Arvidson, P. Gidlund, A. Norberg, and L. Nordberg. 2010. "Development and Implementation of Selective Data Editing at Statistics Sweden." European Conference on Quality in Official Statistics, Helsinki, Finland. Available at: http://q2010.stat.fi/media//presentations/Norberg_et_all__Statistics_Sweden_slutversion.pdf (accessed 5 May 2010).

Anseel, F., F. Lievens, E. Schollaert, and B. Choragwicka. 2010. "Response Rates in Organizational Science, 1995–2008: A Meta-Analytic Review and Guidelines for Survey Researchers." *Journal of Business Psychology* 25: 335–349. DOI: http://dx.doi.org/10.1007/s10869-010-9157-6.

Bargh, J.A., P.M. Gollwitzer, and G. Oettingen. 2010. "Motivation." In *Handbook of Social Psychology*, edited by S.T. Fiske, D.T. Gilbert, and G. Lindzey, 5th ed., 268–316. New York: Wiley.

Baruch, Y. 1999. "Response Rate in Academic Studies — A Comparative Analysis." *Human Relations* 52: 421–438. DOI: http://dx.doi.org/10.1177/001872679905200401.

Baruch, Y. and B.C. Holtom. 2008. "Survey Response Rate Levels and Trends in Organizational Research." *Human Relations* 61: 1139–1160. DOI: http://dx.doi.org/10.1177/0018726708094863.

Bavdaž, M. 2010. "The Multidimensional Integral Business Survey Response Model." *Survey Methodology* 36: 81–93.

Bavdaž, M., ed. 2011. *Final Report Integrating Findings on Business Perspectives Related to NSIs' Statistics*. Deliverable 3.2 of the BLUE-ETS Project. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable3.2.pdf (accessed October 2012).

Berglund, F., G. Haraldsen, and Ø. Kleven. 2013. "Causes and Consequences of Actual and Perceived Response Burden Based on Norwegian Data." In *Comparative Report on Integration of Case Study Results Related to Reduction of Response Burden and Motivation of Businesses for Accurate Reporting*, edited by D. Giesen, M. Bavdaž, and I. Bolko. Deliverable 8.1 of the BLUE-ETS Project. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable8.1.pdf (accessed February 2013).

Bing, M.N., J.M. LeBreton, H.K. Davison, D.Z. Migetz, and L.R. James. 2007. "Integrating Implicit and Explicit Social Cognitions for Enhanced Personality Assessment: A General Framework for Choosing Measurement and Statistical Methods." *Organizational Research Methods* 10: 346–389. DOI: http://dx.doi.org/10.1177/1094428107301148.

Bless, H., G. Bohner, N. Schwarz, and F. Strack. 1990. "Mood and Persuasion: A Cognitive Response Analysis." *Personality and Social Psychology Bulletin* 16: 331–345. DOI: http://dx.doi.org/10.1177/0146167290162013.

Braun, V. and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3: 77–101. DOI: http://dx.doi.org/10.1191/1478088706qp063oa.

Cameron, J., K.M. Banko, and W.D. Pierce. 2001. "Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues." *The Behavior Analyst* 24: 1–44.

Carton, J.S. 1996. "The Differential Effects of Tangible Rewards and Praise on Intrinsic Motivation: A Comparison of Cognitive Evaluation Theory and Operant Theory." *The Behavior Analyst* 19: 237–255.

Coffey, A. and P. Atkinson. 1996. *Making Sense of Qualitative Data: Complementary Research Strategies*. Thousand Oaks, CA: Sage.

Coyne, I.T. 1997. "Sampling in Qualitative Research: Purposeful and Theoretical Sampling; Merging or Clear Boundaries?" *Journal of Advanced Nursing* 26: 623–630. DOI: http://dx.doi.org/10.1046/j.1365-2648.1997.t01-25-00999.x.

Cutcliffe, J.R. 2000. "Methodological Issues in Grounded Theory." *Journal of Advanced Nursing* 31: 1476–1484. DOI: http://dx.doi.org/10.1046/j.1365-2648.2000.01430.x.

Cycyota, C.S. and D.A. Harrison. 2006. "What (Not) to Expect When Surveying Executives: A Meta-Analysis of Top Manager Response Rates and Technique over Time." *Organizational Research Methods* 9: 133–160. DOI: http://dx.doi.org/10.1177/1094428105280770.

Dale, T. and G. Haraldsen eds. 2007. *Handbook for Monitoring and Evaluating Business Survey Response Burdens*, Eurostat. Available at: http://epp.eurostat.ec.

europa.eu/portal/page/portal/research_methodology/documents/HANDBOOK_FOR_MONITORING.pdf (accessed May 2014).

Davis, W.R. and N. Pihama. 2009. "Survey Response as Organizational Behavior: An Analysis of the Annual Enterprise Survey, 2003–2007." New Zealand Association of Economists Conference 2009, 1–16. New Zealand: New Zealand Association of Economists. Available at: http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1832&context=eispapers (accessed May 2014).

Deci, E.L., R. Koestner, and M.R. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effect of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125: 627–668. DOI: http://dx.doi.org/10.1037/0033-2909.125.6.627.

Deci, E.L. and R.M. Ryan. 1980. "The Empirical Exploration of Intrinsic Motivational Processes." In *Advances in Experimental Social Psychology*, Vol. 13, edited by L. Berkowitz, 40–80. New York/London: Academic Press Inc.

Deci, E.L. and R.M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum.

De Leeuw, E. and W. De Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 41–54. New York: Wiley.

Dey, I. 1993. *Qualitative Data Analysis: A User-Friendly Guide for Social Scientists*. London: Routledge Kegan Paul.

Fazio, R.H. and M.A. Olson. 2003. "Implicit Measures in Social Cognition: Their Meaning and Use." *Annual Review of Psychology* 54: 297–327. DOI: http://dx.doi.org/10.1146/annurev.psych.54.101601.145225.

Fredrickson, B.L. 2001. "The Role of Positive Emotions in Positive Psychology. The Broaden-and-Build Theory of Positive Emotions." *American Psychologist* 56: 218–226. DOI: http://dx.doi.org/10.1037/0003-066X.56.3.218.

Gagné, M. and E.L. Deci. 2005. "Self-Determination Theory and Work Motivation." *Journal of Organizational Behavior* 26: 331–362.

Gerber, E.R. 1999. "The View from Anthropology: Ethnography and the Cognitive Interview." In *Cognition and Survey Research*, edited by M.G. Sirken, D.J. Herrmann, S. Schlechter, N. Schwarz, J.M. Tanur, and R. Tourangeau, 214–237. New York: Wiley-Interscience.

Giesen, D. 2012 "Exploring Causes and Effects of Perceived Response Burden." In Proceedings of the Fourth International Conference on Establishment Surveys (ICES IV), Montreal, Canada, 11–14 June 2012. Available at: http://www.amstat.org/meetings/ices/2012/papers/302171.pdf (accessed May 2014).

Giesen, D. and J. Burger. 2013 "Measuring and Understanding Response Quality in the Structural Business Survey Questionnaires." Paper prepared for the European Establishment Statistics Workshop, Nuremberg, Germany, 9–11 September 2013. Available at: http://enbes.wikispaces.com/file/view/Giesen%20Burger%202013.pdf/456104004/Giesen%20Burger%202013.pdf (accessed May 2014).

Groves, R.M., R.B. Cialdini, and M.P. Couper. 1992. "Understanding The Decision to Participate in a Survey." *The Public Opinion Quarterly* 56: 475–495. DOI: http://dx.doi.org/ 10.1086/269338.

Haines, E.L. and K. Sumner. 2006. "Implicit Measurement of Attitudes, Stereotypes, and Self-Concepts in Organizations: Teaching Old Dogmas New Tricks." *Organizational Research Methods* 9: 536–553. DOI: http://dx.doi.org/10.1177/1094428106286540.

Haraldsen, G., J. Jones, D. Giesen, and L.C. Zhang. 2013. "Understanding and Coping with Response Burden." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, G. Haraldsen, J. Jones, and D. Willimack, 219–252. Hoboken: John Wiley & Sons.

Hedlin, D., T. Dale, G. Haraldsen, and J. Jones, eds. 2005 *Developing Methods for Assessing Perceived Response Burden*. A Joint Report of Statistics Sweden, Statistics Norway and the UK Office for National Statistics. Available at: epp.eurostat.ec.europa.eu/pls/portal/url/ITEM/1CA1E9AC26242D43E0440003BA9322F9 (Accessed October 2006).

Hedlin, D., H. Lindkvist, H. Bäckström, and J. Erikson. 2008. "An Experiment on Perceived Survey Response Burden Among Businesses." *Journal of Official Statistics* 24: 301–318.

Janik, F., and S. Kohaut. 2009 "Why Don't They Answer? – Unit Non-Response in the IAB Establishment Panel." FDZ Methodenreport, Nr. 7/2009, Bundesagentur für Arbeit. Available at: http//doku.iab.de/fdz/reporte/2009/MR_07-09.pdf (accessed February 2011).

Jobber, D., J. Saunders, and V.W. Mitchel. 2004. "Prepaid Monetary Incentive Effects on Mail Survey Response." *Journal of Business Research* 57: 347–350. DOI: http://dx.doi.org/10.1016/S0148-2963(02)00280-1.

Johnson, R.E. and L. Steinman. 2009. "The Use of Implicit Measures for Organizational Research: An Empirical Example." *Canadian Journal of Behavioural Science* 41: 202–212. DOI: http://dx.doi.org/10.1037/a0015164.

Kehr, H.M. 2004. "Integrating Implicit Motives, Explicit Motives and Perceived Abilities: The Compensatory Model of Work Motivation and Volition." *Academy of Management Review* 29: 479–499. DOI: http://dx.doi.org/10.5465/AMR.2004.13670963.

Kennedy, J., and P. Phipps. 1995 "Respondent Motivation, Response Burden, and Data Quality in the Survey of Employer-provided training." Annual Meeting of the American Association for Public Opinion Research, May 1995, Ft. Lauderdale, FL. Available at: http://www.bls.gov/osmr/pdf/st950250.pdf (accessed May 2014).

Kruglanski, A.W. 1975. "The Endogenous-Exogenous Partition in Attribution Theory." *Psychological Review* 82: 387–406. DOI: http://dx.doi.org/10.1037/0033-295X.82.6.387.

McClelland, D.C., J.W. Atkinson, R. Clark, and E.L. Lowell. 1953. *The Achievement Motive*. New York: Free Press.

McClelland, D.C. 1985. "How Motives, Skills and Values Determine What People Do." *American Psychologist* 40: 812–825. DOI: http://dx.doi.org/10.1037/0003-066X.40.7.812.

McClelland, D.C., R. Koestner, and J. Weinberger. 1989. "How Do Self-Attributed and Implicit Motives Differ?" *Psychological Review* 96: 690–702. DOI: http://dx.doi.org/10.1037/0033-295X.96.4.690.

MacQueen, K.M., E. McLellan, K. Kay, and B. Milstein. 1998. "Codebook Development for Team-Based Qualitative Analysis." *Cultural Anthropology Methods Journal* 10: 31–36.

Ouellette, J.A. and W. Wood. 1998. "Habit and Intention in Everyday Life: The Multiple Processes by Which Past Behavior Predicts Future Behavior." *Psychological Bulletin* 124: 54–74. DOI: http://dx.doi.org/10.1037/0033-2909.124.1.54.

Pittman, T.S., A.K. Boggiano, and D.N. Ruble. 1983. "Intrinsic and Extrinsic Motivational Orientations: Limiting Conditions on the Undermining and Enhancing Effects of Reward on Intrinsic Motivation." In *Teacher and student perceptions: implications for learning*, edited by J.M. Levine, 319–340. Hillsdale: Lawrence Erlbaum Associates.

Porter, S.R. 2004. "Raising Response Rates: What Works?" In *New Directions for Institutional Research*, 2004: 5–21. Wiley Periodicals, Inc. DOI: http://dx.doi.org/10.1002/ir.97.

Rivière, P. 2002. "What Makes Business Statistics Special?" *International Statistical Review* 70: 145–159. DOI: http://dx.doi.org/10.1111/j.1751-5823.2002.tb00353.x.

Rogelberg, S.G. and J.M. Stanton. 2007. "Introduction: Understanding and Dealing With Organizational Survey Nonresponse." *Organizational Research Methods* 10: 195–209. DOI: http://dx.doi.org/10.1177/1094428106294693.

Romero, E.J. and K.W. Cruthirds. 2006. "The Use of Humor in the Workplace." *Academy of Management Perspectives* 20: 58–69. DOI: http://dx.doi.org/10.5465/AMP. 2006. 20591005.

Rose, D.S., S.D. Sidle, and K.H. Griffith. 2007. "A Penny for Your Thoughts: Monetary Incentives Improve Response Rates for Company-Sponsored Employee Surveys." *Organizational Research Methods* 10: 225–240. DOI: http://dx.doi.org/10.1177/1094428106294687.

Ryan, R.M. and E.L. Deci. 2000. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25: 54–67. DOI: http://dx.doi.org/10.1006/ceps.1999.1020.

Ryan, G.W. and H.R. Bernard. 2003. "Techniques to Identify Themes." *Field Methods* 15: 85–109. DOI: http://dx.doi.org/10.1177/1525822X02239569.

Sandelowski, M., D. Holditch-Davis, and B.G. Harris. 1992. "Using Qualitative and Quantitative Methods: the Transition to Parenthood of Infertile Couples." In *Qualitative Methods in Family Research*, edited by J.F. Gilgum, K. Daly, and G. Handel, 301–322. London: Sage.

Seens, D. 2010. *Analysis of Regulatory Compliance Costs: Part II. Paperwork Time Burden, Costs of Paperwork Compliance, and Paperwork Simplification*. Statistics Canada: Ottawa. Available at: http://www.reducingpaperburden.gc.ca/eic/site/pbri-iafp.nsf/vwapj/December-Decembre2010_eng.pdf/$file/December-Decembre2010_eng.pdf (accessed September 2013).

Seiler, C. 2010. "Dynamic Modelling of Nonresponse in Business Surveys." In *Technical Report Number 093*, Department of Statistics, University of Munich. Available at: http://www.stat.uni-muenchen.de (accessed February 2011).

Seo, M.G., L.B. Feldman, and J.M. Bartunek. 2004. "The Role of Affective Experience in Work Motivation." *Academy of Management Review* 29: 423–439. DOI: http://dx.doi.org/10.5465/AMR.2004.13670972.

Silvester, J. 2008. "Work and Organizational Psychology." In *The SAGE Handbook of Qualitative Research in Psychology*, edited by C. Willig and W. Stainton-Rogers, 489–505. London: Sage.

Snijkers, G., G. Haraldsen, J. Jones, and D. Willimack. 2013. *Designing and Conducting Business Surveys*. Hoboken: John Wiley & Sons.

Stern, P. 1980. "Grounded Theory Methodology its Uses and Applications." *Image* 12: 20–23.

Strack, F. and R. Deutsch. 2004. "Reflective and Impulsive Determinants of Social Behaviour." *Personality and Social Psychology Review* 8: 220–247. DOI: http://dx.doi.org/10.1207/s15327957pspr0803_1.

Strauss, A. and J. Corbin. 1994. "Grounded Theory Methodology: an Overview." In *Handbook of Qualitative Research*, edited by N.K. Denzin and Y.S. Lincoln, 273–285. London: Sage.

Tetlock, P.E. 1985. "Accountability: A Social Check on the Fundamental Attribution Error." *Social Psychology Quarterly* 48: 227–236.

Tomaskovic-Devey, D., J. Leiter, and S. Thompson. 1994. "Organizational Survey Nonresponse." *Administrative Science Quarterly* 39: 439–457.

Tomaskovic-Devey, D., J. Leiter, and S. Thompson. 1995. "Item Nonresponse in Organizational Surveys." *Sociological Methodology* 25: 77–110. DOI: http://dx.doi.org/10.2307/271062.

Torres van Grinsven, V., I. Bolko, M. Bavdaž, and S. Biffignandi. 2011 "Motivation in Business Surveys." In Proceedings of the BLUE-ETS Conference on Business Burden and Motivation in NSI Survey, Statistics Netherlands, Heerlen, 22–23 March 2011, 7–22. Available at: http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf (accessed May 2014).

Uhlmann, E.L., K. Leavitt, J.I. Menges, J. Koopman, M. Howe, and R.E. Johnson. 2012. "Getting Explicit About the Implicit: A Taxonomy of Implicit Measures and Guide for Their Use in Organizational Research." *Organizational Research Methods* 15: 553–601. DOI: http://dx.doi.org/10.1177/1094428112442750.

Wenemark, M., G. Hollman Frisman, T. Svensson, and M. Kristenson. 2010. "Respondent Satisfaction and Respondent Burden among Differently Motivated Participants in a Health-Related Survey." *Field Methods* 22: 378–390. DOI: http://dx.doi.org/10.1177/1525822X10376704.

Wenemark, M., A. Persson, H. Noorlind Brage, T. Svensson, and M. Kristenson. 2011. "Applying Motivation Theory to Achieve Increased Response Rates, Respondent Satisfaction and Data Quality." *Journal of Official Statistics* 27: 393–414.

Willimack, D.K., E. Nichols, and S. Sudman. 2002. "Understanding Unit and Item Nonresponse in Business Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 213–228. New York: Wiley.

Willis, G.D. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications.

Received February 2013
Revised May 2014
Accepted June 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 607–621, http://dx.doi.org/10.2478/JOS-2014-0040

An Adaptive Data Collection Procedure for Call Prioritization

Jean-Francois Beaumont[1], *Cynthia Bocci*[2], *and David Haziza*[3]

We propose an adaptive data collection procedure for call prioritization in the context of computer-assisted telephone interview surveys. Our procedure is adaptive in the sense that the effort assigned to a sample unit may vary from one unit to another and may also vary during data collection. The goal of an adaptive procedure is usually to increase quality for a given cost or, alternatively, to reduce cost for a given quality. The quality criterion often considered in the literature is the nonresponse bias of an estimator that is not adjusted for nonresponse. Although the reduction of the nonresponse bias is a desirable goal, we argue that it is not a useful criterion to use at the data collection stage of a survey because the bias that can be removed at this stage through an adaptive collection procedure can also be removed at the estimation stage through appropriate nonresponse weight adjustments. Instead, we develop a procedure of call prioritization that, given the selected sample, attempts to minimize the conditional variance of a nonresponse-adjusted estimator subject to an overall budget constraint. We evaluate the performance of our procedure in a simulation study.

Key words: Adaptive collection design; nonresponse bias; nonresponse variance; nonresponse weight adjustment; paradata; responsive collection design.

1. Introduction

The focus of this article is the prioritization of calls in the context of Computer-Assisted Telephone Interview (CATI) surveys. We develop an adaptive data collection procedure that attempts to maximize quality given a certain budget. Our procedure is adaptive in the sense that the effort assigned to a sample unit may vary from one unit to another and may also vary during data collection, which requires the use of paradata. Paradata are data about the data collection process, such as response rates by subgroups of the sample at different time points of data collection.

The quality criterion often considered in the literature is the nonresponse bias of an estimator that is not adjusted for nonresponse. Although the reduction of the nonresponse bias is a desirable goal, we believe that it is not a useful criterion to apply at the data

[1] Statistics Canada, 100 Tunney's Pasture Driveway, R.H. Coats Bldg., 16-B Ottawa K1A 0T6, Canada. Email: Francois.Beaumont@statcan.gc.ca
[2] Statistics Canada, 100 Tunney's Pasture Driveway, R.H. Coats Bldg., 18-E Ottawa K1A 0T6, Canada. Email: Cynthia.Bocci@statcan.gc.ca
[3] Université de Montréal, Département de mathématiques et de statistique, Pavillon André Aisenstadt, Case postale 6128, Montréal H3C 3J7, Canada. Email: david.haziza@umontreal.ca
Acknowledgments: The authors are indebted to Joël Bissonnette of Statistics Canada for his programming of the call prioritization procedure used in the simulations. They would also like to thank all the reviewers for their comments, which greatly improved the overall quality of the article.

collection stage of a survey because the bias that can be removed at this stage through an adaptive collection procedure can also be removed at the estimation stage through appropriate nonresponse weight adjustments. For instance, we could consider a collection procedure that prioritizes cases to be interviewed so as to equalize response rates between domains of interest and then use an estimator that is not adjusted for nonresponse. In terms of nonresponse bias, we expect this strategy to be equivalent to using an estimator that adjusts design weights by the inverse of response rates within domains of interest along with a data collection procedure where cases are selected at random. The reasoning is that auxiliary information known for both respondents and nonrespondents is necessary to reduce the nonresponse bias. (Note that the auxiliary information is the domain information in the above example.) Whether this information is used at the data collection stage or not should not make a difference in terms of nonresponse bias as long as it is used at the estimation stage. This is confirmed in our empirical study (see Section 4). Therefore, the quality criterion that we suggest to minimize is the variance of a nonresponse-adjusted estimator conditional on the selected sample. We use the term nonresponse variance for this conditional variance as it emerges only because of nonresponse and has nothing to do with sampling. This variance disappears in the absence of nonresponse.

Before describing our call prioritization procedure in Section 3, we first provide a selected literature review in Section 2. Section 4 presents the results of a simulation study that evaluates the properties of our procedure and a few alternatives. The conclusion is given in Section 5, which includes some suggestions for potential improvement.

2. Literature Review

The literature on adaptive collection designs, sometimes called adaptive survey designs, responsive collection designs, responsive survey designs or simply responsive designs, is fairly recent. In our context, we prefer the terms adaptive collection designs and responsive collection designs as they make it clear that we are concerned with improvements in data collection methods so that any confusion with the different notion of adaptive sampling designs, which are typically used to sample from rare populations, is avoided.

Groves and Heeringa (2006) defined a responsive survey design as one that uses paradata to guide changes in the features of data collection in order to achieve higher quality estimates per unit cost. Three examples of features of data collection are the data collection mode, the use of incentives and the call prioritization procedure. The implementation of responsive designs in practice requires defining what is meant by quality and determining suitable quality indicators. A cost function must also be chosen. There are two other main concepts underlying the framework of Groves and Heeringa (2006): phase and phase capacity. A phase is a period of data collection during which the same set of methods is used. The first phase is used to gather information about data collection features. In subsequent phases, features are modified (e.g., subsampling of nonrespondents, larger incentives, etc.). A given phase is continued until it reaches its phase capacity, which is typically judged by the stability of some indicator (e.g., an estimate) as the phase matures. Axinn et al. (2011) recently evaluated the consequences of implementing responsive design methods. They studied the extent to which a responsive design altered conclusions reached from analyses of multivariate models by comparing

differences between model coefficients obtained using the main phase sample to those obtained from the responsive phase sample. They concluded that the addition of a responsive design phase can add very different people to the respondent pool, creating significant differences in the magnitude of model coefficients.

Schouten et al. (2009) proposed an indicator of nonresponse bias, called R-indicator, as an alternative to response rates. An R-indicator is sometimes chosen as the quality indicator to be used in conjunction with an adaptive collection design. The proposed R-indicator is a function of estimated probabilities of response to the survey and is designed to measure the representativeness of the respondents to the complete sample. It is constructed using the variability of the response probabilities. A large value of the R-indicator is associated with a low variability of the response probabilities. One drawback of this indicator is that it depends on the proper choice of a nonresponse model; that is, a model for the indicators of response to the survey. In particular, the R-indicator depends on the proper choice of explanatory variables used to model the response probabilities. For instance, if no explanatory variable is included in the nonresponse model, the R-indicator is equal to 1, which is the best value it can reach. Thus, a poor choice of explanatory variables may lead to an artificially large value of the indicator yet does not divulge anything about the actual nonresponse bias. Indeed, the nonresponse bias may vary from one variable of interest to another. Since the R-indicator is independent of any of these variables, it can only provide limited information about nonresponse bias. The authors also considered the maximal bias of an estimator that is not adjusted for nonresponse (no adjustment of design weights). This additional measure is related to the R-indicator and depends on the variable of interest. Like the R-indicator, the maximal bias depends on the proper specification of a nonresponse model. Another limitation of the maximal bias is that it is based on an unadjusted estimator that is rarely used in practice. Schouten et al. (2011) extended the notion of an R-indicator to define partial R-indicators designed to evaluate the contribution of a single specified auxiliary variable to the representativeness of the respondents.

Peytchev et al. (2010) investigated an approach to reducing nonresponse bias through case prioritization. They suggested targeting individuals with lower estimated response probabilities. For instance, individuals could be given larger incentives or interviewers could have larger incentives for completing these cases. Their approach is basically equivalent to trying to increase the R-indicator (or achieving a more balanced sample). They also recommended using explanatory variables that are associated with the variables of interest when modelling the response probability so that the R-indicator is also indirectly associated with these variables.

Laflamme and Karaganis (2010) developed and implemented responsive collection designs for CATI surveys at Statistics Canada. Their approach fits well into the Groves and Heeringa (2006) framework. They considered four phases: a planning phase, an initial collection phase and two responsive design phases. The planning phase is conducted before data collection starts. It consists of analyzing previous data, determining strategies, and so on. The initial collection phase is used to evaluate different indicators to determine when the next phase should start. It corresponds to the first phase of the Groves and Heeringa (2006) framework. The two responsive design phases differ in the way cases are prioritized. The goal of the first responsive design phase is to improve response rates by targeting individuals with higher estimated response probabilities. This tends to increase

the number of respondents, which is desirable. The goal of the second responsive design phase is to reduce the variability of response rates between domains of interest, which is essentially equivalent to increasing the R-indicator. This will likely reduce the variability of nonresponse weight adjustments, which is also desirable. Note that objectives of both phases are intuitively appealing but may be contradictory in terms of cases' prioritization. Laflamme and Karaganis (2010) tried to achieve a compromise between these conflicting objectives by separating data collection into two responsive design phases, each one focusing on a single objective. Our approach, described in Section 3, tries to make a compromise by using a single phase with a single objective function (quality indicator).

Lundquist and Särndal (2013) considered alternatives to the R-indicator based on the distance between the mean of respondents and the mean of the full sample for some auxiliary variables. They chose these alternative indicators to evaluate three experimental strategies using data of the Swedish Living Conditions Survey. Each strategy consists of breaking the sample into groups and, using different intervention points, declaring data collection terminated in a group if the response rate is above a certain target. The strategies differ in the number of intervention points and the selected target. These authors noted that their indicators improve as the target decreases. They concluded that data collection costs could be reduced by choosing a lower target response rate and suggested that these cost savings be used to improve other aspects of the survey design.

Schouten et al. (2013) proposed an interesting theoretical framework for adaptive survey designs focused on assigning collection strategies to sample units. It is apparently the first paper to develop some theory on this topic. The authors suggested maximizing quality for a given cost or, equivalently, minimizing cost for a given quality. The framework requires the choice of a quality indicator such as the overall response rate, the R-indicator, the maximal bias, and so on. The authors did not provide any firm recommendation regarding the choice of an appropriate indicator and a cost function. Our approach fits into this framework in the sense that we maximize quality for a given cost. Our approach is also related to the methodology in Choudhry et al. (2011), although they looked at a different problem. They investigated opportunities for improving the data collection process by focusing on interviewer allocation, whereas we focus on call prioritization.

3. Call Prioritization Procedure

In this Section, we develop a procedure for call prioritization in the context of CATI surveys. The reason for the restriction to CATI surveys is that it is easier to come up with a cost function since the overall cost is highly related to the total time used to conduct data collection. Our procedure aims at maximizing quality given a fixed overall budget. As pointed out in Section 1, our quality indicator is the nonresponse variance of a nonresponse-adjusted estimator.

3.1. A Nonresponse-Adjusted Estimator and Its Nonresponse Variance

Let y_i be the value of a variable of interest y for unit i of the finite population U of size N, and $\theta = \sum_{i \in U} y_i$ be the population total to be estimated. Let s be a sample of size n selected from U through some probability sampling design $p(s)$. Let s_r be the set of respondents of size n_r observed at the end of data collection and generated according to

some nonresponse mechanism, $q(s_r|s)$, which depends on the data collection procedures. Denote a nonresponse adjustment cell by the subscript g, for $g = 1, \ldots, G$, where G is the number of cells. These nonresponse adjustment cells are assumed to be known before data collection and are deemed to be homogeneous with respect to the propensity to respond to the survey. For instance, they may be some important domains of interest. Let s_g of size n_g be the sample units falling in cell g and s_{rg} of size n_{rg} be the set of respondents in cell g at the end of data collection. Every sample unit belongs to one and only one cell. We suppose that the estimator that would be used to estimate θ under complete response is the expansion estimator, which in our notation can be written as $\hat{\theta} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{gi}\, y_{gi}$, where y_{gi} is the y-value of unit i in cell g, $w_{gi} = 1/\pi_{gi}$ is its design weight and $\pi_{gi} = \Pr(i \in s_g)$ is its selection probability. The expansion estimator $\hat{\theta}$ is p-unbiased (design-unbiased) for θ in the sense that $E_p(\hat{\theta}) = \theta$. The subscript p indicates that the expectation is evaluated with respect to the sampling design.

Let us denote by $\rho_{gi} = \Pr(i \in s_{rg} | s, i \in s_g)$ the probability that sample unit i in cell g is a respondent to the survey at the end of data collection. It can be interpreted as the proportion of times sample unit i would respond to the survey if data collection could be repeated independently an infinite number of times, always with the same procedures and the same sample s. Obviously, the response probability ρ_{gi} depends on the data collection procedures and, more specifically, on the resources spent to obtain a response from unit i in cell g. In the sequel, we assume uniform nonresponse within cells; that is, all sample units $i \in s$ respond independently of one another and all sample units $i \in s_g$ have the same probability of response to the survey; that is, $\rho_{gi} \equiv \rho_g$ for all $i \in s_g$. The response probability ρ_{gi} may vary from one cell to another but is assumed to be constant within a cell. This is a standard assumption in the survey nonresponse literature. Assuming that ρ_g is known, $g = 1, \ldots, G$, a nonresponse-adjusted estimator of θ is:

$$\tilde{\theta}_A = \sum_{g=1}^{G} \sum_{i \in s_{rg}} \frac{w_{gi}}{\rho_g} y_{gi}.$$

It is q-unbiased for $\hat{\theta}$ in the sense that $E_q(\tilde{\theta}_A | s) = \hat{\theta}$. The subscript q indicates that the expectation is evaluated with respect to the nonresponse mechanism. As a result, the adjusted estimator $\tilde{\theta}_A$ is also pq-unbiased for θ; that is, $E_{pq}(\tilde{\theta}_A) = \theta$. In practice, the response probabilities ρ_g are unknown and must be estimated. A possible q-unbiased estimator is the response rate in cell g, $\hat{\rho}_g = n_{rg}/n_g$. Since $E_q(\hat{\rho}_g | s) = \rho_g$, the response probability ρ_g can be interpreted as the expected response rate in cell g. The use of $\hat{\rho}_g$ leads to the standard nonresponse-adjusted estimator of θ:

$$\hat{\theta}_A = \sum_{g=1}^{G} \sum_{i \in s_{rg}} \frac{w_{gi}}{\hat{\rho}_g} y_{gi}. \tag{1}$$

The estimator $\hat{\theta}_A$ is not q-unbiased for $\hat{\theta}$, unlike $\tilde{\theta}_A$. However, under certain conditions, including a large sample size, the squared nonresponse bias of $\hat{\theta}_A$, $\left\{E_q(\hat{\theta}_A - \hat{\theta}|s)\right\}^2$, is small compared with its nonresponse variance, $\mathrm{var}_q(\hat{\theta}_A | s)$. We make this assumption and consider the nonresponse variance of the adjusted estimator (1), $\mathrm{var}_q(\hat{\theta}_A | s)$, as our quality indicator. We choose to condition on the sample s to define our quality indicator because

data collection procedures have an impact only on the nonresponse mechanism, $q(s_r|s)$, and not on the sampling design, $p(s)$. They have also no effect on the sampling error, $\hat{\theta} - \theta$, and can only reduce the nonresponse error, $\hat{\theta}_A - \hat{\theta}$. The nonresponse variance of the adjusted estimator (1) is approximated through a first-order Taylor linearization by

$$\text{var}_q(\hat{\theta}_A|s) \cong \sum_{g=1}^{G} (\rho_g^{-1} - 1)(n_g - 1)S_{wy,g}^2, \tag{2}$$

where

$$S_{wy,g}^2 = \frac{1}{n_g - 1}\sum_{i \in s_g}(w_{gi}y_{gi} - \hat{\mu}_g)^2 \text{ and } \hat{\mu}_g = \frac{1}{n_g}\sum_{i \in s_g} w_{gi}y_{gi}.$$

Remark 1: The variance $S_{wy,g}^2$ is variable specific and typically unknown. It could be estimated using data from a previous period of the survey (see Subsection 3.5). When there are many variables of interest, a compromise must be made. This is an issue similar to sample allocation in stratified sampling. The typical solution used in practice consists of replacing the variables of interest by a single variable x that is hopefully strongly associated with the main variables of interest. Data reduction techniques such as principal component analysis could possibly be used to obtain such a variable.

Remark 2: The nonresponse-adjusted estimator (1) has a small nonresponse bias if the sample size is large and if uniform nonresponse within cells is a reasonable assumption. When the nonresponse bias is small, it makes sense to consider a data collection procedure that minimizes the nonresponse variance (2). If the uniform nonresponse assumption is not valid, then the nonresponse-adjusted estimator (1) may become substantially biased. However, this nonresponse bias cannot be eliminated at the data collection stage if no additional information on the nonrespondents is used, as evidenced in our simulation study in Section 4 (see response scenario I in Table 1). This is why we ignore the nonresponse bias and focus on the minimization of the nonresponse variance.

Remark 3: Calibration is often performed after nonresponse weight adjustment. By using linearization techniques, it would not be difficult to extend our approach to account for calibration. For simplicity, we restrict our study to the non-calibrated estimator $\hat{\theta}_A$ in (1).

3.2. *The Overall Cost and Its Expectation*

We assume that the overall cost of the survey depends only on $C_{NR,g}$, $C_{R,g}$ and m_{gi}, which are the cost of an unsuccessful attempt in cell g, the cost of an interview in cell g and the total number of attempts at the end of data collection for unit i in cell g, respectively. The overall cost can thus be expressed as:

$$C_{TOT} = \sum_{g=1}^{G} C_{TOT,g},$$

where

$$C_{TOT,g} = \sum_{i \in s_{rg}} \left[(m_{gi} - 1) C_{NR,g} + C_{R,g} \right] + \sum_{i \in s_g - s_{rg}} m_{gi} C_{NR,g}.$$

The expected overall cost is given by

$$\tilde{C}_{TOT} = E_q(C_{TOT}|s) = \sum_{g=1}^{G} \tilde{C}_{TOT,g},$$

where

$$\tilde{C}_{TOT,g} = E_q(C_{TOT,g}|s) = (C_{R,g} - C_{NR,g}) n_g \rho_g + C_{NR,g} \sum_{i \in s_g} \tilde{m}_{gi}$$

and $\tilde{m}_{gi} = E_q(m_{gi}|s)$ is the expected number of attempts made at the end of data collection for unit i in cell g. Suppose the number of calls for unit i in cell g is restricted to be no greater than a certain fixed value, M_{gi}, which is known as the cap on calls for unit i in cell g. Note that we allow the cap on calls to differ between sample units although in practice it is often set to a constant. The expected number of attempts $\tilde{m}_{gi}$ depends on the probability of response at each call attempt for unit i in cell g, denoted by p_{gi}, and the cap on calls, M_{gi}. Note that p_{gi} is different from the probability of response to the survey, ρ_{gi}, introduced in Subsection 3.1. The expected number of attempts $\tilde{m}_{gi}$ depends not only on p_{gi} and M_{gi} but also on the effort made to obtain a response for unit i in cell g, which is itself related to the overall budget and the data collection procedures. The strict derivation of $\tilde{m}_{gi} = E_q(m_{gi}|s)$ is not straightforward. To simplify it, we make the following three assumptions:

i) The response probability p_{gi} is constant from one attempt to the next.
ii) For any given sample unit, response is independent from one attempt to the next.
iii) At the end of data collection, every sample unit is either a respondent or has reached the cap on calls.

Assumption (i) implies that the response probability p_{gi} does not depend on characteristics that vary over time for sample units. Assumption (ii) is more realistic if a certain amount of time is imposed between two successive calls. Assumption (iii) means that a sample unit cannot be a nonrespondent without having reached the cap on calls. It would be satisfied if the overall budget is sufficiently large (and there is no refusal). A consequence of this assumption is that the expected number of attempts $\tilde{m}_{gi}$ is only a function of p_{gi} and M_{gi}. Although we recognize that these three assumptions may not always be satisfied in practice, we believe that they provide a useful approximation to $E_q(m_{gi}|s)$. This is confirmed in our empirical study in Section 4. Using these three assumptions, we obtain:

$$\begin{aligned} \tilde{m}_{gi} &= E_q(m_{gi}|s) \\ &= \left(\sum_{t=1}^{M_{gi}-1} t\, p_{gi} (1 - p_{gi})^{t-1} \right) + M_{gi} (1 - p_{gi})^{M_{gi}-1} = \frac{1}{p_{gi}} \left(1 - (1 - p_{gi})^{M_{gi}} \right). \end{aligned} \tag{3}$$

The algebra to go from the second to the third equation in (3) is simple but tedious and is thus omitted.

Since $\tilde{m}_{gi}$ in (3) is only a function of p_{gi} and M_{gi}, the expected overall cost becomes a linear function of the expected response rates, $\rho_g, g = 1, \ldots, G$:

$$\tilde{C}_{TOT} = \lambda_0 + \sum_{g=1}^{G} \lambda_{1g}\rho_g \tag{4}$$

with $\lambda_0 = \sum_{g=1}^{G} C_{NR,g} \sum_{i \in s_g} \tilde{m}_{gi}$ and $\lambda_{1g} = (C_{R,g} - C_{NR,g})n_g$.

3.3. The Optimization Problem and Its Solution

Our objective consists of finding the target expected response rates, $\rho_{Tg}, g = 1, \ldots, G$, that minimize the nonresponse variance (2), with ρ_g replaced by ρ_{Tg}, subject to the budget constraint, $\lambda_0 + \sum_{g=1}^{G} \lambda_{1g}\rho_{Tg} = K$, for a constant K that represents the overall budget. The solution is:

$$\rho_{Tg} = \sqrt{\frac{(n_g - 1)S^2_{wy,g}}{\delta\lambda_{1g}}} = \sqrt{\left(\frac{n_g - 1}{n_g}\right)\frac{S^2_{wy,g}}{\delta(C_{R,g} - C_{NR,g})}}, \tag{5}$$

where

$$\delta = \frac{\left(\sum_g \sqrt{\lambda_{1g}(n_g - 1)S^2_{wy,g}}\right)^2}{(K - \lambda_0)^2}. \tag{6}$$

If $C_{R,g}$ and $C_{NR,g}$ are constant from one cell to another and if n_g is large, then $\rho_{Tg} \propto S_{wy,g}$, which is a solution similar to the sampling fraction obtained using Neyman allocation in stratified sampling. Moreover, if $S_{wy,g}$, $g = 1, \ldots, G$, are found to be constant then the resulting target expected response rates ρ_{Tg} are also constant and thus maximize the R-indicator. However, our proposed solution does not generally maximize the R-indicator.

Unfortunately, nothing guarantees that the target expected response rates ρ_{Tg} are smaller than 1. If some of the ρ_{Tg}, $g = 1, \ldots, G$, are not smaller than 1, they must be replaced by a value smaller but close to 1 (see Subsection 4.1 for a possible choice).

It might be useful to graph the minimum nonresponse variance, obtained using (2) with ρ_g replaced by ρ_{Tg} in (5), as a function of the overall budget K. The minimum nonresponse variance should decrease as the budget increases. There may be a value of budget above which the minimum nonresponse variance cannot be reduced significantly and it may not be justified to spend more than that value.

3.4. Procedure for the Selection of Cases to Be Interviewed

Once the target expected response rates ρ_{Tg} have been determined, we must find the effort needed to achieve these targets. Let e_{gi} be the maximum effort (in terms of the number of attempts) associated with unit i in cell g. Under assumptions (i) and (ii) and assuming that unit i in cell g will be attempted at most e_{gi} times, its response probability to the survey is $\rho_{gi} = 1 - (1 - p_{gi})^{e_{gi}}$. It is the probability that there is a response in no more than e_{gi} attempts. We now want to find the effort e_{gi} that makes this response probability equal to the target expected response rate ρ_{Tg} for each sample unit. This yields

$$e_{gi} = \frac{\ln(1 - \rho_{Tg})}{\ln(1 - p_{gi})}. \tag{7}$$

Our call procedure consists of selecting cases to be interviewed with probability proportional to the effort e_{gi}. The effort in (7) increases with the target expected response rate and decreases with the response probability at each attempt. A larger response probability at each attempt indicates that this unit is easier to contact and thus requires less effort to reach the target expected response rate. Note that it might sometimes be advisable to trim large values of e_{gi}, which may occur because of small response probabilities p_{gi}, so as to avoid unduly large efforts for some units. This will prevent spending a large portion of the budget for such units, especially if there is no cap on calls.

3.5. *Estimation of $S^2_{wy,g}$ and p_{gi}*

In practice, $S^2_{wy,g}$ and p_{gi} are unknown and must be estimated. In a repeated survey, a natural choice is to use data collected at a previous point in time to obtain estimates of both $S^2_{wy,g}$ and p_{gi}. The estimation of p_{gi} requires the contact history as p_{gi} is the probability of response at a given attempt and not the probability of response to the survey. Paradata such as the number of call attempts should be incorporated in the response probability model if revision of the solution is considered during data collection (see Subsection 3.6). Even though the number of call attempts varies over the data collection period for each sample unit, this does not necessarily invalidate the above assumption (i). It may well be that the true response probability depends on a variable that does not vary over the data collection period but that is not observed and thus cannot be used in the model. At a certain point during data collection, some of the noninterviewed units will have been attempted many times. Such units are likely to be associated with a lower response probability p_{gi}. As a result, the response rate of units with a large number of attempts is expected to be smaller than the response rate of units with a small number of attempts. Therefore, a response probability model that uses the number of attempts should yield estimated response probabilities that become closer to the true response probabilities as data collection progresses than a response probability model that ignores the number of attempts. Other paradata may also be useful, such as information on the data collection mode or the time of the interview.

Estimates of $S^2_{wy,g}$ and p_{gi} are denoted by $\hat{S}^2_{wy,g}$ and $\hat{p}_{gi}$, respectively. Once they are obtained, they must replace $S^2_{wy,g}$ and p_{gi} in the solution of the optimization problem and in the selection of cases. The resulting target expected response rate and effort are denoted $\hat{\rho}_{Tg}$ and $\hat{e}_{gi}$, respectively.

3.6. *Revision of the Solution to the Optimization Problem*

The solution to the above optimization problem is found before data collection starts. However, it may be desirable to revise the solution periodically (e.g., daily) as data collection progresses. Consider a revision at time t whereby the call prioritization procedure is temporarily suspended to allow updating of the estimated response probabilities at an attempt, the target expected response rate and the effort. Updated values of effort are then used in the call prioritization procedure when it resumes.

With R revisions in total, the call prioritization procedure is applied $R+1$ times during the data collection process; that is, at times $t = 0, 1, \ldots, R$. Let $\hat{p}_{gi}^{[t]}$ denote the estimated probability of response at a given attempt at time t. As pointed out in Subsection 3.5, paradata may be used as explanatory variables in the response probability model so that $\hat{p}_{gi}^{[t]}$ is not necessarily constant over the data collection period. The estimated probability $\hat{p}_{gi}^{[t]}$ is used in the above optimization problem, leading to a revised target expected response rate at time t, $\hat{\rho}_{Tg}^{[t]}$.

The revised target expected response rate, $\hat{\rho}_{Tg}^{[t]}$, must then be updated to account for the units that have already responded. Define $n_{rg}^{[t]}$ as the number of respondents in cell g at time t with $n_{rg}^{[0]} = 0$. The actual response rate in cell g at time t is: $\hat{\rho}_{g}^{[t]} = n_{rg}^{[t]}/n_g$. The target expected number of respondents in cell g at time t is $n_g\hat{\rho}_{Tg}^{[t]}$, so that $n_g\hat{\rho}_{Tg}^{[t]} - n_{rg}^{[t]}$ is the expected number of respondents that still remain to be interviewed. The total number of units that have not yet been interviewed at time t is $n_g - n_{rg}^{[t]}$. We thus suggest using the updated target expected response rate,

$$\hat{\rho}_{Tg}^{*[t]} = \frac{n_g\hat{\rho}_{Tg}^{[t]} - n_{rg}^{[t]}}{n_g - n_{rg}^{[t]}} = \frac{\hat{\rho}_{Tg}^{[t]} - \hat{\rho}_{g}^{[t]}}{1 - \hat{\rho}_{g}^{[t]}}, \tag{8}$$

to account for units that have already responded. Using (8), we obtain the updated effort at time t,

$$\hat{e}_{gi}^{*[t]} = \frac{\ln(1 - \hat{\rho}_{Tg}^{*[t]})}{\ln(1 - \hat{p}_{gi}^{[t]})}. \tag{9}$$

Unsurprisingly, less effort is spent in the cells for which the response rate $\hat{\rho}_{g}^{[t]}$ is already close to the revised target expected response rate $\hat{\rho}_{Tg}^{[t]}$. It is even possible that $\hat{\rho}_{g}^{[t]} > \hat{\rho}_{Tg}^{[t]}$, which yields negative values of $\hat{\rho}_{Tg}^{*[t]}$ and $\hat{e}_{gi}^{*[t]}$. In such cases, the effort should be focused on the cells for which the effort $\hat{e}_{gi}^{*[t]}$ is positive.

4. Numerical Example

We simulated the proposed call prioritization procedure to investigate some aspects of performance and compared it with a few alternatives.

4.1. Description of the Simulation Experiment

We used data of the 2005 Workplace and Employee Survey (WES) conducted at Statistics Canada. Our sample, s, consists of 773 business locations in the Atlantic provinces for which a unique identifier, a design weight, stratum information used for cell assignment and gross payroll (the variable of interest y) are available. For illustration purposes, we considered three cells defined using the number of employees in a workplace. The resulting sample sizes for the three groups were $n_1 = 305$, $n_2 = 188$ and $n_3 = 280$.

The cost of a nonresponse attempt, the cost of an interview and the cap on calls were set to $C_{NR,g} = 1$, $C_{R,g} = 25$ and $M_{gi} = 25$, respectively. Note that $C_{NR,g}$ and $C_{R,g}$ can be interpreted as time units in this experiment. Furthermore, the overall budget was set to

$K = 20,000$. The quantity $S^2_{wy,g}$ was estimated using a previous iteration of the survey, the 2003 WES data in the Atlantic provinces.

We considered three response scenarios that differ in the way the true probability of response at an attempt, p_{gi}, is defined. In each scenario, response is generated according to assumptions (i) and (ii). In addition, response is generated independently from one unit to another. The response probability p_{gi} for the three scenarios is given as follows:

(C) Uniform: The probability of response is constant with $p_{gi} = 0.1$, $i \in s$;
(G) Uniform within cells: The probability of response varies by cell with units in the same cell having the same probability (i.e., $p_{gi} = p_g$, $i \in s_g$) with $p_1 = 0.24$, $p_2 = 0.16$, and $p_3 = 0.04$. The average response probability over all units $i \in s$ is equal to 0.15;
(I) Not missing at random: The probability of response, p_{gi}, depends on gross payroll, y_{gi}, and thus varies from one unit to another. It is defined such that the average response probability over all units $i \in s$ is equal to 0.15. The coefficient of correlation between p_{gi} and y_{gi} is 0.67.

Several effort scenarios were also considered: (C) $e_{gi} =$ constant (P) proposed definition in (9) (R) $\hat{e}^{[t]}_{gi} \propto -1/\ln(1 - \hat{p}^{[t]}_{gi})$ (E) $\hat{e}^{[t]}_{gi} \propto (\hat{p}^{[t]}_{gi})^{10}$. The effort scenario R is obtained by using (9) with constant target expected response rates. It is thus designed to maximize the R-indicator of Schouten et al. (2009). The effort scenario E was added to see the effect of maximizing (approximately) the overall response rate.

For effort scenarios P, R, and E, effort at the start of data collection, referred to as initial effort, and revised effort were functions of estimated response probabilities at an attempt. A response probability model was developed using the 2003 WES survey data in the Atlantic provinces. Real call data information for this previous sample was no longer available. Therefore, a single set of respondents was generated for each response scenario using a constant effort and the simulated call history was saved for each sample unit. For each response scenario, we modeled the response probability p_{gi} as a function of the cell g, $g = 1, 2, 3$, and the categorized number of attempts a, $a = 1, 2, 3, 4, 5$. The categorized number of attempts was defined as follows: $a = 1$ for 1 attempt; $a = 2$ when the number of attempts is 2 or 3; $a = 3$ when number of attempts is 4 or 5; $a = 4$ when number of attempts is 6, 7 or 8; and $a = 5$ when the number of attempts is greater than 8. Using the simulated call history of this previous sample, the estimated response probability at an attempt was the response rate within cell and attempt categories; that is, it was computed as

$$\hat{p}_{gi} \equiv \hat{p}(g, a) = \frac{\text{number of responses among observations in cell } g \text{ with attempt category } a}{\text{number of observations in cell } g \text{ with attempt category } a}$$

Thus, there were 15 estimated response probabilities for each of the response scenarios C, G, and I. These probabilities were used to determine the effort for the 2005 WES sample. Since the number of attempts changes over the data collection period, the estimated response probability $\hat{p}_{gi}$ is modified at each revision. At the start of data collection we have $\hat{p}^{[0]}_{gi} \equiv \hat{p}(g, 1)$. There were two revisions in the data collection process: at time $t = 1$ when 1/3 of the total budget was exhausted and at time $t = 2$ when 2/3 of the total budget was exhausted. At times of revision, estimated response probabilities were revised by noting the

cumulative number of attempts at time of revision t and using this value to derive and update the categorized number of attempts at time of revision t, $a^{[t]}$, so that $\hat{p}_{gi}^{[t]} \equiv \hat{p}(g, a^{[t]})$.

There is nothing in our theoretical framework to prevent a value of $\hat{\rho}_{Tg}^{*[t]} \geq 1$ or $\hat{\rho}_{Tg}^{*[t]} < 0$ resulting in an undefined or negative effort. We addressed these problems as follows. If $\hat{\rho}_{Tg}^{*[t]} \geq 1$ then the effort for the units in that cell was set to one plus the maximum effort observed among all other units in s. If during revision $\hat{\rho}_{Tg}^{*[t]} < 0$, then the revised effort for the units in that cell was set to the minimum positive effort observed among all units in s.

Data collection ends when the total budget K is exhausted. At this point, the estimate $\hat{\theta}_A$ can be computed from the realized set of respondents. We generated $B = 5{,}000$ sets of respondents for each response-effort scenario and computed $\hat{\theta}_A^{(b)}$, $b = 1,2,\ldots,B$. The complete data estimate $\hat{\theta} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{gi} y_{gi}$ is fixed in this simulation because sampling is not repeated. Monte Carlo measures of Relative Bias (RB), Mean Squared Error (MSE) and Variance (V) of the estimator $\hat{\theta}_A$ were used to assess the performance of various scenarios. They are given respectively as

$$RB = \frac{(1/\hat{\theta}) \sum_b (\hat{\theta}_A^{(b)} - \hat{\theta})}{B}, \quad MSE = \frac{\sum_b (\hat{\theta}_A^{(b)} - \hat{\theta})^2}{B} \quad \text{and} \quad V = MSE - \left(\hat{\theta} \times RB\right)^2.$$

Then, we computed the Relative MSE (RMSE), which is the ratio of the mean squared error for each response-effort scenario to that obtained for the same response scenario with constant effort:

$$RMSE^{resp,eff} = \frac{MSE^{resp,eff}}{MSE^{resp,C}},$$

where $resp = C$, G, or I, $eff = C$, P, R or E, and $MSE^{resp,eff}$ is the Monte Carlo mean squared error for a given response-effort scenario. Similarly, we computed the Relative Variance (RV) defined as

$$RV^{resp,eff} = \frac{V^{resp,eff}}{V^{resp,C}}.$$

The RV could be termed call prioritization effect by analogy with the notion of design effect, which is used to measure the impact on the variance of a complex sampling design compared with simple random sampling. In addition to the above measures, we computed the average over the 5,000 repetitions of the overall response rate and of the R-indicator.

4.2. Results

Table 1 gives summary statistics for all the response-effort scenarios considered. As expected, the proposed effort scenario was the most efficient for the three response scenarios with the smallest $RMSE^{resp,eff}$ and $RV^{resp,eff}$. Response scenarios C and G yielded approximately unbiased estimates for all effort scenarios. The effort scenario E maximized the overall response rate in all cases. As demonstrated in the (*resp*, *eff*) scenario (G, E), the overall response rate is not necessarily a good indicator of the nonresponse variance or of nonresponse MSE. The effort scenario R is intended to maximize the R-indicator. Our results showed that there does not seem to be any relationship between the R-indicator and the nonresponse bias or variance. The response scenario I leads to large nonresponse biases that

Table 1. Simulation statistics based on 5,000 repetitions for various response and effort scenarios

Scenario					Average overall response rate (%)	Average R-indicator (%)
Response Scenario	Effort Scenario	*RB* (%)	$RMSE^{resp,eff}$ (%)	$RV^{resp,eff}$ (%)		
C	P	−0.118	61.1	61.0	76.1	87.6
C	C	0.044	100.0	100.0	76.1	97.3
C	R	0.024	95.2	95.2	76.1	96.1
C	E	0.034	133.0	133.0	76.1	77.2
G	P	−0.158	53.9	53.8	73.7	75.7
G	C	0.123	100.0	100.0	78.3	64.2
G	R	−0.196	66.2	66.1	74.0	86.9
G	E	0.203	111.7	111.7	79.2	56.6
I	P	11.013	91.8	70.2	77.2	87.2
I	C	11.406	100.0	100.0	79.4	82.1
I	R	11.863	106.0	73.4	78.0	95.3
I	E	10.506	92.0	197.1	80.0	67.6

are similar for all the effort scenarios. This is because missing values are not missing at random and the true response probabilities p_{gi} are not estimated accurately. As expected, none of the effort scenarios can completely eliminate the nonresponse bias in this case. For each of the response scenarios, the different effort scenarios led to similar nonresponse biases. In other words, the use of cells at the data collection stage did not have any significant effect on the nonresponse bias because this information was already used at the estimation stage. It is thus not possible to favor one effort scenario based solely on nonresponse bias. Based on nonresponse variance and MSE, our approach was the best in this simulation study.

Out of curiosity, we ran the response scenario I using the known probabilities of response p_{gi} in the calculation of efforts instead of the estimated response probabilities. This case never occurs in practice but is informative to better understand the reasons for the nonresponse biases in the response scenario I in Table 1. These results are shown in Table 2. The nonresponse bias is reduced significantly for the effort scenario R because this scenario can take advantage of the additional information available when p_{gi} is known. Note that our proposed approach also succeeded in reducing the nonresponse bias even though it is focused on reducing the nonresponse variance. Note also that the information contained in p_{gi} (when it is known) could be used to improve the estimator $\hat{\theta}_A$; for example, by making nonresponse classes homogeneous with respect to p_{gi}. This would result in a reduced nonresponse bias for all the effort scenarios because we would be in a situation similar to the response scenario G in Table 1.

5. Conclusion

We have proposed a call prioritization procedure that attempts to minimize the nonresponse variance of a nonresponse-adjusted estimator subject to an overall budget constraint. Our empirical study showed that it had a smaller mean squared error than alternatives such as maximizing the R-indicator for three different nonresponse mechanisms.

Although we focused on the reduction of the nonresponse variance, we believe that the reduction of the nonresponse bias remains a more important issue in practice. However,

Table 2. Simulation statistics based on 5,000 repetitions for the response scenario I and various effort scenarios using known probabilities of response at an attempt

Scenario						
Response Scenario	Effort Scenario (using known probabilities of response)	*RB* (%)	$RMSE^{resp,eff}$ (%)	$RV^{resp,eff}$ (%)	Average overall response rate (%)	Average R-indicator (%)
I	P	1.862	9.7	113.5	76.4	86.6
I	C	11.406	100.0	100.0	79.4	82.1
I	R	−0.636	14.8	217.2	76.7	97.2
I	E	18.224	241.5	38.8	81.0	71.0

we think that a call prioritization procedure cannot reduce the nonresponse bias to a better extent than a proper nonresponse weight adjustment. This is confirmed in our empirical study, since all the data collection procedures tested led to similar nonresponse biases. Hansen and Hurwitz (1946) proposed an approach to reducing the nonresponse bias at the data collection stage of a survey by randomly selecting a subsample of nonrespondents after a certain point during data collection. Indeed, their approach completely eliminates the nonresponse bias if all the units selected in the subsample respond. The latter condition is not realistic in practice and nonresponse is usually present in the subsample, which is likely to result in nonresponse bias. We conjecture that the method of Hansen and Hurwitz (1946) can achieve some nonresponse bias reduction even if there is some nonresponse in the subsample of nonrespondents. If the method is used, our call prioritization procedure can still be applied within the subsample to reduce the nonresponse variance.

Our procedure could be improved in a number of ways. For instance, we could distinguish two different types of nonrespondents at the end of data collection: those who have refused to respond to the survey and those who have not been contacted. The probability that a given unit $i \in s_g$ is a respondent to the survey at the end of data collection, ρ_g, would then become the product of two different probabilities: the probability that unit i is contacted at the end of data collection, denoted by ρ_g^c, and the probability that unit i responds to the survey once contacted, denoted by $\rho_g^{r|c}$. A call prioritization procedure only has an effect on the contact probability ρ_g^c. Similarly to Subsection 3.3, the goal would be to find the target contact probabilities that minimize the nonresponse variance subject to a budget constraint. The cost function would need to be modified to account for the difference between the cost of a refusal and the cost of a noncontact. Also, no further call attempt is made after a refusal, unlike a noncontact. The probability of response at a given call attempt for a unit $i \in s_g$, p_{gi}, would be modeled in two steps similarly to the probability ρ_g, as described above. Although this idea was not fully developed in our article and still requires further thought, we believe that it would not be too difficult to extend our approach so as to handle different types of nonrespondents.

Further research and improvements remain to be done to make the approach even more useful. Our notion of effort was defined in terms of number of call attempts. It might be possible to extend our definition of effort so as to cover other types of data collection features such as incentives. This definitely requires further thought. Another improvement

would be to restrict the estimated target expected response rate at time t, $\hat{\rho}_{Tg}^{[t]}$, to be no smaller than the achieved response rate at time t, $\hat{\rho}_{g}^{[t]}$, but smaller than 1; that is, $\hat{\rho}_{g}^{[t]} \leq \hat{\rho}_{Tg}^{[t]} < 1$. This inequality constraint would require an iterative algorithm since a closed-form solution to the optimization problem would no longer be possible. Finally, it would be useful to relax the three assumptions made to obtain the expected number of attempts $\tilde{m}_{gi}$. In particular, it should be investigated how to relax assumption (iii) so as to account for the fact that the budget is not unlimited and that the effort e_{gi} has an effect on $\tilde{m}_{gi}$.

6. References

Axinn, W.G., C.F. Link, and R.M. Groves. 2011. "Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior." *Demography* 48: 1127–1149. DOI: http://dx.doi.org/10.1007/s13524-011-0044-1.

Choudhry, G.H., M.A. Hidiroglou, and F. Laflamme. 2011. *Optimizing CATI Workload Simultaneously Minimize Data Collection Cost for Several Surveys*, Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, October 31-November 1, 2011.

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, Series A* 169: 439–457. DOI: http://dx.doi.org/10.1111/j.1467-985x.2006.00423.x.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-response in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529. DOI: http://dx.doi.org/10.1080/01621459.1946.10501894.

Laflamme, F. and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." In *Proceedings of the European Conference on Quality in Official Statistics*, Helsinki, Finland, May 2010. Available at: http://q2010.stat.fi./media//presentations/1_Responsive_design_paper_london_event1_revised.doc (accessed October, 2014).

Lundquist, P. and C.-E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29: 557–582. DOI: http://dx.doi.org/10.2478/jos-2013-0040.

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias in Surveys through Case Prioritization." *Survey Research Methods* 4: 21–29.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response through Adaptive Survey Designs." *Survey Methodology* 39: 29–58.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the representativeness of survey response." *Survey Methodology* 35: 101–113.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics* 27: 1–24.

Received January 2013
Revised February 2014
Accepted July 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 623–649, http://dx.doi.org/10.2478/JOS-2014-0041

Measuring Representativeness of Short-Term Business Statistics

Pim Ouwehand[1] *and Barry Schouten*[2]

Short-term statistics (STS) are important early indicators of economic activity. The statistics are obligatory for all EU countries and also serve as input to national accounts. In most countries, short-term Statistics are based on business surveys. However, in recent years a number of countries have gradually replaced their business surveys with business VAT registry data. An important question is whether these surveys and registries are representative of the populations and whether representativity is stable in time. We apply R-indicators and partial R-indicators to measure the representativity of both kinds of data sources. We find large differences between different months of the year and between the two data sources. We discuss dual frame approaches that optimize the accuracy of STS statistics.

Key words: Business surveys; registry data; survey nonresponse.

1. Introduction

Short-term business statistics (STS) provide early indicators of economic activity in the EU countries. These statistics are produced on a monthly basis and represent estimated total revenue for various business sizes (in terms of number of employees) and types of economic activity (according to NACE classification of business activities, an abbreviation of 'Nomenclature statistique des activités économiques dans la Communauté européenne'). The STS estimates are mostly based on business surveys. However, an increasing number of countries are starting to include Value Added Tax (VAT) registry data or even to use registry data to replace business surveys entirely. In the Netherlands, registry data is used because legislation prohibits surveying economic indicators that can be derived from registry data with sufficient accuracy. The prerequisite for the use of registry data, hence, is a constraint on quality, which is to some extent left ambiguous. In this article, we investigate an important aspect of quality: the representativeness of the business data that form the input to the STS. We do so by applying a new set of indicators that has recently been proposed and that can supplement more traditional measures such as the unit and quantity response rates.

Both business surveys and business registry data suffer from nonresponse and thus may not be completely representative of the population. Although participation in STS business

[1] Department of Methodology, Statistics Netherlands, PO Box 4000, 2270JM Den Haag, The Netherlands. Email: powd@cbs.nl
[2] Department of Methodology, Statistics Netherlands, PO Box 4000, 2270JM Den Haag, The Netherlands. Email: bstn@cbs.nl

surveys is obligatory by law, some of the businesses do not respond or respond too late. Reporting business data to the VAT register is also obligatory, but the VAT register was not set up to serve statistical needs and, as a consequence, reporting deadlines do not meet the STS deadlines. Some of the reports are still missing when the STS is produced. Furthermore, the Tax Authorities allows smaller businesses to report at a lower frequency than larger businesses. Smaller businesses have to apply for permission to do so, but it can be presumed that this option has a considerable impact on the accuracy of STS statistics.

The nonresponse error is an influential component of the total estimation error. Nonresponse leads to missing data, which in turn may lead to biased estimators of population parameters. The response to business surveys and business registry data should therefore be representative of the population. Although the feature of representativeness is often discussed and debated, it is seldom defined with mathematical rigor. Little and Rubin (2002) provide a clear definition of three missing data mechanisms that underlie inferences about a population parameter of a certain variable. Nonresponse is Missing-Completely-at-Random (MCAR) for a certain variable, say revenue, when the nonresponse is independent of that variable. Nonresponse is Missing-at-Random (MAR) for a variable conditional on a specified set of covariates, when the nonresponse is independent of the variable given the covariates. All other nonresponse is called Not-Missing-at-Random (NMAR). Most business statistics implicitly assume a Missing-at-Random mechanism conditional on business size and type of activity.

Schouten et al. (2009) gave explicit definitions for representative response and for conditionally representative response and introduced quality indicators that measure deviations from these two properties. They have labelled the indicators generally as representativeness indicators, or R-indicators. Response is termed representative for a set of covariates when the propensities to respond are equal over the classes formed by these covariates. Response is termed conditionally representative for one set of covariates conditionally on another set of covariates, when the response propensities for the first set are equal within the classes formed by the second set. The two definitions are closely related to the missing data mechanisms: When response is representative for a set of covariates X, then it is MCAR for all variables in X. When response is conditionally representative for X given Z, then it is MAR for all variables in X given Z. The indicators are based on the estimated variation in response probabilities and have been extensively tested on social survey data. The indicators serve four purposes: comparison of representativeness over surveys, comparison of representativeness of a survey in time, monitoring of representativeness during data collection, and optimization of data collection designs. The choice of covariates depends on the purpose of the indicators, but clearly always excludes the survey variables themselves. Therefore, indicators cannot be used to extrapolate conclusions about MCAR, MAR or NMAR mechanisms beyond those of the selected covariates and one should always mention the selected covariates in order to avoid such conclusions. The rationale behind the indicators is, however, that they measure process quality: The stronger the deviation from representative response on relevant covariates, the more one should worry about nonrepresentative response on survey variables. In our case study for the STS, the available covariates are strongly related to the main survey variables. An extensive exposition and discussion of representativity is given in the papers by Kruskal and Mosteller (1979 a, b and c).

To produce statistics more efficiently, less labour intensively, and of higher quality, Statistics Netherlands is replacing part of its surveys with registry data, particularly for small and medium-sized enterprises. However, there is a clear difference in missing data mechanisms between these two sources of data, which is based on the reporting schedule (monthly, quarterly, annually) and lateness of the VAT data.

The main underlying question of this article is whether VAT data can lead to the same accuracy of STS statistics as survey data or whether a dual frame approach is required. Of course, this question has many angles, of which representativeness of response is just one, but an important one in our view. In order to investigate representativeness, we focus on two purposes of the indicators: comparison of STS over time and monitoring during data collection. This is done for both business survey data and business VAT registry data. The monitoring and adjustment of the collection process based on R-indicators is clear for STS, but less clear for VAT data, since these latter data are not collected via a survey. However, the collection process can be influenced in a less direct way, by agreeing with the Tax Authorities when data is sent. The detailed research questions are:

- How representative are survey and registry data with respect to relevant business characteristics?
- How does representativeness evolve in time, that is, over months and during data collection?
- What groups need to be targeted to improve representativeness of survey and registry data?
- How can survey and registry data be optimally combined in a dual frame approach?

The answer to the fourth question is dependent on the answer to the third question; only if both data sources attract different respondents can they complement each other. We will show that VAT and the STS survey indeed have different underrepresentations of businesses.

To evaluate the representativeness of response and be able to compute the R-indicators, we linked various registries to the business survey and VAT registry of 2007. The VAT registry data for 2006 were linked to both data sets. The Tax Authorities registry of wages and the type of economic activity as derived by the Chamber of Commerce were linked to the VAT data as well. We linked similar variables from the business population register as maintained by Statistics Netherlands to the business survey.

In Section 2, we provide a short background with respect to representativeness and representativeness indicators. In Section 3, we describe the STS data sources and the available business characteristics. We answer the four research questions in Section 4 and end with a discussion in Section 5.

2. How to Measure Representativeness?

In this section, we briefly revisit the definitions of representative response and of so-called representativeness indicators or R-indicators. These measures were introduced by Schouten et al. (2009) and Schouten et al. (2011). We do not give a detailed statistical account of their statistical properties but refer to Shlomo et al. (2012) for details.

In this section, we also link R-indicators and unit response rates to quantity response rates, which are more common measures of nonresponse error in business surveys (see the recent review by Thompson and Oliver 2012).

Throughout this section, we illustrate the concepts using a simplified example. Consider a simulated business population stratified into four disjoint subpopulations defined by crossing two characteristics: type of economic activity (NACE) in two categories and activity status in previous calendar year (yes or no reported VAT). The sizes of the four groups in the population are: NACE Type 1 business and not active in previous year = 33%, NACE Type 1 business and active in previous year = 17%, NACE Type 2 business and not active in previous year = 17%, and NACE Type 2 business and active in previous year = 33%. Table 1 contains the unit response rates over the first six months for the four subpopulations. Also given are the average monthly revenues of businesses in the four subpopulations, which we take as constant over the six months for the sake of simplicity. The unit response rates for the four subpopulations are consistently different, with each response pattern remaining fairly consistent over the observed months. In the following sections, we evaluate the representativeness of the response over time in the example.

2.1. *Overall Representativeness – R-indicators*

In daily survey practice, the term 'representativeness' is often used as a desirable property of response, but without a rigorous definition. Schouten et al. (2009) therefore propose a definition of representative response. They call a response representative when response probabilities are equal for all population units, or, in other words, when the population units all show exactly the same response behaviour. A natural measure of deviation from representative response given the definition is the standard deviation of response probabilities. Schouten et al. (2009) transform the standard deviation, so that it takes values between 0 (fully nonrepresentative) and 1 (fully representative), and call it a representativeness indicator or R-indicator. The rationale behind R-indicators is that they are a relevant measure that can be monitored, evaluated and compared over different surveys or registry data, and that are complementary to the unit response rate. In Subsection 2.4, we show how the unit response rate and the R-indicator relate to the quantity response rate.

We introduce some notation. Let X be a vector consisting of auxiliary variables, for example, number of employees, reported VAT in a previous year and economic activity

Table 1. Monthly unit response rates and average monthly revenue for subpopulations based on type of economic activity (NACE) and activity status in the previous year (reported VAT > 0)

Type	Status	Jan	Feb	Mar	Apr	May	Jun	Average revenue
1	Not active	65%	72%	71%	69%	71%	65%	100
1	Active	92%	88%	92%	92%	89%	85%	300
2	Not active	62%	66%	65%	66%	66%	60%	200
2	Active	91%	89%	90%	89%	88%	85%	400

Table 2. Monthly R-indicators with respect to economic activity and status

	Jan	Feb	Mar	Apr	May	Jun
R(X)	0.726	0.809	0.779	0.778	0.807	0.781

(NACE). Let the response propensity function $\rho_X(x)$ be defined as the probability of response given that $X = x$. A response to a survey is called representative with respect to X when response propensities are constant for X, that is, when $\rho_X(x)$ is a constant function.

The R-indicator for X is defined as the standard deviation $S(\rho_X)$ of the response propensities transformed to the [0,1] interval by

$$R(X) = 1 - 2S(\rho_X). \tag{1}$$

When all propensities are equal, the standard deviation is zero and hence fully representative response is represented by a value of 1 for the indicator. A value of 0 indicates the largest possible deviation from representative response.

Table 2 provides the R-indicator values for the example of Table 1 based on the two auxiliary variables 'type of economic activity' and 'activity status'. It shows that the indicator for January is considerably lower than for the other months. Hence, in January the variation in the subpopulation response propensities is largest and the businesses show the most diffuse response behaviour.

2.2. *Disentangling Nonrepresentative Response – Partial R-indicators*

In order to locate the sources of deviations from representative response, Schouten et al. (2011) introduce partial R-indicators. Partial R-indicators perform an analysis of variance decomposition of the total variance of response probabilities into between and within variances. The between and within variance components help to identify variables that are responsible for a large proportion of the variance. The partial R-indicators are linked to a second definition called conditional representative response, defined as a lack of within variance. The resulting between and within components are termed unconditional and conditional partial R-indicators.

Again we introduce some notation. Let Z be an auxiliary variable not included in X, for example, the region in which a business is located. Let $\rho_{X,Z}(x, z)$ be the probability of response given that $X = x$ and $Z = z$. The response to a survey is called conditionally representative with respect to Z given X when conditional response propensities given X are constant for Z, that is, when $\rho_{X,Z}(x, z) = \rho_X(x)$ for all z. Hence, when the response propensities over country regions are the same for businesses employing the same type of economic activity, then response for region is conditionally representative given economic activity.

The square root of the between variance $S_B(\rho_{X,Z})$ for a stratification based on Z is called the unconditional partial R-indicator. It is denoted by $P_u(Z)$ and it holds that $P_u(Z) \in [0, 0.5]$. So values of $P_u(Z)$ close to 0 indicate that Z does not produce variation in response propensities, while values close to 0.5 represent a variable with maximal impact on representativeness.

For categorical variables the between variance can be further decomposed to the category level in order to detect which categories contribute most. Let Z be a categorical variable with categories $k = 1, 2, \ldots, K$ and let Z_k be the 0–1 variable that indicates whether $Z = k$ or not. For example, Z represents the region of a country and Z_k is the indicator for area k. The partial R-indicator for category k is defined as

$$P_u(Z,k) = \sqrt{\frac{N_k}{N}}\left(\frac{1}{N_k}\sum_U Z_k \rho_{X,Z}(x_i,z_i) - \frac{1}{N}\sum_U \rho_{X,Z}(x_i,z_i)\right) \tag{2}$$

with N_k the number of population units in category k. It follows that $P_u(Z,k) \in [-0.5, 0.5]$. So a value close to 0 implies that the category subpopulation shows no deviation from average response behaviour, while values close to -0.5 and 0.5 indicate maximal underrepresentation and overrepresentation respectively. The category-level indicators are the category components in the total between variance.

The logical counterpart to the unconditional partial R-indicator is the conditional partial R-indicator. It considers the other variance component: the within variance. The conditional partial R-indicator for Z given X, denoted by $P_c(Z|X)$, is defined as the square root of the within variance $S_W(\rho_{X,Z})$ for a stratification based on X. Again it can be shown that $P_c(Z|X) \in [0, 0.5]$, but now the interpretation is conditional on X. A value close to 0 means that the variable does not contribute to variation in response propensities in addition to X, while large values indicate that the variable brings in new variation. When X is type of economic activity and Z is region, then $P_c(Z|X) = 0$ means that one should focus on economic activity when improving response representativeness, as region does not add any variation.

Again for categorical variables Z, the within variance can be broken down to the category level. The category-level conditional partial R-indicator for category k is

$$\boldsymbol{P}_c(\boldsymbol{Z},\boldsymbol{k}|\boldsymbol{X}) = \sqrt{\frac{1}{N-1}\sum_U Z_k(\rho_{X,Z}(x_i,z_i) - \rho_X(x_i))^2}. \tag{3}$$

Unlike the unconditional indicators, the conditional indicators do not have a sign. A sign would have no meaning as the representation may be different for each category of X. For instance, in some categories a certain economic activity may have a positive effect on response while in others it may have a negative effect. The conditional partial R-indicator for Z is always smaller than the unconditional partial R-indicator for that variable; the impact on response behaviour is to some extent removed by accounting for other characteristics of the population unit.

Table 3 shows the partial R-indicators for the two variables in the example of Table 1; type of economic activity and activity status. As expected, January shows larger values for the partial indicators. However, after conditioning it follows that the extra contribution to selective response in January comes mostly from activity status. In all months, the activity status is the dominant source of selective response.

Table 3. Monthly unconditional and conditional partial R-indicators for type of economic activity (NACE) and activity status (reported VAT in the previous year >0)

		Jan	Feb	Mar	Apr	May	Jun
Type of activity	P_u	0.04	0.02	0.02	0.02	0.02	0.03
	P_c	0.01	0.02	0.02	0.01	0.02	0.01
Activity status	P_u	0.14	0.09	0.11	0.11	0.10	0.11
	P_c	0.13	0.09	0.11	0.11	0.10	0.11

2.3. Representativeness and Nonresponse Bias

R-indicators can be interpreted in terms of nonresponse bias through the variance of response propensities. Consider the standardized bias of the design-weighted, unadjusted response mean $\hat{\bar{y}}_r$ of an arbitrary variable y, say total revenue. The standardized bias of the mean can be bounded from above by

$$\frac{|B(\hat{\bar{y}}_r)|}{S(y)} = \frac{|Cov(y, \rho_Y)|}{\rho_U S(y)} = \frac{|Cov(y, \rho_{\aleph})|}{\rho_U S(y)} \leq \frac{S(\rho_{\aleph})}{\rho_U} = \frac{1 - R(\aleph)}{2\rho_U}, \tag{4}$$

with ρ_U the unit response rate (or average response propensity) and $\aleph$ some 'super' vector of auxiliary variables providing full explanation of nonresponse behaviour.

Clearly, the propensity function $\rho_{\aleph}$ is unknown. Since R-indicators are used for the comparison of the representativeness of response in different surveys or the same survey over time, the interest lies in the general representativeness of a survey, that is, not the representativeness with respect to single variables. Therefore, as an approximation for (4) is used:

$$B_m(X) = \frac{1 - R(X)}{2\rho_U}. \tag{5}$$

B_m is the maximal (standardized) bias for all variables that are linear combinations of the components of X. For other variables, (5) does not provide an upper bound to the bias. The choice of X, therefore, is very important, but even for relevant X, (5) cannot be extrapolated to all survey target variables. If the selected set of variables in X is correlated with the survey variables, then (5) is informative as a quality indicator. If it is not correlated with the survey variables, then it has limited utility.

A useful graphic display of unit response rates and response representativeness is given by so-called response-representativity functions. Ideally, one would like to bound the R-indicator from below, that is, to derive values of the R-indicator that are acceptable and values that are not. If the R-indicator takes a value below some lower bound, then measures to improve response are paramount. Response-representativity functions can be used for deriving such lower bounds for the R-indicator. They are a function of a threshold γ and the unit response rate ρ_U. The threshold γ represents a quality level. The functions are defined as

$$RR(\gamma, \rho) = 1 - 2\rho_U \gamma, \tag{6}$$

and follow by demanding that the maximal bias given by (5) is not allowed to exceed the prescribed threshold γ, that is, from taking $B_m(X) = \gamma$. For STS, a reasonable threshold γ

can be set by considering the final response obtained at the end of data collection when unit response rates are very high.

Figure 1 presents an RR-plot for the example given in Table 1. The pair of values for January is the only set that is above the 15% level. All other months are between the 10% and 15% levels.

2.4. *R-Indicators, Unit Response Rates and Quantity Response Rates*

A measure commonly used in business statistics is the quantity response rate. It is the ratio between the quantity reported by the respondents and the quantity that would be reported if all sample units were respondents. The quantity response rate is different for each study variable Y. We denote it by ρ_Q and suppress the dependence on Y. The application to business statistics is natural; businesses have diverse revenues and often a small number of businesses make up most of the total revenue. For a useful and recent discussion we refer to Thompson and Oliver (2012).

The quantity response rate is defined as

$$\rho_Q = \frac{y_r}{y_n} = \frac{\sum_{i=1}^{n} d_i r_i y_i}{\sum_{i=1}^{n} d_i y_i}, \tag{7}$$

with d_i the design weight for business i, r_i the 0–1 response indicator for business i, and y_i the value of the study variable for business i. Hence, y_r and y_n denote, respectively, the design-weighted response and sample totals. In Appendix A we show that for large sample sizes, the expected value of (7) is approximately equal to

$$E\rho_Q = \rho_U + \frac{\mathrm{cov}\,(y, \rho_Y)}{\bar{y}_N}, \tag{8}$$

with $\bar{y}_N$ the population mean. We may view (8) as the population representation of the quantity response rate which is estimated by (7). From (8) we can conclude that the quantity response rate is equal to the unit response rate whenever there is no linear relation

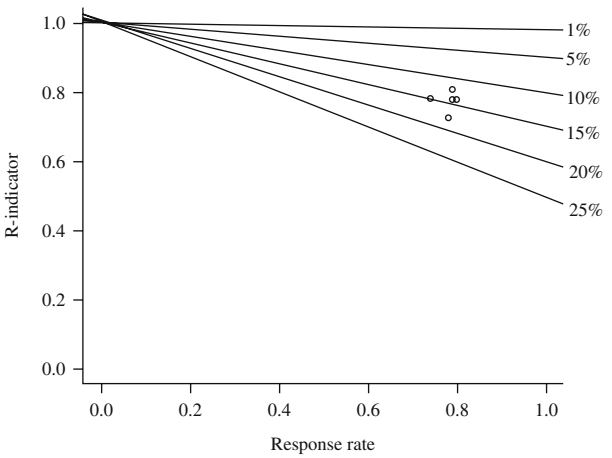


Fig. 1. RR-plot for six months. The thresholds γ are 1%, 5%, 10%, 15%, 20% and 25% (from top to bottom)

between the quantity under study and response propensities. With similar arguments to (4) it can be shown that

$$E\rho_Q \in \rho_U \pm \frac{(1 - R(\aleph))S(y)}{2\bar{y}_N}, \tag{9}$$

so that the R-indicator appears as a component in lower and upper limits to the quantity response rate for auxiliary variables.

The quantity response rate in (7) is an (unbiased) estimator for (8) but can only be computed for variables that are not subject to nonresponse themselves, that is, variables that are auxiliary and can be linked to the sample. For survey variables, the denominator in (7) is unknown and needs to be estimated. It is usually estimated by imputing the nonrespondents or weighting the respondents. The denominator in (7) is then replaced by an estimator that employs auxiliary information, usually taken from the same set of available auxiliary variables that are input to R-indicators. As a result the estimated quantity response rate may be biased itself. Furthermore, when new response comes in during data collection this bias may change and the estimator must be updated retrospectively. Consequently, quantity response rate patterns that are computed when data collection is completed may look different from quantity response rate patterns that are computed in real time during data collection. As a result, and somewhat confusingly, the quantity response rate is not necessarily monotone increasing and may decrease through some periods of data collection; the estimated sample total may become larger when new response comes in.

In this article, we estimate the denominator of (7) using a poststratification estimator. The population is stratified into H subpopulations, $h = 1, 2, \ldots, H$, based on an auxiliary variable, say Z, and the sample mean per stratum is estimated by the design-weighted response mean per stratum. The quantity response rate estimator is then defined as

$$\rho_Q = \frac{y_r}{y_{post}} = \frac{y_r}{\sum_{h=1}^{H} N_h \bar{y}_{r,h}}, \tag{10}$$

with N_h the size of stratum h in the population and

$$\bar{y}_{r,h} = \frac{\sum_{i \in h} d_i r_i y_i}{\sum_{i \in h} d_i r_i} \tag{11}$$

the design-weighted response mean in stratum h.

Appendix A shows that the expected value of (10) can be approximated by

$$E\rho_Q = \frac{\rho_U \bar{y}_N + \operatorname{cov}(y, \rho_Y)}{\bar{y}_N + \sum_{h=1}^{H} \frac{N_h}{N} \frac{\operatorname{cov}_h(\varepsilon, \rho_Y)}{\rho_{U,h}}}, \tag{12}$$

where ε represents the residuals in the poststratification.

When the residuals show no correlation to the response propensities, that is, when the poststratification provides unbiased estimators of the stratum means, then (12) equals (8). If there is a nonzero correlation, then the denominator is biased and (12) and (8) are different. Assuming that the study variable only takes non-negative values, it is possible to

derive lower and upper limits to (12) that are expressed in terms of unit response rates and unconditional partial R-indicators

$$E\rho_Q \in \frac{\rho_U \bar{y}_N}{\bar{y}_N - \sum_{h=1}^{H} \frac{N_h}{N} \frac{S_h(\varepsilon)|P_U(Z,h)|}{\rho_{U,h}}} \pm \frac{(1 - R(\aleph))S(y)/2}{\bar{y}_N - \sum_{h=1}^{H} \frac{N_h}{N} \frac{S_h(\varepsilon)|P_U(Z,h)|}{\rho_{U,h}}}. \quad (13)$$

In Table 4, we show the two response rates for the example of Table 1. As expected, the quantity response rate is always higher as businesses with larger revenues have higher response probabilities (see Table 1). Both rates are relatively stable over the months, except for June that has smaller response rates. The simultaneous drop of the rates for June indicates that this drop is not strongly related to revenue.

2.5. *The Utility and Limitations of R-Indicators*

R-indicators and partial R-indicators can be useful tools to supplement unit and quantity response rates, but they also have limitations. We discuss both here.

In the setting of STS, the quantity response rate would be computed for total business revenue, the key variable. As a single indicator, the complement of the quantity response rate represents the total revenue that is still missing. In conjunction with the unit response rate however, it allows for more elaborate conclusions. The height of the quantity response rate relative to the unit response rate tells whether larger or smaller businesses are overrepresented. A difference in slope between the two rates can provide information on the evolution of these representations; for example, when the quantity response rate grows faster than the unit response rate, then it is likely that bigger businesses have responded better over that time window. The utility of the R-indicator, in addition to unit and quantity response rates, is that it quantifies over- and underrepresentation, it allows for a multivariate view on multiple business characteristics, and it can in theory be estimated without bias both after and during data collection. The R-indicator and partial R-indicators are designed to have a multivariate view. The R-indicator measures the simultaneous deviation from representative response for a range of variables and allows any particular variable to be zoomed in on. The unconditional partial R-indicators do just what quantity response rates are doing: show the impact on single variables. The conditional partial R-indicators allow for a search for the strongest variables in a multivariate context, which is what quantity response rates are lacking; they do not account for multicollinearity.

It is important to stress that the R-indicator values depend on the vector of auxiliary variables X. For different selections of X, the R-indicator attains different values and the (partial) R-indicators do not allow for statements about NMAR nonresponse outside the selected vector of variables. Therefore the selection is a crucial and influential part of

Table 4. Monthly unit and quantity response rates

	Jan	Feb	Mar	Apr	May	Jun
ρ_U	78%	79%	80%	79%	79%	74%
ρ_Q	84%	83%	85%	84%	83%	79%

the analysis. The purpose of the indicator determines the selection of the auxiliary variables that are used. When multiple surveys are compared, it is essential that representativeness is evaluated in terms of generally available and relevant characteristics, such as type of economic activity or business size. For the other three purposes mentioned in Section 1, it is important to select characteristics that are closer to the survey topics and key variables. In the case of short-term statistics, it is paramount to have variables that relate to the revenue of a business. We return to this issue in Section 3.

The response propensity function ρ_X is unknown, and needs to be estimated from the survey response data. A consequence of the estimation of the propensities is that R, P_u, P_c and B_m need to be estimated as well. Schouten et al. (2011) and Shlomo et al. (2012) propose estimators for these population parameters and derive analytic approximations to their standard errors and bias. The estimators replace population means with design-weighted sample and response means and response propensities with estimated propensities. Propensities are estimated by means of general linear models such as linear regression, logistic regression, or probit regression. The resulting estimators have a standard error and indicator values need to be evaluated along with their precision. On the website www.risq-project.eu code in SAS and R is available to compute indicators and their standard errors. To allow for comparison it is crucial that the set of auxiliary variables and the link function, for example, linear or logistic, are kept fixed. Hence, variables are selected beforehand based on their relevance to the survey variables and are always included in the models when monitoring or comparing nonresponse.

Since only response propensities for X need to be estimated, the models for nonresponse cannot be misspecified in terms of omitted variables and in theory response propensities can be estimated without bias. However, since sample sizes are always limited in practice, some interactions between the variables may have to be omitted and/or some categories of variables may have to be merged. In order to enable comparison over surveys and over time, such adaptations need to be applied beforehand to all data sets under study. As a result, the models for nonresponse may be viewed as misspecified for the selected variables and leading to biased estimators for response propensities. It is therefore not enough to provide the variable names when presenting indicators; their classification also needs to be specified.

The R-indicator, variable-level and category-level partial R-indicators together form a set of tools that can be used to search effectively for population subgroups that need to be targeted in data collection. A strategy is given by Schouten et al. (2012):

1. Compute the R-indicator for different time periods.
2. When strong differences are found in Step 1, assess the unconditional variable-level partial R-indicators for all auxiliary variables; the variables that have the highest scores have the strongest single impact on representativity of response. They are also the strongest candidates to be monitored and analysed more closely and subsequently to be involved in design changes and data collection interventions.
3. Assess the conditional variable-level partial R-indicators for all auxiliary variables; the conditional values are needed in order to check whether some of the variables are strongly collinear. If indicator scores remain high, then the strongest variables are selected. If indicator scores vanish by conditioning, then it is sufficient to focus only

on a subset of the variables. A low conditional indicator value implies that the corresponding variable is conditionally representative.

4. Repeat Steps 2 and 3 but now for the category-level partial R-indicators and for the selected auxiliary variables only; the subgroups that need to be targeted in design changes are those categories that have large negative unconditional scores and large conditional scores.

This strategy is used in Subsection 4.3, where we identify the business groups that influence representativeness the most.

3. Short-Term Statistics

3.1. Survey and Registry Data

The traditional way of collecting data for business statistics is to send questionnaires to a sample of enterprises. To produce statistics more efficiently, less labour intensively, and with higher quality, Statistics Netherlands is replacing part of its surveys with registry data, particularly for small and medium-sized enterprises. Apart from costs, a strong incentive for the use of registry data is business response burden. The use of VAT data reported to the Tax Authorities would reduce the burden to enterprises as they have to provide data only once. Yet another advantage of registry data is their sheer size. Registry data aim at a full enumeration of the population. As a consequence, the number of observations is much larger than for regular business surveys.

However, in both surveys and registers part of the data is missing at the time when statistics need to be produced. For the VAT registry data this is particularly the case for monthly statistics (Vlag and Van den Bergen 2010). Although both sources of data are subject to missing data, the missing data mechanisms are very different. In a survey, typically some of the enterprises in the sample do not respond to the questionnaire, or have to be prompted several times. At Statistics Netherlands, however, the enterprises are not targeted in a specific way and data collection is therefore uniform. Registers, on the other hand, may not be complete due to regulations about reporting of enterprises to register holders and time delays in reporting.

The data sets used represent turnover data for both Retail trade and Manufacturing industries for 2007. Turnover refers to the invoice value of sales to third parties of goods and services produced within a company. The VAT register is linked to the employment register containing wages, so that these can be used as auxiliary information. About 75% of the VAT units could be linked to wages from the employment register. For the smallest enterprises (<€2,500 VAT) this was about 60%; for the larger enterprises this was at least 80%. For VAT, we selected all VAT units that were obliged to report their VAT. The number of records is given in Table 5.

The VAT data includes records for companies reporting on a monthly, quarterly or annual basis. The reporting frequency depends on the amount of VAT a company is expected to report, or is based on individual requirements made by the Tax Authorities. If the VAT of a company lies below €1,883 per year, they can report on an annual basis. If it exceeds €15,000 per quarter, they should report on a monthly basis. Most companies

Table 5. Sample and register size

	Retail trade		Manufacturing	
	VAT	*Survey*	*VAT*	*Survey*
January	124,602	7,852	59,346	5,393
June	126,158	7,871	60,229	5,381
July	127,568	7,727	61,023	5,355
December	128,212	7,864	61,521	5,078

report on a quarterly basis (Van Delden and Aelen 2008, and Slootbeek and Van Bemmel 2010).

In the case of the VAT register, companies are required to report 25 days after a reporting period has ended, and statistics are produced 30 days after that period. However, some companies do not report within 30 days. For the STS survey, companies are given the same deadline for responding. The STS sample is a stratified random sample of all enterprises where strata are business size classes. The design weights also depend on the NACE category at the highest level, that is, between the Retail and Manufacturing industries but not within these business types.

3.2. *Auxiliary Variables in the Computation of the R-indicators*

The comparison was made for four different months with very distinct characteristics of VAT data: January, June, July, and December. The data for January includes only companies reporting on a monthly basis. In June, we have companies reporting on a monthly and on a quarterly basis. July, again, only includes companies reporting on a monthly basis, but unlike January is not at the beginning of the year. December includes companies reporting on a monthly, quarterly, and annual basis.

Ideally, we would like to compare the R-indicators for both types of data using the same auxiliary variables. However, the VAT and survey data sets do not share the exact same set of auxiliary variables. This is caused by the difference in population frames as used by Statistics Netherlands and the Tax Authorities. For VAT data, we can use the current year's monthly wages records, the previous year's VAT records (for the same month), and a business classification (enterprise groups according to NACE classification of 1974). For survey data we can use business size, business classification (economic activity according to NACE classification of 1993) and VAT of the previous year (for the same month). Table 6 presents an overview.

Table 6. Available variables and their number of categories

Variable	# categories
VAT($t-12$)	9
Wages(t)	10
Business size	9
NACE 2-digit (1974) Manufacturing	20
NACE 3-digit (1974) Retail trade	18
NACE subsection (1993) Manufacturing	12
NACE 3-digit (1993) Retail trade	7

The current year's wages resemble business size. Business size is a classification of the number of employees which can be expected to be proportional to the wages. It is, however, not the *same* variable so that a direct comparison is hampered. The two business classifications also show a clear resemblance but are not exactly the same. This leads to a problem when we want to combine these specific VAT and STS data sets. Tables, however, are available that link the codes of both classifications. For the majority of codes there is a direct translation between the classifications. However, some codes in one classification may be divided into two or more codes in the other system. For this, heuristic solutions are available.

A second difference between the data sets is the units for which turnover is recorded. Tax units do not completely match survey units, especially when larger businesses are concerned. The differences in variables and units imply that some care is needed in the comparison of absolute values of indicators. However, what can be compared is the patterns of representativeness over months and in time.

For both Retail trade and Manufacturing, we tested a model based on the VAT register and a model based on STS survey data. Table 7 presents an overview of the models. For the moment we ignore the type of economic activity.

Since VAT data should replace surveys, we compute the representativity of the response through time, as additional survey or VAT data becomes available. We compare the representativity for both types of data and compare representativity to the unit response rate. Since companies are required to report 25 days after a reporting period has ended, and statistics are produced 30 days after that period, we computed both unit response rate and R-indicator 25, 26, 27, 28, 29, 30, and 60 days after a reporting period had finished.

4. Results

4.1. What is the Representativeness of STS Based on Survey and Registry Data?

We first computed the representativity at 25 days after the end of the reporting period has finished. This is currently the deadline for companies to report their VAT, and the moment at which the production of statistics commences.

Table 8 and Table 9 show unit response rates, quantity response rates, and R-indicators for both industries and all months. For VAT, the unit response rate is the number of units that have reported VAT as a proportion of all units in the register (i.e., units that should report their VAT on either a monthly, quarterly or annual basis). For survey data, the unit response rate is the proportion of units in the sample that have responded. The quantity response rate is the proportion of total turnover available at a certain time point. For VAT, these proportions are calculated as the sum of turnovers of reporting units divided by turnover of all units in the register. For survey data, the proportions are calculated as the

Table 7. Models used for the estimation of response propensities

Model	Data set used	Specifications
VAT	VAT register	VAT($t-12$) + Wages(t)
STS	STS survey data	VAT($t-12$) + Business size

Table 8. VAT data: Unit response rates, quantity response rates and R-indicators for four months, 25 days after the reporting period

Industry	Month	Response rate		R-indicator
		Unit	Quantity	
Retail trade	January	0.20	0.57	0.68
	June	0.64	0.74	0.74
	July	0.15	0.41	0.76
	December	0.48	0.42	0.85
Manufacturing	January	0.26	0.65	0.54
	June	0.67	0.57	0.61
	July	0.18	0.46	0.68
	December	0.49	0.34	0.80

sum of turnovers of responding units divided by turnover of all units in the sample. For survey data, the quantity response rate is thus calculated retrospectively.

When we look at the results for VAT data for both Retail trade and Manufacturing, the unit and quantity response rates clearly vary from month to month due to the types of businesses that respond. January and July have lower response rates than June and December. For STS data, response rates show less variation since there is less variation in business types reporting than for VAT data. Despite variation in response rates, the representativity shows less variation. Apparently, the additional enterprises that respond in some months do not make response more representative.

4.2. *How Does Representativeness Evolve in Time?*

The results in the previous subsection focus on a single time lag only. In this subsection we will discuss how response and representativity change during data collection.

In Figures 2 and 3, we present graphs of the unit response rate, the quantity response rate and R-indicator for Retail trade using the VAT model and the STS model, respectively. The graphs show results for January, June, July, and December 2007. In all graphs, indicators are computed after 25, 26, 27, 28, 29, 30, and 60 days of data collection.

Table 9. STS data: Unit response rates, quantity response rates and R-indicators for four months, 25 days after the reporting period

Industry	Month	Response rate		R-indicator
		Unit	Quantity	
Retail	January	0.67	0.71	0.89
	June	0.71	0.70	0.90
	July	0.64	0.71	0.93
	December	0.73	0.65	0.91
Manufacturing	January	0.63	0.67	0.93
	June	0.67	0.72	0.94
	July	0.64	0.69	0.93
	December	0.72	0.76	0.92

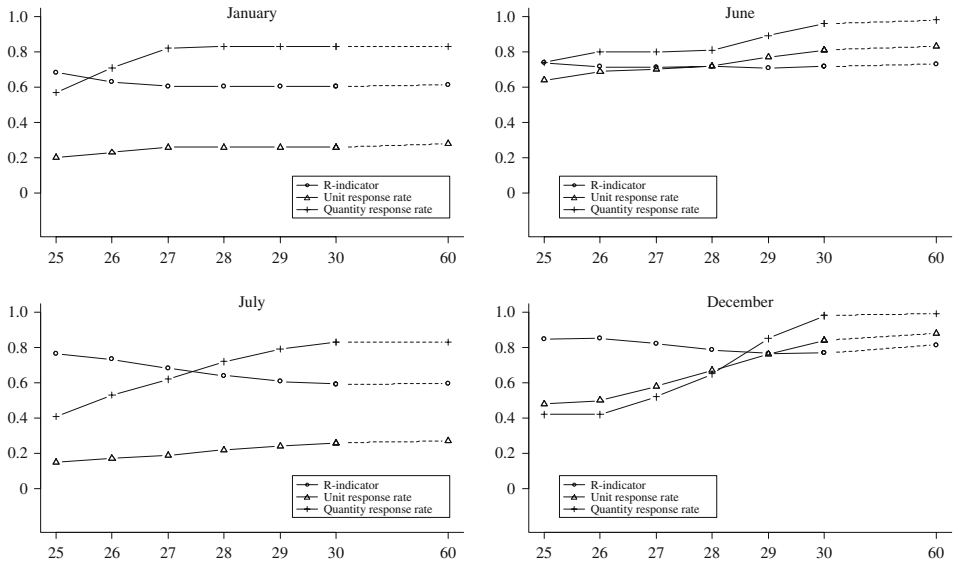


Fig. 2. Unit response rates, quantity response rates and R-indicators based on VAT data for Retail trade (VAT model), for four months

The figures show that for both the VAT and STS model, and all four months under investigation, both the unit response rate and quantity response rate increase as the data collection period progresses. For VAT data, the quantity response rate for July and December approaches 100% (since all companies must report), while for survey data it is relatively stable after 25 days. The response patterns for VAT in January and July are quite

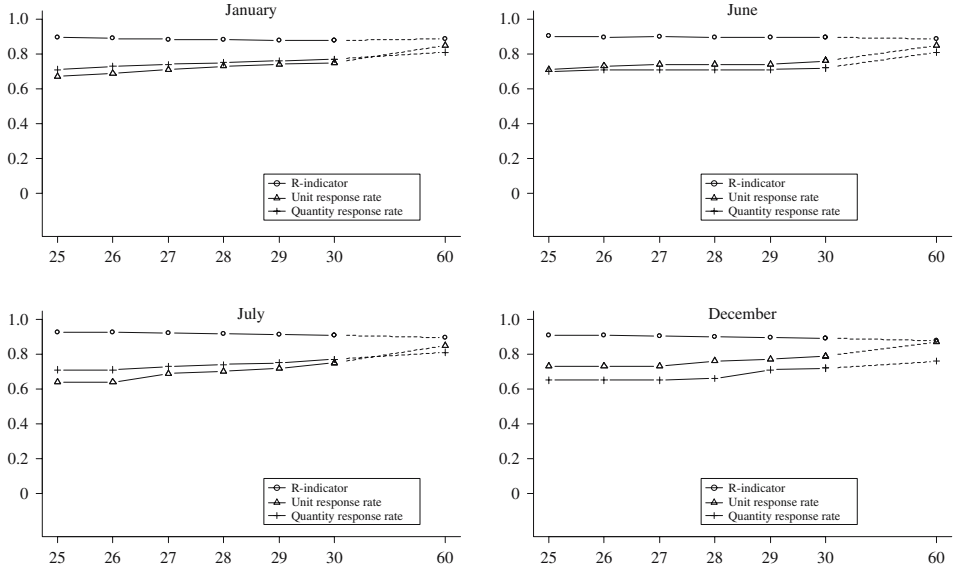


Fig. 3. Unit response rates, quantity response rates and R-indicators based on survey data for Retail trade (STS model), for four months

similar, since these consist of monthly reporters only. Likewise, the patterns for July and December are similar, since these also include quarterly reporters.

In some cases there is a clear difference between the development of the unit response rate and the quantity response rate. This is an indication of a bias in the response. For January and for July (in case of VAT data), these two lines are far apart, meaning that only a relatively small number of companies have reported a large portion of total turnover. This indicates that large companies are overrepresented. At the same time, at the beginning of data collection, the slope of the quantity response rate of January (between 25 and 27 days) and July (between 25 and 30 days) is steeper than that of the unit response rate. In these periods, the number of companies reporting increases only slightly, while the amount of turnover reported increases significantly. This shows that the composition of the response is changing, and this is reflected in the change in the representativity indicators as well. They may change only slightly as the unit response rate increases, or may even decline. Generally, the R-indicators drop as data collection proceeds and there is only a slight increase after 30 days of data collection.

We conclude that the contrast between reporting and nonreporting units increases. The additional response between 25 and 30 days is thus not as representative of the population as the initial response. Hence, waiting longer than 25 days before producing statistics based on VAT data does not make the data more representative.

For the survey data, the difference between the four months is only small. As was mentioned above, in our dataset we only have companies taking part in surveys on a monthly basis. It is only in July that the unit response rate is slightly lower than in other months, which may be due to seasonal effects in Retail trade, such as holidays. Representativity, however, is not different from other months.

In Figures 4 and 5, we present RR-plots of the unit response rate and the R-indicator for VAT and survey data for Retail trade using the VAT model and STS model, respectively.

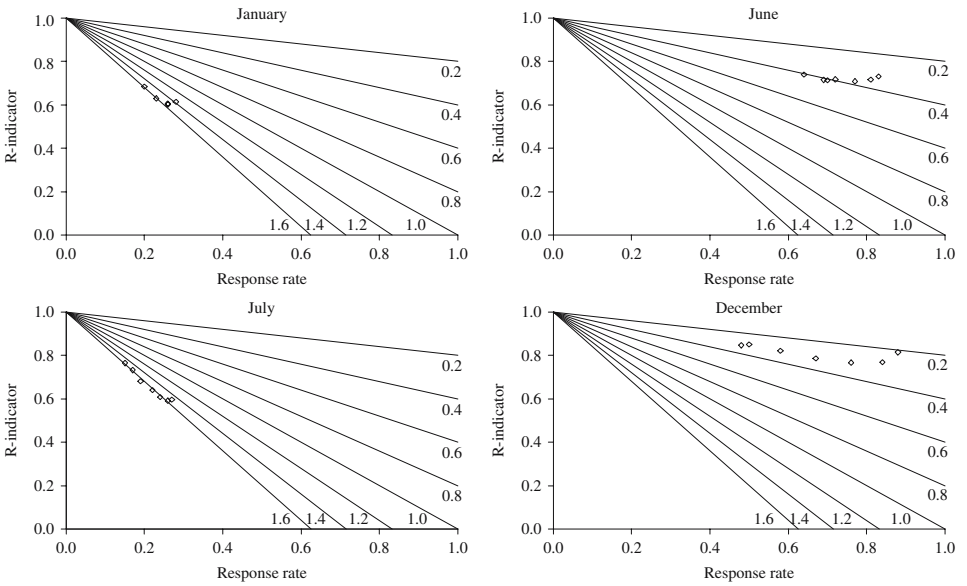


Fig. 4. RR, based on VAT data for Retail trade (VAT model), for four months

The straight lines in the plots represent the maximal bias levels of 0.2, 0.4, . . . 1.6. The plots confirm the previous analyses. The STS survey data show stable patterns over the months. During data collection the maximal bias level remains almost constant. For the VAT data, however, the maximal bias levels vary considerably over the months. Periods with only monthly reporters have a higher maximum bias than other periods.

In summary, the main difference between representativeness of response to surveys and to register holders is the stability over time and during data collection. We conclude that for VAT there is no improvement in the R-indicator and no improvement in the maximal bias when data collection is continued between 25 and 30 days. Since it is crucial for the editing and imputation of business data to start as early as possible, we recommend starting these activities at 25 days after the end of the reference month. For VAT data one must, however, rely much more strongly on nonresponse adjustment methods in months with only monthly reporters and, equally important, be aware that comparability over months is weaker.

4.3. What Groups of Businesses to Target?

In this section we deal with the important question of how we can improve the representativeness of STS and VAT. To answer this question, we first need to identify the subpopulations that impact representativeness most. Second, the data collection design needs to be adapted in such a way that these subpopulations receive more attention.

In the previous sections we restricted ourselves to two auxiliary variables: VAT of the previous reporting year and business size. For the VAT data we used total wages as a proxy for business size. In addressing subpopulations, we now add the type of economic activity, see Table 10, as a variable to the assessment of representativeness.

With the exception of Manufacturing in STS, the R-indicator values decrease only marginally when type of economic activity is added. However, for the Manufacturing

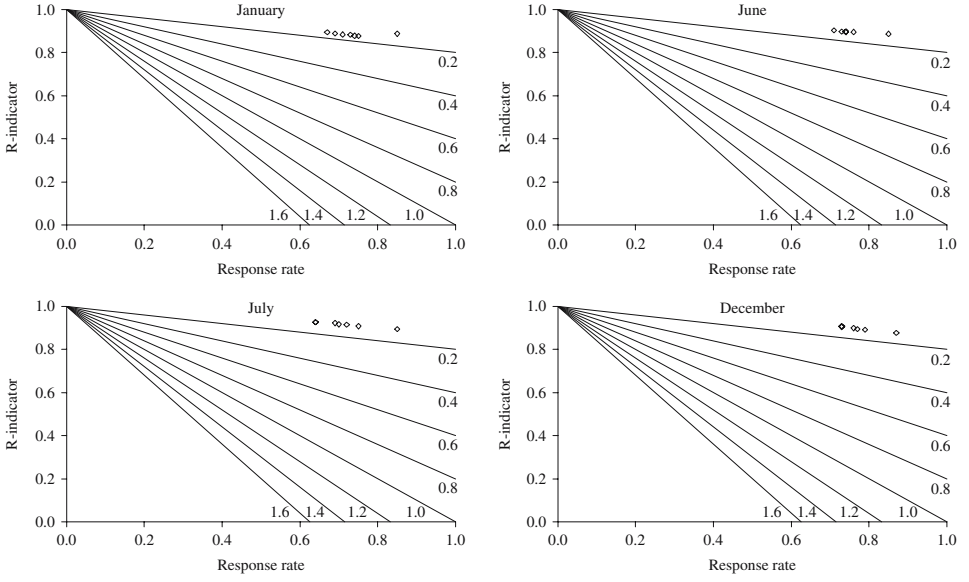


Fig. 5. RR, based on survey data for Retail trade (STS model), for four months

Table 10. Extended models used for the estimation of response propensities

Model	Specifications
Extended VAT model	
Manufacturing	VAT ($t-12$) + wages (t) + NACE-2 digit (1974)
Retail trade	VAT ($t-12$) + wages (t) + NACE-3 digit (1974)
Extended STS model	
Manufacturing	VAT($t-12$) + Business size + NACE subsection (1993)
Retail trade	VAT($t-12$) + Business size + NACE -3 digit (1993)

industry the drop of the STS R-indicator is almost 0.1 and hence the variable provides additional deviation from representative response. Table 11 presents an overview of the variable level partial R-indicators for the auxiliary vector with and without type of economic activity.

From Table 11 we conclude that the extended models do not alter the impact of VAT ($t-12$) and business size or wages (t). We can conclude that type of economic activity plays an almost separate, independent role in representativeness. For this reason, in the remainder of this section, we shall consider the extended models only.

Next, let us explore the dependence of the partial impact of the variables on the data collection month. Table 12 contains the partial R-indicator values for January, June, July and December.

From Table 12 we conclude that the STS representativeness is relatively stable over months and over variables. For VAT, however, the months present quite different pictures. The table also demonstrates that in December, generally, the impact of all variables has reduced considerably. One exception is the impact of wages (t) for Manufacturing, which is strongest in June and comparable to January in December. Furthermore, Table 12 shows that for STS the strongest impact comes from VAT ($t-12$) for Retail trade, and from type of economic activity for Manufacturing. For VAT the strongest impact comes from VAT

Table 11. Unconditional and conditional partial R-indicators without (small) and with (extended) type of economic activity for January after 25 days of data collection

			Unconditional		Conditional	
	Type	Variable	Small	Extended	Small	Extended
STS	Retail	VAT ($t-12$)	0.051	0.051	0.048	0.046
		Business size	0.022	0.022	0.013	0.013
		Activity	–	0.029	–	0.025
	Manufacturing	VAT ($t-12$)	0.023	0.023	0.024	0.023
		Business size	0.028	0.029	0.029	0.028
		Activity		0.038		0.036
VAT	Retail	VAT ($t-12$)	0.152	0.152	0.114	0.117
		Wages (t)	0.110	0.109	0.043	0.051
		Activity	–	0.074	–	0.088
	Manufacturing	VAT ($t-12$)	0.224	0.225	0.213	0.207
		Wages (t)	0.081	0.081	0.039	0.038
		Activity	–	0.057	–	0.028

Table 12. Unconditional and conditional partial R-indicators for the extended model for January, June, July and December

			Unconditional				Conditional			
			Jan	Jun	Jul	Dec	Jan	Jun	Jul	Dec
STS	Retail	VAT ($t-12$)	0.051	0.045	0.034	0.038	0.046	0.039	0.032	0.031
		Business size	0.022	0.026	0.016	0.032	0.012	0.016	0.013	0.023
		Activity	0.029	0.031	0.024	0.028	0.024	0.026	0.022	0.023
	Manufacturing	VAT ($t-12$)	0.023	0.014	0.017	0.022	0.022	0.015	0.019	0.022
		Business size	0.029	0.026	0.030	0.032	0.028	0.025	0.029	0.030
		Activity	0.038	0.039	0.039	0.040	0.036	0.037	0.036	0.039
VAT	Retail	VAT ($t-12$)	0.152	0.129	0.114	0.074	0.117	0.105	0.090	0.061
		Wages (t)	0.109	0.075	0.076	0.045	0.051	0.020	0.036	0.018
		Activity	0.074	0.043	0.053	0.026	0.088	0.033	0.061	0.020
	Manufacturing	VAT ($t-12$)	0.225	0.132	0.159	0.063	0.207	0.084	0.148	0.039
		Wages (t)	0.081	0.173	0.051	0.094	0.038	0.139	0.027	0.078
		Activity	0.057	0.052	0.040	0.032	0.028	0.024	0.016	0.021

($t - 12$) for Retail trade only. For Manufacturing, the same is true with the exception of June and December, where wages (t) is strongest. We added more detail to the evaluation, in line with the proposed guidelines in Subsection 2.4, for variable VAT ($t - 12$) in STS Retail trade, VAT Retail trade and VAT Manufacturing and for variable type of economic activity in STS Manufacturing. For reasons of brevity, we here omit the detailed analysis of variable wages (t) in VAT Manufacturing, but refer the reader to Ouwehand and Schouten (2011).

Table 13 shows that the lack of availability of VAT ($t - 12$) has a negative impact on representativeness in all cases. It also shows that the impact does not decrease after conditioning on the other variables. When VAT of the previous year is not available, then in most cases it concerns newcomers, that is, businesses that launched at some point during the year under consideration. It is not surprising that these businesses are bad responders as they are still starting up and may not have all reporting procedures in order. For VAT, the smaller businesses in terms of revenue also perform worse. This effect was anticipated as small businesses report VAT annually. The values for VAT are smoothed when they are conditioned on wages (t) and type of economic activity; part of the impact of revenue is compensated for by these variables. Surprisingly, for STS Retail trade there is little difference between businesses given that they were active one year ago; the values over the different wage categories are almost constant.

Figure 6 plots the category-level partial R-indicators for type of economic activity in STS Manufacturing. As expected, the unconditional and conditional values are almost identical in an absolute sense: The variable has an orthogonal impact on the other variables. It must be noted here that one group of businesses stands out negatively: the businesses that manufacture food products (NACE 15 and 16). The businesses that manufacture chemicals and chemical products (NACE 23 and 24) perform best.

In sum, with respect to VAT, small businesses and newcomers deserve more attention, while for STS Retail trade it is the newcomers that should be targeted in the data collection. Finally, for STS Manufacturing more effort is needed for specific NACE categories.

Table 13. Categorical partial R-indicators for VAT ($t - 12$) in STS Retail trade, VAT Retail trade and VAT Manufacturing for January

	STS retail		VAT retail		VAT manufacturing	
	Pu	Pc	Pu	Pc	Pu	Pc
$<$2.5k	0.012	0.009	−0.045	0.034	−0.079	0.073
2.5k–10k	0.011	0.010	−0.046	0.034	−0.071	0.079
10k–20k	0.013	0.010	−0.008	0.015	−0.007	0.018
20k–30k	0.016	0.013	0.020	0.019	0.030	0.027
30k–50k	0.005	0.003	0.044	0.031	0.059	0.053
50k–100k	0.012	0.010	0.070	0.047	0.092	0.085
100k–200k	0.005	0.006	0.064	0.044	0.089	0.080
$>$200k	−0.005	0.005	0.081	0.062	0.126	0.108
Not available	−0.041	0.038	−0.034	0.039	−0.047	0.040

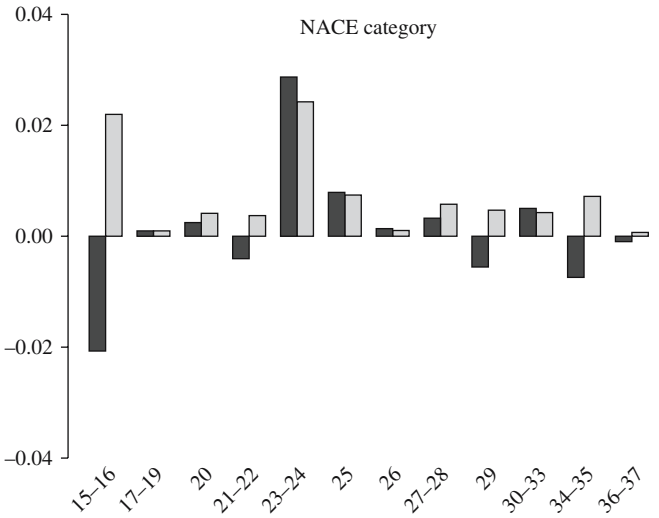


Fig. 6. Categorical partial R-indicators for type of economic activity in STS Manufacturing for January. Black columns represent unconditional and grey columns conditional values. A description of the categories can be found in Appendix B.

4.4. How to Combine the STS Survey and VAT into a Dual Frame Approach?

An important next step is a change of design to obtain higher unit response rates for the underrepresented groups. In this case, three dual-frame approaches can be adopted: VAT-based statistics, STS-based statistics and a combination of STS and VAT. The first and second approaches assume that VAT or STS, respectively, is the primary input to statistics and the other source is used only to supplement types of businesses that are strongly underrepresented. The third approach is a hybrid design, in which both sources are treated as equal. This approach is pragmatic and uses the source per type of business that performs best. For all approaches, however, the explicit targeting of data collection to business units needs to take into account costs and the response burden of data collection. Therefore, representativeness should be optimized subject to constraints on costs and the number of requests for revenue data. Such designs are termed adaptive survey designs (Wagner 2008; Schouten et al. 2013). These designs have begun to emerge in social surveys and may also be applicable to business data collection.

There are three complications to a dual-frame approach that need mentioning first. The first complication is formed by the population frames of the STS survey and the VAT register. Although they are essentially based on the same underlying frame that is maintained by the Dutch Chamber of Commerce, the frames used by Statistics Netherlands and the Tax Authorities are different. This difference applies mostly to larger businesses, where the two offices use different criteria to cluster economic activity. These criteria are logical from their respective operations and perspectives, but a nuisance to any method that combines the two frames. For smaller businesses there is a one-to-one correspondence for virtually all business units, but for larger business units there could be *n* to *1*, *1* to *m* or even *n* to *m* correspondences. As a result, linkage of the two frames cannot be performed without dividing or combining business units. Clearly, when a dual frame approach is applied, complex decision rules are needed for the larger businesses.

The second complication is a conceptual difference in the STS variables themselves. The definition of total revenue and its components is not fully harmonised across the two sources, again for the same operational reasons. This difference is more severe again for the larger businesses. The third complication lies in the classification of businesses. The survey and VAT population frames have different sets of additional, auxiliary variables, as mentioned above. These variables are used to classify businesses. Since there is no one-to-one correspondence between the two frames, transformation rules need to be applied in order to link auxiliary variables from one frame to the other. In summary, it can be concluded that any dual frame approach will need to find methodological solutions for the larger businesses.

We first look at STS Retail trade. These businesses are mostly smaller and both frames have a strong correspondence. Here, it is anticipated that the above-mentioned complications play only a minor role. In Subsection 4.3, we concluded that the smallest businesses are underrepresented in the VAT and that both STS survey and VAT have an underrepresentation of newcomers. Hence, for newcomers no approach will be satisfactory and there is no suitable hybrid approach. The only solution is the development of special invitation letters and instructions and guidance to raise response rates of newcomers in the STS survey. For the small businesses, the STS survey can be conducted to supplement or replace VAT. In STS-based statistics, there is no reason to employ VAT. In VAT-based statistics, the STS survey can be conducted to supplement response for small businesses and, if successful nonresponse reduction methods can be developed, also for newcomers. A hybrid approach would employ STS for small businesses and newcomers and VAT for all other businesses.

For STS Manufacturing, the picture is very different as it consists of larger businesses. Here, frame differences and conceptual differences may complicate a dual-frame approach. The conceptual differences imply that the three approaches are likely to cause method effects. We concluded in Subsection 4.3 that specific NACE categories have a lower representation in the STS survey, while for VAT no specific types of businesses are underrepresented. Hence, VAT-based statistics and a hybrid approach do not employ STS survey data and coincide, but STS-based statistics may employ VAT for these NACE categories. Because of the method effects, STS-based statistics should use a stable design in order to maintain comparability in time.

When adopting a dual frame approach, the focus is on design. Even when more effort is made to raise the response rates of underrepresented businesses, it is likely that some businesses will have lower response rates than others. Apart from a change of design, one may therefore in addition use the VAT records of the previous reporting period to adjust for nonresponse in either the VAT or the STS survey of the current reporting period. Such adjustment is termed nowcasting in economic studies. In nowcasting, the frame differences again pose problems but conceptual differences are not an issue; VAT of the previous reporting period is merely used as a predictor.

5. Discussion

This article compared the unit response rate, quantity response rate and representativity of the STS survey and VAT data over several months and during data collection. Both data

sources can be used to produce monthly short-term business statistics. However, Statistics Netherlands intends to replace part of its survey efforts with data from administrative registers. To this end, the available data should, of course, lead to accurate statistics. An important data quality aspect that is assumed to be a good predictor of accuracy is the representativity of the data. In this article, we therefore compared the two data sources with respect to representativity, as measured by the R-indicator.

In our comparison, we focused completely on nonresponse error and ignored measurement and sampling errors. Clearly, the STS survey response has a bigger sampling error than the VAT data as the Tax Authorities records are a full enumeration of enterprises in the Netherlands. Measurement errors were conjectured to play an important role as well. However, there is little empirical evidence in favour of survey or administrative data. A complete comparison of both data sources should also account for these errors.

In our comparison, we answered three research questions. They regard the representativeness of survey and registry data per industry, per month, and through time, but also regard the enterprise groups that need to be targeted to improve representativeness of response. The main question underlying these investigations is the more general issue of whether STS statistics should be based on a dual-frame approach using both register and survey data.

The representativeness of survey data and register data is quite different over the months. The results indicate that the unit response rate for both Retail trade and Manufacturing is substantially lower for VAT than for STS, due to the nature of the collection method. However, the R-indicator for VAT can still be relatively high even in months of low response rates. This shows that the unit response rate alone is not sufficient for assessing data quality.

During data collection, and more specifically between 25 and 30 days after the end of the reference month, the unit response rates increased, as could be expected. Representativity, however, is not in line with the unit response rate patterns: It may change only slightly as the unit response rate increases, or it may even decline. From this we conclude that the contrast between reporting and nonreporting units may increase as data reporting proceeds. Hence, waiting longer before producing statistics based on VAT data does not make the data more representative.

The findings for the R-indicators are in line with the combined patterns of unit response rates and quantity response rates. Whenever quantity response rates showed a different increase from unit response rates, the R-indicator also changed. The strong feature of the R-indicator is that it quantifies over- and underrepresentation, it allows for a simultaneous assessment on multiple auxiliary variables and it can be estimated without bias after and during data collection. The quantity response rates in this article were computed retrospectively, but could normally not be estimated during or shortly after data collection without bias.

In summary, the main difference between representativeness of response to surveys and to register holders is the stability over time and during data collection. The survey data are more stable in time and during data collection.

Representativity patterns may differ from subpopulation to subpopulation. We found that in VAT small businesses and newcomers deserve more attention, while for STS Retail

trade it is the newcomers that should be targeted in the data collection. Finally, for STS Manufacturing more effort is needed for specific NACE categories.

Future research is required. Our study had some limitations with respect to the data set used. It used a specific set of auxiliary variables, only focused on a period of two years and on two industries. The auxiliary variables were not the same for the two data sources (caused by the difference in population frames), so the absolute values of the R-indicator could not be compared. It is important, therefore, that our results are replicated on other years and industries and in other countries.

Appendix A: Approximations to the Expected Quantity Response Rate

We restrict ourselves to a first-order Taylor approximation of (7) and (10). For the expected value of a ratio of two random variables, this leads to the ratio of the expected values of the two random variables. This is a crude approximation, but we merely want to show how the various indicators relate to each other for large sample sizes. The STS survey sample sizes are indeed large and VAT is a full enumeration of the population.

Assuming that the population is large and $(N-1)/N \approx 1$, for the numerator and denominator of (7), respectively, we arrive at

$$E\sum_{i=1}^{n} d_i r_i y_i = \sum_{i=1}^{N} \rho_i y_i = N \operatorname{cov}(y, \rho_Y) + N\rho_U \bar{y}_N, \tag{A.1}$$

$$E\sum_{i=1}^{n} d_i y_i = \sum_{i=1}^{N} y_i = N\bar{y}_N. \tag{A.2}$$

For the denominator of (10), we first rewrite as

$$\sum_{h=1}^{H} N_h \frac{\sum_{i\in h} d_i r_i y_i}{\sum_{i\in h} d_i r_i} = \sum_{h=1}^{H} N_h \frac{\sum_{i\in h} d_i r_i (y_i - \bar{y}_h)}{\sum_{i\in h} d_i r_i} + \sum_{h=1}^{H} N_h \bar{y}_h, \tag{A.3}$$

and define residual $\varepsilon_i = y_i - \bar{y}_h$ for unit i. The expectation of a weighted stratum response mean can be approximated (again using a first-order Taylor expansion) by

$$E\frac{\sum_{i\in h} d_i r_i \varepsilon_i}{\sum_{i\in h} d_i r_i} = \frac{\operatorname{cov}_h(\varepsilon, \rho_Y)}{\rho_{U,h}}, \tag{A.4}$$

since the stratum residual means $\bar{\varepsilon}_h$ are equal to zero. In (A.4) $\rho_{U,h}$ is the unit response rate in stratum h and $\operatorname{cov}_h(y, \rho_Y)$ is the covariance between response propensities and residuals within stratum h.

Using (A.3) and (A.4) the expectation of the denominator of (10) is approximated as

$$E\left(\sum_{h=1}^{H} N_h \frac{\sum_{i\in h} d_i r_i \varepsilon_i}{\sum_{i\in h} d_i r_i} + \sum_{h=1}^{H} N_h \bar{y}_h\right) = \sum_{h=1}^{H} N_h \frac{\operatorname{cov}_h(\varepsilon, \rho_Y)}{\rho_{U,h}} + N\bar{y}_N. \tag{A.5}$$

Appendix B: NACE categories

15–16 : manufacture of food products
17–19 : manufacture of apparel, leather, leather products, and footwear
20 : manufacture of wood, and wood and cork products, except furniture
21–22 : manufacture of paper and paper products, and printing and reproduction of recorded media
23–24 : manufacture of chemicals and chemical products
25 : manufacture of rubber and plastic products
26 : manufacture of other nonmetallic mineral products
27–28 : manufacture of basic metals and manufacture of fabricated metal products, except machinery and equipment
29 : manufacture of machinery and equipment
30–33 : manufacture of computers, electronic and optical products
34–35 : manufacture of transport equipment
36–37 : manufacture of furniture

6. References

Kruskal, W. and F. Mosteller. 1979a. "Representative Sampling I: Non-Scientific Literature." *International Statistical Review* 47: 13–24. DOI: http://dx.doi.org/10.2307/1403202.

Kruskal, W. and F. Mosteller. 1979b. "Representative Sampling II: Scientific Literature Excluding Statistics." *International Statistical Review* 47: 111–123. DOI: http://dx.doi.org/10.2307/1402564.

Kruskal, W. and F. Mosteller. 1979c. "Representative Sampling III: Current Statistical Literature." *International Statistical Review* 47: 245–265. DOI: http://dx.doi.org/10.2307/1402647.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics*. New York: John Wiley & Sons.

Ouwehand, P. and B. Schouten. 2011. Representativity of VAT and Survey Data for Short Term Business Statistics. CBS discussion paper 201106. Available at: http://www.cbs.nl/nl-NL/menu/methoden/onderzoek-methoden/discussionpapers/archief (accessed date).

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Survey Response." *Journal of Official Statistics* 27: 231–253.

Schouten, B., J. Bethlehem, K. Beulens, Ø. Kleven, G. Loosveldt, K. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring and Improving Representativeness of Survey Response Through R-indicators and Partial R-indicators." *International Statistical Review* 80: 382–399. DOI: http://dx.doi.org/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology* 39: 29–58.

Shlomo, N., C. Skinner, and B. Schouten. 2012. "Estimation of an Indicator of the Representativeness of Survey Response." *Journal of Statistical Planning and Inference* 142: 201–211. DOI: http://dx.doi.org/10.1016/j.jspi.2011.07.008.

Slootbeek, M. and K. van Bemmel. 2010. *Onderzoek Naar Overstappers/Blijvers*, Internal report, Statistics Netherlands.

Thompson, K.J. and B.E. Oliver. 2012. "Response Rates in Business Surveys: Going Beyond the Usual Performance Measure." *Journal of Official Statistics* 28: 221–237.

Van Delden, A. and F. W. L. Aelen. 2008. Redesigning the Chain of Economic Statistics at Statistics Netherlands: STS Statistics as an Example. IAOS Conference 'Reshaping Official Statistics', 14–16 October 2008, Shanghai. Available at: http://www.stats.gov.cn/english/specialtopics/IAOSconference/200811/w020130912531695358471.doc (accessed September 30, 2014).

Vlag, P. and D. van den Bergen. 2010. The Use of VAT for Short Term Statistics: Some Quality Aspects, Statistics Netherlands. In Proceedings ESSnet seminar on administrative data – Rome, 18–19 March 2010. Available at: http://essnet.admindata.eu/Document/Getfile?objectid=5147 (accessed September 30, 2014).

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*. Ann Arbor, MI: University of Michigan, Ph.D. thesis.

Received November 2012
Revised April 2014
Accepted June 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 651–674, http://dx.doi.org/10.2478/JOS-2014-0042

Does the Length of Fielding Period Matter? Examining Response Scores of Early Versus Late Responders

Richard Sigman[1], *Taylor Lewis*[2], *Naomi Dyer Yount*[1], *and Kimya Lee*[2]

This article discusses the potential effects of a shortened fielding period on an employee survey's item and index scores and respondent demographics. Using data from the U.S. Office of Personnel Management's 2011 Federal Employee Viewpoint Survey, we investigate whether early responding employees differ from later responding employees. Specifically, we examine differences in item and index scores related to employee engagement and global satisfaction. Our findings show that early responders tend to be less positive, even after adjusting their weights for nonresponse. Agencies vary in their prevalence of late responders, and score differences become magnified as this proportion increases. We also examine the extent to which early versus late responders differ on demographic characteristics such as grade level, supervisory status, gender, tenure with agency, and intention to leave, noting that nonminorities and females are the two demographic characteristics most associated with responding early.

Key words: FEVS; employee surveys; employee satisfaction; employee engagement; fielding period.

1. Introduction

Employee surveys are used by government and private establishments worldwide (Kraut 1996). Many organizations use employee surveys as a cost-effective way to gauge the extent to which employees' beliefs and perceptions are in line with the organization's mission and goals. These surveys can convey employee morale, and they can also provide direct, actionable information about employee satisfaction and engagement, intent to leave, and training needs. A distinct advantage of employee surveys is that they may alert management to budding problems before they become serious and prevent the loss of an organization's most important asset, their employees.

However, along with these advantages, there are also unique challenges associated with employee surveys. Since employee surveys are voluntary, nonresponse and the effect it can have on estimates is always a concern (Rogelberg and Stanton 2007). Indeed, response rates to employee surveys have declined over the past few decades (Baruch and Holtom 2008), as they have for surveys in general (de Leeuw and de Heer 2002). A longer period

[1] Westat, 1600 Research Blvd, Rockville, MD 20850 U.S.A. Email: RichardSigman@westat.com and NaomiYount@westat.com
[2] U.S. Office of Personnel Management, 1900 E Street, NW, Washington, DC 20415, U.S.A. Email: Taylor.Lewis@opm.gov and Kimya.Lee@opm.gov
Acknowledgments: The opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect those of the U.S. Office of Personnel Management.

of data collection may boost response rates, but comes at the costs of less timely data and higher administrative costs (e.g., following up with nonrespondents, staffing survey support centers). Faced with the unfortunate reality of stagnant or reduced data collection budgets, many survey managers find themselves questioning whether the fielding period could be shortened without adversely affecting the quality of data produced.

A natural way to evaluate a shortened fielding period is to compare the response patterns and demographic profiles for some definition of "early" versus "late" respondents. Studies with this goal in mind have a long and rich history in the survey research literature. Some of the many examples include Baur (1947), Newman (1962), Mayer and Pratt (1966), Gannon et al. (1971), Filion (1975), and Bates and Creighton (2000). In terms of demographics, these studies have found that early respondents tend to be older (Filion 1975), nonminority (Mayer and Pratt 1966), and female (Gannon et al. 1971), and of a higher education level or socioeconomic status (Newman 1962; Mayer and Pratt 1966).

With respect to attitudinal measures captured as part of a self-administered employee survey, the literature is much less robust, but a few examples are Pace (1939), Schwirian and Blaine (1966), Ellis et al. (1970), Green (1991), and Borg and Tuten (2003). Arguably the most pervading theme is that few noteworthy differences are found. For instance, Pace (1939), Green (1991), and Borg and Tuten (2003) essentially concluded there were no significant differences for questions asking about various dimensions of job satisfaction, whereas Schwirian and Blaine (1966) found early respondents tended to be more satisfied, although differences were slight.

As is generally the case with establishment surveys, a feature of the employee survey response timing studies identified above is that the target populations are often highly specialized. For example, Pace (1939) studied recent college graduates, Schwirian and Blaine (1966) studied members of the United Automobile Workers union, Green (1991) studied teachers, and Borg and Tuten (2003) studied employees of two German advanced technology companies. It is unclear whether these findings generalize to other employee populations. To the best of our knowledge, there has never been any research aimed specifically at our target population of interest, employees of the United States federal government. This article offers one such contribution, as we examine data from the U.S. Office of Personnel Management's (OPM) 2011 Federal Employee Viewpoint Survey (FEVS).

Section 2 provides some background about the FEVS. The remainder of the article utilizes 2011 FEVS data to determine the effects of reducing the length of the FEVS data collection period to two weeks. Section 3 provides a comparison of the demographic characteristics of early responders versus late responders. Section 4 compares early-responder estimates with all-responder estimates. Section 5 contains our conclusions.

2. Background About the FEVS

OPM conducts the FEVS to collect data on U.S. federal government employees' opinions of whether, and to what extent, conditions that characterize successful organizations are present in their agency, focusing on critical drivers of employee satisfaction, engagement, commitment, and retention. Results from the survey enable OPM and agency managers to take positive steps that have a direct effect on the workplace, such as developing policies

and action plans that improve agency performance. The 95-item questionnaire consists of eight topic areas: personal work experiences, work unit, agency, supervisor/team leader, leadership, satisfaction, work/life, and demographics. Demographic items include location of employment (headquarters vs. field), supervisory status, gender, ethnicity/race, age, grade, federal employment tenure, and agency tenure. In addition, the survey includes items capturing intent to leave the organization and plans to retire. OPM administered the FEVS for the first time in 2002 and repeated biennially through 2010, when it began to be administered annually.

The sample frame is constructed from a personnel database maintained by OPM that contains information on over 2,000,000 federal civilian employees. For the 2011 FEVS, the total sample size was 560,084, consisting of full-time, permanent employees from 83 agencies on board as of September 2010. A total of 1,114 strata were formed by the cross-classification of (1) organizational subgroup (e.g., bureaus or offices within a larger agency) and (2) supervisory status (nonsupervisors, supervisors, and executives). Note that some degree of stratum collapsing was performed (e.g., if the executive stratum within a given organizational subgroup contained only a few individuals, it was collapsed with the supervisor stratum) and not all three supervisory strata are present within all organizational subgroups. Stratum sample sizes were initially calculated to achieve a $+/-5\%$ margin of error within each, accounting for nonresponse, though some agencies requested a full census of their workforce.

The FEVS is primarily a web-based, self-administered survey, but a limited number of people (less than 5,000) without Internet access are provided with a paper version of the instrument. Electronically surveyed individuals are sent an initial email invitation to participate that contains a hyperlink to the survey site with a unique respondent key embedded. Time stamps for each response are recorded, and weekly reminder emails are sent only to those who have not completed their survey. The response rate for FEVS 2011 was 48%, calculated according to the RR3 formula defined by the American Association for Public Opinion Research (AAPOR 2009).

To mitigate the potential biases attributable to unequal probabilities of selection across strata and uneven patterns of nonresponse, a three-stage procedure is implemented to develop and append a weight to each respondent's survey record (Kalton and Flores-Cervantes 2003). First, a base weight equaling the reciprocal of the probability of selection is calculated for all sampled employees. Second, weighting cells are formed independently for each agency, within which base weights of nonrespondents are shifted to respondents. The sample frame variables used to form these weighting cells include supervisory status, sex, minority status, age group, length of service as a federal employee, and workplace location (headquarters vs. field office). The free, SAS-callable %search macro developed by researchers at the University of Michigan (http://www.isr.umich.edu/src/smp/search/) is employed to partition each agency's sample into cells, with the goal of differentiating the response probabilities as much as possible across cells. The %search macro is based on techniques discussed in Sonquist et al. (1974). Third, respondent weights are raked such that the raked weights aggregate to frame totals for the sampling strata and for raking cells defined by agency, gender, and minority status.

Aside from demographics, most survey items are attitudinal, using five-point response scales ranging from "Strongly Agree" to "Strongly Disagree," sometimes with a "Do Not

Know" or "No Basis to Judge" option provided. A common calculation for summarizing FEVS responses is to compute an item's *percent positive* estimate. This is found by dichotomizing the response scale into positive responses and nonpositive responses (e.g., a positive response would consist of those answering "Strongly Agree" or "Agree"). Nonsubstantive answers such as "Do Not Know" are treated as missing. The percent positive estimate is simply the weighted portion of positive responses relative to all substantive responses. Jacoby and Matell (1971) have found that converting multi-level Likert-item data to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity.

Percent positive estimates for certain thematically-linked survey items are averaged to form indices. There are six such indices reported at the governmentwide and agency levels: four Human Capital Assessment and Accountability Framework (HCAAF) Indices, an Employee Engagement Index, and a Global Satisfaction Index. This article focuses on the last two, which were of particular interest because they were first developed and reported following the 2011 administration of the FEVS (OPM 2011). The 15-item Employee Engagement Index is comprised of three subindices: Leaders Lead, Supervisors, and Intrinsic Work Experience. Each of the subindices is composed of five items. Employee engagement can be defined in numerous ways (Macey and Schneider 2008). For the purposes of this study, employee engagement is defined as ". . . passion and commitment – the willingness to invest oneself and expend one's discretionary effort to help the employer succeed" (Erickson 2005, 14). While the FEVS does not directly measure employee engagement, the 15 items making up the index are items representing work conditions or perceptions that would lead one to be engaged. The Global Satisfaction Index is composed of four items, addressing employees' satisfaction with their job, pay, organization, plus their willingness to recommend their organization as a good place to work.

3. Comparison of Early and Late Responders

In the 2011 FEVS, agencies had staggered fielding periods from April to May 2011 ranging from three to nine weeks in the field. For the purposes of this article, we define an *early* responder as one who completed the survey within the first two weeks after the initial email invitation to participate was sent. We also considered several other definitions of an early responder (e.g., first half of the field period, first month). However, after observing that the agency-specific percentages of early responders using these alternative definitions often constituted nearly 100 percent of the final set of responders, whether early or late, we felt the two-week threshold allowed for more meaningful comparisons. We also felt this offered a degree of standardization, considering how agencies were given some flexibility in setting their survey launch and close dates. In fact, about one in four agencies evaluated in this study had a fielding period lasting one month or less, which partially explains our initial point. Lastly, we note that because their response times were not precisely captured, in this study we excluded data for the small subset of paper survey responders.

Many of the 83 agencies participating in FEVS 2011 were small and did not include demographic items on their survey. In order to only compare estimates with stable standard errors, and to be able to compare demographic profiles, we restricted our analysis to the 30 agencies for which at least 1,000 responses were obtained and demographic questions

were included. Appendix A lists these agencies with a few other distributional statistics regarding their respective fielding periods (e.g., response rates, length in the field). In this article, estimates and figures labeled as *governmentwide* refer only to these 30 agencies, making up 98% of the target population and 253,285 of the total 266,376 electronically completed surveys for all participating agencies, or 95% of all responses. A completed survey is defined as an individual who answered at least one-quarter of the 84 core nondemographic survey items. As can be gathered from Appendix A, approximately 59% of the 253,285 respondents in this study completed the survey in the first two weeks, but the figure varies widely by agency: ranging from 43% early respondents to 86%.

3.1. Comparing Demographic Profiles of Early and Late Responders

Table 1 presents a governmentwide comparison of certain unweighted demographic distributions of early and late respondents. The largest difference found in Table 1 is how minorities are much more likely to respond after the first two weeks. While minorities make up 31% of early respondents, they constitute 39% of late respondents. Females are more likely to respond early, although the discrepancy is slightly smaller, with a difference of 3.2 percentage points. Responders 60 years of age or older are also more likely to

Table 1. Governmentwide demographic distributions for early and late responders

Demographic	Value	All	Early	Late	Difference between early and late
Age	<40	21.1	20.9	21.3	−0.4
	40–59	65.9	65.5	66.3	−0.8
	60+	13.1	13.5	12.4	1.1
Agency tenure	<5 years	30.8	31.6	29.5	2.1
	6–20 years	41.1	40.7	41.7	−1
	20+ years	28.1	27.7	28.8	−1.1
Intent to leave	Stay	71.0	69.9	72.7	−2.8
	Retire	6.4	6.5	6.1	0.4
	Leave	22.6	23.5	21.2	2.3
Minority status	Nonminority	65.7	69.0	61.0	8.0
	Minority	34.3	31.0	39.0	−8.0
Pay category	Federal wage system	3.5	3.3	3.7	−0.4
	GS 1–6	5.0	5.4	4.5	0.9
	GS 7–12	39.1	40.1	37.7	2.4
	GS 13–15	44.7	43.5	46.4	−2.9
	SES, SL, and other	7.8	7.7	7.8	−0.1
Gender	Male	52.5	51.2	54.4	−3.2
	Female	47.5	48.8	45.6	3.2
Supervisory status	Nonsupervisor/ Team leader	72.7	73.4	71.7	1.7
	Supervisor/manager	25.4	24.9	26.1	−1.2
	executive	1.9	1.7	2.2	−0.5
Location	Headquarters	41.7	40.8	43.1	−2.3
	Field	58.3	59.2	56.9	2.3

respond early than late. These findings seem to agree with the literature (Mayer and Pratt 1966; Gannon et al. 1971; Filion 1975), but this is not true of all demographics investigated. For example, Newman (1962) found respondents of higher socioeconomic status tended to respond earlier, while we find somewhat conflicting results. Those in Grades 13–15 within the General Schedule (GS) pay scale were more likely to be late responders than early responders. Furthermore, those within the GS 7–12 ranges were more likely to be early responders than late responders. (A higher grade with respect to the GS pay scale for U.S. federal government employees is associated with higher pay. For more information, see www.opm.gov/oca.) Two other differences worth mentioning are that employees with lower intentions to leave their current position are more likely to be late responders and that employees working at the agency's headquarters are more likely to be late responders as compared to employees working in a field office.

These demographic distributions were also examined for each agency. For brevity, none of those tables are given in this article, but several of the general findings noted above prevailed. For example, minority respondents were more likely to respond after the first two weeks in all 30 agencies. Those intending to leave were more likely to respond within two weeks in all but four agencies. The gender disparity was not found to be universal across agencies, however, as we found females were more likely to respond early in 17 out of 30 agencies. The other demographic comparisons were also mixed on an agency–by-agency basis.

4. Comparison of Early Responder and All-Responder Estimates

The previous section compared the demographic characteristics of employees responding before or after the first two weeks of data collection. This section discusses the effects on the survey estimates by reducing the FEVS data collection period to two weeks. In particular, we compare the survey estimates that would be published if the FEVS data collection period were shortened to two weeks versus the estimates based on the full data collection period. We call the differences between these estimates the *early-minus-all estimate differences*. Positive differences signify that the early responders are more positive than all responders, while negative differences signify the opposite, that early responders were more negative than all responders. Subsection 4.1 expresses the early-minus-all estimate difference in terms of the prevalence of late responders and the survey characteristics of employees responding before and after the first weeks of data collection. Subsections 4.2 and 4.3 use 2011 FEVS data to assess early-minus-all estimate differences at the governmentwide and agency levels, respectively. Subsections 4.4 and 4.5 further explore the agency-level results by examining relationships between early-minus-all estimate differences and agency-level characteristics and between early-minus-all differences and the levels of early-responder and all-responder estimates.

4.1. Differences in Estimates Due to Reducing the Fielding Period

Survey nonresponse can be modeled deterministically or stochastically. A deterministic model would assume that a population of individuals consists of N_{early} individuals who always respond early, N_{late} individuals who always respond late, and N_{never} individuals who never respond. We will refer to estimates obtained when the survey is *not* reduced in

length as *all-responder estimates* and to estimates obtained when the data collection period is reduced in length as *early-responder estimates*. Further assume that the estimates of interest are estimates of population means or proportions. Under a deterministic model for nonresponse, the early-minus-all estimate differences are estimates of

$$E^{(i)} = \bar{X}^{(i)}_{early} - \bar{X}^{(i)}_{all},$$

where

$\bar{X}^{(i)}_{early}$ = population mean for survey item i for *early* responders, and
$\bar{X}^{(i)}_{all}$ = population mean for survey item i for *all* responders.

Since

$$\bar{X}^{(i)}_{all} = \frac{N_{early}\bar{X}^{(i)}_{early} + N_{late}\bar{X}^{(i)}_{late}}{N_{early} + N_{late}},$$

where $\bar{X}^{(i)}_{late}$ is the population mean for survey item i for all responders that are not early responders, it follows that

$$E = \bar{X}^{(i)}_{early} - \frac{N_{early}\bar{X}^{(i)}_{early} + N_{late}\bar{X}^{(i)}_{late}}{N_{early} + N_{late}} = r_{late}(\bar{X}^{(i)}_{early} - \bar{X}^{(i)}_{late}), \quad (1)$$

where

$$r_{late} = \frac{N_{late}}{N_{early} + N_{late}}$$

is the prevalence of late responders among all responders, that is, the expected proportion of all responders that are not early responders.

4.2. *Governmentwide Early-Minus-All Estimate Differences*

For all 30 agencies, the early-responder data were used to compute early-responder weights using the same procedures used in the 2011 FEVS all-respondent dataset. These weights were then used to calculate the percent positive estimates for the early responders. The early-minus-all estimate differences for the indices and sub-indices were computed by subtracting the index (or subindex) for all responders from the corresponding index (or subindex) for early responders. Both the early-responder weights and the all-responder weights contain adjustments for nonresponse calculated within nonresponse-adjustment cells defined by sampling-frame variables. This eliminates ignorable nonresponse biases (Little and Rubin 2002) associated with variables for which the missing-at-random assumption holds within the defined nonresponse-adjustment cells but it does not eliminate nonignorable nonresponse biases or additional ignorable nonresponse biases associated with variables not on the sampling frame. Hence, the early-minus-all estimate differences estimate not only the quantity defined in terms of population parameters by Equation (1) but also include differences in nonignorable nonresponse biases between early and late respondents.

Table 2 contains the governmentwide early-responder estimates, the all-responder estimates, and the early-minus-all estimate differences for the Employee Engagement and

Table 2. Governmentwide early-minus-all estimate differences for (sub)indices and associated percent positive items

(Sub-)indices and items	Early-responder estimates (%)		All-responder estimates (%)		Early-minus-all difference (%)	
	Indices	Items	Indices	Items	Indices	Items
Employee Engagement	**65.24**		**66.63**		**− 1.39**	**− 1.39**
leaders lead (LE)	**54.57**		**56.34**		**− 1.76**	**− 1.76**
Q53. In my organization, leaders generate high levels of motivation and commitment in the workforce.		43.31		45.27		− 1.96
Q54. My organization's leaders maintain high standards of honesty and integrity.		55.51		57.16		− 1.65
Q56. Managers communicate the goals and priorities of the organization.		63.20		64.65		− 1.45
Q60. Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor/team leader?		55.88		57.76		− 1.87
Q61. I have a high level of respect for my organization's senior leaders.		54.96		56.85		− 1.89
Intrinsic work experience (IN)	**70.46**		**71.75**		**− 1.29**	
Q03. I feel encouraged to come up with new and better ways of doing things.		57.88		59.56		− 1.68
Q04. My work gives me a feeling of personal accomplishment.		72.50		73.80		− 1.30
Q06. I know what is expected of me on the job.		79.14		80.19		− 1.04
Q11. My talents are used well in the workplace.		58.63		60.54		− 1.90
Q12. I know how my work relates to the agency's goals and priorities.		84.16		84.66		− 0.51

Table 2. Continued

(Sub-)indices and items	Early-responder estimates (%)		All-responder estimates (%)		Early-minus-all difference (%)	
	Indices	Items	Indices	Items	Indices	Items
Supervisors (SP)	**70.68**		**71.81**		**− 1.13**	
Q47. Supervisors/team leaders in my work unit support employee development.		65.81		67.01		−1.20
Q48. My supervisor/team leader listens to what I have to say.		74.38		75.23		−0.85
Q49. My supervisor/team leader treats me with respect.		79.44		80.16		−0.72
Q51. I have trust and confidence in my supervisor.		65.92		67.25		−1.33
Q52. Overall, how good a job do you feel is being done by your immediate supervisor/team leader?		67.84		69.40		−1.56
Global satisfaction (GL)	**64.90**		**66.21**		**− 1.31**	
Q40. I recommend my organization as a good place to work.		67.44		69.14		−1.70
Q69. Considering everything, how satisfied are you with your job?		69.20		70.74		−1.54
Q70. Considering everything, how satisfied are you with your pay?		62.07		62.42		−0.36
Q71. Considering everything, how satisfied are you with your organization?		60.89		62.52		−1.64
Minimum	54.57	43.31	56.34	45.27	−1.76	−1.96
Median	64.90	64.50	66.21	65.83	−1.31	−1.54
Maximum	70.68	84.16	71.81	84.66	−1.13	−0.36

Global Satisfaction indices and the associated 19 items. All of the estimate differences were negative, ranging from −1.96 percent to −0.36 percent, indicating that overall the early responders were more negative than all responders. The median governmentwide early-minus-all estimate difference across the 19 items was −1.54 percent. All of the index differences were also negative, ranging from −1.76 percent to −1.13 percent.

Across the 19 items and five (sub)indices, smaller percent positive values were associated with more negative early-minus-all estimate differences. For the 19 items, the Pearson correlation between early-minus-all estimate difference and the early-responder estimates was 0.72; between early-minus-all estimate differences and all-responder estimates it was 0.70. For the five (sub)indices, the Pearson correlation with early-minus-all estimate differences was 0.96 for (sub)indices based on early-responder estimates and was 0.95 for those based on all-responder estimates. Because governmentwide early-minus-all differences are negative, the positive correlation between early estimates and early-minus-all differences indicates that across items and (sub)indices, as percent positive values get larger the difference between early estimates and all-responder estimates moves closer to zero. In the next section we investigate these relationships across individual agencies.

4.3. *Agency-Level Early-Minus-All Estimate Differences*

Table 3a displays summary statistics for agency-level early-minus-all estimate differences for the five (sub)indices. The minimum and median-level early-minus-all estimate differences across agencies are negative, whereas the maximum early-minus-all estimate difference across agencies is positive. The median agency-level early-minus-all estimate difference across agencies ranges from −2.00 percent (for Leaders Lead) to −1.12 (for Supervisors). The relationship between early-responder estimates and all-responder estimates found at the governmentwide level was also seen at the agency level; however, as shown in Table 3a, there are some agencies that did not exhibit this pattern, rather early responders were more positive than late responders for some agencies.

The 30 box plots in Figure 1 help to uncover why some early-minus-all differences are positive and why others are negative. The box plots show the distributions of early-minus-all estimate differences across the two indices and three subindices for each agency. Though not shown, for each agency we also produced a box plot indicating the distribution of early-minus-all estimate differences across the 19 percent positive items. Within each agency, the range of the estimated early-minus-all difference for the indices and

Table 3a. Agency-level early-minus-all estimate differences for five (sub)indices

	Agency-level early-minus-all difference (%)				
(Sub-)index	Minimum	Mean	Median	Maximum	Skewness
Employee engagement	−3.93	−1.39	−1.39	0.19	−0.47
Leaders lead	−4.40	−1.80	−2.00	0.42	0.02
Intrinsic work experiences	−3.92	−1.21	−1.13	0.40	−0.80
Supervisors	−3.49	−1.16	−1.12	1.10	−0.30
Global satisfaction	−5.92	−1.49	−1.28	0.30	−1.52

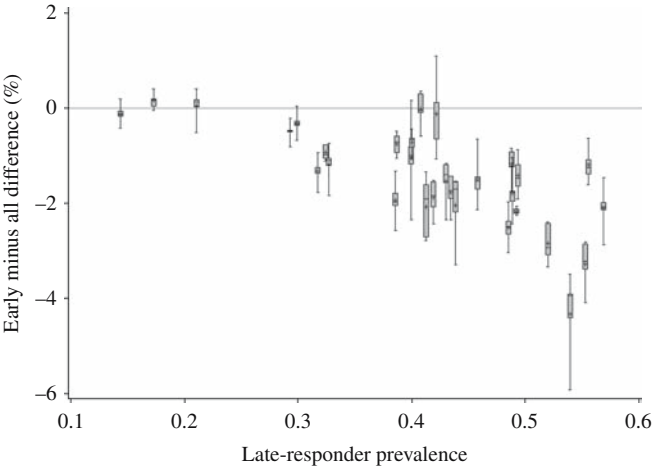


Fig. 1. "Skeletal" box plots by agency (in increasing order of agency's prevalence of late responders) indicating distributions of estimated early-minus-all estimate differences across percent positive indices. The end of the lower whisker is the minimum, and the end of the upper whisker is the maximum.

subindices was much smaller than the range of the early-minus-all estimate differences for the associated percent positive items. For both (sub)indices and percent positive items, when an agency has a higher prevalence of later responders, the early-minus-all estimate differences are more negative. In other words, as an agency's proportion of employees responding after two weeks increases, the percent positive estimates computed for that agency from all respondents are increasingly higher than the corresponding estimates computed from employees who responded in the first two weeks.

4.4. Relationships between Agency-Level Early-Minus-All Estimate Differences and Agency-Level Characteristics

This section investigates relationships between agency-level early-minus-all estimate differences and agency-level characteristics. Because of the tendency for agencies with a larger prevalence of late responders to have early-minus-all estimate differences that are more negative, we first investigated if an agency's prevalence of late responders could be predicted from its demographic characteristics. In particular, we developed an agency-level linear regression model for predicting an agency's prevalence of late responders. An alternative modeling approach would have been to use logistic regression, in which the logistic transform of the prevalence of late responders is modeled. However, over the observed range of prevalence values − 14.4 to 56.0 percent – the logistic transformation is accurately approximated by a linear relationship. The independent variables, calculated as unweighted means from the 2011 FEVS sampling frame data, described the following agency characteristics:

- Minority: Percentage of agency employees that are minorities,
- Gender: Percentage of agency employees that are male,
- Location: Percentage of agency employees assigned to the field,

- Supervisory Status: Percentage of agency employees that are not supervisors or managers, and
- Federal Government Tenure: Agency average of employees' length of federal service in years.

We also had available each agency's average age of its employees, but it was highly correlated with length of federal service or tenure, so to avoid multicollinearity we did not include it as an independent variable. We estimated an agency's prevalence of late responders by using the all-responder weights to compute the weighted mean of a variable equal to 1.0 for late responders and equal to 0.0 for early responders. Table 3b contains summary statistics for the agency characteristics.

Using the estimated prevalences of late responders and the associated independent variables for the 30 agencies, we calculated unweighted regression coefficients for an agency-level model containing an intercept and only linear terms involving the independent variables. Though the detailed results are not shown here, the R^2 for the developed prediction model was 0.286. The only regression coefficient that was statistically significant was the linear coefficient for the percentage of agency employees that are male ($p = 0.013$). If the agency prevalence of late responders is expressed as a percentage, this estimated regression coefficient equals 0.56. Since this regression coefficient was positive, agencies with a larger proportion of males had a larger proportion of their employees who reported later or after two weeks from the start of data collection. In particular, if two agencies differ by ten percentage points in their proportion of males, then the agency with the larger proportion of males is predicted to have a prevalence of late responders that is 5.6 percentage points greater than the agency with a smaller proportion of males.

Next, we developed a set of models that predicted the agency-level early-minus-all estimate differences for the (sub)indices from the agency-level characteristics listed in Table 3b. Equation 1 suggested that if the early-minus-all estimate difference was the dependent variable, then the independent variables should all include r_{late}, an agency's prevalence of late responders. Alternatively, in order to reduce heteroscedasticity one can transform the prediction models by dividing both sides by a power of r_{late}. Following the suggestion of Carrol and Ruppert (1988, 34), we assessed the need for such

Table 3b. Summary statistics for agency characteristics (n = 30)

Characteristic	Minimum	Mean	Median	Maximum	Skewness
Late-responder prevalence (%)	14.4%	40.9%	42.1%	56.0%	−0.78
Male prevalence (%)	32.2%	53.4%	55.4%	73.4%	−0.15
Minority prevalence (%)	20.3%	36.9%	32.8%	77.0%	1.34
Proportion located in the field (%)	11.6%	69.0%	74.2%	97.4%	−0.97
Proportion nonsupervisors (%)	76.8%	85.8%	85.7%	91.0%	−0.57
Average length of service (years)	10.7	16.4	16.7	20.3	−0.54

transformations by computing the Spearman rank correlation between the squared residuals and the predicted values produced by each model. The Spearman correlations for the untransformed models were between −0.04 and 0.09. We concluded that transformations were not needed. For example, when both sides were divided by r_{late} the Spearman correlations were between 0.02 and 0.30.

The independent variables for the untransformed models are listed in the first column of Table 4. Though Equation 1 suggested each of these models should not contain an intercept, we initially included an intercept in order to calculate the associated R^2 values. The first row of Table 4 contains the unadjusted R^2 values, ranging from 0.74 to 0.78, and the associated root-mean-square errors for prediction for each model when an intercept is included. In each model, the intercept was not significantly different from zero. We then re-estimated the regression coefficients for models not containing intercepts. Columns 2 through 5 of Table 4 contain the estimated coefficients, and those that are significantly different from zero ($p \leq 0.05$) are highlighted.

The coefficients for $(r_{late})^2$ were statistically significant in all five models and coefficients for r_{late} were statistically significant in three of the five models. All other coefficients were not significant, except that the interactions of the minority percentage with r_{late} or $(r_{late})^2$ were significant in models for the Employee Engagement Index and one of its subindices (Intrinsic Work Experience) and for the Global Satisfaction index. In addition, the interaction of the average length of federal service with r_{late} was significantly different from zero only in the model for the Supervisors subindex. These models show that across the (sub)indices, the agency prevalence of late responders and also the square of this prevalence, along with interactions with the prevalence of minorities and length of service with the agency, were significant predictors of the difference in early−minus-all estimates. Based on these findings of significant predictors, we performed additional analyses to investigate the behavior of the developed models.

Table 5 examines predicted early-minus-all estimate differences for an "average" agency (i.e., $d() = 0$ in the Table 4 coefficient expressions) and the effect of an increase in minority prevalence for the three indices or subindices in which the agency-percentage-of-minorities coefficients were statistically significant. These three models predict the early-minus-all differences for Intrinsic Work Experience, Employee Engagement, and Global Satisfaction. The Supervisors subindex was modeled separately as it had different predictors. The different rows of Table 5 correspond to different levels of the prevalence of late responders. The rows at the top of Table 5 have a low prevalence of late responders – that is, nearly all of the agency's responding employees respond during the first two weeks of data collection. The rows at the bottom of Table 5 have a high prevalence of late responders – that is, a large proportion of the agency's responding employees respond after the first two weeks of data collection.

Columns 2 through 4 of Table 5 contain the results of using the models to predict the early-minus-all estimate difference for different values of the prevalence of late responders for an "average" agency – that is, for an agency in which all of its demographic characteristics are equal to the unweighted all-agency average of the demographic characteristics. For a particular value of an agency's prevalence of late responders, the predicted values of early-minus-late estimate differences are very close to each other across the four indices and subindices.

Table 4. Results for modeling early-minus-all estimate differences for (sub)indices. Highlighted coefficients are statistically significant ($p \leq 0.05$)

	Model coefficients for predicted early-minus-all difference of:				
	Intrinsic experience	Leaders lead	Supervisors	Employee engagement	Global satisfaction
Model with intercept					
Unadjusted R^2	0.77	0.74	0.75	0.78	0.77
Adjusted R^2	0.61	0.56	0.58	0.63	0.61
Root MSE	0.59	0.64	0.72	0.58	0.77
Model without intercept					
Root MSE	0.59	0.63	0.73	0.58	0.75
Independent variables*					
r_{late}	4.58	4.58	5.49	4.88	3.81
$(r_{late})^2$	−17.51	−21.06	−19.36	−19.31	−17.02
r_{late}[d(% minority)]	0.68	0.67	0.71	0.68	0.82
r_{late}[d(% male)]	0.14	0.19	0.13	0.15	0.25
r_{late}[d(% field)]	−0.03	−0.03	0.01	−0.02	−0.01
r_{late}[d(% non-supervisors)]	−0.40	0.15	−0.79	−0.35	−0.31
r_{late}[d(avg. length of service)]	−0.47	0.05	−2.05	−0.82	−0.64
$(r_{late})^2$[d(% minority)]	−1.54	−1.44	−1.54	−1.51	−1.97
$(r_{late})^2$[d(% male)]	−0.23	−0.18	−0.19	−0.20	−0.57
$(r_{late})^2$[d(% field)]	0.04	0.06	−0.02	0.03	0.03
$(r_{late})^2$[d(% non-supervisors)]	1.28	0.21	1.81	1.10	1.27
$(r_{late})^2$[d(avg. length of service)]	0.53	0.02	3.89	1.48	0.36

*r_{late} = agency prevalence of late responders.
d( ) = difference of agency demographic statistic from unweighted average over all agencies.

Table 5. Model predictions for agency-level early-minus-all estimate differences for Intrinsic Work Experience (IN), Employee Engagement (EE), and Global Satisfaction (GL) (sub)indices

	Predicted early-minus-all estimate difference for "average" agency (%):			Predicted additive effect of increase in minority prevalence*		
Prevalence of late responders	IN	EE	GL	IN	EE	GL
0.10	0.3	0.3	0.2	0.3	0.3	0.3
0.20	0.2	0.2	0.1	0.4	0.4	0.4
0.30	−0.2	−0.3	−0.4	0.3	0.3	0.3
0.40	−1.0	−1.1	−1.2	0.1	0.2	0.1
0.50	−2.1	−2.4	−2.4	−0.2	−0.2	−0.4
0.60	−3.6	−4.0	−3.8	−0.7	−0.7	−1.1

*Predicted effect on early-minus-all estimate difference of a +5 percentage points difference in agency minority percentage from average minority percentage (%)

The predicted early-minus-all estimate differences for an "average" agency decrease to zero and then become more negative as the proportion of late responders increases. In particular, note that when an agency's prevalence of late responders is less than 30 percent, the predicted early-minus-late differences for an "average" agency are positive. This indicates that in such agencies the early responders have higher average scores for the modeled indices and subindices than do all responders. On the other hand, when an agency's prevalence of late responders is 30 percent or greater the entries in columns 2 through 4 for an "average" agency are negative. This indicates that in these agencies the early responders have lower average scores for the modeled indices and subindices. These two results suggests that at some point in time in an "average" agency's data collection period there may be a peak in the average value of the modeled indices and subindices among employees responding at this point in time. For agencies with a low prevalence of late responders, two weeks into the data collection period occurs after the peak, so the average of the early responders exceeds the average of the late responders, and hence the

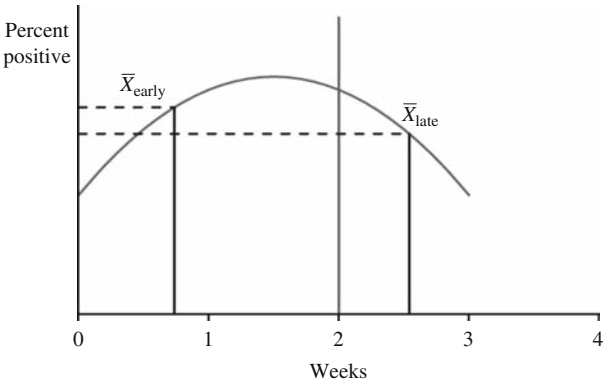


Fig. 2a. Possible explanation for a positive early-minus-all estimate difference. Because prevalence of late responders is low, the peak of percent positive for responses by time occurs less than two weeks into the data collection period. The overall percent positive for responses for early responders is greater than the overall percent positive responses for late responders, which produces a positive early-minus-all estimate difference.

early-minus-all estimate difference is positive (see Figure 2a). On the other hand, for agencies with a high prevalence of late responders, two weeks into the data collection period occurs before the peak, so the average of the early responders is less than the average of the late responders, and hence the early-minus-all estimate difference is negative (see Figure 2b).

Columns 5 through 7 of Table 5 predict the additive effect on early-minus-all estimate differences resulting from an agency's minority percentage differing by +5 percentage points from the unweighted all-agency average minority percentage. These results predict the difference in early-minus-all estimate differences between a particular agency and an "average" agency, where the particular agency's minority percentage differs by +5 percentage points from the minority percentage for the "average" agency. These results predict that the particular agency's early-minus-all estimate differences will be more positive for a low prevalence of late responders and will be more negative for a high prevalence of late responders. In particular, for agency prevalences of late responders less than 50 percent, the predicted early-minus-all differences in columns 5 through 7 are positive. This indicates that among those agencies in which 50 percent or fewer of the agency's responding employees responded in the first two weeks, the agencies with a higher proportion of minorities compared to the "average" agency will have more positive early-minus-all differences in the modeled indices and subindices than the "average" agency. For agency prevalences of late responders of 50 percent or greater, however, the predicted early-minus-all estimate differences in columns 5 through 7 of Table 5 are negative. This indicates that among agencies in which 50 percent or fewer of the agency's responding employees responded in the first two weeks, the agencies with higher proportions of minorities compared to the "average" agency will have more negative early-minus-all estimate differences in the modeled indices and subindices.

Table 6 examines the predicted early-minus-all estimate differences for an "average" agency and the additive effect of an increase in agency-average length of service for the Supervisors subindex model. Column 2 of Table 6 contains the predictions for different prevalences of late responders for an "average" agency – that is, an agency in which all of its demographic characteristics are equal to the unweighted all-agency average of the

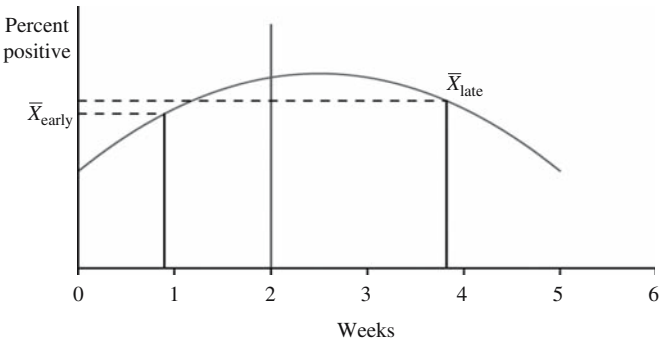


Fig. 2b. Possible explanation for a negative early-minus-all estimate effect. Because the prevalence of late responders is high, the peak of percent positive for responses by time occurs more than two weeks into the data collection period. The overall percent positive for responses for early responders is less than the overall percent positive for responses for late responders, which produces a negative early-minus-all estimate difference.

Table 6. Model predictions for agency-level early-minus-all estimate differences for supervisors subindex

Prevalence of late responders	Predicted early-minus-all estimate difference for "average" agency (%):	Predicted additive effect of increase in agency average federal length of service*
0.10	0.4	−0.8
0.20	0.3	−1.3
0.30	−0.1	−1.3
0.40	−0.9	−1.0
0.50	−2.1	−0.3
0.60	−3.7	0.9

* Effect of +5 years difference in agency average federal length of service (LOS) from average LOS averaged over all agencies (%)

demographic characteristics. For a particular value of an agency's prevalence of late responders, the predicted values of the early-minus-all estimate differences in Column 2 of both Table 5 and Table 6 are very close to each other, with the predicted values becoming more negative as the prevalence of late responders increases. Column 3 of Table 6 contains the predicted effects on early-minus-all estimate differences resulting from an agency's average of employees' length of federal service differing by +5 years from the unweighted all-agency average of employee's length of federal service. Note that the sum of columns 2 and 3 is negative. This suggests that for agencies in which the average length of federal service differs by at least +5 years from that for the "average" agency, the point in time of the peak value of the Supervisors subindex for responding employees occurs after the first two weeks of data collection, or maybe there is no peak, with the average percent positive of responding employees increasing with time.

4.5. Relationships between Agency-Level Early-Minus-All Estimate Differences and Levels of Early-Responder and All-Responder Estimates

In Subsection 4.2, we observed that smaller percent positive values for the governmentwide indices and subindices for both early responders and all responders were associated with more negative early-minus-all estimate differences. To determine if this also held at the agency level, we developed an agency-level model for each (sub)index to predict early-minus-all estimate differences from early-responder estimates as well as a model to predict early-minus-all estimate differences from all-responder estimates. The right-hand side of these models contained an intercept and either the early-responder estimate or the all-responder estimate multiplied by a slope coefficient:

$$E^{(i)} = intercept^{(i)}_{early} + slope^{(i)}_{early}\bar{X}^{(i)}_{early} + error^{(i)}_{early}$$

and

$$E^{(i)} = intercept^{(i)}_{late} + slope^{(i)}_{late}\bar{X}^{(i)}_{late} + error^{(i)}_{late}.$$

Table 7. Slope coefficients and R^2 values for predicting early-minus-all differences

	Slope coefficients		Adjusted R^2 values		
(Sub-)Index	Early-responder estimate	All-responder estimate	Agency characteristics	Early-responder estimate	All-responder estimate
Employee engagement	0.13	0.12	0.63	0.42	0.26
Leaders lead	0.11	0.10	0.56	0.42	0.28
Intrinsic work experiences	0.14	0.11	0.61	0.33	0.14
Supervisors	0.12	0.09	0.58	0.26	0.09
Global satisfaction	0.11	0.09	0.61	0.29	0.13

The intercepts and slope coefficients were statistically significant ($p \leq 0.05$) in all of the models for predicting early-minus-all estimate differences from early-responder or all-responder estimates.

Table 7 contains the values of the estimated slope coefficients and compares the R^2 values to those for the models discussed in the preceding section in which early-minus-all differences were predicted from agency characteristics. The estimated slope coefficients for predicting early-minus-all estimate differences from early-responder estimates are between 0.11 and 0.14, with the largest being for Intrinsic Work Experience. Hence, on average if two agencies' early-responder estimates for Intrinsic Work Experience differ by five percentage points, then their corresponding all-responder estimates will differ by $5*(1 + 0.14) = 7$ percentage points.

The estimated slope coefficients for predicting early-minus-all estimate differences from the all-responder estimates are slightly smaller, ranging between 0.09 and 0.12, with largest being for Employee Engagement. Hence, on average, if two agencies' all-responder estimates for Employee Engagement differ by eight percentage points, then their corresponding early-responder estimate will differ by $8*(1 - 0.12) = 7$ percentage points. Tables 4, 5, and 6 indicate that the variation across agencies in early-minus-all differences can be explained by variation in agency characteristics. Table 7, on the other hand, indicates that variation across agency in early-minus-all differences can be explained by the variation in agency-level estimates calculated from early responders or, alternatively, in the estimates calculated from all responders.

5. Conclusions

This article explored the impact of shortening the fielding period of the FEVS using the results from a subset of 30 agencies participating in the 2011 FEVS. If the FEVS data collection period were to be shortened to two weeks and no other changes were made to the timing of FEVS survey administration activities, the analyses conducted suggest that the response rate, the demographic profile of respondents, and the survey estimates for the Employee Engagement and Global Satisfaction indices could change significantly.

By shortening the survey fielding period, fewer employees would have the chance to respond. The number of completed surveys for the 2011 FEVS would have been reduced by approximately 41 percentage points (ranging from 14 percent to 57 percent across agencies). However, it is unclear whether a reduction of this magnitude would be observed in practice in future FEVS administrations. One potential reason is that sampled employees receive a barrage of tailored notifications indicating the fielding period is about to end, which generally results in a surge of completed surveys. This study artificially shortened the fielding period without attempting to account for the effect of a pending deadline. Further research could explore ways to model and incorporate this effect into the process of estimating early-minus-all percent positive differences.

The demographic profiles of those who responded in the first two weeks (early responders) were significantly different from late responders. Early responders were more likely to be nonminority employees, female employees, older employees, or employees who intend to leave their current position for another job either within or outside the government. The late responders were more likely to be higher-grade employees, supervisors, executives, male employees, and younger employees.

In addition to demographic profile differences, shortening the fielding period results in a decrease of governmentwide percent positive estimates and associated indices, with changes in percent positive estimates ranging between -1.76 percent and -1.13 percent depending on the index. However, the relationship is not straightforward and uniform. These differences are influenced by the apportionment of early/late responders and the prevalence of longer tenured and/or nonminority employees. If an agency has a higher proportion of early responders, it tends to have higher percent positive estimates, on average. As the share of late responders increases, the percent positive estimates tend to be lower, as calculated from only the early responders. This translates to the indices, which are simple averages of the percent positive estimates. For example, with the Global Satisfaction and Employee Engagement indices as well as the Intrinsic Work Experience subindex, there was an additive impact associated with the proportion of minority employees in the agency. If an agency has a lower proportion of late responders and a higher proportion of minorities, the early responders will tend to yield higher average percent positive estimates and (sub)index scores. For the Supervisors subindex, an opposite relationship was found for length of service in the federal government. Almost regardless of the proportion of late responders, if an agency has an average length of federal services $+5$ years from the average agency, the early responders will tend to have lower average percent positive estimates on the indices. Lastly, further analysis showed that an item's percent positive estimate itself seems to impact the magnitude of the early-minus-all difference. Specifically, the smaller the percent positive estimate, the more negative the difference.

Despite many of the factors discussed above falling outside the survey sponsor's locus of control, one general best practice recommendation appears to emerge from scrutinizing the data in Appendix A. Although not explicitly stated elsewhere in the article, there is clearly a positive association between an agency's prevalence of early responders and its overall response rate. Therefore, it seems plausible that efforts to boost the overall response rate could, in turn, boost the portion of employees who respond promptly, thereby tempering some of the noted item score differentials and reintroducing the

possibility of a shortened fielding period. From our own practical experience conducting employee surveys, we find that the agencies consistently generating higher response rates are those in which senior officials aggressively publicize the survey via internal agency correspondence and other pertinent media outlets to reach as many employees as possible in the weeks immediately preceding the survey launch. Because many of these surveys are recurring, as is the case with the FEVS, another critical element is to communicate specific actions taken as a result of a prior survey administration. This helps foster a sense of employee empowerment in taking the survey, a belief that the feedback provided will be used to drive organizational change. An item can be included on the survey instrument to help gauge the organization's success in this regard. For example, an item was added to the FEVS instrument in 2010, asking employees about their agreement with the statement "I believe the results of this survey will be used to make my agency a better place to work."

In conclusion, this article presents evidence, based on the 2011 FEVS survey administration, that reducing the field period to two weeks would have ramifications for the response rates, the demographic profile for those responding during that time frame, and the attitudinal measures and aggregates thereof estimated by the survey. Although it was the first known comparison of early versus late responders in a self-administered employee survey of our population of interest, U.S. government employees, it follows a long tradition of similar analyses in the survey methodology literature. In addition, there is a wide body of survey literature on the causes and correlates of the decision to respond to a survey (Groves and Couper 1998), attempts to tailor contact attempts to maximize response (Kreuter 2013; Weeks 1987; Wagner 2013), as well as sociological (Dillman et al. 2009) and psychological (Groves et al. 2000) factors associated with survey participation. We feel that studies such as ours would benefit greatly if this literature were expanded to not only explain the dichotomy of whether one *ultimately* ignores or answers the call to participate in a survey, but the point during the fielding period when one makes his or her final decision.

Appendix A. Federal employee viewpoint survey (FEVS) 2011 fielding period lengths and distributional statistics of early versus late respondents for the 30 agencies analyzed in this study

Agency	Fielding period lengths (weeks)	Total count of respondents	Overall response rate[2]	Count of early[1] respondents	Percent of early respondents	Early respondent response rate
Broadcasting Board of Governors	7	1,089	65.8	511	46.9	30.9
Department of Agriculture	5	14,375	54.1	9,687	67.4	36.5
Department of Commerce	9	18,069	53.9	7,776	43.0	23.2
Department of Education	6	2,891	72.3	1,693	58.6	42.3
Department of Energy	4	5,509	39.2	3,742	67.9	26.6
Department of Health and Human Services	6	23,092	39.5	13,259	57.4	22.7
Department of Homeland Security	5	15,499	49.1	8,179	52.8	25.9
Department of Housing and Urban Development	5	5,365	62.2	3,184	59.3	36.9
Department of Justice	9	21,488	53.2	12,856	59.8	31.8
Department of Labor	4	7,475	48.8	5,162	69.1	33.7
Department of State	6	2,422	43.0	1,410	58.2	25.0
Department of the Interior	5	7,051	51.2	4,338	61.5	31.5
Department of the Treasury	8	17,985	65.1	10,977	61.0	39.7
Department of Transportation	8	9,811	67.0	5,660	57.7	38.7
Department of Veterans Affairs	4	13,703	44.9	11,822	86.3	38.7
Environmental Protection Agency	8	8,585	51.8	3,918	45.6	23.6
Equal Employment Opportunity Commission	8	1,252	52.1	710	56.7	29.5
General Services Administration	3	2,491	51.1	2,006	80.5	41.2
National Aeronautics and Space Administration	10	9,239	53.8	4,136	44.8	24.1
National Archives and Records Administration	8	1,854	69.2	1,059	57.1	39.5
Nuclear Regulatory Commission	5	2,612	67.7	1,570	60.1	40.7
Office of Personnel Management	6	3,462	73.5	2,449	70.7	52.0
Small Business Administration	4	1,619	70.2	1,349	83.3	58.5
Social Security Administration	4	7,069	55.0	5,111	72.3	39.8

Appendix A. Continued

Agency	Fielding period lengths (weeks)	Total count of respondents	Overall response rate[2]	Count of early[1] respondents	Percent of early respondents	Early respondent response rate
U.S. Agency for International Development	4	1,243	39.0	747	60.1	23.4
U.S. Army Corps of Engineers	9	3,897	28.7	1,989	51.0	14.6
U.S. Department of the Air Force	9	8,757	34.2	4,713	53.8	18.4
U.S. Department of the Army	9	14,911	32.4	7,604	51.0	16.5
U.S. Department of Defense Fourth Estate	9	8,003	36.2	4,235	52.9	19.2
U.S. Department of the Navy	9	12,469	39.7	6,532	52.4	20.8
Totals		253,287	49.7	148,384	58.6	29.1

Notes:
[1]An early respondent is defined as one who completes the survey within the first two weeks of the agency fielding period
[2]Response rate calculated using RR3 of the American Association of Public Opinion Research (AAPOR)

6. References

American Association for Public Opinion Research. 2009. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Sixth Edition*. Available at: http://www.aapor.org/Content/NavigationMenu/ResourcesforResearchers/StandardDefinitions/StandardDefinitions2009new.pdf (accessed January 27, 2014).

Baruch, Y. and B.C. Holtom. 2008. "Survey Response Rate Levels and Trends in Organizational Research." *Human Relations* 61: 1139–1160. DOI: http://dx.doi.org/10.1177/0018726708094863.

Bates, N. and K. Creighton. 2000. "The Last Five Percent: What Can We Learn from Difficult/Late Interviews?" In Proceedings of the Section on Government Statistics: American Statistical Association, August 13, 2000. 120–125. Alexandria, VA: American Statistical Association.

Baur, E.J. 1947. "Response Bias in a Mail Survey." *Public Opinion Quarterly* 11: 595–600. DOI: http:/dx.doi.org/10.1086/265895.

Borg, I. and T. Tuten. 2003. "Early versus Later Respondents in Intranet-Based, Organizational Surveys." *Journal of Behavioral and Applied Management* 4: 134–145.

Carroll, R.J. and D. Ruppert. 1988. *Transformation and Weighting in Regression*. New York, NY: Chapman and Hall.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3rd ed. Hoboken, NJ: Wiley.

Filion, F. 1975. "Estimating Bias Due to Nonresponse in a Mail Survey." *Public Opinion Quarterly* 39: 482–492. DOI: http://dx.doi.org/10.1086/268245.

Gannon, M.J., J.C. Nothern, and S.J. Carroll. 1971. "Characteristics of Nonrespondents Among Workers." *Journal of Applied Psychology* 55: 586–588. DOI: http://dx.doi.org/10.1037/h0031907.

Green, K.E. 1991. "Reluctant Respondents: Differences Between Early, Late, and Nonresponders to a Mail Survey." *The Journal of Experimental Education* 59: 268–276. DOI: http://dx.doi.org/10.1080/00220973.1991.10806566.

Groves, R.M. and M. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York, NY: Wiley. DOI: http://dx.doi.org/10.1002/9781118490082.index.

Groves, R.M., E. Singer, and A. Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." *Public Opinion Quarterly* 64: 299–308. DOI: http://dx.doi.org/10.1086/317990.

De Leeuw, E. and W. de Heer. 2002. "Trends in Household Survey Nonresponse: a Longitudinal and International Comparison." In *Survey Nonresponse*, edited by Robert M. Groves, et al. New York, NY: Wiley. DOI: http://dx.doi.org/10.1093/poq/nfl033.

Ellis, R.A., C.M. Endo, and J.M. Armer. 1970. "The Use of Potential Nonrespondents for Studying Nonresponse Bias." *Pacific Sociological Review* 13: 103–109. DOI: http://dx.doi.org/10.2307/1388313.

Erickson, T.J. 2005. "The 21st Century Workplace: Preparing for Tomorrow's Employment Trends Today." (Testimony submitted before the U.S. Senate Committee on Health, Education, Labor, and Pensions, May 26, 2005). Available at: http://www.help.senate.gov/hearings/index.cfm?year=2005&month=05 (accessed December 2012).

Jacoby, J. and M.S. Matell. 1971. "Three-Point Likert Scales are Good Enough." *Journal of Marketing Research* 8: 495–500. DOI: http://dx.doi.org/10.2307/3150242.

Kalton, G.F. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19: 81–97.

Kraut, A.I. 1996. *Organizational Surveys: Tools for Assessment and Change*. San Francisco, CA: Jossey-Bass.

Kreuter, F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley. DOI: http://dx.doi.org/10.1002/9781118596869.ch1.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Second ed. New York, NY: Wiley. DOI: http://dx.doi.org/10.1002/sim.1697.

Macey, W.H. and B. Schneider. 2008. "The Meaning of Employee Engagement." *Industrial and Organizational Psychology* 1: 3–30. DOI: http://dx.doi.org/10.1111/j.1754-9434.2007.0002.x.

Mayer, C.S. and R.W. Pratt., Jr. 1966. "A Note on Nonresponse in a Mail Survey." *Public Opinion Quarterly* 30: 637–646. DOI: http://dx.doi.org/10.1086/267461.

Newman, S.W. 1962. "Differences between Early and Late Respondents to a Mailed Survey." *Advertising Research* 2: 37–39.

Pace, R.C. 1939. "Factors Influencing Questionnaire Returns from Former University Students." *Journal of Applied Psychology* 23: 388–397. DOI: http://dx.doi.org/10.1037/h0063286.

Rogelberg, S.G. and J.M. Stanton. 2007. "Understanding and Dealing with Organizational Survey Nonresponse." *Organizational Research Methods* 10: 195–209. DOI: http://dx.doi.org/10.1177/1094428106294693.

Schwirian, K.P. and H.R. Blaine. 1966. "Questionnaire-Return Bias in the Study of Blue-Collar Workers." *Public Opinion Quarterly* 30: 656–663. DOI: http://dx.doi.org/10.1086/267463.

Sonquist, J.A., E.L. Baker, and J.N. Morgan. 1974. *Searching for Structure*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

U.S. Office of Personnel Management. 2011. *2011 Federal Employee Viewpoint Survey: Governmentwide Management Report*. Washington, DC: OPM. Available at: http://www.fedview.opm.gov/2011/Published (accessed January 27, 2014).

Wagner, J. 2013. "Adaptive Contact Strategies in Telephone and Face-to-Face Surveys." *Survey Research Methods* 7: 45–55. DOI: http://dx.doi.org/10.1002/9781118596869.ch7.

Weeks, M.F. 1987. "Optimal Call Scheduling for a Telephone Survey." *Public Opinion Quarterly* 51: 540–549. DOI: http://dx.doi.org/10.1086/269056.

Received December 2012
Revised September 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 675–700, http://dx.doi.org/10.2478/JOS-2014-0043

The Utility of Nonparametric Transformations for Imputation of Survey Data

Michael W. Robbins[1]

Missing values present a prevalent problem in the analysis of establishment survey data. Multivariate imputation algorithms (which are used to fill in missing observations) tend to have the common limitation that imputations for continuous variables are sampled from Gaussian distributions. This limitation is addressed here through the use of robust marginal transformations. Specifically, kernel-density and empirical distribution-type transformations are discussed and are shown to have favorable properties when used for imputation of complex survey data. Although such techniques have wide applicability (i.e., they may be easily applied in conjunction with a wide array of imputation techniques), the proposed methodology is applied here with an algorithm for imputation in the USDA's Agricultural Resource Management Survey. Data analysis and simulation results are used to illustrate the specific advantages of the robust methods when compared to the fully parametric techniques and to other relevant techniques such as predictive mean matching. To summarize, transformations based upon parametric densities are shown to distort several data characteristics in circumstances where the parametric model is ill fit; however, no circumstances are found in which the transformations based upon parametric models outperform the nonparametric transformations. As a result, the transformation based upon the empirical distribution (which is the most computationally efficient) is recommended over the other transformation procedures in practice.

Key words: Missing data; multiple imputation; empirical CDF; kernel density; ARMS; Markov chain Monte Carlo.

1. Introduction

Missing data are a particularly common and particularly troublesome problem in establishment surveys. A large portion of the statistical literature has been devoted to the analysis of data that contain missing values, and as a result a myriad of approaches exist. Pertinent techniques include calibration weighting (Kott and Chang 2010) and the EM algorithm (Dempster et al. 1977); however, imputation (for a summary, see Rubin 1987) is often the preferred method for handling missing data since it yields a completed dataset on which classical tools for analysis may be applied. Additionally, multiple (or repeated)

[1]Associate Statistician, RAND Corporation, Pittsburgh, PA 15213 U.S.A. Email: mrobbins@rand.org
Acknowledgments: The author acknowledges partial funding from USDA grant #58-3AEU-2-0065, from the Cross-Sector Research in Residence Program between the National Institute of Statistical Sciences (NISS) and National Agricultural Statistics Service (NASS), and from the University of Missouri Research Board. The author acknowledges and thanks Sujit Ghosh, Barry Goodwin, Joshua Habiger, Darcy Miller and Kirk White for their contributions to the research project associated with the work presented here. The author thanks the editoral staff and anonymous reviewers whose helpful comments greatly improved the article. The views expressed are those of the author and do not necessarily represent the views of RAND, NASS or the USDA.

imputation (Rubin 1996) may be used to quantify imputation error. Despite the ubiquity of missing data problems and methodology designed to address them, existing imputation algorithms have many drawbacks, largely with respect to robustness and computational efficiency.

Multivariate imputation techniques tend to be fairly restrictive with respect to the types of model assumptions. Techniques that impute via a multivariate normal model (Schafer 1997; Robbins et al. 2013) are popular and theoretically justified. Techniques that use fully conditional specification (a.k.a. SRMI, as outlined in Raghunathan et al. 2001), which is implemented in several software packages including IVEware (Raghunathan et al. 2002), MICE (Van Buuren and Oudshoorn 1999), and mi (Su et al. 2011), can be used to create imputations in data that contain categorical and discrete variates but lack theoretical justification due to the use of a potentially incompatible Gibbs sampler. However, each of the aforementioned procedures is best suited to sample (i.e., draw) imputations for continuous variables from a normal distribution.

Multivariate techniques that do not sample imputations for continuous items under Gaussian assumptions are relatively sparse. Algorithms which employ fully conditional specification can be modified so that imputations are generated via a conditional modeling/sampling technique known as predictive mean matching (PMM, Little 1988). PMM is a nearest-neighbor procedure; imputations are sampled from observed data values. However, PMM is computationally burdensome in comparison to its Gaussian counterparts and thus can have little utility in high dimensional settings. The IRMI algorithm (Templ et al. 2011) is similar in structure to SRMI-type procedures with the added functionality of estimating conditional models through robust regression; however, steps are not taken to ensure that imputations are sampled from the true conditional distribution, which implies that IRMI imputations will likely distort complex distributional characteristics (further justification for this claim is provided in Section 5). To increase the robustness of traditional normality-based methods, many authors recommend the use of marginal transformations of continuous variates prior to the application of imputation methodology. For example, Raghunathan et al. (2001) suggest a power transformation, whereas Robbins et al. (2013) suggest a density-based transformation (specifically, a skew-normal density is used).

The practicality of the aforementioned procedures is muddled by their computational complexity. The growing ubiquity of multiple imputation, the prevalence of iterative sampling techniques (e.g., Markov chain Monte Carlo) for imputation, and the high dimensional nature of modern statistical analyses result in algorithms that mandate a substantial computational burden. Such issues become increasingly problematic under the guise of the benefits provided by the use of a wide-ranging imputation model (Robbins and White, Forthcoming).

Here, the transformation-based schemes of Robbins et al. (2013) are extended, resulting in the introduction of robust techniques for transformation. In particular, a transformation based on the kernel density is suggested. Woodcock and Benedetto (2009) use a kernel density to generate data values for the purpose of creating a public use dataset from confidential data. Additionally, a fully empirical transformation (which uses a modified empirical distribution) is presented here. The empirical transformation yields a hot-deck

(or nearest-neighbor) technique that may be applied jointly with commonly used multivariate imputation algorithms (such as IVEware, MICE or mi) in a very computationally efficient manner. The proposed methodologies yield simple tools which uphold the ability to preserve complex distributional structures provided by PMM while maintaining the computational efficiency of techniques which mandate Gaussian assumptions.

In this article, imputations for a widely-used data product are generated via the aforementioned transformation techniques. The marginal and multivariate efficacy of the resulting imputations, as well as the inadequacies of imputations generated using a fully parametric model, are illustrated. Specifically, in Section 2, the dataset that will be used throughout, and the technique that will be used to generate imputations (following transformation), are introduced. The robust methods of transformation are presented in Section 3, and data analysis is provided in Section 4. Further, Section 5 presents a simulation study (performed using real and synthetic data) that illustrates the effectiveness of the proposed transformation schemes. The article concludes by providing comments and practical advice in Section 6.

2. The ARMS and Associated Imputation Technique

In June 2009, a research project commenced with the goal of creating a new imputation method for the US Department of Agriculture's (USDA) Agricultural Resource Management Survey (ARMS). Partial findings of the research project are outlined in Robbins and White (2011), Robbins et al. (2013) and Robbins and White (Forthcoming); this article relates additional findings of the project. Although the methodologies presented here are widely applicable, the problem of interest is motivated here through a discussion of the ARMS and its recently developed imputation technique.

ARMS data are a key source of information for congressional decisions that allocate billions of dollars in farm subsidies (Robbins et al. 2013). The survey provides the USDA's most comprehensive view of the American farm household; ARMS data contain 30,000–40,000 units (observations) with 1,000–2,000 items (variables). The ARMS has a multiphase, dual-frame, stratified, probability-weighted sampling design. Design weights are calibrated, and the calibrated weights are used to calculate key survey indications (U.S. Department of Agriculture 2011). Calibration of design weights also accounts for unit nonresponse; the rate of unit nonresponse tends to hover around 30% (National Research Council 2008). Analyses presented herein use data from the 2010 ARMS.

Aside from being high dimensional, ARMS data have a complex distributional structure – the majority of ARMS variables have semicontinuous distributions. To elaborate, a portion of units will report a zero for a given variable, whereas the responses for the remaining observations for that variable are sampled from some strictly positive and (theoretically) continuous distribution.

The new ARMS imputation procedure handles semicontinuous variables via a commonly used mixture model (see Javaras and van Dyk 2003, for example). Specifically, a semicontinuous variable $\boldsymbol{Y}$ is broken down into two latent variables, $\boldsymbol{B}$ and $\boldsymbol{Y}^*$, where $\boldsymbol{B}$ is an indicator variable denoting whether or not $\boldsymbol{Y}$ is positive, and $\boldsymbol{Y}^*$ is a strictly

continuous variable that indicates the positive portion of $\boldsymbol{Y}$. The imputation algorithm treats $\boldsymbol{Y}^*$ as missing whenever $\boldsymbol{Y}$ is missing or 0. All semicontinuous ARMS variables are transformed in this manner, and all ARMS variables with missing values are assumed to be semicontinuous. See Su et al. (2011) for an example of an extant procedure that utilizes similar approaches for handling semicontinuous data. Another key characteristic of the missingness in ARMS data is that all missing values are assumed to be positive. Thus, $\boldsymbol{B}$ is fully observed for all variables.

The positive portions of ARMS variables (i.e., the $\boldsymbol{Y}^*$s) tend to be highly skewed. Since all imputation procedures that are practical in high-dimensional settings link variables through a multivariate normal model, each $\boldsymbol{Y}^*$ is transformed in order to achieve normality. Letting $\boldsymbol{X}$ (which is theoretically Gaussian) represent a transformed version of $\boldsymbol{Y}^*$, Robbins et al. (2013) provide the following procedural outline of the algorithm for imputation in ARMS data:

1. Break each semicontinous variable $\boldsymbol{Y}$ into $\boldsymbol{B}$ and $\boldsymbol{Y}^*$ (observed 0s are treated as missing).
2. Transform: $\boldsymbol{Y}^* \Rightarrow \boldsymbol{X}$ for each variable.
3. Impute: Find $\hat{\boldsymbol{X}}$ (the imputed version of $\boldsymbol{X}$) for each variable.
4. Untransform: $\hat{\boldsymbol{X}} \Rightarrow \hat{\boldsymbol{Y}}$ (the imputed version of $\boldsymbol{Y}$) for each variable (values that are originally observed as 0 are reset to 0).

The imputed data also undergo an editing process to ensure that imputations satisfy all data constraints prior to release. Most variables are not subject to such constraints, and the editing process does not damage the quality of the imputations with regards to analytic properties.

Robbins et al. (2013) focus on Step 3 above. For that purpose, they introduce a dynamic imputation procedure, the so-called iterative sequential regression (ISR) method, that builds a multivariate (normal) model for the $\boldsymbol{X}$s (and respective covariates) through a sequence of conditional linear models while allowing flexibility in the form of each conditional model. For the purpose of transformation, they apply a skew-normal model (Azzalini 1985) to the logged versions of the $\boldsymbol{Y}^*$s. It had been established that such a transformation is sufficient for the majority of ARMS variables (Miller et al. 2010). However, for certain ARMS variables (and surely data from most any other survey) such a model is insufficient.

As a result, the focus here turns to Steps 2 and 4 above: the mechanisms for transformation. We present robust nonparametric methods for transformation that will retain the applicability of the ISR procedure while ensuring that imputations preserve the marginal structure of complex survey variables (as will be illustrated in the sections that follow). It is emphasized that the methods presented in the following are widely applicable; these techniques may be applied to any data that contain theoretically continuous (or semicontinuous) variables and may be applied in conjunction with a wide array of imputation procedures.

To help illustrate the applicability of the methodology presented here to general imputation problems, statistical analyses that require the specific ARMS design are not the focus here. Regardless, the survey design is not expected to have a substantial influence on the choice of transformation scheme.

In this article, the standard errors of estimators derived using imputed data are adjusted for imputation error via multiple imputation (MI, Rubin 1987;1996). MI involves the generation of multiple datasets which have been imputed independently of one another; imputations are presumed to have been randomly sampled from the posterior distribution of the missing data given the observed data. Rubin's rules for combining information across datasets have been provided in a number of references (including the two given above). The validity of MI inferences in settings where complex survey data are used has been called into question frequently (Kott 1995; Fay 1996; Kim et al. 2006). Although MI has demonstrated utility for analysis of ARMS data (Robbins and White, Forthcoming), MI is used here primarily due to its simplicity and effectiveness in comparing imputation error across datasets imputed via differing methods.

3. Transformation Techniques

Let the length-n vector $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ denote a survey variable, where n is the sample size (i.e., number of experimental units). To develop a transformation scheme that attains normality, consider the fact that any continuous random variable with a known cumulative distribution function (CDF) can be transformed into a standard normal variate. Specifically, let X be any scalar random variable with known CDF $F(x)$, and let

$$T(x) = \Phi^{-1}(F(x)) \tag{1}$$

represent the transformation function, where $\Phi(\cdot)$ denotes the standard normal CDF, then

$$T(X) \sim N(0, 1)$$

It is noted that when variables are transformed via (1) and then linked through a multivariate normal distribution (which is the model used for imputation here), the resulting model may be considered a Gaussian copula (Nelsen 2009).

The impasse with respect to application of the above transformation scheme is the fact that in practical circumstances, the CDF $F(x)$ tends to be unknown. Thus, in order to apply the above transformation to the positive portions survey, it is necessary to first develop a manner for determining (or approximating) the CDF of these positive portions. As mentioned above, a log-skew-normal model suffices for the majority of ARMS variables. That is, in accordance with (1); Robbins et al. (2013) suggest that if

$$T_1(y) = \Phi^{-1}(F(y|\hat{\xi}, \hat{\omega}, \hat{\alpha})) \tag{2}$$

then $T_1(\log Y_i)$ should have (or approximately have) a standard normal distribution for all relevant i. In the above, $\{\hat{\xi}, \hat{\omega}, \hat{\alpha}\}$ represent consistent estimators of the skew-normal parameters. Clearly, such marginal transformations provide no general implication that joint normality will be obtained; however, Robbins et al. (2013) illustrate rigorously that for ARMS data multivariate normality is (adequately) achieved through marginal transformation to normality. It is noted that these conclusions also hold when the nonparametric transformations proposed herein are used.

As was also mentioned above, the transformation in (2) is inadequate for certain ARMS variables. For instance, labor variables, where the response indicates the number of weekly

Table 1. List and description of ARMS variables pertinent to this study. The number of positive and observed (n_{obs}) values and the number of missing values (n_{mis}) is provided for each listed variable. Within the simulation study of Subsection 5.1, additional missingness is imposed in the variables marked with an asterisk

Name	Description	n_{obs}	n_{mis}
P758*	Operator's expenditure for hired labor	9,354	0
P764*	Operator's wage expenditure for operator	1,296	0
P784*	Contractor's expenditure for contract labor	151	0
P828*	Operator's on-farm labor (in hrs/wk) for Jan.–Mar.	19,285	1,296
P829*	Operator's on-farm labor (in hrs/wk) for Apr.–Jun.	19,342	1,438
P830	Operator's on-farm labor (in hrs/wk) for Jul.–Sept.	19,274	1,474
P831	Operator's on-farm labor (in hrs/wk) for Oct.–Dec.	19,114	1,517
P832*	Spouse's on-farm labor (in hrs/wk) for Jan.–Mar.	8,991	533
P833*	Spouse's on-farm labor (in hrs/wk) for Apr.–Jun.	9,298	513
P834	Spouse's on-farm labor (in hrs/wk) for Jul.–Sept.	9,298	529
P835	Spouse's on-farm labor (in hrs/wk) for Oct.–Dec.	9,097	559
P884	Estimated value of farm credit stock on Dec. 31	4,273	1,025
P952*	Operator and spouse off-farm labor	10,462	1,081

hours worked, tend to observe onerous marginal distributions. Names and descriptions of ARMS variables that will be discussed in this study are given in Table 1. Names of ARMS variables are formed by placing a "P" in front of the numeric item code seen on the survey questionnaire.

As an example, Figure 1 provides a histogram of log(P829) with the best-fitting skew-normal density curve. Only positive responses for this variable are included in this graph (and similar plots that follow). A scatter plot of log(P829) and log(P830) is also provided in the figure to illustrate the bivariate dispersion of the data points. Likewise, only units that report positive values for both variables are plotted in this graph (and similar ones that follow). These labor variables are analyzed on the log scale because logged values are closer to being Gaussian than the untransformed values.

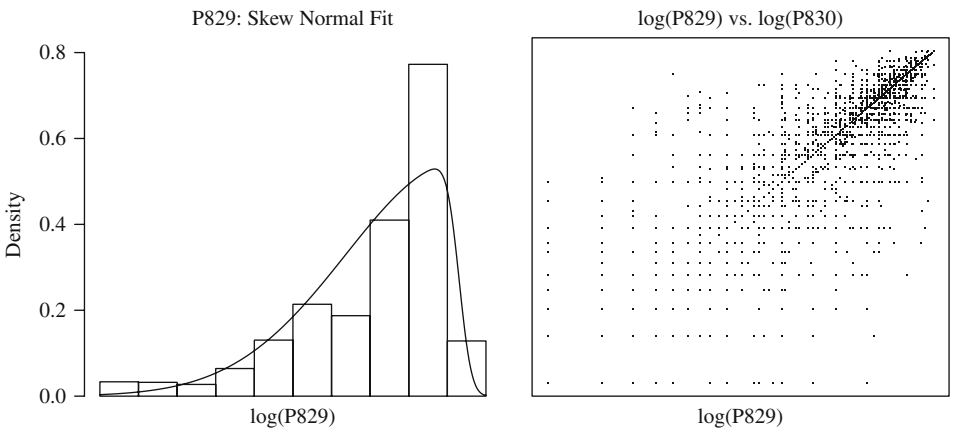


Fig. 1. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right). The left plot has the best fitting skew-normal density curve overlaid. Axis values are suppressed to avoid disclosure where necessary

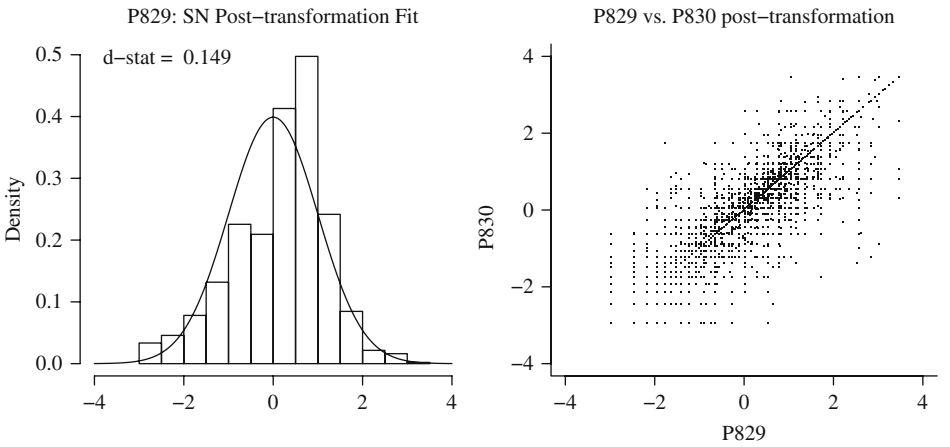


Fig. 2. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following skew-normal transformation. The left plot has the standard normal density curve overlaid

To illustrate further the specific deficiencies of the skew-normal (SN) transformation for the labor variables, Figure 2 provides a histogram log(P829) and a scatter plot of log(P829) versus log(P830), all following skew-normal transformation. As the transformed data should observe a standard normal distribution, the standard normal density is plotted over the histogram of the transformed data. Additionally, a Kolmogorov-Smirnov (KS) test under the assumption of a standard normal distribution is applied to the transformed values shown in the left graph in Figure 2, and the distance statistic (d-stat) is given in the upper-left corner of the plot. Labor variables such as P829 tend to have repeating values, which makes the KS test theoretically inappropriate, but such results are given here and in further plots for a comparison of goodness of fit.

The power (or Box-Cox) transformation is often applied within imputation procedures (e.g., Raghunathan et al. 2001). However, the Box-Cox transformation show no increase in utility over the log-skew-normal transformation described above; therefore it is not discussed further. A more robust transformation scheme is clearly warranted. Accordingly, nonparametric models for $F(x)$ are considered.

3.1. Transformation Via the Kernel Density

Next, consider the Gaussian kernel, which is used to estimate the probability density function (PDF). Similarly, Woodcock and Benedetto (2009) use kernel densities for marginal transformation to normality. The kernel density (using a Gaussian kernel) of $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - Y_i}{h}\right),$$

where $h > 0$ is a bandwidth parameter, and $\phi(\cdot)$ represents the standard normal PDF. The CDF of $\boldsymbol{Y}$ may be approximated with

$$\hat{F}_h(y) = \int_{-\infty}^{y} \hat{f}_h(x)dx = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{y - Y_i}{h}\right).$$

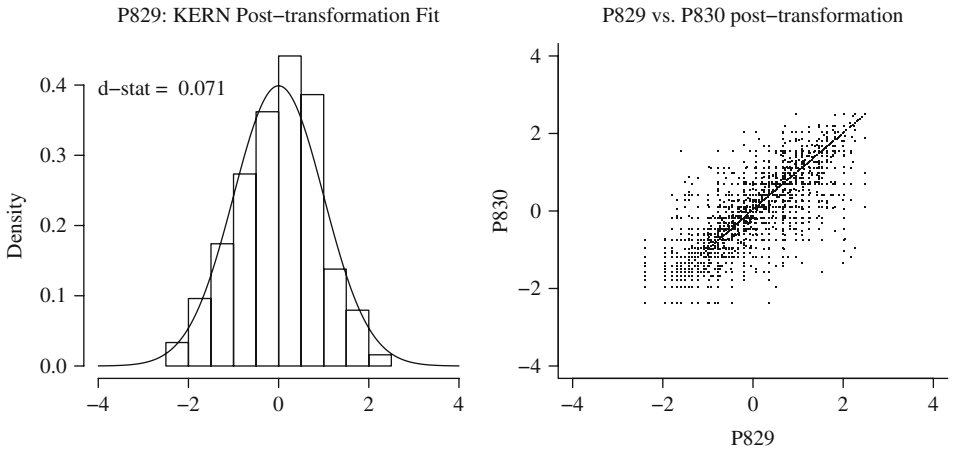


Fig. 3. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following kernel-density transformation. The left plot has the standard normal density curve overlaid

Therefore, the kernel-density transformation for $\boldsymbol{Y}$ is

$$T_2(y) = \Phi^{-1}(\hat{F}_h(y)), \tag{3}$$

and $T_2(Y_i)$ should appear to have been sampled from a standard normal distribution.

Figure 3 provides a histogram log(P829) and a scatter plot of log(P829) versus log(P830), all following the kernel-density (KERN) transformation. Clearly, the figure provides an instance where the kernel density offers a transformation to normality that is superior to that of the skew-normal family – the plots indicate that normality assumptions appear reasonable (in both the univariate and multivariate sense).

Selection of the bandwidth parameter, h, in kernel-density functions is a well-studied issue (Silverman 1986; Sheather and Jones 1991; Scott 2009). Selection algorithms often return small values of h for ARMS variables; such choices of h fail to adequately differentiate the KERN transformation from the EMP transformation described below. To avoid this issue, a bandwidth parameter of $h = 0.2$ is used whenever the KERN transformation is applied to ARMS data herein; this value offers adequate smoothing for the ARMS variables used.

3.2. *Transformation Via the Empirical Distribution*

The empirical distribution function of $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}'$ is now considered:

$$\tilde{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\},$$

where $\mathbf{1}\{A\}$ is the indicator of event A. We, however, focus on

$$\bar{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{1}\{Y_i < y\} + \frac{1}{2}\mathbf{1}\{Y_i = y\}\right),$$

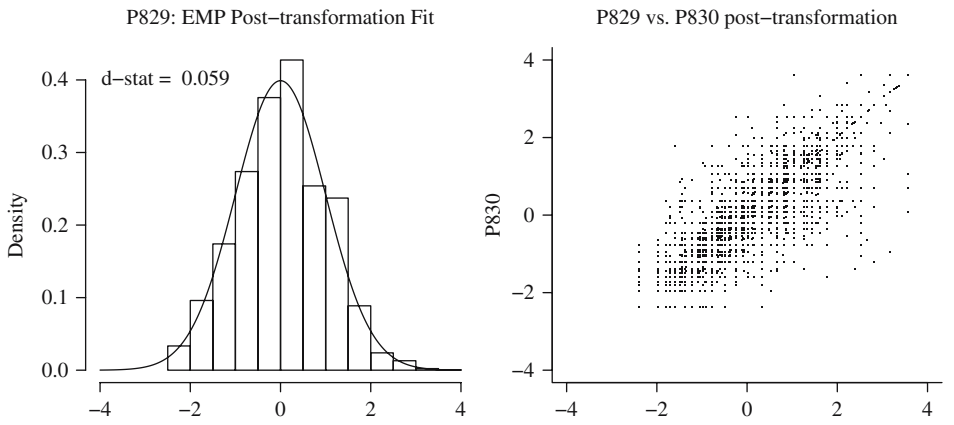


Fig. 4. Histogram of log(P829) (left) and scatter plot of log(P829) versus log(P830) (right) following empirical distribution transformation. The left plot has the standard normal density curve overlaid

since $\bar{F}(y) = lim_{h \to 0} \hat{F}_h(y)$ whenever $y \in \mathbb{R}$. Particularly, $\bar{F}(y)$ is preferable to $\tilde{F}(y)$ in cases where n is small or where $\boldsymbol{Y}$ contains repeating values (which is common for theoretically continuous portions of ARMS items). The empirical distribution (EMP) transformation for $\boldsymbol{Y}$ is

$$T_3(y) = \Phi^{-1}(\bar{F}(y)), \tag{4}$$

and $\tilde{\boldsymbol{Y}} = \{T_3(Y_1), \ldots, T_3(Y_n)\}$ should appear to have been sampled from a standard normal distribution. Note that $T_3(y)$ does not exist if $\bar{F}(y) = 0$ or 1. However, for all $y \in \boldsymbol{Y}$, $\bar{F}(y) \in (0, 1)$, meaning the observed values can be transformed via (4) without issue. Nonetheless, it is recommended to set $\bar{F}(y) = 1/(2n)$ if $y < min_i\{Y_i\}$, and $\bar{F}(y) = (2n - 1)/(2n)$ if $y > max_i\{Y_i\}$.

Figure 4 provides a histogram of P829 and a scatter plot of P829 versus P830, all following the EMP transformation. Repeating values of P829 prevent the EMP transformation from achieving exact normality. Regardless, the figure indicates that the EMP transformation is also clearly superior to the SN transformation in the circumstances illustrated here.

Since $\bar{F}(y) \xrightarrow{a.s.} F(y)$, the transformation in (4) is preferable when there is enough observed data to ensure that the empirical data provide a sufficient scope of the full distribution (including, most importantly, the tails).

3.3. Untransformation

Let $\boldsymbol{X}$ represent a transformed version of $\boldsymbol{Y}$ following application of one of the aforementioned schemes. Imputations will then be created for $\boldsymbol{X}$, resulting in $\hat{\boldsymbol{X}}$, an imputed version of the transformed data. However, the imputations must be "untransformed" (i.e., returned to their original scale). If a transformation of the type in (1) has been applied to $\boldsymbol{Y}$, the following inverse transformation may be applied to the imputed values:

$$T^{-1}(z) = F^{-1}(\Phi(z)), \quad \text{for } z \in (-\infty, \infty), \tag{5}$$

where $F^{-1}(u)$, for $u \in (0, 1)$, represents the inverse of the $F(y)$, for $y \in (-\infty, \infty)$. The CDF found using skew-normal assumptions, $F(y|\xi, \omega, \alpha)$, and the CDF found using

the kernel density, $\hat{F}_h(y)$, are both continuous and one-to-one mappings defined over $\mathbb{R}$. Thus, their respective inverses, $F^{-1}(u|\xi, \omega, \alpha)$ and $\hat{F}_h^{-1}(u)$, exist for $u \in (0, 1)$.

Let $\hat{x}$ represent an imputation for $\boldsymbol{X}$, which represents the transformed version of $\boldsymbol{Y}$. If the skew normal transformation seen in (2) was applied to this variable, $\hat{x}$ can be untransformed by calculating

$$\hat{y} = T_1^{-1}(\hat{x}) = F^{-1}(\Phi(\hat{x})|\hat{\xi}, \hat{\omega}, \hat{\alpha}),$$

and if kernel transformation seen in (3) was applied, the inversion requires the calculation of

$$\hat{y} = T_2^{-1}(\hat{x}) = \hat{F}_h^{-1}(\Phi(\hat{x})).$$

Computations involving the above two expressions (the latter, in particular) can be quite intensive.

The empirical CDF, $\bar{F}(y)$, is neither continuous nor one-to-one. Thus, its inverse, $\bar{F}^{-1}(y)$, does not exist, and (5) is not directly applicable. Hence, inversion of the empirical distribution transformation works as follows. Let $\boldsymbol{U} = \{U_1, \ldots, U_n\}$, where

$$U_i = \bar{F}(Y_i) \quad \text{for } i = 1, \ldots, n.$$

Note that the U_i should resemble uniform variates. For $\hat{x} \in (-\infty, \infty)$, let $u_x = \Phi(\hat{x})$, and after setting

$$i_x = \underset{i}{\operatorname{argmin}}|U_i - u_x|,$$

untransform imputations in variables requiring the empirical transformation by calculating

$$\hat{y} = T_3^{-1}(\hat{x}) = Y_{i_x}.$$

Inverting the empirical distribution in this manner ensures that any imputation of values in variables transformed using (4) will be sampled directly from observed values. Accordingly, an imputation method that utilizes the empirical method can be considered a "hot-deck" technique (Little 1988; Little and Rubin 2002). The EMP transformation is also advantageous due to its computational simplicity. However, the KERN transformation scheme is very demanding computationally (as it requires numeric integration).

4. Analysis of Imputed Data

The ISR algorithm of Robbins et al. (2013) is applied to the complete 2010 ARMS dataset using the imputation model described therein, where only the transformation technique is varied. For instance, five completed datasets were independently created (in the vein of multiple imputation) where the skew-normal (SN) transformation in (2) is used for all variables requiring transformation. This process is then repeated using the kernel-density (KERN) transformation in (3) and the empirical distribution (EMP) transformation in (4). Discussion is limited to the ARMS variables described in Table 1. The table also lists the number of positive and observed values (n_{obs}) and the number of missing values (n_{mis}) for each variable.

The full imputation model includes many additional variables beyond those listed in Table 1. Many of these variables contain missing values; the others are used as fully observed covariates. The respective transformation scheme is applied to all continuous or semi-continuous variables within the imputation algorithm. A list of variables included in the full model is given in Robbins et al. (2011).

To begin, analysis of marginal data characteristics of the variables (which contain missingness) in Table 1 is considered. Table 2 provides the unweighted sample mean ($\bar{x}$) and sample standard deviation (s) of the nonzero values, in addition to the between imputation variance (B) and upper bound (U_{MI}) and lower bound (L_{MI}) for the 95% confidence interval (as found using Rubin's combining rules for multiple imputation) for the population mean of each variable. The reported values of $\bar{x}$ and s represent the mean of their respective values when calculated in each of the five imputed datasets. Table 2 presents the results in "cells", where the top, middle, and bottom value in each cell is the respective estimate found using the SN, KERN and EMP transformations, respectively. Table 2 indicates that the choice of transformation method may result in differing values of means and variances. The discrepancies do not appear to be substantial, although it is noted that they are not explained by randomness in the imputations alone. Further, the lack of influence of the transformation type is likely due to relatively small missingness rates. It is also noted that other quantities (e.g., a 90% quantile) may be more heavily influenced by the transformation technique; however, the objective here is to present statistics that are of practical relevance.

To further examine marginal characteristics of imputations, discussion is now restricted to the variable P884. This variable is of particular interest because a large portion of positive and observed responses take on a single value (the specific value may not be disclosed here). This phenomenon is illustrated by the histogram of the positive and observed values of P884 which is provided in the left plot in Figure 5. The middle plot shows the positive and observed values of P884 following the EMP transformation. The plot provides visual evidence that the EMP transformation imposes "separation" between values that are frequently repeated and neighboring values. This separation ensures that there is a relatively high probability that an imputed value will equal the repeating value. For instance, 16.6% of positive and observed responses for P884 take on the frequently occurring value, and 9.3% of all EMP imputations take on that value (whereas 0% of SN and KERN imputations take on the value). The right plot in Figure 5 provides kernel-density plots of observed and imputed values (for each of the three transformation schemes), which further illustrates the need for a nonparametric transformation procedure.

There are alternative approaches for imputing P884. For instance, a three-level mixture model which includes two indicator variables (the first one indicating the occurrence of an observation equaling zero and the second indicating the observation taking on the frequently occurring value) may be more appropriate. However, such a procedure would have to enable the second indicator variable to have missing values (since it is not known whether or not the missing values of P884 take on the frequently occurring value). Therefore, the use of the marginal transformations (as opposed to higher-level mixture models) permits the convenience of a multivariate normal imputation model while producing high-quality results.

Table 2. Summary statistics for imputation of various ARMS variables. The top, middle and bottom values in each cell are calculated using SN, KERN and EMP imputations, respectively

	P828	P829	P830	P831	P832	P833	P834	P835	P884	P952
$\bar{x}$	36.42	47.37	46.45	42.41	19.84	23.21	23.39	21.51	46400	59400
	36.47	47.43	46.50	42.44	19.82	23.22	23.41	21.50	44600	59500
	36.47	47.47	46.55	42.48	19.80	23.19	23.40	21.48	45300	59300
s	0.0317	0.0406	0.0401	0.0376	0.0415	0.0492	0.0497	0.0452	6.69e6	5.70e5
	0.0316	0.0404	0.0399	0.0374	0.0415	0.0496	0.0502	0.0454	5.25e6	6.20e5
	0.0316	0.0405	0.0399	0.0375	0.0411	0.0492	0.0499	0.0450	5.44e6	5.90e5
B	6.55e-4	1.44e-3	3.43e-3	1.52e-3	2.05e-3	8.68e-3	1.33e-3	1.92e-3	6.09e5	6.71e4
	6.27e-4	2.10e-3	9.81e-4	1.57e-3	2.54e-3	2.74e-3	2.57e-3	1.53e-3	1.03e6	2.22e4
	1.74e-4	3.42e-3	1.62e-3	4.11e-3	1.31e-3	1.98e-3	1.39e-3	2.12e-3	8.92e5	3.01e4
L_{MI}	36.07	46.97	46.04	42.04	19.42	22.77	22.95	21.08	40900	57800
	36.11	47.03	46.10	42.05	19.40	22.77	22.96	21.07	39600	58000
	36.12	47.06	46.14	42.07	19.40	22.74	22.95	21.05	40200	57700
U_{MI}	36.77	47.78	46.87	42.80	20.25	23.65	23.84	21.93	51900	60900
	36.82	47.84	46.90	42.83	20.23	23.68	23.87	21.92	49700	61100
	36.82	47.89	46.95	42.88	20.21	23.63	23.84	21.91	50300	60800

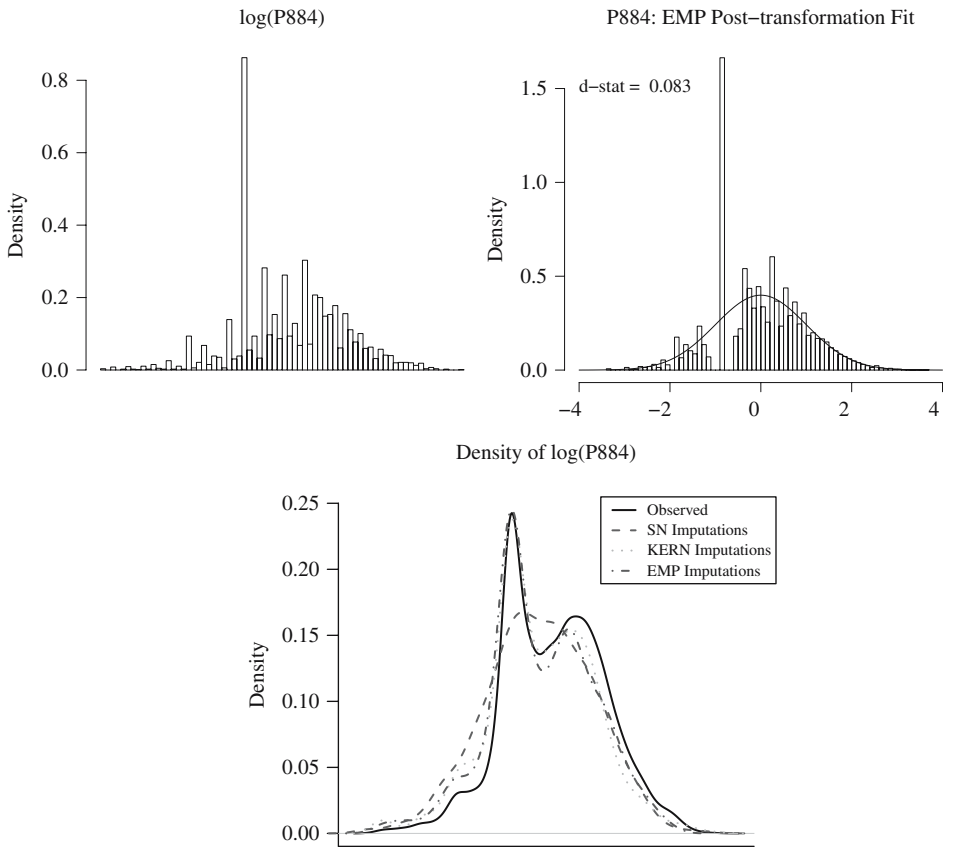


Fig. 5. Histogram of positive and observed values of log (P884) before (top left) and after (top right) an empirical transformation and densities for observed and imputed values of P884 (bottom)

To monitor the multivariate influence of imputations sampled using the various transformation schemes, consider scatter plots. Figure 6 provides scatters plots of log(P829) versus log(P830) for each of the three transformations where pairwise positive and observed pairs are marked with an '×' and imputed pairs are marked with a '+'. Lines of best fit for observed and imputed pairs are also included. Plots are given on the log scale in order to emphasize the differences between methods. The plots appear to indicate that bivariate extremes are underimputed, which may (partially) be a result of imputed values tending to be smaller than observed values for both variables in the plots. This phenomenon is to be expected for the labor variables; data indicate that "hobby" farmers, who are less likely to work on-farm full time, are more likely to refuse response for labor items. Regardless, the EMP transformation is clearly the most likely to preserve the underlying bivariate structure.

To further gauge the multivariate quality of the imputations, consider an econometric model motivated by the following. Farm operators often pursue off-farm sources of income; the on- and off-farm labor decisions of farmers have been well scrutinized in the economic literature. Economic theory suggests that the amount of time a farm operator (and the operator's spouse) choose to work on the farm is heavily influenced by factors

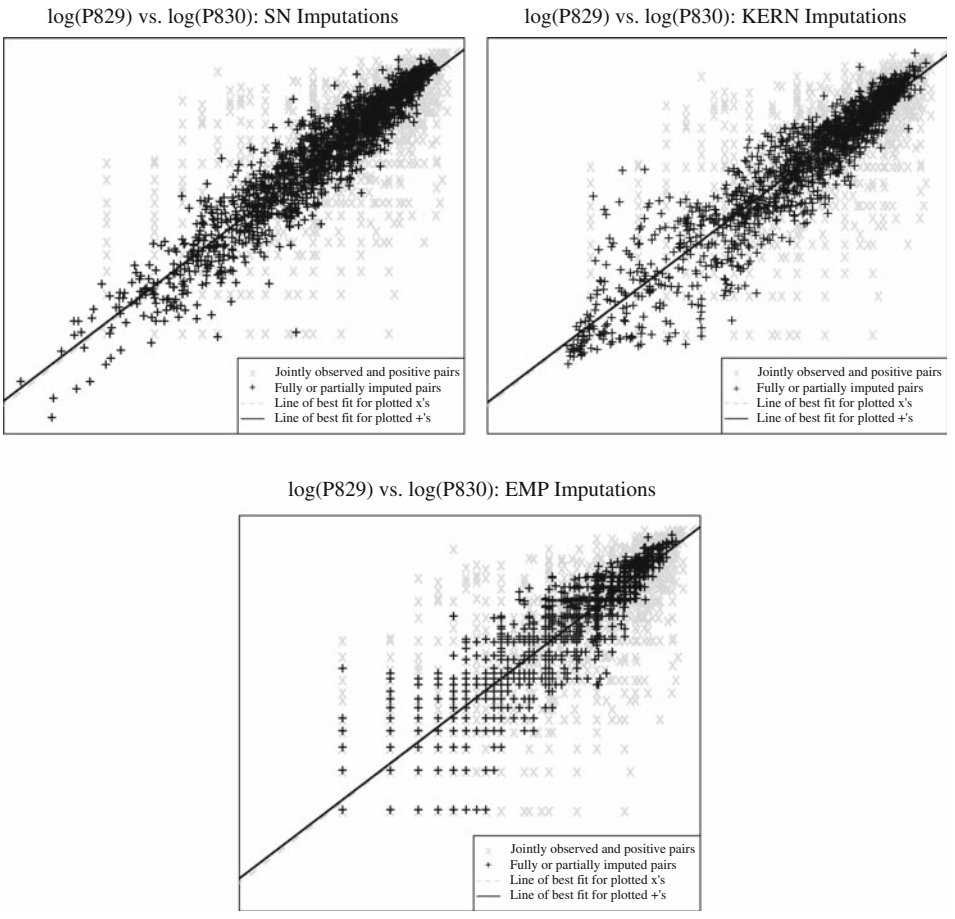


Fig. 6. Scatter plots of imputed and observed pairs of log (P829) and log (P830) for the various transformation schemes

such as the hours worked off farm by the operator and spouse, an on-farm wagerate, off-farm wage rate, the operator's age and level of education, and so forth. Econometric models investigating this concept have been considered by Huffman (1980); Sumner (1982); Huffman and Lange (1989); Mishra and Holthausen (2002) and Kwon et al. (2006) among many others. Here, consider the following linear model:

$$\text{OPHR} = \beta_0 + \beta_1 \text{OPOFFHR} + \beta_2 \text{OFFRATE} + \beta_3 \text{P1242} + \beta_4 \boldsymbol{Z} + \varepsilon, \qquad (6)$$

where $\boldsymbol{Z}$ represents a set of additional categorical covariates and ε is a mean zero error term. In the above, OPHR is the number of on-farm hours worked weekly by the farm operator (calculated as the average of P828, P829, P830 and P831). Likewise, OPOFFHR is the number of hours worked off-farm by the farm operator. OFFRATE is calculated as P952/(OPOFFHR + SPOFFHR) where SPOFFHR is the number of hours worked off farm by the operator's spouse. That is, OFFRATE represents the combined off-farm wage rate for the operator and spouse. Further, P1242 is the operator's age. Estimated values of coefficients are found using least squares while isolating to units that report nonzero values

Table 3. Summary information for the econometric model. The top, middle and bottom values in each cell are calculated using SN, KERN and EMP imputations, respectively

	β_0	β_1	β_2	β_3
Coefficient	2237	−0.432	−8.02e-3	−8.96
	2224	−0.427	−8.05e-3	−9.08
	2240	−0.433	−7.06e-3	−8.90
se(*Coef*.)	9733	2.90e-4	1.96e-6	1.575
	9688	2.89e-4	1.94e-6	1.569
	9642	2.87e-4	1.90e-6	1.559
B	775.5	2.08e-5	7.51e-7	0.1794
	843.7	3.14e-5	7.09e-7	0.2711
	876.6	1.38e-5	8.45e-7	0.1106
L_{MI}	2034	−0.467	−1.14e-2	−11.60
	2039	−0.462	−1.14e-2	−11.80
	2037	−0.468	−1.11e-2	−11.46
U_{MI}	2440	−0.397	−4.60e-3	−6.329
	2445	−0.391	−4.68e-3	−6.357
	2443	−0.400	−4.12e-3	−6.349

for all pertinent variables. A model similar to (6) which involves the hours worked on farm by the spouse was also considered throughout this study, but the findings are redundant and thereby omitted.

Table 3 provides results for these two models. The format of this table is similar to that of Table 2, as are the findings: The choice of transformation method may have a noticeable (but in this case not substantial) impact on the estimations found using econometric modeling.

5. A Simulation Study

This section presents simulation analyses which evaluate the efficacy of the proposed transformation techniques. Ideally, all assessments would be performed using real data, since synthetic data are not guaranteed to adequately mimic the complex structures encountered in practice – the motivation behind the proposed techniques is to capture such structures. Accordingly, when possible, evaluations are performed with observed ARMS data; in circumstances where such analyses do not offer sufficiently clear conclusions; a small-scale study using entirely synthetic data is used to inform the discussion.

5.1. Simulations Involving ARMS Data

A preferable technique for simulation involving real data would be to draw a sample of respondents from the observed units while treating the full dataset as a population from which population parameters can be ascertained; implementations of this scheme are seen

in Reiter (2005) and Manrique-Vallier and Reiter (2014). However, there are not enough available data for this approach to be feasible within the ARMS. ARMS data are high dimensional; nonetheless, the effective sample size (the number of positive values) can be quite small for some variables. Instead, a jackknife-type study is executed here.

As setup, a completed ARMS dataset is created using the imputation scheme outlined in Robbins et al. (2011). Specifically, the full-scale ISR algorithm and model are used in conjunction with various transformation schemes. It is not feasible to use complete cases only since there are an insufficient number of complete cases. This single completed dataset is used to create all of the benchmark values required within the simulation study. Next, missingness is randomly imposed in eight of the ARMS variables according to a probabilistic model. Imputations are then created for these newly missing values and the values of desired metrics as found using the imputed data are compared to values found using the original benchmark dataset. It is worth noting that the rate of missingness that is imposed will vastly exceed the original rate of missingness in ARMS data. The eight variables in which holes are poked are marked in Table 1 with an asterisk; some of these variables originally contained missingness, whereas others did not.

In addition to the eight variables in which missingness is imposed, there are 18 additional variables used as covariates within the imputation model for ISR. The imposed missingness is completely at random (MCAR, in the terminology of Little and Rubin 2002). Specifically, any positive value is imposed as missing with a probability of 0.5, and the occurrence of imposed missingness is independent across all values. Since the imposed rate of missingness is far higher than the missingness rate in the original dataset, the influence of imputations within the benchmark study should be filtered out. The performance of ISR with density transformations has been analyzed in great detail under other missingness mechanisms (e.g., MAR and NMAR – for details, see the supplemental material of Robbins et al. 2013). Analyses under MAR and NMAR are not expected to yield information regarding the influence of the transformation type beyond what is learned under MCAR missingness; for brevity, only MCAR is considered in these ARMS-based simulations. Since ISR is iterative (as it is a form of Markov chain Monte Carlo), each completed dataset is sampled using a burn-in period of 200 iterations.

The goal is to assess the potential for bias (in any point and interval estimates calculated from the ARMS data) caused by the choice of transformation method. The performance of the methodology is measured in terms of the relative change of a metric post imputation. Missingness is randomly imposed in the completed benchmark dataset 100 different times. Each time missingness is imposed, imputations are independently created five times (in the vein of multiple imputation) for each method. The methods used are as follows.

1. SN – The skew-normal transformation of (2) is used for all variables.
2. KERN – The kernel-density transformation of (3) is used for all variables.
3. EMP – The empirical distribution transformation of (4) is used for all variables.
4. EMPABB – EMP with an approximate Bayesian bootstrap.

The transformation schemes discussed in Section 3 will result in imputations that understate variability due to the fact each transformation scheme requires that any variable's CDF, $F(x)$, be treated as known despite the fact that $F(x)$ is, in fact, estimated.

To address this issue, Woodcock and Benedetto (2009) suggest an approximate Bayesian bootstrap (ABB), where $F(x)$ is estimated using a bootstrapped pool of observations as opposed to the actual pool of observations. Here, ABB is used together with the EMP method, resulting in EMPABB as above.

Let $\boldsymbol{\chi}$ denote the benchmark dataset, and let $\boldsymbol{\chi}_k^{[d]}$ denote the d^{th} completed dataset ($d = 1, \ldots, 5$) as imputed for the k^{th} artificially incomplete dataset (where $k = 1, \ldots, 100$). Finally, let $\theta(\cdot)$ denote a metric of interest (where the argument represents the dataset used to compute the metric). The percent change in the metric is computed via

$$\Delta\theta(k) = 100\left(\frac{\bar{\theta}_k - \theta(\boldsymbol{\chi})}{\theta(\boldsymbol{\chi})}\right),$$

where $\bar{\theta}_k = \sum_{d=1}^{5} \theta(\boldsymbol{\chi}_k^{[d]})/5$. Results are presented in the form of box plots of the 100 values of $\Delta\theta(k)$.

Metrics tracked in this simulation study include the sample mean and standard error of the sample mean as calculated over the *nonzero* values of each variable in which missingness is imposed in addition to the regression coefficients in (6) and their respective standard errors. Note that the standard error of a sample mean equals the sample standard deviation times a constant (i.e., $n_{\text{obs}}^{-1/2}$). Covariances were also monitored but yielded results that mimic those of the regression coefficients (accordingly, those results are omitted from the discussion). Confidence intervals for the sample means and regression coefficients can be calculated using Rubin's combining rules for MI, although the details are omitted here.

Findings are shown in Figure 7 for P784, P829, β_1 and β_2. The results indicate that for certain variables (e.g., P829) whose marginal distributions cannot be modeled with an appropriate parametric density, biases in basic marginal characteristics may be induced if one does not utilize a nonparametric transformation. Further, the nonparametric transformations result in imputations that appear to adequately preserve the quantities studied here (though there may be evidence of a moderate decrease in the variance of P784 caused by the nonparametric methodology). Likewise, there does not appear to be an advantage to using the EMPABB method in place of the EMP method.

Finally, since the empirical distribution transformation is designed to handle repeating values, it has the potential to be applied to variables that are binary or ordinal (though not strictly categorical with more than two categories). However, such efficacy of the transformation for such a purpose has not been thoroughly investigated.

Of interest is P784; this variable was included in this study since it has a particularly low number of positive and observed values (151 in the true dataset and thereby approximately 75 prior to imputation within the simulation study – see Table 1). Parametric and nonparametric transformations (when the former are well fit) are expected to perform equivalently on large samples (wherein sufficient data are available to adequately approximate the CDF under all transformation types); discrepancies between transformations are anticipated to be most visible when there are few observations available. To that end, it is noted that the SN transformation results in a substantially wider confidence interval for the mean of the nonzero observations of P784 (approximately three

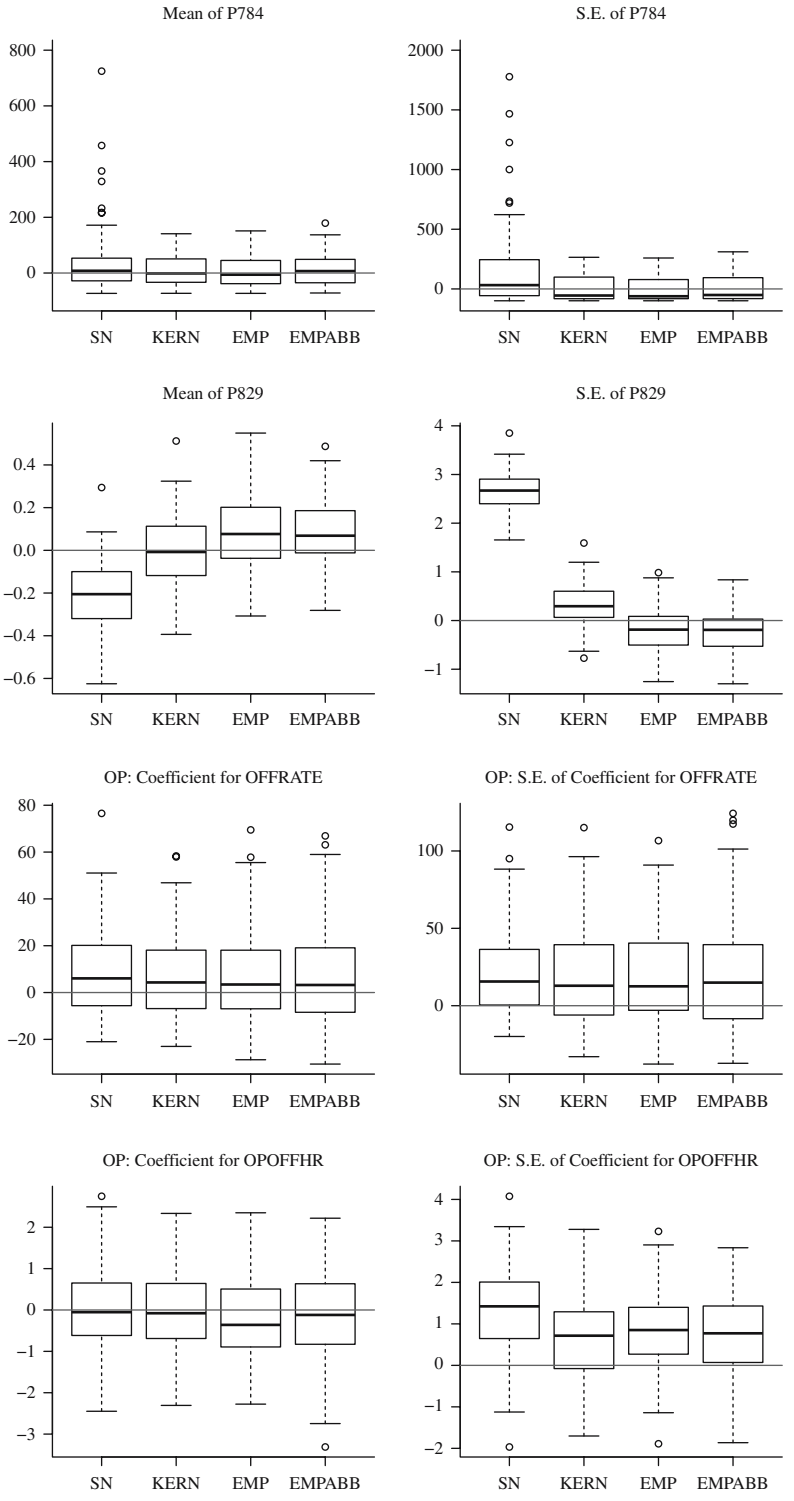


Fig. 7. Box plots of % change in various metrics

to four times wider on average than the KERN and EMP transformations) within the simulations used to generated Figure 7. Since the SN model seems appropriate for this variable (the KS test yields a p-value of 0.801 when a skew-normal distribution is assumed), since it seems unreasonable to assume that 75 observations can sufficiently quantify a CDF, and since Figure 7 implies that the nonparametric transformations may decrease the variance of this variable, it is suggested that the SN transformation is more appropriate than the nonparametric transformations for P784.

Ideally, comparisons to predictive mean matching (PMM, Little 1988) could have been presented in this study. PMM is a popular technique that builds a predictive model for imputations through regression, and then samples imputations from observed data – making it similar to (and useful in the same settings as) the methods presented here. However, direct comparisons to PMM within the simulations above (wherein such comparisons would be most useful due to the unknown distributional structure of ARMSdata) cannot be made here due to computational constraints. For instance, one iteration of ISR takes 1.15 seconds, and one iteration of MICE with PMM takes 15 minutes when run on the group of variables used above. These computations are executed on a 64-bit Windows machine with a 3.3 GHz processor and 8.0 GB of RAM.

To summarize, the above study helps to verify the efficacy of the proposed methodology on real data, but it has some notable shortcomings. For instance, it is desirable to investigate the comparative performance of the proposed techniques against other methods such as PMM, and to present results for a variety of missingness structures. Many of these shortcomings are the consequence of computational issues. Furthermore, the above simulations leave unanswered the question as to whether or not a parametric transformation is preferable in settings involving small samples. A small-scale study involving fully synthetic data is thus presented below.

5.2. *Simulations Involving Synthetic Data*

The small scale of the following simulation study (only two variables are used for various sample sizes) makes it computationally feasible to consider a variety of methods and missingness mechanisms. Specifically, the four transformation techniques mentioned above (SN, KERN, EMP, and EMPABB) are used in conjunction with ISR. As needed, skew-normal MLEs are used, and the kernel bandwidth parameter is estimated via the method of Sheather and Jones (1991). Further, PMM is considered (while used in conjunction with `mice`) as well as IRMI (Templ et al. 2011); no transformation is used when these methods are applied.

Data are generated as follows. Let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ represent a random sample from a skew-normal distribution with parameters $\xi = 4$, $\omega = 2$ and $\alpha = -2$. Additionally, let $\tilde{\boldsymbol{X}} = \{\tilde{X}_1, \ldots, \tilde{X}_n\}$ represent the version of $\boldsymbol{X}$ that has been transformed in accordance with (2) while using the true parameter values, and define $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$, where $Y_i = 1 + 0.5\tilde{X}_i + \varepsilon_i$ for $i = 1, \ldots, n$, and where $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is a random sample of length n from a standard normal distribution.

Missingness is imposed in the values of $\boldsymbol{X}$ through the following mechanisms. Under MCAR missingness, each observation of $\boldsymbol{X}$ is missing with probability 0.5. For MAR missingness, X_i is missing with a probability equal to $1/(1 + exp(-\tilde{Y}_i))$, where $\tilde{Y}_i$

represents a standardized version of Y_i. NMAR missingness was also considered, but the results are excluded for brevity since they provided no additional information regarding the choice of transformation scheme beyond what is learned from the other mechanisms. Imputations in $\boldsymbol{X}$ are generated via the techniques mentioned above; the elements of $\boldsymbol{Y}$ are not transformed at any point. Further, $m = 5$ imputed datasets are created, and no burn-in period is necessary since missingness is restricted to one variable. MI point and interval estimates are generated for a handful of parameters, and the entire process is replicated independently 1,000 times for various values of n.

For a given imputation method, missingness mechanism, and value of n, let $\hat{\theta}_k$ denote the MI point estimate of a generic parameter θ calculated following the k^{th} replication ($k = 1, \ldots, 1,000$). The percent bias in the multiple imputation estimate of θ is approximated by calculating $\bar{\Delta}\theta = 100\sum_{k=1}^{1,000}[(\hat{\theta}_k/\theta) - 1]/1,000$. Similarly, the sequence of 1,000 values of $\hat{\theta}_k$ can be tested to see if the percent bias is statistically nonzero. Further, the empirical coverage of the MI interval estimate of θ is calculated via the portion of the 1,000 replications in which the true value of θ is contained within its 95% confidence interval.

First, we consider the basic univariate parameters $\mu = \mathrm{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$; results are given in Table 4. All transformation methods offer strong performance in terms of bias and coverage for these parameters, as does the PMM procedure. However, the IRMI procedure shows some evidence of bias and observes poor coverage for these simple quantities. It appears that all methods induce a small amount of bias (which mostly disappears with increasing n) under MAR missingness; the fact that this bias tends to be negative is a consequence of the form of the function that generates the MAR missingness. Moreover, the results imply that the use of the approximate Bayesian bootstrap does not improve the results. Finally (and most importantly), all transformation schemes appear to offer equivalent performance.

In order to provide parallels to the log-skew-normal distributions that positive portions of ARMS data observe, we also study summary statistics of the transformed variable $U_i = \exp(X_i)$. Specifically, we use multiple imputation to develop point and interval estimates of $\gamma = \mathrm{E}[U_1]$ and $\nu^2 = \mathrm{Var}(U_1)$ by applying Rubin's combining rules to the sequence $\{\hat{U}_1, \ldots, \hat{U}_n\}$, where $\hat{U}_i$ represents a version of U_i that contains imputations of missingvalues. The ability of an imputation algorithm to preserve such quantities is a strong indication that the distribution of the imputed data matches that of the actual data had they been fully observed (since γ and ν^2 follow from the specificform of the MGF of X_1). Results for these two quantities are shown in Table 5. The table indicates that IRMI imputations provide biased estimates of γ and ν^2 under all missingness mechanisms. This observation is not surprising, since IRMI does not take steps to ensure that the full distributional structure is captured in the imputation process. Although all methods are more imprecise in their estimation of γ and ν^2 than of μ and σ^2, Tables 4 and 5 both yield the same conclusions regarding the comparative performance of the techniques.

In summary, the key findings of the simulation studies presented in this subsection are that all methods involving transformation are comparable to PMM and that the choice of transformation technique does not have a significant influence on bias or coverage probabilities. The latter finding is noteworthy because the SN method, which is ideally suited to this setting, shows no gains over the nonparametric methods, whereas the

Table 4. Empirical bias and coverage probabilities (the latter are in parentheses) of the point estimates and 95% confidence intervals (found using MI) of two parameters involving the synthetic random variable X. An asterisk indicates that the bias is statistically nonzero at the 0.01 significance level

		% bias and % coverage for $\mu = \mathrm{E}[X]$					
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	0.51 (91.3)	0.45 (91.9)	−0.16 (91.2)	−0.08 (90.6)	−0.34 (90.5)	2.22* (68.7)
	100	0.27 (93.2)	0.12 (93.7)	0.25 (93.5)	0.71* (92.4)	0.18 (93.2)	1.95* (70.4)
	250	−0.15 (92.3)	0.15 (93.5)	0.00 (94.6)	0.26 (94.0)	−0.15 (95.1)	1.66* (67.7)
	500	0.02 (93.7)	−0.16 (95.3)	−0.16 (93.3)	−0.03 (94.7)	0.12 (94.0)	1.54* (68.9)
	1,000	−0.02 (93.7)	0.01 (94.1)	−0.10 (94.3)	0.02 (94.5)	0.01 (95.2)	1.49* (62.0)
	2,500	0.09 (94.7)	0.00 (94.5)	0.01 (94.0)	−0.04 (94.4)	0.09 (93.4)	1.50* (54.2)
MAR	50	−2.37* (90.1)	−3.42* (89.4)	−3.13* (90.0)	−2.12* (90.3)	−2.19* (90.0)	−0.65 (68.3)
	100	−1.31* (92.0)	−1.40* (89.4)	−2.02* (89.2)	−1.25* (91.8)	−2.33* (91.7)	0.21 (66.2)
	250	−0.95* (91.2)	−0.59* (93.5)	−0.50* (92.6)	−0.55* (91.6)	−0.72* (91.5)	0.96* (67.2)
	500	−0.41* (91.5)	−0.28 (92.3)	−0.35* (91.7)	−0.38* (91.9)	−0.34* (91.4)	0.94* (63.5)
	1,000	−0.31* (91.6)	−0.26* (93.5)	−0.39* (93.2)	−0.19 (92.9)	−0.36* (92.4)	1.27* (62.3)
	2,500	−0.09 (92.8)	−0.11 (93.0)	0.04 (92.1)	0.05 (92.0)	−0.19* (91.3)	1.29* (55.6)
		% bias and % coverage for $\sigma^2 = \mathrm{Var}(X)$					
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−5.85* (81.9)	−6.60* (81.6)	−5.74* (82.1)	−7.50* (78.7)	−4.78* (81.2)	−40.4* (35.4)
	100	−2.95* (85.0)	−3.26* (86.2)	−3.07* (86.6)	−5.20* (84.6)	−3.29* (85.2)	−40.3* (18.8)
	250	−1.44* (87.5)	−1.24* (88.8)	−1.19* (88.3)	−2.12* (86.6)	−1.21* (91.0)	−40.0* (2.90)
	500	0.01 (91.3)	−0.45 (90.7)	−0.12 (91.6)	−0.70 (90.1)	−0.94* (89.2)	−39.9* (0.20)
	1,000	−0.20 (90.3)	0.01 (91.6)	−0.49 (92.5)	−0.27 (90.6)	−0.28 (91.8)	−40.5* (0.00)
	2,500	−0.22 (91.0)	−0.05 (91.9)	−0.11 (91.3)	0.07 (90.9)	−0.21 (92.6)	−40.2* (0.00)
MAR	50	−5.96* (83.0)	−6.57* (82.3)	−7.05* (80.4)	−9.31* (75.2)	−7.95* (78.7)	−38.8* (38.8)
	100	−3.63* (86.9)	−3.93* (85.5)	−4.26* (85.3)	−4.97* (83.0)	−4.25* (85.3)	−38.7* (24.2)
	250	−2.10* (88.0)	−1.83* (87.6)	−1.48* (88.0)	−1.70* (87.1)	−0.90 (88.6)	−38.2* (6.50)
	500	−1.39* (87.1)	−1.02* (87.4)	−1.50* (88.3)	−1.42* (86.7)	−0.69 (86.5)	−37.7* (0.40)
	1,000	−0.37 (90.0)	−0.68* (88.4)	−0.36 (87.5)	−0.69* (86.8)	−0.91* (85.5)	−37.0* (0.00)
	2,500	−0.24 (90.0)	−0.45* (90.2)	−0.37 (89.1)	−0.50* (88.3)	−0.17 (87.7)	−37.2* (0.00)

Table 5. Empirical bias and coverage probabilities (in parentheses) of the point estimates and 95% confidence intervals of two parameters involving the synthetic random variable $U = \exp(X)$. An asterisk indicates that the bias is statistically nonzero at the 0.01 significance level

% bias and % coverage for $\gamma = \mathrm{E}[U]$ where $U = \exp(X)$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−0.34 (84.2)	−0.44 (82.9)	−1.67 (82.5)	−2.97* (80.6)	−0.11 (81.8)	−18.5* (53.2)
	100	0.30 (86.5)	−0.76 (86.8)	−0.38 (88.1)	−0.30 (86.9)	−0.29 (86.6)	−19.5* (48.8)
	250	−0.89 (89.0)	0.00 (90.0)	−0.32 (89.9)	−0.36 (88.0)	−0.92 (88.8)	−20.3* (29.8)
	500	0.25 (90.8)	−0.39 (92.1)	0.05 (92.2)	−0.41 (89.0)	0.03 (92.3)	−20.9* (12.0)
	1,000	0.01 (91.8)	0.26 (91.0)	0.14 (93.1)	−0.02 (92.0)	0.01 (93.2)	−21.3* (1.40)
	2,500	0.11 (92.1)	0.01 (91.8)	0.18 (93.6)	−0.09 (92.5)	0.14 (93.2)	−21.1* (0.00)
MAR	50	−7.94* (73.1)	−10.1* (73.1)	−8.67* (70.6)	−8.65* (71.7)	−8.88* (72.2)	−25.3* (40.6)
	100	−4.88* (79.8)	−5.23* (77.6)	−6.99* (74.3)	−5.41* (75.1)	−7.16* (75.4)	−26.0* (28.7)
	250	−3.89* (78.9)	−3.11* (81.3)	−1.86* (80.8)	−2.66* (79.9)	−1.94* (81.7)	−25.4* (14.5)
	500	−1.81* (81.3)	−1.76* (83.3)	−1.53* (80.4)	−2.51* (81.8)	−1.38* (83.0)	−25.8* (4.70)
	1,000	−1.39* (84.3)	−1.12* (83.7)	−1.52* (84.1)	−1.02* (82.2)	−1.88* (82.1)	−25.1* (0.70)
	2,500	−0.36 (86.4)	−0.76* (84.7)	−0.30 (84.9)	−0.45 (83.2)	−0.78* (84.8)	−25.3* (0.00)
% bias and % coverage for $\nu^2 = \mathrm{Var}(U)$ where $U = \exp(X)$							
	n	SN	KERN	EMP	EMPABB	PMM	IRMI
MCAR	50	−11.1* (43.2)	−5.13 (43.8)	−8.34 (44.2)	−17.8* (34.1)	1.06 (43.9)	−45.6* (4.50)
	100	−2.55 (52.2)	−3.71 (48.3)	−3.12 (51.5)	−7.06 (43.1)	−0.97 (50.4)	−44.1* (1.80)
	250	−5.22 (57.5)	1.46 (58.0)	−1.35 (56.1)	−8.18* (52.3)	−4.34 (56.6)	−43.6* (0.80)
	500	−0.82 (63.3)	−0.73 (64.0)	0.68 (62.4)	−5.05* (55.5)	2.40 (62.2)	−45.2* (0.30)
	1,000	2.74 (64.5)	2.11 (66.9)	1.30 (66.7)	−1.57 (63.5)	−0.36 (67.8)	−46.0* (0.10)
	2,500	−0.74 (72.1)	−0.54 (71.1)	−0.08 (72.0)	−0.18 (65.4)	0.01 (71.0)	−44.9* (0.00)
MAR	50	−26.1* (34.5)	−29.3* (33.9)	−22.1* (32.8)	−29.1* (25.0)	−33.4* (33.2)	−53.5* (5.40)
	100	−19.4* (41.9)	−20.0* (39.5)	−23.4* (38.4)	−17.9* (35.0)	−20.2* (39.4)	−63.1* (1.60)
	250	−15.1* (45.0)	−12.5* (45.0)	−6.10 (45.6)	−16.8* (43.9)	−7.43 (49.8)	−63.5* (0.10)
	500	−6.26 (49.7)	−9.49* (52.4)	−2.86 (52.0)	−15.6* (45.9)	−4.77 (51.1)	−65.3* (0.10)
	1,000	−8.89* (53.3)	4.87 (52.3)	−6.85* (54.3)	−6.41* (49.2)	−10.2* (50.5)	−64.7* (0.00)
	2,500	0.06 (56.4)	−4.06* (58.2)	−1.50 (57.9)	−5.23* (51.6)	−4.17 (57.4)	−65.3* (0.00)

nonparametric methods will certainly provide higher efficacy in settings where the skew-normal assumption is violated. Figure 7 shows that the nonparametric methods yield a decrease in the variance of P784, and Tables 4 and 5 implicate that all methods may have decreased variability in items with small sample sizes. This decrease is not seen by the SN method in Figure 7, perhaps because the skew-normal distribution does not adequately capture the tails of the distribution of P784 (which also helps to explain the outlying values for this variable and transformation method in Figure 7).

6. Comments

Nonparametric transformation of survey data prior to imputation provides a straightforward manner through which unique marginal data characteristics can be preserved throughout the imputation process – such transformations are also shown to maintain multivariate aspects. The empirical transformation described above has the added advantage that imputations are drawn from observed data, which makes a method that utilizes it a nearest neighbor-type technique, and which also increases the probability that complex underlying data structures (that are common in establishment surveys) are maintained. Further, the empirical transformation is advantageous due to its computational simplicity.

The evaluations presented in this article did not unveil circumstances in which a transformation based upon a parametric model (i.e., the skew-normal distribution) is clearly preferable to the nonparametric methods. Further, no settings were found in which a transformation based upon a kernel density outperformed the transformation based upon an empirical distribution – the latter is more computationally efficient. In light of the above, the recommendation is that in practical circumstances the empirical distribution transformation be used when possible (however, further evaluations beyond those presented here may be needed to support this conclusion). With any transformation method, the practitioner should always investigate the validity of the posttransformation multivariate model (a joint normal distribution was used here) prior to generating imputations.

As an additional comment, it is noted that the nonparametric methods are applied here while exclusively using ISR (Robbins et al. 2013). ISR has the restriction that variables with missing values be sampled from continuous distributions. However, the nonparametric transformations are applicable in conjunction with any imputation technique which applies normality assumptions to continuous variables. For instance, these transformations could be employed with IVEware (Raghunathan et al. 2002) or MICE (Van Buuren and Oudshoorn 1999), which include capabilities for imputation of categorical variables.

Furthermore, it is also possible to use the methods discussed here for simulation of fully or partially synthetic datasets for the purposes of data confidentiality (Rubin 1993; Reiter 2002; Raghunathan et al. 2003). Woodcock and Benedetto (2009) use a kernel-density transformationfor this purpose, and it is noted that the empirical transformation has such utility if it is acceptable for synthetic values to be sampled from the observed data.

Finally, we note that one may use the EMP transformation technique for imputation of ordinal or binary variables (though not for categorical variables with more than two categories) since the method samples imputations from the set of observed values. However, the performance of the EMP method for this purpose has not yet been examined thoroughly.

7. References

Azzalini, A. 1985. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* 12: 171–178.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society* Series B 39: 1–38.

Fay, R.E. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91: 490–498. DOI: http://dx.doi.org/10.1080/01621459.1996.10476909.

Huffman, W.E. 1980. "Farm and Off-Farm Work Decisions: The Role of Human Capital." *Review of Economics and Statistics* 62: 14–23.

Huffman, W.E., and M.D. Lange. 1989. "Off-Farm Work Decisions of Husbands and Wives: Joint Decision Making." *The Review of Economics and Statistics* 71: 471–480. DOI: http://dx.doi.org/10.2307/1926904.

Javaras, K.N., and D.A. van Dyk. 2003. "Multiple Imputation for Incomplete Data with Semicontinuous Variables." *Journal of the American Statistical Association* 98: 703–715. DOI: http://dx.doi.org/10.1198/016214503000000611.

Kim, J.K., J.M. Brick, W.A. Fuller, and G. Kalton. 2006. "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Society* Series B 68: 509–521. DOI: http://dx.doi.org/10.1111/j.1467-9868.2006.00546.x.

Kott, P.S. 1995. *A Paradox of Multiple Imputation*. Tech. rep., National Agricultural Statistics Service, Fairfax, VA. Presented at the Joint Statistical Meetings, August 1995, Orlando, FL

Kott, P.S., and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105: 1265–1275. DOI: http://dx.doi.org/10.1198/jasa.2010.tm09016.

Kwon, C.-W., P. Orazem, and D.M. Otto. 2006. "Off-Farm Labor Supply Responses to Permanent and Transitory Farm Income." *Agricultural Economics* 34: 59–67. DOI: http://dx.doi.org/10.1111/j.1574-0862.2006.00103.x.

Little, R.J.A. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 6: 287–296. DOI: http://dx.doi.org/10.1080/07350015.1988.10509663.

Little, R.J.A., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Manrique-Vallier, D., and J.P. Reiter. 2014. "Bayesian Multiple Imputation for Large-Scale Categorical Data With Structural Zeros." *Survey Methodology* 40: 125–134.

Miller, D., M. Robbins, and J. Habiger. 2010. "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey." In Proceedings of the JSM, Section on Survey Research Methods: American Statistical Association. Alexandria, VA, 816–823.

Mishra, A.K., and D.M. Holthausen. 2002. "Effect of Farm Income and Off-Farm Wage Variability on Off-Farm Labor Supply." *Agricultural and Resource Economics Review* 31: 187–199.

National Research Council. 2008. *Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey*. Washington, D.C.: The National Academies Press.

Nelsen, R.B. 2009. *An introduction to Copulas*. New York: Springer.

Raghunathan, T., J. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27: 85–95.

Raghunathan, T.E., P.W. Solenberger, and J. van Hoewyk. 2002. *Iveware: Imputation and Variance Estimation Software*. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Raghunathan, T., J. Reiter, and D. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16.

Reiter, J.P. 2002. "Satisfying Disclosure Restrictions With Synthetic Data Sets." *Journal of Official Statistics* 18: 531–544.

Reiter, J.P. 2005. "Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study." *Journal of the Royal Statistical Society* Series A 168: 185–205. DOI: http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

Robbins, M.W., S.K. Ghosh, B. Goodwin, J.D. Habiger, D. Miller, and T.K. White. 2011. *Multivariate Imputation Methods for Addressing Missing Data in the Agricultural Resource Management Survey (ARMS)*. A NISS/NASS collaborative research project, National Agricultural Statistics Service/National Institute of Statistical Sciences.

Robbins, M.W., and T.K. White. 2011. "Farm Commodity Payments and Imputation in the Agricultural Resource Management Survey." *American Journal of Agricultural Economics* 93: 606–612. DOI: http://dx.doi.org/10.1093/ajae/aaq166.

Robbins, M.W., S.K. Ghosh, and J.D. Habiger. 2013. "Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey." *Journal of the American Statistical Association* 108: 81–95. DOI: http://dx.doi.org/10.1080/01621459.2012.734158.

Robbins, M.W., and T.K. White. Forthcoming. "Direct Payments, Cash Rents, Land Values, and the Effects of Imputation in U.S. Farm-Level Data." *Agricultural and Resource Economics Review*.

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

Rubin, D.B. 1993. "Discussion of Statistical Disclosure Limitation." *Journal of Official Statistics* 9: 461–468.

Rubin, D.B. 1996. "Multiple Imputation After 18 + Years." *Journal of the American Statistical Association* 91: 473–489. DOI: http://dx.doi.org/10.1080/01621459.1996.10476908.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall/CRC.

Scott, D.W. 2009. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Vol. 383. New York: Wiley.

Sheather, S.J., and M.C. Jones. 1991. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation." *Journal of the Royal Statistical Society* Series B 53: 683–690.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. Vol. 26. New York: CRC Press.

Su, Y.-S., M. Yajima, A.E. Gelman, and J. Hill. 2011. "Multiple Imputation with Diagnostics (mi) in r: Opening Windows into the Black Box." *Journal of Statistical Software* 45: 1–31.

Sumner, D.A. 1982. "The Off-Farm Labor Supply of Farmers." *American Journal of Agricultural Economics* 64: 499–509. DOI: http://dx.doi.org/10.2307/1240642.

Templ, M., A. Kowarik, and P. Filzmoser. 2011. "Iterative Stepwise Regression Imputation Using Standard and Robust Methods." *Computational Statistics & Data Analysis* 55: 2793–2806. DOI: http://dx.doi.org/10.1016/j.csda.2011.04.012.

U.S. Department of Agriculture. 2011. *Farm Production Expenditures 2010 Summary*. Washington, D.C.

Van Buuren, S., and C.G.M. Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. Leiden: TNO Preventie en Gezondheid. For associated software see http://www.multiple-imputation.com (accessed October 21, 2014).

Woodcock, S.D., and G. Benedetto. 2009. "Distribution-Preserving Statistical Disclosure Limitation." *Computational Statistics and Data Analysis* 53: 4228–4242. DOI: http://dx.doi.org/10.1016/j.csda.2009.05.020.

Received November 2012
Revised September 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 701–719, http://dx.doi.org/10.2478/JOS-2014-0044

Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey

Morgan Earp[1], *Melissa Mitchell*[2], *Jaki McCarthy*[3], *and Frauke Kreuter*[4]

Increasing nonresponse rates in federal surveys and potentially biased survey estimates are a growing concern, especially with regard to establishment surveys. Unlike household surveys, not all establishments contribute equally to survey estimates. With regard to agricultural surveys, if an extremely large farm fails to complete a survey, the United States Department of Agriculture (USDA) could potentially underestimate average acres operated among other things. In order to identify likely nonrespondents prior to data collection, the USDA's National Agricultural Statistics Service (NASS) began modeling nonresponse using Census of Agriculture data and prior Agricultural Resource Management Survey (ARMS) response history. Using an ensemble of classification trees, NASS has estimated nonresponse propensities for ARMS that can be used to predict nonresponse and are correlated with key ARMS estimates.

Key words: Nonresponse bias; propensity scores; classification trees; ensemble trees.

1. Introduction

In many ongoing surveys, response rates are declining or now require more resources to maintain (Curtin et al. 2005; Groves and Couper 1998; Stussman et al. 2005; Brick and Williams 2009). Reduced response rates can lead to nonresponse bias when response propensities are correlated with characteristics of interest (i.e., something we are trying to measure) or vary by subdomain (Wagner 2012). This can be a particularly serious problem for establishment surveys, because unlike household surveys, many establishment population distributions are highly skewed (Petroni et al. 2004). Thus severe nonresponse bias can occur if sample units that contribute to the estimates more than others are less likely to respond (Groves et al. 2002; Phipps and Toth 2012; Powers et al. 2006; Thompson 2009). For example, according to the 2007 Census of Agriculture, only 0.3 percent of farms had total sales of five million dollars or more, but they accounted for

[1] Bureau of Labor Statistics – Office of Survey Methods Research, PSB Suite 1950, 2 Massachusetts Avenue, NE Washington District of Columbia 20212, U.S.A. Email: earp.morgan@bls.gov
[2] USDA – National Agricultural Statistics Service, Fairfax, Virginia, U.S.A. Email: Melissa.Mitchell@nass.usda.gov
[3] USDA – National Agricultural Statistics Service, Fairfax, Virginia, U.S.A. Email: Jaki.McCarthy@nass.usda.gov
[4] University of Maryland – JPSM, 1218 Lefrak Hall, College Park, MD 20742, Maryland 20742, U.S.A. Email: fkreuter@survey.umd.edu
Acknowledgments: We would like to thank the Guest Editor and Associate Editor for the Special Issue of papers from ICES-IV, and the anonymous reviewers for their comments.

27.9 percent of total sales in the U.S. (U.S. Department of Agriculture 2007). If these operations failed to respond, it would greatly impact the estimates of total sales (and items strongly related to total sales).

Traditionally, survey methodologists use two approaches for dealing with nonresponse error. One focuses on increasing participation, for example through incentives, notification letters, or personal enumeration (Dillman 1978; Groves and Couper 1998), whereas the other tries to address potential nonresponse bias through weighting adjustment (Kalton and Flores-Cervantes 2003) or imputation (Little and Rubin 2002). However, both approaches have drawbacks. Extra efforts are costly and increased response rates could mean that only more of the same types of establishments are brought into the respondent pool, leaving nonresponse bias unchanged, or worse, increased (Groves 2006). Weighting adjustments or calibration to known population totals can be effective at reducing the nonresponse bias of the variables used in the calibration models (or for variables that are highly correlated with these covariates), but may fail to address potential nonresponse bias in other key estimates in large multipurpose surveys (Earp et al. 2010). Likewise, business programs that use imputation instead of weighting to account for unit nonresponse, for example through the use of administrative data, can induce additional bias if the imputation models are poor (Luzi et al. 2007) or if a high proportion of units in a subdomain of the imputation base have missing data (Thompson and Washington 2013).

In response to these drawbacks, survey methodologists recently started employing responsive design strategies (Groves and Heeringa 2006) that tailor fieldwork efforts to respondents with different response propensities. In order to do this, it is particularly important to identify and target the low propensity groups whose nonresponse is most likely to induce nonresponse bias and to increase their response rates. Successful examples of such approaches are the National Survey of Family Growth (Axinn et al. 2011), and several CATI surveys done by Statistics Canada (Mohl and Laflamme 2007; Laflamme and Karaganis 2010).

In this article, we present a study that assesses a method for identifying such low response propensity groups in a large-scale farm survey. Specifically, we present an application that uses an ensemble of classification trees to model establishment survey nonresponse on the Agricultural Resource Management Survey (ARMS) in relation to multiple farming characteristics collected by the 2002 Census of Agriculture. Both the Census of Agriculture and ARMS are conducted by the National Agricultural Statistics Service (NASS), making it easy to link the data and use Census data as a proxy both in modeling characteristics of nonresponse and assessing the relationship between modeled nonresponse propensity scores and nonresponse bias. To evaluate our proposed methods, we linked units from a later ARMS sample (containing missing values) to their 2007 Census of Agriculture data on numerous common agricultural characteristics using the Census data as a proxy for ARMS respondent and nonrespondent characteristics. Consequently, we can compare the relative difference of the mean between respondents and nonrespondents on several key estimates across varying response propensity groups.

In Section 2, we provide background information on the ARMS data used in the case study. In Section 3, we introduce the classification tree methodology and describe its application to the ARMS data. Section 4 presents our results. We conclude in Section 5 with a brief discussion and ideas for future research.

2. Background on the ARMS

The ARMS collects calendar year economic data from agricultural producers nationwide that describe the financial performance and operational characteristics of farm households. These data are used to inform the U.S. Farm Bill and are used extensively for analysis by the United States Department of Agriculture's Economic Research Service (ERS) to understand the financial performance and household characteristics of farms. The ARMS is conducted in three phases. Phase I screens for potential samples for Phases II and III using a mail questionnaire. Phase II collects data on cropping practices and agricultural chemical usage. Phase III (also referred to as ARMS III) collects detailed economic information about the agricultural operation as well as the operator's household. ARMS III data (referred to as ARMS from here on) are primarily collected through personal interviews and mail questionnaires. There are multiple versions of the ARMS questionnaire. Some versions are administered by mail and personal interview and some by personal interview only. In addition, there are several commodity-specific versions of the questionnaire that vary by year: Examples include organic agriculture, apples, and poultry.

The ARMS is a probability sample, drawn from both a list and an area frame. The list frame is stratified by farm total sales, farm type, and farm region; the area frame is stratified by land use (U.S. Department of Agriculture 2012). Units are selected from the stratified frame using a Sequential Interval Poisson (SIP) design. By utilizing SIP, NASS is able to decrease the probability of sample selection for operations previously sampled for ARMS and other NASS surveys and thus reduce respondent burden (Miller et al. 2010). Note that sample design weights were not used in the creation of the tree models discussed in Subsection 3.2.1, since the purpose was to model the expected response rates specifically for ARMS using the ARMS sample design, and not to model the expected response rates for the entire originating population of farms (Phipps and Toth 2012).

ARMS response rates (see Table 1) have been fairly stable over the years but consistently fall below the target of 80 percent; studies below 80 percent are required to complete a nonresponse bias analysis according to the standards issued by the U.S. Office of Management and Budget in 2006 (United States Executive Office of the President 2006).

Table 1. ARMS response rates 2000–2008

Year	Sample size	Response rate (%)
2000	17,903	63
2001	13,313	64
2002	18,219	74
2003	33,861	63
2004	33,908	68
2005	34,937	71
2006	34,203	68
2007	31,924	70
2008	36,388	66

3. Methodology

3.1. Classification and Ensemble Trees

Often, logistic regression models are used to relate covariates to nonresponse and to compare response rates across subgroups (Axinn et al. 2011; Johansson and Klevmarken 2008; Johnson et al. 2006; Abraham et al. 2006; Little and Vartivarian 2005; Nicoletti and Peracchi 2005; Lepkowsi and Couper 2002; Little 1986; Rosenbaum and Rubin 1983). In many applications, however, classification trees are considered easier to specify and interpret, specifically with regard to interaction effects (Phipps and Toth 2012; Schouten 2007; Schouten and de Nooij 2005). Moreover, classification tree models also have the added benefits of

1. automatically detecting significant relationships and interaction effects without prespecification, thus reducing the risk of selecting the wrong variables or other model specification errors;
2. identifying the variables that are correlated with the target variable, along with the optimal breakpoints within these variables for maximizing their correlation;
3. identifying hierarchical interaction effects across numerous variables and summarizing them using a series of simple rules;
4. incorporating missing data into the model and assessing whether missingness on a given variable is related to the target; and
5. creating a series of simple rules that are easy to interpret and use for identifying and describing subgroups with higher propensities.

While using classification trees provides some advantages over logistic regression, the results from a single tree are also considered to be potentially unstable. Therefore, it is recommended that trees be modeled and validated using independent data. As is typical in classification tree modeling, the dataset used in the creation of our trees was randomly split into three independent sets. An initial training subset of the data is used to grow the tree, and an independent subset is subsequently used to validate the results of the initial model. Finally, a third subset of the data can be used to further test the reliability of the model. This guards against overfitting the model.

A classification tree considers all input variables (independent variables) and grows branches using input variables that demonstrate significant relationships with the target (dependent variable), while also considering interaction effects among the various inputs. Classification tree models work by segmenting the data using a series of simple rules. Each rule assigns an observation to a subgroup, or "segment," based on the value of one predictor variable. The rules are applied sequentially, resulting in a hierarchy of segments within segments (cf., interaction effects in a logistic regression model). The rules are chosen to subdivide cases into segments that have the largest difference with respect to the target variable, which in this case is nonresponse. Thus the rule selects both the variable and the best breakpoint to separate the resulting subgroups maximally. The breakpoints also take into consideration whether data are missing for an item and either uses a surrogate item (something closely related) or classifies missing data into whichever group is most similar in terms of the target. If the observations that have missing data are distinctly different from those not missing data, then the tree will break the item on the

missing classification. These rules make it easier to describe the likely nonrespondents specifically and to identify what characteristics contribute to nonresponse. By itself, a propensity score helps predict the likely nonrespondents and identifies which inputs in the model are positively or negatively related to nonresponse. However, the propensity scores do not provide a specific description of who the nonrespondents are, whereas this is explicit in the classification tree model (Phipps and Toth 2012).

The break points of variables are found using significance testing or reduction in variance criteria. Significance tests (based on F- or chi-square tests) use the *p*-value as the stopping rule. In the application described in Subsection 3.2.1, interval variables were assessed using F-test criteria, and nominal level variables were assessed using a chi-square test, where the best split is the one with the smallest *p*-value (SAS 2009). Bonferroni adjustments are applied to the *p*-value before selecting the split to ". . .mitigate the bias towards inputs with many values" (Neville 1999, 18). Ordinal variables were assessed using entropy, which measures the reduction in variance. The same variables may appear multiple times throughout a tree to introduce further segmentation.

After the initial split, the resulting leaves are considered for splitting using a recursive process that ends when no leaves can be split further (SAS 2009). A leaf can no longer be split when there are too few observations, the maximum depth (hierarchy of the tree) has been reached, or no significant split can be identified.

Using a single classification tree approach, the best initial splitting variable is chosen and significant subsequent splits are selected based on the initial split. However, if the initial splitting variable is chosen based on the significance level using only the training data, it may not actually be the ideal initial splitting variable given all the data; furthermore, it is important to recognize that the effect of subsequent splits is not considered when choosing the optimal initial split. The initial split selected directly affects the optimality of variables considered for subsequent splits. Although one split may be optimal for maximizing the dichotomy at a given level of the tree, there is no guarantee that given subsequent splits, a tree selecting the optimal initial split will correctly identify the greatest number of observations with the target.

To mitigate this, ensemble tree models are used instead. Ensemble trees grow multiple trees each with varying initial splits. With varying initial splits, each of the trees within the ensemble is capable of identifying different (but possibly overlapping) subgroups with high occurrences of the target. Using the ensemble of classification trees results in a more accurate, stable, powerful, and generalizable model than using a single classification tree (Breiman 1998; Dietterch 2000; Matignon 2008). An ensemble tree can either use voting or the average of the propensity scores across all the trees to identify those likely to exhibit the target (SAS 2009). We utilized the average propensity score across all of the trees, since we were interested in the overall propensity to respond as opposed to nonresponse classification.

3.2. ARMS Application

3.2.1. Building the Initial Model

To evaluate the classification tree procedure described in Subsection 3.1, 2002 Census of Agriculture data were matched to all available ARMS 2000–2008 sample units. These data were then used to construct classification trees predicting ARMS non-

respondents and to estimate their nonresponse propensities. 78 percent ($n = 185{,}767$) of all ARMS cases sampled for data collection between 2000 and 2008 had 2002 Census of Agriculture data available.

The dependent (target) variable for our model was ARMS nonresponse. Operations responding to the ARMS were coded "0" and those not responding were coded "1". The classification trees described in this study explored the relationship between key agriculture characteristics collected on the 2002 Census for the ARMS 2000–2008 samples and nonresponse.

All of the classification trees were created using a randomly selected subset of the data (66,876 of 167,190 farms), which is referred to as the training data. The same training data were used for all trees in the ensemble tree. The results were evaluated and tested using the remainder of the data (93,314 farms). The average squared error of the model applied to the training, validation, and test data was equivalent (average squared error = .34), indicating that the model performed equally well on the training dataset used to create the model and on the two independent validation and test datasets.

Using an ensemble tree approach, we grew multiple trees, forcing each one to initially split on one of the 70 of 71 variables significantly related to nonresponse ($p < .20$); Table A1 in the Appendix provides the list of studied variables. We set the minimum number of observations for a leaf to five, the maximum depth of the tree to six, and the significance level to 0.20. A liberal criterion is used to assess the significance of main effects, since classification trees are primarily interested in the subsequent interaction effects and use independent sources of data to evaluate the results. According to Uther and Veloso (1998, 4), "In the tree based learning literature, it is well known that stopping criteria often have to be weak to find good splits hidden low in the tree."

A popular form of an ensemble tree model called random foresting randomly selects subsets of variables to split on, since in most cases it is not possible to grow all possible trees (Banfield et al. 2007). In our case, we did not grow all possible trees, but we explored all initial splitting variables. We forced each of the 70 variables to be used as an initial splitting variable, so that we could ensure that each of these variables was considered at least once in the overall model. This was important for us in being able to assure operational and field staff that each of the variables in Table A1 were tested in relation to nonresponse. While some of these variables may not be as strongly related to nonresponse as total sales or total acres operated, they are still important to NASS in terms of nonresponse bias. For example, by forcing Tree 66 to split on acres of certified organic farming, we were able to model the relationship between number of certified organic acres and nonresponse. Only significant initial and subsequent splits were retained in our model. After the initial split, all significant subsequent splits were detected automatically using the splitting algorithm described above. Out of 71 initial forced splits shown in Table A1, only one was considered nonsignificant – whether the farm operator was Native Hawaiian or Pacific Islander.

Each tree identified unique subsets of likely nonrespondents based on varying initial splits, and therefore provided unique indicators and thus probabilities of nonresponse. This resulted in a richer and more inclusive model that included not only the characteristics we knew were related to nonresponse, such as total sales and total acres operated, but also the gender of the operator and the number of female operators, which we previously did not know were related to nonresponse. The overall ensemble tree propensity score for each

sample unit was estimated by taking the average of all 70 individual tree nonresponse propensities for that unit. The average propensity score from multiple trees with varying significant splits is considered to be more accurate and generalizable than those taken from an individual tree (Bauer and Kohavi 1999; Breiman 1998). The segmentation rules for all 70 trees were saved into a score code that could be used to estimate nonresponse propensities of future ARMS samples using their 202 census data.

3.2.2. Evaluating the Model for Nonresponse Predictive Power

Once the ensemble model was created, we evaluated the model using the 2009 ARMS sample, a completely independent ARMS sample which had not been used in creating any of the trees. By pulling the 2002 Census data for the 2009 ARMS sample, we were able to apply the model specification rules to the 2009 sample and evaluate the predictive power of the ensemble tree nonresponse propensity scores using a logistic regression model. The logistic regression model specified the ARMS 2009 nonresponse as the dependent variable and the ARMS ensemble tree nonresponse propensity score as the independent variable, controlling for Census 2007 total sales and total acres operated. By controlling for total sales and total acres operated, we could determine whether the ensemble tree propensity scores were significantly correlated with future ARMS nonresponse beyond just farm size and total sales. The logistic regression analysis was run using the 21,969 of the 34,429 operations for which we had both 2002 and 2007 Census data; 2002 data were needed to generate the nonresponse propensities and 2007 data were necessary as the proxy data for the 2009 sample. Census 2002 data were available for 24,264 (70%) of the ARMS 2009 sampled operations, and Census 2007 data were available for 27,830 (81%) of the ARMS 2009 sampled operations; both Census 2002 and 2007 data were available for 64 percent of all operations sampled for the 2009 ARMS.

3.2.3. Evaluating the Model for Nonresponse Bias Predictive Power

If the nonresponse propensities are correlated with 2009 ARMS nonresponse beyond just measures of farm size, they can be used to classify the 2009 sample into nonresponse subgroups with similar response propensities. According to Eltinge and Yansaneh (1997), nonresponse propensity score deciles are considered to be more stable than the individual propensity scores, and therefore can be used to distinguish less likely respondents from more likely respondents. Using deciles, we classified the ARMS 2009 sample into ten groups using their nonresponse propensity scores. We then compared the ten nonresponse propensity groups on key estimates (by using their Census 2007 data as a proxy of key ARMS estimates for this sample) (see Table 3). Finally, we plotted the relative difference of the mean (and median) as shown in Equation 1 for all ten deciles in order to determine if the groups least likely to respond might contribute substantively more to nonresponse bias on the studied characteristics than those more likely to respond.

$$\text{Relative Difference of the Mean} = \frac{\bar{y}_c - \bar{y}_o}{\bar{y}_o} \tag{1}$$

where,

$\bar{y}_c =$ Class Mean
$\bar{y}_o =$ Overall Mean of Full Sample Results.

4. Results

Figure 1 demonstrates a weak positive relationship between the ensemble tree nonresponse propensities and probability of ARMS 2009 nonresponse given an operation's modeled nonresponse propensity score, value of Census 2007 total sales, and Census 2007 total acres operated.

Table 2 shows that even though the relationship between the ensemble tree nonresponse propensity score and 2009 ARMS nonresponse appeared weak, it was still a significant predictor of 2009 ARMS nonresponse, even after controlling for the operation's 2007 total sales and total acres operated; indicating that ARMS nonresponse is not completely explained by farm value and size.

Having evaluated our classification tree nonresponse propensities on an "independent" dataset, we then grouped the nonresponse propensity scores for the ARMS 2009 sample into deciles so that we could distinguish between operations expected to be more likely versus less likely to respond. Lower classes were expected to have lower rates of nonresponse and higher classes were expected to have higher rates of nonresponse. Figure 2 shows that the percent of nonrespondents within each class generally increases from Class 1 (C1) (the group most likely to respond) through Class 10 (C10) (the group least likely to respond), although counter to expectation the group with the highest predicted nonresponse propensities did not have the highest nonresponse rate.

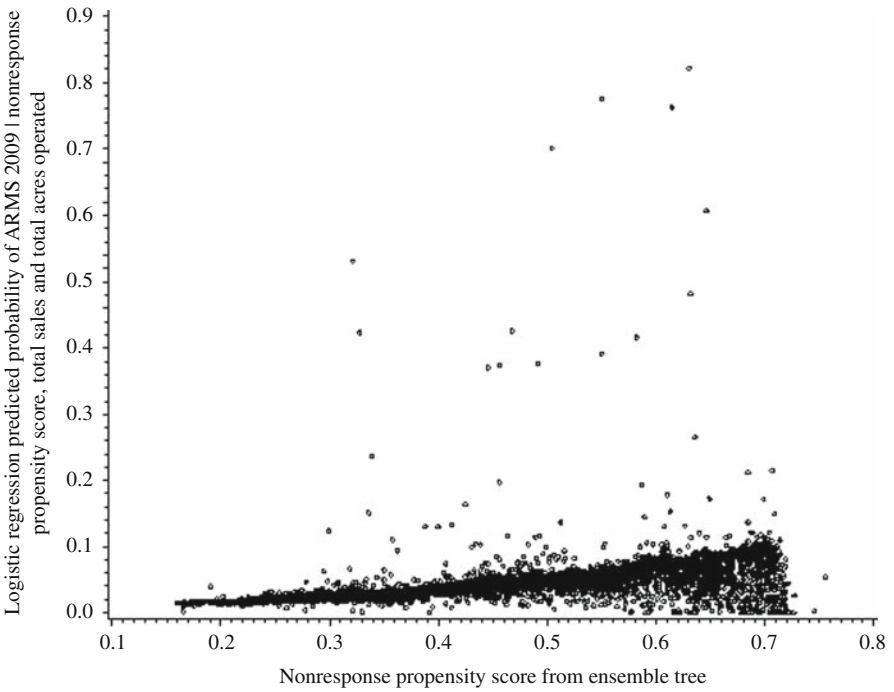


Fig. 1. Plot of the logistic regression predicted probability of 2009 ARMS nonresponse given the ensemble tree nonresponse propensity score, 2007 total sales, and 2007 total acres operated, by the ensemble tree nonresponse propensity score

Table 2. Logistic regression model fit statistics

Analysis of maximum likelihood estimates					
Predictor	β	SE β	Wald's χ^2 ($df = 1$)	p	e^{β} Odds Ratio
Constant	−4.77	0.14	1191.55	<.0001	
Propensity score	3.76	.34	118.99	<.0001	42.93
Total sales	−9.02−08	2.11E-08	18.35	<.0001	1.00
Total acres operated	2.0E-05	3.19E-06	40.67	<.0001	1.00

Finally, we compared the 14 key agricultural estimates (again, using their 2007 Census data as a proxy) across the ten nonresponse propensity classes to see whether these estimates varied by class. Table 3 provides the mean value of the 14 key estimates by the ten nonresponse propensity classes, Class 11 (C11) identifies the ARMS 2009 sampled operations that were missing Census 2002 data and therefore have missing nonresponse propensity estimates, but were not missing Census 2007 data. This allowed us to assess how operations missing nonresponse propensity scores compared to those not missing nonresponse propensity scores. Using the overall mean and the class means shown in Table 3, we calculated the relative difference of the mean for each class shown in Figure 3 (Särndal 2011).

Given the significant correlations between the modeled nonresponse propensity scores, ARMS 2009 nonresponse, and the key estimates shown in Table A2, we expected to see a relationship between nonresponse propensity classes and relative difference of the mean.

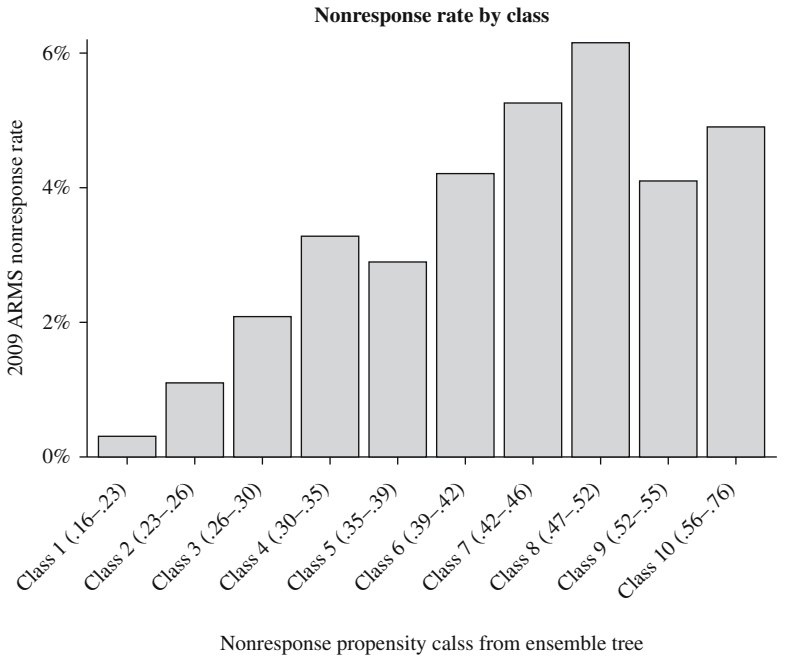
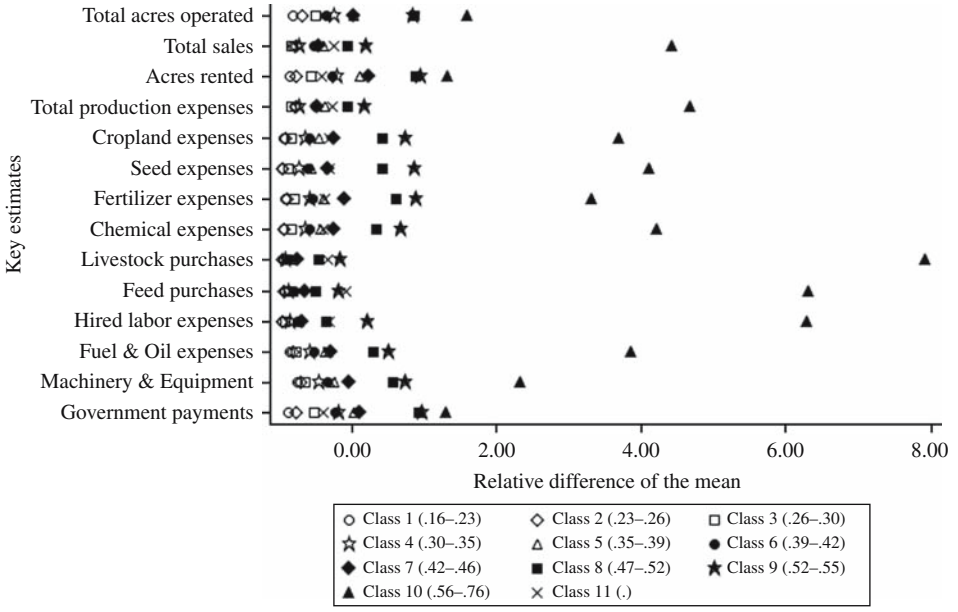


Fig. 2. ARMS 2009 nonresponse rate by ensemble tree nonresponse propensity class

Table 3. Key estimate means by nonresponse propensity class. (Note that the deciles were created for all ARMS 2009 sampled operations with 2002 Census data (n = 24,264) and since not all sampled operations also had 2007 Census data, the ns for the deciles are not all equal, but are very close.)

		Nonresponse propensity classes										
Key estimates	Overall	C1 (.16–.23)	C2 (.23–.26)	C3 (.26–.30)	C4 (.30–.35)	C5 (.35–.39)	C6 (.39–.42)	C7 (.42–.46)	C8 (.47–.52)	C9 (.52–.55)	C10 (.56–.76)	C11 (.)
Total acres operated	1,536	311	489	764	1,166	1,586	998	1,565	2,855	2,844	3,992	952
Acres of land rented	772	116	179	341	618	854	564	950	1,458	1,502	1,791	465
Seed expenses	37,359	1,730	1,999	5,111	10,519	16,873	14,962	25,366	53,272	69,670	190,396	25,701
Fertilizer expenses	50,597	4,765	4,929	10,562	21,291	31,046	23,763	45,683	81,720	95,061	217,912	32,519
Chemical expenses	32,877	2,214	2,128	5,498	12,162	18,333	14,059	24,546	44,349	54,783	171,250	20,491
Feed expenses	110,157	8,830	6,961	12,887	12,681	22,160	23,647	39,969	56,465	88,685	804,299	101,551
Hired labor expenses	96,288	5,131	5,021	7,517	14,474	26,752	24,781	31,127	63,732	116,288	701,271	67,082
Fuel & oil expenses	32,942	5,015	6,656	7,568	14,048	20,874	15,982	23,386	42,801	49,942	159,622	22,544
Machinery & equipment value	319,088	79,207	92,313	119,093	172,158	239,160	215,385	307,331	501,652	554,382	1,066,464	229,141
Total government payments	15,577	2,137	3,568	7,663	12,806	16,064	11,991	17,452	29,977	30,646	35,830	9,690
Crop expenses	127,985	9,477	9,939	22,394	45,701	69,254	54,059	97,000	181,969	222,969	598,429	88,920
Livestock expenses	92,153	5,977	3,921	9,395	7,976	15,734	12,964	23,101	49,761	77,755	822,388	61,539
Total sales	1,049,540	209,400	255,876	183,701	282,365	637,756	512,065	566,066	996,348	1,254,680	5,702,015	788,303
Total production expenses	797,476	165,106	188,632	140,530	213,950	508,571	400,236	422,024	754,359	933,910	4,520,589	586,415
n	27,830	2,449	2,445	2,438	2,432	2,426	2,419	2,411	2,414	2,410	2,419	3,566



Relative difference of the mean = [(class mean – overall mean)/overall mean]

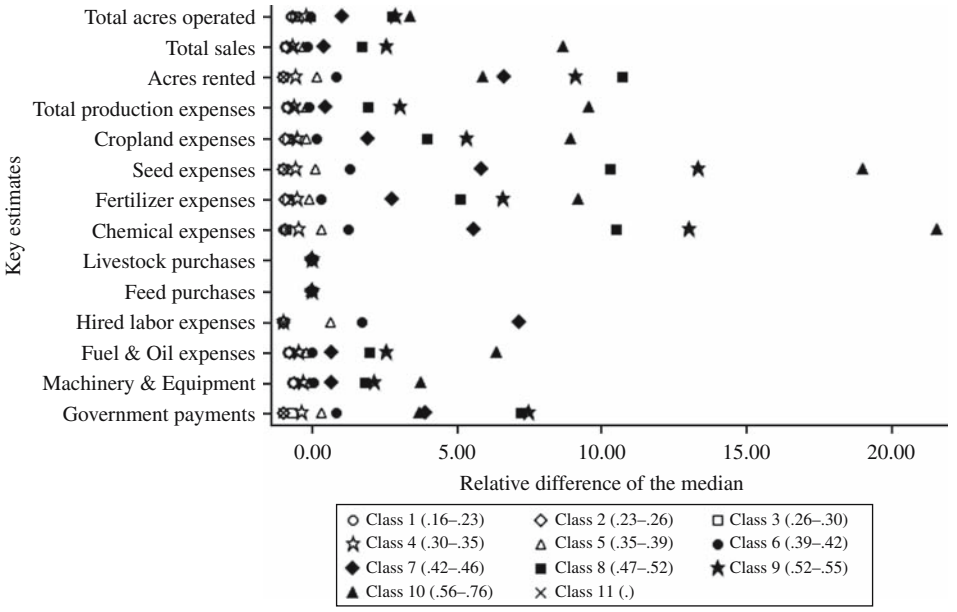
Fig. 3. Relative difference of the mean for key estimates by nonresponse propensity class

The relative difference of the mean as plotted in Figure 3 indicates that the group least likely to respond in terms of their propensity score (Class 10) also poses the greatest threat in terms of relative difference of the mean. Without response from this group, all 14 key estimates would be underestimated. In order to mitigate the potential impact of a few extreme values in any class, we also show the same comparison for the medians by class (see Figure 4). This may be a problem particularly for highly skewed establishment populations. These results are comparable to the estimate means.

Note that the relative difference of the class median for hired labor expenses was extremely high for Classes 8 ($62.86), 9 ($112.68) and 10 ($476.47) and has been omitted from the chart for clarity. Furthermore, the overall median and almost all of the class medians were zero for both livestock and feed expenses, indicating zero relative difference; however, in the few instances where the class median was not zero (livestock: Class 10 ($3,125); feed: Class 1 ($1,000), Class 2 ($89), and Class 3 ($23)) we were unable to estimate the relative difference of the median since the overall median was zero and would have resulted in dividing the class medians by zero.

While Class 11 operations were missing Census 2002 data and thus propensity scores, the relative difference for Class 11 did not stand out in comparison to the other classes shown in Figure 3 or 4.

These results demonstrate that by using an ensemble of classification trees with Census of Agriculture data, we created nonresponse propensity scores that were significantly correlated with future ARMS nonresponse and with all 14 key agricultural estimates from the ARMS (see Table A2). The farms classified into the lowest expected response propensity had the greatest relative difference of the mean and therefore posed the greatest potential threat in terms of both nonresponse and nonresponse bias.



Relative difference of the median = [(class median – overall median)/overall median]

Fig. 4. Relative difference of the median for key estimates by nonresponse propensity class

5. Discussion

This article presents a procedure that uses an ensemble of classification trees to produce robust nonresponse propensity estimates. By examining the individual trees used to create the average nonresponse propensity, we can easily identify the characteristics of various types of nonrespondents. These models not only considered the most obvious and significant predictors of nonresponse in the studied program, but they also identified the rare and yet important variables that are also related to nonresponse. The resulting average nonresponse propensity scores from all the trees may not be greatly influenced by these less predictive or important variables, but they are at least considered given the forced initial-split method we used, which is important to operational and field office staff.

While the logistic regression model's pseudo R^2 (McFadden 1974) was low ($= 0.03$), this may be partly due to the fact that the nonresponse rate was much lower for those operations that had both Census 2002 data and Census 2007 data than for the overall ARMS 2009 sample. Operations with both 2002 and 2007 Census data had a nonresponse rate of 0.03 ($n = 21{,}969$) compared to 0.32 for the overall ARMS 2009 sample ($n = 34{,}429$). Had we been able to estimate propensity scores and had proxy data for the entire 2009 ARMS sample, the propensity scores might have been more strongly related to future ARMS nonresponse.

The results of the study might have differed had we included sample design weights. Using sample design weights and or calibration weights in the models would allow development of prediction models that identify nonrespondent characteristics and estimate nonresponse propensities for the entire population, instead of being restricted to the

selected sample (Phipps and Toth 2012). We may consider including sample or calibration weights in a future model to gain a more general understanding. Given that Class 10 had the greatest relative difference of the mean for all 14 key estimates, NASS may consider using adaptive design efforts to increase the likelihood of response for operations that fall into this class, and potentially Classes 9 and 8 as well depending on funding.

While we did not evaluate whether using varying forced initial splits works as well as random forests, this method did provide us with a level of control that made our methods and results easy to explain to operational and field staff. This was important given that this was only the third operational use of classification trees at NASS, and the first in relation to survey nonresponse. In a future article, we would like to explore the performance of this method in comparison to random foresting. We would be specifically interested in assessing the relative difference of specialty crops and rare operator characteristics such as being female, since we believe this may be a potential strength of using initial forced splits.

The ensemble tree method of modeling survey nonresponse introduced in this article can be helpful in identifying and describing characteristics of influential nonrespondents in other surveys. It provides a tool that allows the researcher to assess the impact of multiple establishment characteristics and interaction effects on nonresponse. Classification trees provide a series of simple rules that can be used to describe specific characteristics of likely nonrespondent subgroups to operational and field staff. The modeled nonresponse propensities can then be used to create nonresponse subgroups. These subgroups can then be used to evaluate the potential impact on survey estimates, or as inputs to adaptive design strategies targeting different data collection strategies to different subgroups of a sample.

Appendix

Table A1. Census of agriculture operational characteristic variables in ranking order of initial split significance

Rank	Variable name
1	Total sales not under production contract (NUPC)
2	Total value of products sold + government payments
3	Total production expenses
4	The number of hired workers employed more than 150 days
5	Machinery and equipment value in Dollars
6	Acres of cropland harvested
7	Cropland acres
8	Total reported acres of crops harvested
9	Acres of land owned
10	State
11	Total acres operated
12	The number of hired workers employed less than 150 days
13	Any migrant workers Y/N
14	Total cattle and calf inventory
15	Total expenditures
16	Farm type code
17	Type of organization
18	Percent of principle operator's income from the farm operation
19	Computer used for the farm business Y/N

Table A1. Continued

Rank	Variable name
20	Acres of all other land
21	Principal occupation of principle operator is farming Y/N
22	Total government payments
23	ARMS III production region (Atlantic, South, Midwest, Plains, or West)
24	Acres of land rented from others
25	Any hired manager Y/N
26	Operation had internet access Y/N
27	Number of households sharing in net farm income
28	Acres of all irrigated hay and forage harvested
29	Number of days principle operator worked off farm
30	Total fruit acres
31	Total acres of vegetables
32	Acres of woodland pasture
33	Principal operator's age
34	Acres of woodland not in pasture
35	Number of operators
36	Acres on which manure was applied
37	Acres of permanent pasture & rangeland
38	Acres of all hay and forage harvested
39	Total poultry inventory
40	Partnership registered under state law Y/N
41	Acres of cropland used for pasture
42	Total hog and pig inventory
43	Principal operator lives on operation Y/N
44	Percent of operators that are women
45	Acres of cropland for which all crops failed
46	Acres of cropland in summer fallow
47	ARMS III questionnaire version
48	Total sales under production contract (UPC)
49	Total citrus acres
50	Nursery indicator Y/N
51	Principal operator's sex
52	Principal operator – race, black
53	Acres of land rented to others
54	Operation farm tenure (1 = full owner, 2 = part owner, or 3 = tenant)
55	Number of persons living in principle operator's household
56	Acres of cropland idle or used for cover crops
57	Have other farm Y/N
58	Principal operator – race, white
59	Sheep and lamb indicator Y/N
60	Year principal operator began this operation
61	Number of women operators
62	Other livestock animals
63	Agriculture on indian reservations Y/N
64	Principal operator – race, american indian
65	Acres of Christmas trees and Short rotation woody crops
66	Acres of certified organic farming
67	Possible duplicate Y/N

Table A1. Continued

Rank	Variable name
68	Principal operator is of Spanish origin Y/N
69	Principal operator – race, Asian
70	Aquaculture indicator Y/N
71	Principal operator – race, native Hawaiian, or Pacific Islander[5]

[5] Not significant at the 0.20 level.

Table A2. Pearson & Point biserial correlation matrix of nonresponse propensity score, indicator of 2009 ARMS response, and key estimates

		Nonresponse propensity score	2009 ARMS nonrespondent
Nonresponse propensity score		1.00	0.08
	p		<.0001
	n	24,264	24,264
2009 respondent		0.08	1.00
	p	<.0001	
	n	24,264	34,429
Total acres operated		0.20	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Acres of land rented from others		0.15	−0.02
	p	<.0001	0.00
	n	21,969	27,830
Seed expenses		0.23	−0.02
	p	<.0001	0.00
	n	21,969	27,830
Fertilizer expenses		0.35	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Chemical expenses		0.30	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Feed expenses		0.20	−0.01
	p	<.0001	0.18
	n	21,969	27,830
Labor expenses		0.29	−0.03
	p	<.0001	<.0001
	n	21,969	27,830
Fuel & oil expenses		0.37	−0.04
	p	<.0001	<.0001
	n	21,969	27,830
Machinery & equipment value		0.44	−0.04
	p	<.0001	<.0001
	n	21,969	27,830
Total government payments		0.26	−0.04
	p	<.0001	<.0001
	n	21,969	27,830

Table A2. Continued

		Nonresponse propensity score	2009 ARMS nonrespondent
Total sales		0.34	−0.03
	p	<.0001	0.00
	n	21,969	27,830
Livestock		0.12	−0.01
	p	<.0001	0.16
	n	21,969	27,830
Crop expenses		0.35	−0.02
	p	<.0001	<.0001
	n	21,969	27,830
Total Production expenses		0.30	−0.03
	p	<.0001	<.0001
	n	21,969	27,830

6. References

Abraham, K.G., A. Mailand, and S.M. Bianchi. 2006. "Nonresponse in the American Time Use Survey. Who is Missing from the Data and How Much Does it Matter?" *Public Opinion Quarterly* 70: 676–703. DOI: http://dx.doi.org/10.1093/poq/nfl037.

Axinn, W., C. Link, and R. Groves. 2011. "Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior." *Demography* 48: 1127–1149. DOI: http://dx.doi.org/10.1007/s13524-011-0044-1.

Banfield, E., L.O. Hall, K.W. Bowyer, and W.P. Kegelmeyer. 2007. "A Comparison of Decision Tree Ensemble Creation Techniques." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29: 173–180. DOI: http://dx.doi.org/10.1109/TPAMI.2007.250609.

Bauer, E. and R. Kohavi. 1999. "An Emperical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." *Machine Learning* 36: 105–132. DOI: http://dx.doi.org/10.1023/A:1007515423169.

Breiman, L. 1998. "Arcing Classifiers (with discussion)." *Annals of Statistics* 26: 801–849.

Brick, J.M. and D. Williams. 2009. "Reasons for Increasing Nonresponse in U.S. Household Surveys." Paper presented at the Workshop of the Committee on National Statistics, Washington, DC, December 14.

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Last Quarter Century." *Public Opinion Quarterly* 69: 87–98. DOI: http://dx.doi.org/10.1093/poq/nfi002.

Dietterich, T.G. 2000. "Ensemble Methods in Machine Learning." In Proceedings of the Multiple Classifier Systems: First International Workshop, MCS 2000, June 21–23, Cagliari, Italy. Available at: http://www.eecs.wsu.edu/~holder/courses/CptS570/fall07/papers/Dietterich00.pdf (accessed August 2014).

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

Earp, M., J. McCarthy, E. Porter, and P. Kott. 2010. "Assessing the Effect of Calibration on Nonresponse Bias in the 2008 ARMS Phase III Sample Using Census 2007 Data." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: http://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2010/earp-2010_jsm_paper_arms_calibration.pdf (accessed August 2014).

Eltinge, J.L. and I.S. Yansaneh. 1997. "Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the US Consumer Expenditure Survey." *Survey Methodology* 23: 33–40.

Groves, R. 2006. "Nonresponse Rates and the Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. DOI: http://dx.doi.org/10.1093/poq/nfl033.

Groves, R. and M. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R.M., D. Dillman, J.L. Eltinge, and R.J. Little. 2002. *Survey Nonresponse*. New York: Wiley.

Groves, R. and S. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A: Statistics in Society* 169: 439–457. DOI: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Johansson, F. and A. Klevmarken. 2008. "Explaining the Size and Nature of Response in a Survey on Health Status and Economic Standard." *Journal of Official Statistics* 24: 431–449.

Johnson, T.P., I.K. Cho, R.T. Campbell, and A.L. Holbrook. 2006. "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey." *Public Opinion Quarterly* 70: 704–719. DOI: http://dx.doi.org/10.1093/poq/nfl032.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19: 81–97.

Laflamme, F. and M. Karaganis. 2010. "Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada." In Proceedings of the European Quality Conference: Helsinki, Finland.

Lepkowsi, J.M. and M.P. Couper. 2002. "Nonresponse in the Second Wave of Longitudinal Household Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little. New York: Wiley and Sons.

Little, J. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *Journal of the American Statistical Association* 77: 237–250.

Little, R. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Luzi, O., T. De Waal, B. Hulliger, M. Di Zio, J. Pannekoek, D. Kilchmann, and C. Tempelman. 2007. *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*. Italian Statistical Institute ISTAT.

Matignon, R. 2008. *Data Mining Using SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.

McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press.

Miller, D., M. Robbins, and J. Habiger. 2010. "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: https://www.amstat.org/sections/srms/proceedings/y2010/Files/306438_56491.pdf (accessed August 2014).

Mohl, C. and F. Laflamme. 2007. "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: https://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000421.pdf (accessed August 2014).

Neville, P. 1999. *Decision Trees for Predictive Modeling*. Carey, NC: SAS Institute, Inc.

Nicoletti, C. and F. Peracchi. 2005. "Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel." *Journal of the Royal Statistical Society* Series A: 168: 763–781. DOI: http://dx.doi.org/10.1111/j.1467-985X.2005.00369.x.

Petroni, R., R. Sigman, D. Willimack, S. Cohen, and C. Tucker. 2004. "Response Rates and Nonresponse in Establishment Surveys – BLS and Census Bureau." *Federal Economic Statistics Advisory Committee*, 1–50.

Phipps, P. and D. Toth. 2012. "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data." *Annals of Applied Statistics* 6: 772–794. DOI: http://dx.doi.org/10.1214/11-AOAS521.

Powers, R., J. Eltinge, and M. Cho. 2006. "Evaluation of the Detectability and Inferential Impact of Nonresponse Bias in Establishment Surveys." In Proceedings of the Joint Statistical Meetings: American Statistical Association. Alexandria, VA: American Statistical Association. Available at: http://www.bls.gov/ore/pdf/st060130.pdf (accessed August 2014).

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. DOI: http://dx.doi.org/10.1093/biomet/70.1.41.

Särndal, C.-E. 2011. "The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation." *Journal of Official Statistics* 27: 1–21.

SAS Institute Inc. *Enterprise Miner 6.2 Help and Documentation*. Cary, NC: SAS Institute Inc., 2009.

Schouten, B. 2007. "A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption." *Journal of Official Statistics* 23: 51–68.

Schouten, B. and G. de Nooij. 2005. *Nonresponse Adjustment Using Classification Trees*. CBS, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/1245916E-80D5-40EB-B047-CC45E728B2A3/0/200501x10pub.pdf (accessed August 2014).

Stussman, B., J. Dahlhamer, and C. Simile. 2005. "The Effect of Interviewer Strategies on Contact and Cooperation Rates in the National Health Interview Survey." Federal Committee on Statistical Methodology, Washington, DC

Thompson, K.J. 2009. "Conducting Nonresponse Bias Analysis for Two Business Surveys at the US Census Bureau: Methods and (Some) Results." In Proceedings of the Section on Survey Research Methods: American Statistical Association Alexandria, VA: American Statistical Association. Available at: http://www.scs.gmu.edu/~wss/wss100922linebackpaper.pdf (accessed August 2014).

Thompson, K.J. and K.T. Washington. 2013. "Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study." *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=2991.

United States Department of Agriculture. 2012. *2012 Agricultural Resource Management Survey – Phase III Cost and Returns Report Survey Administration Manual*. Washington, DC: US Department of Agriculture.

United States Department of Agriculture. 2007. *2007 Census of Agriculture*. Washington, DC: US Department of Agriculture. Available at: http://www.agCensus.usda.gov/Publications/2007/Full_Report/ (accessed August 2014).

United States Executive Office of the President. 2006. *Office of Management and Budget Standards and Guidelines for Statistical Surveys*. Washington, DC: U.S. Executive Office of the President. Available at: http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf (accessed August 2014).

Uther, W.T.B. and M.M. Veloso. 1998. "Tree Based Discretization for Continuous State Space Reinforcement Learning." In Proceedings of AAAI-98, the Fifteenth National Conference on Artificial Intelligence: 769–774. Available at: http://www.cs.cmu.edu/~mmv/papers/will-aaai98.pdf (accessed August 2014).

Wagner, J. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76: 555–575. DOI: http://dx.doi.org/10.1093/poq/nfs032.

Received December 2012
Revised August 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 721–747, http://dx.doi.org/10.2478/JOS-2014-0045

Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey

Mary H. Mulry[1], *Broderick E. Oliver*[2], *and Stephen J. Kaputa*[3]

In survey data, an observation is considered influential if it is reported correctly and its weighted contribution has an excessive effect on a key estimate, such as an estimate of total or change. In previous research with data from the U.S. Monthly Retail Trade Survey (MRTS), two methods, Clark Winsorization and weighted M-estimation, have shown potential to detect and adjust influential observations. This article discusses results of the application of a simulation methodology that generates realistic population time-series data. The new strategy enables evaluating Clark Winsorization and weighted M-estimation over repeated samples and producing conditional and unconditional performance measures. The analyses consider several scenarios for the occurrence of influential observations in the MRTS and assess the performance of the two methods for estimates of total retail sales and month-to-month change.

Key words: Outlier; Winsorization; M-estimation.

1. Introduction

In survey data, an observation is considered influential if its value is correct but its weighted contribution has an excessive effect on an estimated total or period-to-period change. To be clear, our focus is on influential values that remain after all the data have been verified or corrected, so these unusual values are true and not the result of reporting or recording errors. Failure to "treat" such influential observations may lead to substantial over- or under-estimation of survey totals, which in turn may lead to overly large increases or decreases in estimates of change.

The presented research was motivated by a request from the methodologists and subject matter experts who supervise the U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) to find a method that improves or replaces current methodology for identifying and treating influential values. New methodology would need to use the influential observations, but in a manner that assures their contribution does not have an excessive effect on the monthly totals or an adverse effect on the estimates of month-to-month

[1] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: mary.h.mulry@census.gov
[2] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: broderick.e.oliver@census.gov
[3] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: stephen.kaputa@census.gov

Acknowledgments: The authors thank Katherine Jenny Thompson for her useful suggestions. The authors are very grateful to Jean-Francois Beaumont for providing the SAS code for the M-estimation algorithm and for all his advice and consultations. The authors thank Lynn Weidman, Eric Slud, Scott Scheleur, William Davie, Jr., Paul Smith, two referees, and the associate editor for their helpful comments on earlier versions of this manuscript. This article is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

change. The tight time schedule for producing MRTS estimates monthly means that the preference is for a new methodology for detecting and treating influential values that is automated, but is implemented in a manner that allows for a final (manual) review. Therefore, the objective of this research is to find an automated statistical procedure to replace the current subjective procedure performed by analysts.

Each month, the MRTS surveys a sample of about 12,000 retail businesses with paid employees to collect data on sales and inventories. The MRTS is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. Moreover, significant changes in levels are important to monetary and budgetary decision makers, economists, business analysts, and economic researchers in assessing the health of the economy, and in making corporate investment decisions. The MRTS sample design is typical of business surveys, employing a one-stage stratified sample with stratification based on major industry, further substratified by the estimated annual sales. The sample design requires the sampling rates to be higher in the strata with the larger units than in the strata with the smaller units and companies that have been determined to comprise a large portion of the total are included with certainty. The sample is selected every five years after the Economic Census and then updated as needed with a quarterly sample of births (new businesses) and removal of deaths (businesses no longer in operation). MRTS publishes Horvitz-Thompson estimates of totals, as well as month-to-month change. Because of its typical sample design and characteristic data, the results that we obtain by studying the program in detail should be applicable to other similar programs.

In the MRTS, when an influential observation appears in a month's data, the current corrective procedures depend on whether the subject-matter experts believe the observation is a one-time phenomenon or a permanent shift. If the influential value appears to be an atypical occurrence for the business, then the influential observation is replaced with an imputed value. If the influential value persists for a few months and appears to represent a permanent change, then methodologists adjust its sampling weight using principles of representativeness or move the unit to a different industry when the nature of the business appears to have changed (Black 2001). Prior to influential value detection, the MRTS processing already includes running the algorithm by Hidiroglou and Berthelot (1986), often called the HB edit, to identify (and – on occasion – treat) within-imputation-cell outliers and create the imputation base (Hunt et al. 1999). Treatment of influential values is the final step of the estimate review process. Hence, the methods described here are developed to complement, not replace, the HB edit.

The research reported in this article builds on several previous studies on methods of addressing influential values in the MRTS. Initial work (Mulry and Feldpausch 2007a) examined a variety of outlier detection and treatment methods from the literature on empirical data from one month of a volatile MRTS industry with an obvious influential value. Of the considered methods, Clark Winsorization (Clark 1995) and M-estimation (Beaumont and Alavi 2004; Beaumont 2004) emerged as the most promising. This study examined several methods, including a second type of Winsorization that developed the cut-off value for the observations by stratum (Kokic and Bell 1994) (instead of specifying an individual cut-off value for each observation as in Clark Winsorization) and a combination of robust estimation and reverse calibration to address influential values

(Ren and Chambers 2003; Chambers and Ren 2004). Mulry and Feldpausch (2007a) concluded that the MRTS data was too volatile for the other methods, which may perform very well in other situations. One might also consider the robust estimators studied by Hulliger (1995) or Farrell and Salibian-Barrer (2006) for other applications.

Subsequent work (Mulry and Feldpausch 2007b) with 38 months of empirical MRTS data from the same industry confirmed the potential for both methods (Clark Winsorization and weighted M-estimation) to address influential values in MRTS data. The infrequent appearance of influential values in empirical data made it difficult to evaluate the performance of the considered methods with respect to relative magnitude of identified influential observation(s) or to examine the statistical properties of the considered methods over repeated samples. Consequently, Mulry and Oliver (2009) conducted a simulation study and presented some preliminary but inconclusive results.

The focus of this article is the use of simulation methodology to investigate these two robust statistical methods for identifying and treating influential observations: Clark Winsorization (Clark 1995) and M-estimation (Beaumont and Alavi 2004; Beaumont 2004). In a sample survey setting, robust methods are especially appealing since they are valid for a variety of probability distributions and therefore are less sensitive to model misspecifications. This is especially important for economic data that generally have skewed populations where the assumption of a normal distribution, or even symmetry, is unlikely to hold.

Building on past research, we developed simulation methodology to obtain decisive results about the statistical properties of Clark Winsorization and weighted M-estimation when applied to data like that collected for industries in the MRTS. The methodology includes simulation of a stationary time series for the population data and the development of performance measures. This simulation examines the effectiveness of the methodologies when seasonal effects are *not* present to illuminate the properties of the methods.

This article describes the simulation methodology and includes performance results for Clark Winsorization and M-estimation in several scenarios for influential values. Both methods were designed for totals estimates, but the most important measure for MRTS is month-to-month change. Therefore, our analysis emphasizes the simulation's estimates of relative bias for estimates of total sales and month-to-month change, both when an influential value is present and when it is not. Additional evaluation criteria include the number of true and false detections.

2. Detection and Treatment Methods

In this section, we present the studied methods. Subsection 2.1 describes the Clark Winsorization methodology for modifying an influential value, and Subsection 2.2 discusses the M-estimation methodology that provides the choice of adjusting the influential value or its weight. Figure 1 illustrates how Clark Winsorization and M-estimation adjust an influential observation.

Before describing the methods, we first introduce the notation. For the i^{th} business in a survey sample of size n for the month of observation t, Y_{ti} is the characteristic of interest (revenue in our application), w_{ti} is its survey weight (which may be equivalent to the

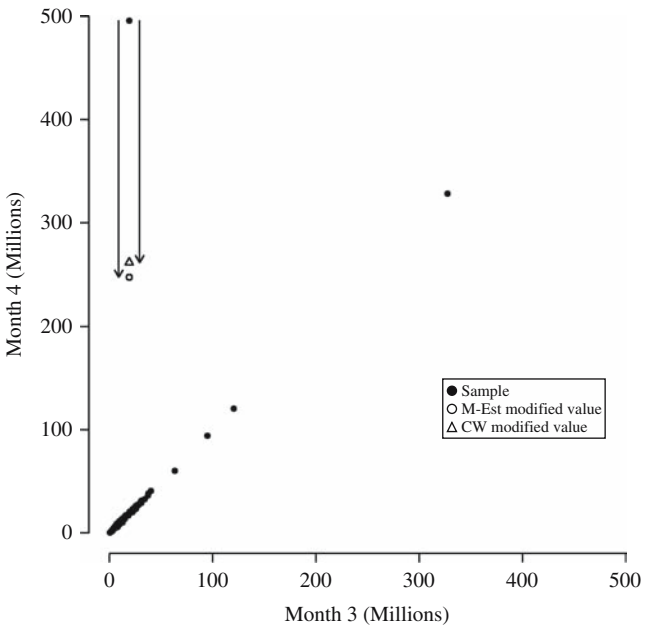


Fig. 1. Illustration of an influential value and its adjustments from Clark Winsorization and weighted M-estimation

inverse probability of selection but can include poststratification, generalized regression, or calibration adjustments), and X_{ti} is a variable highly correlated with Y_{ti}, such as the previous month's collected revenue or the frame revenue value. Note that the more general formulations allow X to be a vector of auxiliary variables. We restrict our analysis to a single covariate and set X_{ti} equal to the unit's previous month's revenue, paralleling the MRTS ratio imputation and outlier-detection (HB edit) procedures. The total monthly revenue Y_t is estimated by

$$\hat{Y}_t = \sum_{i=1}^{n} w_{ti} Y_{ti}.$$

In MRTS, the missing data treatment is imputation (Thompson and Washington 2013), and consequently, the survey weight w_{ti} is the design weight. For ease of notation, hereafter we suppress the t index. Both Clark Winsorization and weighted M-estimation methodologies use a comparison to the prior month's value to detect observations with influential values in the current month.

2.1. Clark Winsorization

Winsorization procedures replace extreme values with less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization methods offer adjustments for the observed influential value but could be used to derive an adjustment for the survey weight if that is needed instead. Winsorization procedures may

be one-sided or two-sided, but the method developed by Clark (1995) and described by Chambers et al. (2000) is one-sided.

The general form of the one-sided Winsorized estimator of the total is designed for large values and is written as

$$\hat{Y}^* = \sum_{i=1}^{n} w_i Z_i \quad \text{where} \quad Z_i = \min\{Y_i, K_i + (Y_i - K_i)/w_i\}. \tag{1}$$

Detection of observation i as an influential value by Clark Winsorization occurs when $Z_i \neq Y_i$. To implement the method, Clark suggests approximating the K_i that minimizes the mean squared error under the general model by $K_i = \mu_i + L(w_i - 1)^{-1}$, using a general model where the Y_i are characterized as independent realizations of random variables with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \sigma_i^2$. To estimate μ_i and L, Clark's approach builds on a method developed by Kokic and Bell (1994) that derived a K for each stratum rather than for each individual unit.

Chambers et al. (2000) suggest using the results of a robust regression to obtain the estimate of μ_i as bX_i where b is the regression coefficient and X_i is the auxiliary variable (the previous month's observed revenue in our application). We used the least median of squares (LMS) robust regression method (Rousseeuw 1984; Rousseeuw and Leroy 1987) because other robust regression methods that we considered, including the least median trimmed (LMT), appeared too sensitive in that they flagged many non-influential values (Mulry and Feldpausch 2007a). To estimate L, the Clark Winsorization first uses the estimate of μ_i to estimate weighted residuals

$$D_i = (Y_i - \mu_i)(w_i - 1) \text{ by } \hat{D}_i = (Y_i - bX_i)(w_i - 1),$$

which are sorted in decreasing order $\hat{D}_{(1)}, \hat{D}_{(2)}, \ldots \hat{D}_{(n)}$. The Clark method finds the last value of k, called k^*, such that $(k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}$ is positive, and then estimates L by $\hat{L} = (k^* + 1)^{-1} \sum_{j=1}^{k^*} \hat{D}_{(j)}$. Last, the estimate of K_i is formed by $\hat{K}_i = bX_i + \hat{L}(w_i - 1)^{-1}$, which is used to determine the values of Z_i for the estimate of the total $\hat{Y}^*$.

2.2. *Weighted M-Estimation*

M-estimators (Huber 1964) are robust estimators that come from a generalization of maximum likelihood estimation. The application of M-estimation examined in this investigation is regression estimation. The weighted M-estimation technique proposed by Beaumont and Alavi (2004) uses the Schweppe version of the weighted generalized technique (Hampel et al. 1986, 315–316). The estimator of the total using this approach is consistent for a finite population since it equals the finite population total when a census is conducted (Särndal et al. 1992, 168).

A key assumption of the M-estimation approach is that y_i given x_i is distributed under the prediction model m with

$$E_m[y_i|x_i] = x_i'\beta \text{ and } V_m[y_i|x_i] = v_i\sigma^2. \tag{1.1}$$

In our application, y_i is the current month's value; x_i is the previous month's value, and the regression model does not include an intercept. With retail trade, the regression of current

month's sales on the previous month's sales tends to go through the origin (Huang 1984). We use the unbiased sampling weights w_i to maintain parallel estimation with the MRTS.

Briefly, the method estimates $\hat{B}^M$, which is implicitly defined by

$$\sum_{i \in S} w_i^*(\hat{B}^M)(y_i - x_i\hat{B}^M)\frac{x_i}{v_i} = 0 \tag{2}$$

where

$$v_i = \lambda x_i$$
$$w_i^* = w_i\psi\{r_i(\hat{B}^M)\}/r_i(\hat{B}^M)$$
$$r_i(\hat{B}^M) = h_i e_i(\hat{B}^M)/Q\sqrt{v_i}$$
$$e_i(\hat{B}^M) = y_i - x_i\hat{B}^M$$

and Q is a constant that is specified. The variable h_i is a weight that may or may not be a function of x_i. The variable λ, possibly a constant, is chosen to ensure the correct specification of the form of the variance in the underlying prediction model.

Section 4 contains a discussion of the investigation to determine the settings for these parameters.

The role of the function ψ is to reduce the influence of units with a large weighted residual $r_i(\hat{B}^M)$. We focus on two choices for the function ψ, Type I and Type II Huber functions, and describe their one- and two-sided-forms. The one-sided Type I Huber function is

$$\psi\{r_i(\hat{B}^M)\} = \left\{ \begin{array}{l} r_i(\hat{B}^M), r_i(\hat{B}^M) \leq \varphi \\ \varphi, \text{otherwise} \end{array} \right\} \tag{4}$$

where φ is a positive tuning constant. This form is equivalent to a Winsorization of $r_i(\hat{B}^M)$. Detection of observation i as an influential value by M-estimation with the Huber I function occurs when $r_i(\hat{B}^M) > \varphi$. In the two-sided Huber I function $r_i(\hat{B}^M)$ is replaced by its absolute value $\left|r_i(\hat{B}^M)\right|$.

The weight adjustment corresponding to the Type I Huber function ψ above is

$$w_i^*(\hat{B}^M) = \left\{ \begin{array}{l} w_i, r_i(\hat{B}^M) \leq \varphi \\ \dfrac{\varphi}{r_i(\hat{B}^M)}, \text{otherwise} \end{array} \right\} \tag{5}$$

an undesirable feature of using the Type I Huber function is that the unit's adjusted weight may be less than one if the influential value is very extreme, thereby not allowing the influential value to represent itself in the estimation. The Type II Huber function ψ ensures that all adjusted units are at least fully represented in the estimate. The one-sided Type II Huber function is

$$\psi\{r_i(\hat{B}^M)\} = \left\{ \begin{array}{l} r_i(\hat{B}^M), r_i(\hat{B}^M) \leq \varphi \\ \dfrac{1}{w_i} r_i(\hat{B}^M) + \dfrac{(w_i - 1)}{w_i}\varphi, \text{otherwise} \end{array} \right\} \tag{6}$$

where φ is a positive tuning constant. Detection of observation i as an influential value by M-estimation with the Huber II function occurs when $r_i(\hat{B}^M) > \varphi$. In the two-sided Type II Huber function $r_i(\hat{B}^M)$ is replaced by its absolute value $\left|r_i(\hat{B}^M)\right|$. This form is equivalent to a Winsorization of $r_i(\hat{B}^M)$, cf. the Type I Huber function.

An interesting feature of using the one-sided Type II Huber function in the M-estimation method is that the parameters can be set to mimic the assumptions of the Clark Winsorization outlined in Subsection 2.1 (Beaumont 2004). However, the results will not be identical because the method used to estimate $\hat{B}^M$ is different.

Solving for $\hat{B}^M$ requires the Iteratively Reweighted Least-Squares algorithm in many circumstances, although for certain choices of the weights and variables, the solution is the standard least-squares regression estimator.

The weight adjustment for the Type II Huber function above is

$$w_i^*(\hat{B}^M) = \left\{ \begin{array}{l} w_i, r_i(\hat{B}^M) \leq \varphi \\ 1 + (w_i - 1)\dfrac{\varphi}{r_i(\hat{B}^M)}, \text{otherwise} \end{array} \right\}. \tag{7}$$

The adjusted value corresponding to the Type II Huber function is

$$y_i^* = \frac{1}{w_i} y_i + \frac{(w_i - 1)}{w_i} \left\{ x_i \hat{B}^M + \frac{\sqrt{v_i}}{h_i} Q\varphi \right\}. \tag{8}$$

We use an adjusted value Beaumont and Alavi (2004) derived by using a weighted average of the robust prediction $x_i\hat{B}^M$ and the observed value y_i of the form

$$y_i^* = a_i y_i + (1 - a_i) x_i \hat{B}^M \text{ where } a_i = \frac{w_i^*(\hat{B}^M)}{w_i}. \tag{9}$$

Beaumont (2004) finds an optimal value of the tuning constant φ by deriving and then minimizing a design-based estimator of the mean-square error via numerical analysis. Unlike the Clark Winsorization algorithm, the Beaumont version of M-estimation does not require a model to hold for all the data, or for the influential value, in particular, and therefore relies on less stringent assumptions.

Since the algorithm is an iterative procedure, convergence is not guaranteed. Failure of convergence appears to be more problematic with the use of two-sided Huber functions than with one-sided Huber functions. Section 4 contains more discussion of the possible consequences when convergence is not achieved.

3. Methodology

3.1. Research Approach

To assess how well M-estimation and Clark Winsorization identify and treat influential values in MRTS data, we conduct a simulation study using different – but realistic – influential value scenarios, considering detection and treatment effects on estimates of totals and of current-to-prior period change.

To do this, we generated two separate time-series populations of monthly sales data, modeled from two MRTS industries with different natures. We generate a stationary time series for each industry to avoid potential confounding of the influential value detection methods and other patterns such as trends or seasonality. Industry 1 has monthly sales of approximately 46.1 billion and one of the most volatile industries. Industry 2 has a more stable pattern and has monthly sales of approximately 2.5 billion. The sample sizes in our simulations are 1,161 for Industry 1 and 147 for Industry 2. Subsection 3.2 describes the procedure used to generate these simulated populations.

Our simulation evaluation approach is two-fold: an *unconditional analysis* where a small subset of the samples (replicates) contain an induced influential value and the majority do not; and a *conditional analysis* that employs only the subset of samples that contain the induced influential value. The objective of the unconditional analysis is to evaluate the performance of Clark Winsorization and M-estimation over a realistic survey setting, where it is not expected that each sample will include an influential value. The objective of the conditional analysis is to evaluate the respective performance of each approach when the sample does contain an influential value.

In practice, the most common scenario pertaining to influential values is an observation whose measurement is much higher than previous measurements and whose high weight greatly amplifies its impact on the estimates. Failure to address this scenario properly can have far-reaching consequences in interpreting the state of the economy, so we focus on this scenario.

3.2. *Simulation Methodology*

Recall that the MRTS is a stratified sample, with strata defined by unit size within industry where the measure of size is sales. An exploratory empirical analysis of the simulated data for both studied industries confirmed that the stratum-level means differ by within-industry-strata as shown in the examples in Figure 2, and that a realistic within-stratum prediction model is given by the stationary series.

$$\hat{y}_{hi,t} = \beta_h \hat{y}_{hi,t-1} + \varepsilon_{hi}, \varepsilon_{hi} \sim (0, \sigma_{hi}^2), t > 1$$

where h indexes the strata as illustrated in the examples in Figure 3.

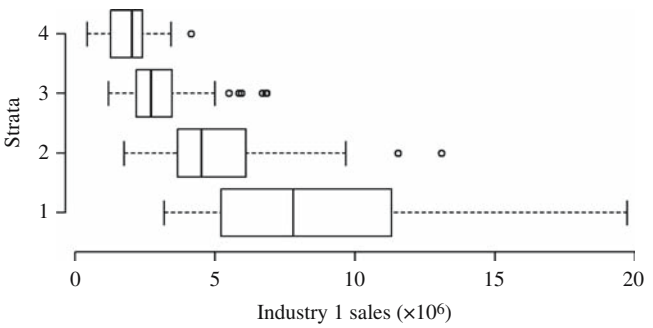


Fig. 2. Stratum-level Box-plots for simulated retail trade Industry 1

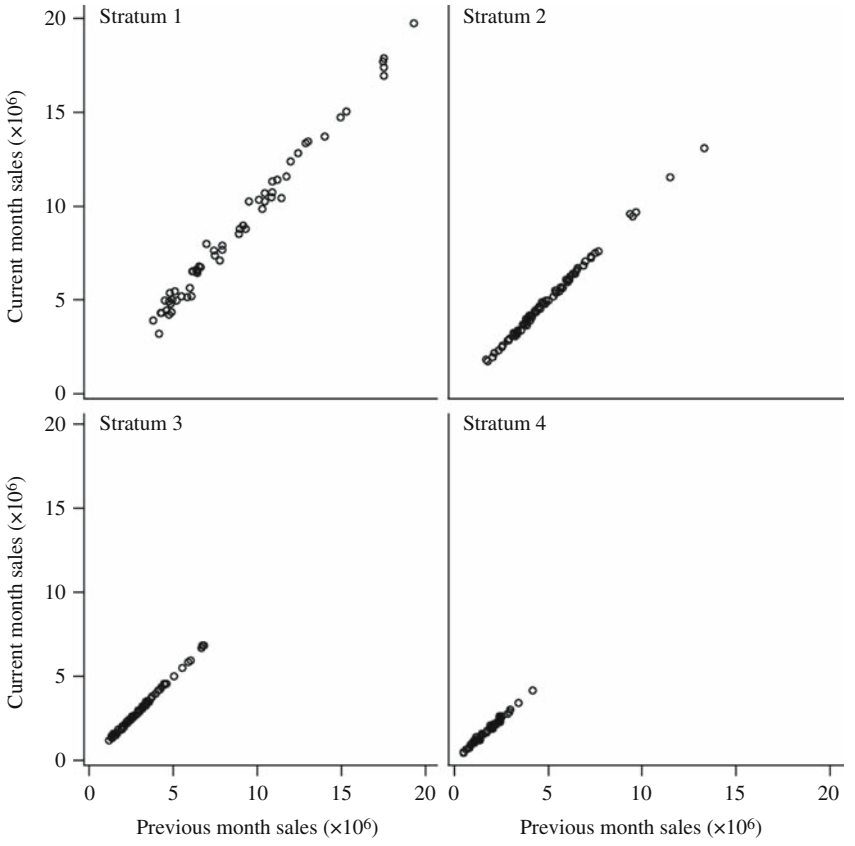


Fig. 3. Scatter plots of current month to previous month sales at the stratum level for simulated retail trade Industry 1

In the notation provided in Subsection 2.2, the "true" prediction model for the simulated data is $E_m\left[y_{hi,t}|y_{hi,t-1}\right] = y'_{hi,t-1}\beta_{h,t-1}$ and $V_m\left[y_{hi,t}|y_{hi,t-1}\right] = \sigma_h^2$, so that $v_h \equiv 1$ within stratum.

To obtain a series 20 months in length, we generated the population for the first month and then generated the next 19 months as a stationary time series essentially as a forecast going forward from Month 1. The population data for the first month were generated using the SIMDAT algorithm (Thompson 2000) with modeling cells equal to sampling strata and population size equal to the original frame size in each cell. The stationary time series was generated using historical standard errors and autocorrelations to develop the AR(1) model within stratum for Months 2 to 20 given by

$$y_t - m = \Phi^*(y_{t-1} - m) + a_t, \quad \text{for } t = 2, \ldots, 20 \tag{10}$$

where

$y_1 - m = 0$ and m is the series mean,

$a_t \sim N(0, \sigma^2)$ (white noise process where s is estimated empirically by the observation for the unit in the first month times the median of percent difference between observations in the historical first and second months),

Φ = the sample-based estimate of lag one autocorrelation for the selected industry.

The time series algorithm written in SAS creates an AR(1) series so that each new observation is set equal to Φ times the previous value $+$ a_t, where a_t is generated from the $N(0, \sigma^2)$ distribution. The initial value of the series is set to zero so that each subsequent point has an expected value of zero – which is necessary for series to be stationary. After all 20 observations for a unit have been created, the initial value (first month value) is added to them so that this number is actually the mean over the time series (in short, it shifts the mean from zero to the first month value).

Generating the series in this manner assures that each of the two populations (one for each industry) is a stationary series within strata, but not at the industry level. Our simulated population data follow directly from the stratification model and mimic the conditions under which the influential observation procedures would be implemented (i.e., after micro-data automatic editing/imputation and HB outlier detection). However, the stratification model diverges greatly from the prediction models assumed by Clark Winsorization (industry-level models, with one population model describing the industry data) and by M-estimation (also, industry-level, with the underlying weighted regression model using the v_i term to account for expected increasing variability with unit size). The funnel shape of the plot in Figure 4 illustrates how the variance of the observations of the retail trade industry data increases as the values of the observations increase. However, Figure 5 illustrates that neither the assumption $v_i = 1$ nor the assumption $v_i = x_i$ for the v_i in the prediction model in Equation (1.1) fits the data well at industry-level, but at the same time, both assumptions appear to have comparable weaknesses. Therefore, we defer the choice of the setting for v_i until we view the detection error rates as defined later in this section and discussed further in Subsection 4.1.

To assess the statistical properties of each influential value treatment method (M-estimation and Clark Winsorization), we induce an influential value into the

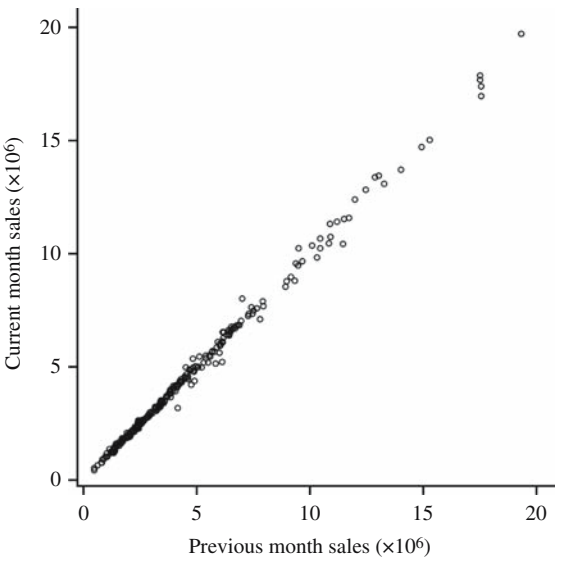


Fig. 4. Industry-level scatter plot of current month to previous month sales for simulated retail trade Industry 1

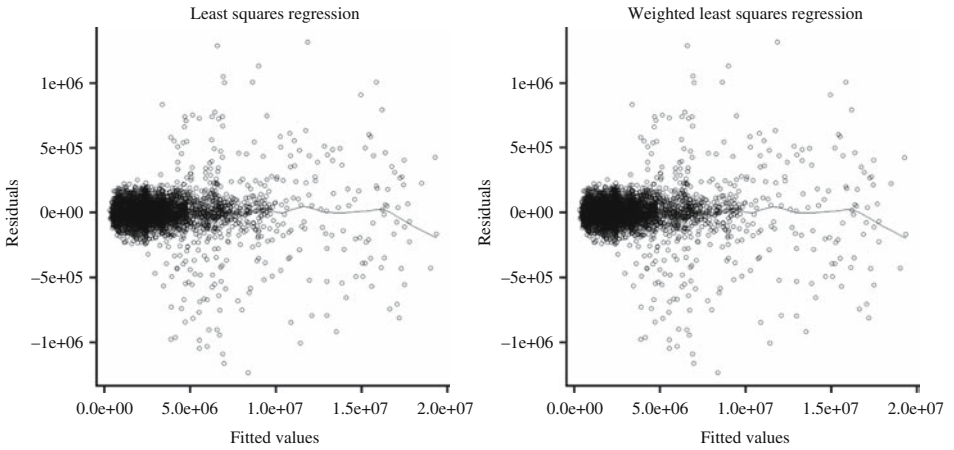


Fig. 5. Residual versus Fitted Values with LOESS curve from models for predicting Industry current month sales using previous month sales with Least Square Regression corresponding to $v_i = 1$ (left) and Weighted Least Square Regression where the weight $= 1/x_i$ corresponding to $v_i = x_i$ (right)

population in Month 4. The choice of Month 4 allows gauging the performance in the months before as well as after the influential value appears which is particularly important for estimates of month-to-month change. The induced influential value does not have an undue effect on the population total, but does have undue influence on the estimated population total if selected in sample. The details of constructing the time series for the population follow using Industry 1 for illustration; the same procedure generated the Industry 2 population.

First, we generate a time series for the Industry 1 population of length 20 months using the methodology described in the first paragraph of this section. We let $Y_1, Y_2, \ldots, Y_{20}$ represent the population totals for this stationary series.

Next, we create one influential unit in the population in Month 4 in a stratum with a sampling rate of approximately 1/50 by adding eight million to the unweighted value of a randomly selected unit in this stratum. Hence, the population total for Month 4 is now eight million larger than its initial value. This influential value *does not* have an undue effect on the population at approximately 46.1 billion in Month 4, but it can have an undue influence on the estimated population total if selected in sample since its weighted value is 400 million larger than its initial weighted value. With this design, we can expect the unit to be selected for one of every 50 samples and when selected, increase the estimated total by about one percent. The induced influential value in the simulation is based on influential values that occurred during the 38 months of the MRTS examined in Mulry and Feldpausch (2007b).

After creating the population time series, we select stratified simple random without replacement (SRS-WOR) samples of size comparable to the MRTS sample from Month 1 until 200 of these samples contain the unit that has the induced influential value in Month 4. The choice of 200 samples was a function of the processing requirements for M-estimation because the required number of samples to achieve 200 with the influential value was quite large and the algorithm had to be run on the total number of samples in

the unconditional analysis. For Industry 1, the necessary number of samples is 10,742, and for Industry 2, the necessary number of samples is 11,931. By requiring the same unit to be included in all samples in the conditional analysis, we effectively reduce the size of the probability sample by one, but continue to give the influential value its stratum weight. This results in a small bias in the months without the induced influential value, and the magnitude of the bias is a function of how close the unadjusted unit's value is to the stratum mean in these months.

In each independent sample, we apply the M-estimation and Clark Winsorization algorithms to Month 2 using Month 1 as the auxiliary data and then continue to apply both methods to each month through Month 20 using the previous month as the auxiliary data. Modified values in a given month are used as auxiliary data in the next month. After repeating these procedures on each independent sample, we conduct the two analyses mentioned in Subsection 3.1, a conditional analysis that uses only the 200 samples with the influential value and an unconditional analysis using all the samples.

3.3. *Estimators and Evaluation Criteria*

To define the estimators, we first need some notation:

δ = u for the unconditional analysis,
c for the conditional analysis.
$S(\delta)$ = the total number of samples selected for analysis δ
$S(u)$ = 10,742 for the unconditional analysis in Industry 1
11,931 for the unconditional analysis in Industry 2
$S(c)$ = 200 for the conditional analysis in Industry 1 and Industry 2
ε = the outlier detection method
m = M-estimation
w = Clark Winsorization, none for the untreated estimate
Y_t = the true population total of the simulated data for month t
$\hat{Y}_{t,i}$ = the untreated estimate of Y_t for month t in sample i
$\hat{Y}^{\varepsilon}_{t,i}$ = the treated estimate of Y_t for month t in sample i with ε = M-estimation or Clark Winsorization.

The mean of the simulated values for month t, analysis δ, method ε is an estimate of Y_t

$$\hat{Y}^{\varepsilon}_{t}(\delta) = \frac{\sum_{i=1}^{S(\delta)} \hat{Y}^{\varepsilon}_{t,i}}{S(\delta)}.$$

The population values of the change are:

$\frac{Y_t}{Y_{t-1}}$ = true month-to-month change for the simulated data in month t, $t = 2$ to 20.

The estimates of this change are:

$\frac{\hat{Y}^{\varepsilon}_{t}(\delta)}{\hat{Y}^{\varepsilon}_{t-1}(\delta)}$ = estimate of month-to-month change for month t, analysis δ, method ε.

Now, let E^{ε}_{t} be a month t true population value, namely Y_t (total sales) or $\frac{Y_t}{Y_{t-1}}$ (month-to-month change). Also, let $\widehat{E}^{\varepsilon}_{tl}(\delta)$ be the estimate of total sales or month-to-month change

for month t, analysis δ, method ε from replicate i. Then the relative bias (RB) of $\widehat{E}_t^{\varepsilon}(\delta)$ is

$$\text{RB} = \frac{\sum_{i=1}^{S(\delta)} \left[\frac{100(\widehat{E}_{ti}^{\varepsilon}(\delta) - E_t^{\varepsilon})}{E_t^{\varepsilon}} \right]}{S(\delta)}. \tag{11}$$

We expect that the RB of the treated estimate is less than or equal to the RB of the untreated estimate in most circumstances.

The relative root mean square error (RRMSE) of $\widehat{E}_t^{\varepsilon}(\delta)$ is

$$\text{RRMSE} = \sqrt{\frac{\sum_{i=1}^{S(\delta)} \left[\frac{100(\widehat{E}_{ti}^{\varepsilon}(\delta) - E_t^{\varepsilon})}{E_t^{\varepsilon}} \right]^2}{S(\delta)}}. \tag{12}$$

We expect that the RRMSE of the treated estimate is less than or equal to the RRMSE of the untreated estimate since the methods minimize MSE.

Mirroring Thompson and Sigman (1999), to evaluate the outlier detection performance of each method, we view each application as a hypothesis test, in which the null hypothesis is "the data item's value is *not* an influential value". One rejects the null hypothesis when the item's value is flagged as influential. Under this framework, two types of errors can occur:

- **Type I error rate** equals the percentage of observations that were *not induced* influential values that were designated as influential (false positive). If a method adjusts values that are not induced influential values, then the Type I error rate will be positive.
- **Type II error rate** equals the percentage of *induced* influential values that were not detected (false negative). The Type II error rate applies only to samples containing the induced influential value. So, the Type II error rate is equal to 0 in Months 1–3 and 5–20 since no influential values were induced in these months.

4. Results

In this section, we examine the simulation results regarding the performance of the two treatments and the quality of the estimates they produce. The Clark Winsorization algorithm does not require parameter settings, but the M-estimation algorithm does. First, we investigate the settings of the parameters for the M-estimation algorithm to determine which options produce the best estimates. Then we use those settings for M-estimation in the comparison with Clark Winsorization. As we will see in the simulation results, the choices of the M-estimation parameter settings affect whether the algorithm converges in some situations and therefore are important. For the Winsorization, we developed the software in SAS. For the M-estimation, we used SAS software developed by Jean-Francois Beaumont (personal communication), with minor modifications.

4.1. M-estimation Algorithm Settings

The M-estimation algorithm discussed in Subsection 2.2 requires settings for Q, h_i, v_i, the function ψ, and an initial value of the tuning constant φ. We use the default settings of

$Q = 1$ and $h_i = (w_i - 1)\sqrt{x_i}$, but explore different settings for the other parameters, as summarized in Table 1. We also consider whether to include the observations selected with certainty in fitting the regression line.

Our investigation considers two values of the weighting parameter for the residuals $v_i = \lambda x_i$ namely $v_i = x_i$ and $v_i = 1$. The choice $v_i = 1$ corresponds to $\lambda = 1/x_i$ so that $V_m\left[y_i|x_i\right] = \sigma^2$ (equal variances) and the choice $v_i = x_i$ corresponds to $\lambda = 1$ so that $V_m\left[y_i|x_i\right] = x_i\sigma^2$. Ideally, the choice of the setting for v_i should be a data-driven decision because v_i essentially specifies the variance of the model errors underlying the regression estimator for M-estimation. In our (realistic) setting, neither $v_i = x_i$ nor $v_i = 1$ provide a good model for the studied *industry* level estimates from the MRTS data. Indeed, this model misspecification is an inherent challenge with economic data.

Notice that when we used the default settings for Q and h_i along with setting $v_i = x_i$ for all units in sample, $r_i = (w_i - 1)(y_i - x_i\hat{B}^M)$ has the same form as $\hat{D}_i$ in the Clark Winsorization. However, recall that the b in the Winsorization estimation method and the $\hat{B}^M$ in the M-estimation method are not usually going to be equal because they use different estimation methods. With $Q = 1$ and $h = (w_i - 1)\sqrt{x_i}$ (the default settings), setting $v_i = 1$ tends to give the residuals for large weighted values of x_i more influence in fitting the M-estimation regression line than when $v_i = x_i$.

The M-estimation algorithm detects and adjusts influential values through finding an optimal value of the tuning constant φ, which is the cut-off value for the weighted regression residuals. The user sets an initial value for the tuning constant φ, and the algorithm finds the value of φ that minimizes the mean squared error (MSE). Setting the algorithm parameters in a manner appropriate for the MRTS data requires considerable investigation. We consider two options for the function ψ, the one-sided Huber I and II functions described in Subsection 2.2 and two options for the initial value of φ, one high and the other low. After exploring the application of M-estimation to samples that included and excluded the units selected with certainty, we found little difference and included the certainty units in our simulation. The units selected with certainty contribute to fitting the regression line but cannot be designated as influential because $r_i(\hat{B}^M)$ equals zero for a certainty unit with the default setting $h_i = (w_i - 1)\sqrt{x_i}$.

Selecting the high and low initial values of φ for the simulation depends on the data for the industry. If there are no weighted residuals larger than the initial value of φ, the M-estimation algorithm runs for only one iteration and does not offer any adjustments.

Table 1. M-estimation algorithm parameters

Parameter	Parameter function	Values
Q	Constant	1 (default)
h_i	Unit weight	$(w_i - 1)\sqrt{x_i}$ (default)
v_i	Model error underlying regression estimator	1 or x_i
ψ	ψ function	Huber I or Huber II
φ	Tuning constant (determines starting point for critical region)	User provides initial value and program calculates optimal value

Therefore, for low initial φ we choose a value that tended to be lower than the highest weighted residual in a sample since we wanted the algorithm always to run in the simulation. For the high initial φ, we want only to assure that the algorithm detects the induced influential value when it appears in Month 4. Consequently, we choose a value that is lower than the weighted residual for the induced influential value but higher than the weighted residuals for the other values. For Industry 1, the low initial φ is 4.8 million and the high value initial value is 150 million. The low and high initial values of φ for Industry 2 are 1.5 million and 150 million, respectively.

Table 2 summarizes the results for Type I and Type II errors for the parameter settings for Industry 1 and Industry 2 and offers results for the different parameter settings and functions using Type I and Type II errors as the evaluation criteria. A Type I error (false positive) may occur in all the months in all the samples, but a Type II error (false negative) may occur only in Month 4 of the 200 samples with the induced influential value in Month 4.

Both settings for the parameter v_i display some Type I errors when the initial setting of φ is the low value of 4.8 million while there are no Type I errors when the initial φ is the large value of 150 million. The Type I errors occur because the algorithms for Clark Winsorization and M-estimation when the initial φ is low (4.8 million) make small adjustments to several observations to achieve the minimum MSE although the reduction in MSE is small.

Remember that neither $v_i = 1$ nor $v_i = x_i$ is an appropriate error model for the simulated data for either of the two industries. The Type I and Type II errors are very similar for the two choices of the function ψ, Huber I and Huber II, when the same high or low initial φ is used in the unconditional analysis. The Type II error rate for $v_i = 1$ is zero for both options for the initial φ in Month 4 for Industry 1 for both Huber I and Huber II. However, when $v_i = 1$ for Industry 2, the Type II error rate is 0.0065 for the high initial φ, and 0.04 for Huber I and 0.05 for Huber II for the low initial φ. The Type II error rate when $v_i = x_i$ is always zero for all combinations of the options.

Table 2. Summary of M-estimation results for the unconditional analysis with Industry 1 and Industry 2 data in the scenario of one high influential value for two settings of the parameters v_i, two settings of the initial φ, and two options for the function ψ

	ψ function			
v_i	Huber I	Huber II	Type I error	Type II error
x_i	Option 1	Option 2	• Small Type I error rate when initial φ small at 4.8 million • No Type I errors when initial φ large at 150 million	Industry 1 rate: zero Industry 2 rate: zero
1	Option 3	Option 4	• Very small Type I error rate when initial φ small at 1.5 million • No Type I errors when initial φ large at 150 million	Industry 1 rate: zero Industry 2 rates: • when initial φ small, 0.04 for Huber I, 0.05 for Huber II • when initial φ large, 0.0065 for Huber I & II

Since there is some Type II error when $v_i = 1$ and none when $v_i = x_i$, and the two settings produce about the same results regarding Type I error, we decided to pursue only $v_i = x_i$.

4.2. *One High Influential Value*

4.2.1. Industry 1 Estimates and Quality

First, we focus on the simulation results for Industry 1, the more volatile of the two simulated industries and the larger of the two (in terms of sample size and total sales). We show results for only the Huber II function ψ because results for Huber I and Huber II functions are approximately equal. Since the M-estimation algorithm is an iterative procedure, convergence is not guaranteed. We used the default convergence criterion of a difference of 0.001 between the current and previous iterations and did not explore other options. In this simulation, the algorithm did not converge for about two percent of the samples in the unconditional analysis. Usually a researcher puts a limit on the number of iterations that the algorithm may run. We chose a limit of five iterations. When the limit is reached, the program choses the larger of the last two values of φ. The results for the performance measures include the consequences of this choice. In the conditional analysis, the algorithm converged for Month 4 in all 200 samples, and the convergence properties in other months were similar to those in the same months in the rest of samples in the unconditional analysis.

The relative bias estimates of total sales in Months 2 to 7 in the unconditional and conditional analyses are shown in Table 3 while Table 4 shows the RRMSE estimates for the same months. The population value of total sales in these months varies slightly around \$46.1 billion. Tables 3 and 4 only show the results involving Months 2 through 7 because the results for the rest of the 20 months parallel those involving Month 7. This is to be expected since the series is stationary and only Month 4 has an induced influential value.

In the unconditional analysis, the untreated estimate of the total for Month 4 has a relative bias of 0.012 percent, corresponding to approximately \$4.6 million, and an even smaller relative bias in the other months, corresponding to $-\$1.7$ million to $-\$3.6$ million. Since the reported estimates of total sales are in millions, this level of bias does appear in the reported estimates and is within the survey sampling error where the coefficient of variation is approximately two percent. In Month 4, the treated estimates do reduce the bias even further, with M-estimation with a high initial φ having the lowest absolute relative bias. In the other months, estimates of total from M-estimation with a high initial φ have a relative bias equal to that of the untreated because no observations are adjusted in those months. However, in months other than Month 4, Clark Winsorization and M-estimation with the low initial φ tend to introduce additional negative relative bias, about -0.01 percent, because they tend to trim about 0.5 percent of the observations to achieve a minimum MSE. Interestingly, Table 4 shows that the three methods produce estimates of total sales for Month 4 with approximately the same RRMSE of 1.261 in the unconditional analysis. Since Table 3 shows that Clark Winsorization and M-estimation with a low initial φ have more relative bias than M-estimation with a high initial φ, we

Table 3. Relative bias (percent) for one high influential value scenario (Industry 1, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	−0.005	−0.005	−0.015	−0.015	0.246	0.246	0.236	0.236
3	−0.004	−0.004	−0.016	−0.015	0.256	0.256	0.244	0.244
4	0.012	0.003	−0.008	−0.008	1.166	0.716	0.716	0.721
5	−0.006	−0.006	−0.018	−0.018	0.237	0.237	0.222	0.225
6	−0.005	−0.005	−0.016	−0.016	0.233	0.233	0.222	0.222
7	−0.007	−0.007	−0.018	−0.018	0.238	0.238	0.226	0.226
Month-to-Month change								
2 to 3	0.001	0.001	−0.001	−0.0001	0.010	0.010	0.008	0.008
3 to 4	0.016	0.007	0.008	0.008	0.908	0.460	0.471	0.475
4 to 5	−0.017	−0.009	−0.010	−0.010	−0.918	−0.476	−0.491	−0.493
5 to 6	0.001	0.001	0.002	0.002	−0.004	−0.004	<0.001	−0.003
6 to 7	−0.002	−0.002	−0.002	−0.002	0.004	0.004	0.004	0.005

Table 4. RRMSE (percent) for one high influential value scenario (Industry 1, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	1.255	1.255	1.255	1.255	1.229	1.229	1.227	1.227
3	1.257	1.257	1.257	1.257	1.223	1.223	1.220	1.220
4	1.267	1.261	1.261	1.261	1.675	1.400	1.400	1.403
5	1.257	1.257	1.257	1.257	1.221	1.221	1.218	1.218
6	1.255	1.255	1.256	1.256	1.233	1.233	1.231	1.231
7	1.256	1.256	1.256	1.256	1.222	1.222	1.219	1.219
Month-to-Month change								
2 to 3	0.106	0.106	0.106	0.105	0.097	0.097	0.097	0.097
3 to 4	0.165	0.125	0.126	0.126	0.914	0.471	0.482	0.486
4 to 5	0.167	0.128	0.129	0.128	0.925	0.489	0.503	0.505
5 to 6	0.108	0.108	0.108	0.107	0.109	0.109	0.109	0.109
6 to 7	0.108	0.108	0.108	0.108	0.108	0.108	0.107	0.107

conclude that these estimates achieve a comparable RRMSE by reducing the variance through trimming several observations. We are observing a classic bias versus variance trade-off and since the bias is a small component of the RRMSE, changes to the variance have a larger impact.

When we turn to month-to-month change in the unconditional analysis, the induced influential value in Month 4 causes a positive bias in the untreated estimate of change from Months 3 to 4 and a negative bias of comparable size in the untreated estimate of change from Months 4 to 5. All the treated estimates reduce the relative bias by about half in the change from Months 3 to 4 and from Months 4 to 5. The treatments reduce the RRMSE in the untreated estimate by about 24 percent. As with the estimates of total, the relative bias and RRMSE for the untreated and treated estimates of change are comparable in the months not involving Month 4.

For the conditional analysis, Table 3 shows that the relative bias is approximately equal for all the estimates of total sales in Months 2, 3, and 5 to 7. In Month 4, the relative bias in both versions of M-estimation and Clark Winsorization is approximately 60 percent of the relative bias in the untreated estimate. Recall that the simulation design introduces a small amount of bias in the conditional analysis. Table 4 shows that Clark Winsorization and both versions of M-estimation produce estimates with approximately 84 percent of RRMSE for the untreated estimate in Month 4, but the RRMSEs are comparable in the other months.

In the conditional analysis in Table 3, we see that untreated and treated estimates of change from Months 3 to 4 have a positive relative bias and an approximately offsetting negative relative bias for the change from Months 4 to 5. The relative bias for the estimates of change that do not involve Month 4 is very small and does not appear in estimates of change which are reported in tenths of percent. When Month 4 is involved, the untreated estimates of change would be apparent in the reported estimates. All treatments reduce the relative bias by approximately one-half with M-estimation with a high initial φ having slightly less relative bias than Clark Winsorization and M-estimation with a low initial φ. The treatments also reduce RRMSE in the untreated estimates of change by about one-half with M-estimation with a high initial φ having the lowest as shown in Table 4. Apparently, the trimming by the latter two methods to reduce the variance in the estimates of total sales creates additional bias in the estimates of change when Month 4 is involved. Clark Winsorization and M-estimation with a low initial φ appear to have some residual effect in the estimate of change from Months 5 to 6 since each has a lower relative bias than the untreated estimate and M-estimation with a high initial φ. However, the RRMSEs of all four estimates of change are approximately equal.

4.2.2. Industry 2 Estimates and Quality

Now we turn our attention to the simulation results for Industry 2, which has a less volatile pattern of change and a smaller sample size than Industry 1. The population value of total sales in these months is about $2.5 billion and the sample size is 147.

The patterns in the performance measures for the unconditional analysis for Industry 2 shown in Tables 5 and 6 are very similar to the results for Industry 1. The effect of the induced influential value in Month 4 is larger because its size relative to the population total is larger as is the effect of adjusting it. The M-estimation algorithm converged for all

Table 5. Relative bias (percent) for one high influential value scenario (Industry 2, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
Month	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	0.007	0.007	−0.028	−0.044	0.103	0.103	0.068	0.053
3	0.010	0.010	−0.029	−0.075	0.173	0.173	0.120	0.089
4	0.330	0.172	0.132	0.122	19.021	9.607	9.600	9.737
5	0.006	0.006	−0.060	−0.053	0.079	0.079	−1.307	0.024
6	0.011	0.011	−0.034	−0.050	0.191	0.191	0.133	0.127
7	0.011	0.011	−0.030	−0.047	0.134	0.134	0.094	0.079
Month-to-Month change								
2 to 3	0.004	0.004	<0.001	−0.030	0.071	0.071	0.053	0.037
3 to 4	0.319	0.162	0.161	0.197	18.829	9.424	9.475	9.647
4 to 5	−0.272	−0.151	−0.175	−0.160	−15.922	−8.697	−9.957	−8.856
5 to 6	0.005	0.005	0.026	0.003	0.112	0.112	1.461	0.104
6 to 7	<0.001	<0.001	0.004	0.003	−0.056	−0.056	−0.038	−0.048

Table 6. RRMSE (percent) for one high influential value scenario (Industry 2, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
Month	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	2.783	2.783	2.783	2.784	2.620	2.620	2.616	2.618
3	2.773	2.773	2.773	2.772	2.593	2.593	2.588	2.587
4	3.712	3.043	3.043	3.051	19.196	9.948	9.942	10.075
5	2.770	2.770	2.775	2.772	2.609	2.609	2.912	2.608
6	2.770	2.770	2.771	2.771	2.615	2.615	2.612	2.611
7	2.782	2.782	2.782	2.783	2.607	2.607	2.604	2.603
Month-to-Month change								
2 to 3	0.277	0.277	0.276	0.293	0.287	0.287	0.283	0.289
3 to 4	2.454	1.251	1.257	1.281	18.838	9.432	9.482	9.654
4 to 5	2.080	1.160	1.320	1.181	15.928	8.704	9.970	8.863
5 to 6	0.276	0.276	0.339	0.283	0.287	0.287	1.555	0.294
6 to 7	0.271	0.271	0.269	0.277	0.281	0.281	0.273	0.283

samples with the high initial φ, but with the low initial φ, each month experienced a failure to converge in approximately ten percent of the samples. However, this does not appear to change the pattern observed in the unconditional analysis for Industry 1. Yet, M-estimation with the low initial φ experienced convergence problems in Month 5 in 106 of the 200 samples in the conditional analysis, which is the focus of this section. The reason for the failure to converge is a combined effect of a low influential value in Month 5 that is the consequence of an induced very high influential value in Month 4 and the small sample size in Industry 2.

In Month 4 in the conditional analysis, M-estimation with a high initial φ reduces the relative bias in the untreated estimate of total sales by 49 percent. The reduction in relative bias using M-estimation with a low initial φ is 50 percent while for Clark Winsorization the reduction is 49 percent. Viewing the results for the estimates of total sales for the other months, the relative bias and RRMSE from M-estimation with a high initial φ equal those for the untreated. In months other than Month 5, M-estimation with a low initial φ reduces the relative bias in the untreated estimate by 30 to 34 percent while Clark Winsorization achieves reductions ranging from 35 to 49 percent. Both methods appear to be trimming as in their application to Industry 1 although the percentage reductions are greater than seen for Industry 1. However, Month 5 is different – the relative bias for M-estimation with a low initial φ has a much bigger absolute value than the untreated and is negative which makes the *RRMSE* twelve percent higher than the untreated.

When we turn to month-to-month change, we see more anomalies when Month 5 is involved. First, for estimates involving neither Month 4 nor Month 5, the relative bias for M-estimation with a high initial φ equals the relative bias for the untreated while the trimming by M-estimation with a low initial φ and Clark Winsorization achieves reductions of 17 to 48 percent, but the RRMSEs for all four estimates are comparable. The relative bias in the untreated estimate of change for Months 3 to 4 continues to offset the relative bias for Months 4 to 5. All three treatments achieve a reduction of approximately 50 percent in RRMSE of the untreated estimate of change from Months 3 to 4. For the change from Months 4 to 5, both M-estimation with a high initial φ and Clark Winsorization reduce the relative bias by about 45 percent while M-estimation with a low initial φ produces a 30 percent reduction. The reductions in the RRMSE for the untreated estimate are comparable to the percentage reductions in the relative bias for the three treatments. For the change from Months 5 to 6, the relative bias in the untreated and M-estimation with a high initial φ are equal but slightly larger than Clark Winsorization. However, the relative bias for M-estimation with a low initial φ is 1.461 percent, an order of magnitude higher than for the other three estimates.

An examination of the data provides insight about what happens with the M-estimation algorithm when using the low initial φ in Month 5 for some samples with the induced influential value in Month 4. The algorithm identifies and treats the influential value in Month 4. However, in Month 5 the sample unit returns to a range closer to its value in Month 3. In some samples, but not all, the Month 5 value is small enough to create an unusually large negative weighted regression residual as illustrated in Figure 6.

Because the version of the M-estimation algorithm used in the simulations uses a one-sided Huber II function ψ, it does not treat unusually low values, and therefore, the MSE

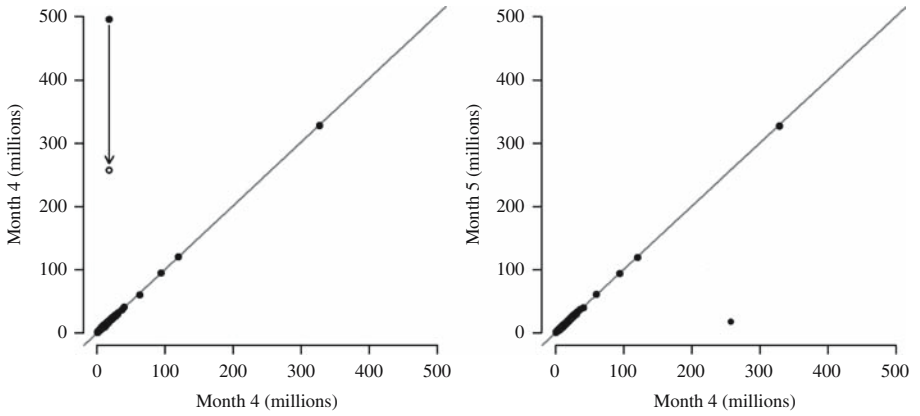


Fig. 6. Scatterplots of Month 4 versus Month 3 (left) and Month 5 versus Month 4 (right) with robust regression line when applying M-estimation with a low initial φ in a sample from Industry 2. The unusually high influential value in Month 4 was adjusted but not enough to avoid producing an unusually low influential value in Month 5 when the unit returned to its routine range

can be a strictly decreasing function of φ, which causes the algorithm not to converge. In the case of a strictly decreasing MSE, the algorithm does not converge by the limit on the number of iterations (five in our study) and instead selects the larger φ of the last two iterations, which is usually very small. This small φ causes the program to flag many observations in the sample as influential and to adjust them in over 50 percent of the samples. When using the one-sided version of the M-estimation algorithm, the adjustments reduce only observations larger than their previous month's values and thereby introduce a negative bias in the estimates of total sales. If the limit for the number of iterations increases beyond five, in some applications the algorithm converges to a local minimum that is usually very small. Therefore, increasing the number of iterations does not solve the problem.

To gauge whether a two-sided function ψ would perform better than a one-sided function ψ with a low initial value of φ, we applied the M-estimation algorithm to Months 4 and 5 to the 200 replicates that contained the influential value, but also found convergence problems. In Month 4, the algorithm failed to converge for eleven samples, but 96 of the 189 that achieved convergence produced a final value of φ that was very small and therefore, not helpful because it designated a large number of observations as influential. Results in Month 5 also were problematic since the algorithm did not converge for 39 samples and of the 161 achieved that convergence, 21 converged to nearly zero. In one other sample where the algorithm converged, it flagged more than ten percent of the observations as influential, which we consider to be many.

The samples with convergence problems caused by the induced high influential value returning to its routine range and producing a particularly low residual (Figure 6) illustrate the situation where the most desirable option probably is no adjustment. With the high initial φ setting, no residual is larger than the initial φ so the M-estimation algorithm does not run for any of the samples, and therefore, it produces no adjustment, and achieves the desirable option. This highlights the importance of choosing the initial φ to be a value low

enough that an observation with a larger weighted residual requires an adjustment, but high enough for the algorithm not to run when no adjustment is needed.

5. Summary

Our investigation finds both weighted M-estimation and Clark Winsorization to be effective in identifying and treating influential values; however, each method has advantages and disadvantages that may affect a decision about which to employ. Although the simulation procedure was designed to produce data similar to the Census Bureau's MRTS, the studied problem and context are broadly applicable to other programs.

A big advantage of Clark Winsorization is the ease of implementation of its straightforward formulas. By design, the method identifies and treats only influential values that are unusually high so it does not identify or treat values that are influential because they are unusually low. However, the major concern in economic surveys regarding influential values usually is the occurrence of high ones. When an influential value is present, Clark Winsorization always identifies it and offers an adjustment.

On the other hand, the Clark Winsorization trims about 0.5 percent of the observations when no influential value is present in the sample, introducing adjustments that achieve a very small reduction in MSE for estimated totals and month-to-month change. The trimming increases the bias of the Winsorized estimate over that obtained with M-estimation with a high initial φ. Since the Clark Winsorization trimming reduces the variance in the treated estimates, the RRMSEs of the two studied methods are comparable. The trimming is also disadvantageous because the staff usually researches whether observations flagged as influential are accurate. The tight time schedule for production of monthly estimates requires avoiding unnecessary investigations. However, in some situations, the ease of implementation of Clark Winsorization and the protection that it offers against unusual influential values could outweigh the small amount of bias introduced by trimming a few falsely identified observations by a small amount. These would be situations where knowledge of the population is limited and/or where verification of values designated as influential could be restricted to focus only on those with treated values exhibiting large changes relative to the remainder of the units.

The weighted M-estimation methodology identifies and treats both high and low influential values. Our investigation focused on high influential values because they usually are the major concern in the studied programs although low influential values do occur and can introduce bias. The M-estimation algorithm has flexibility in setting parameters to make assumptions appropriate for the underlying data. In addition, weighted M-estimation with a high value of the initial tuning constant φ performed the best overall of the three options considered.

An attractive feature of M-estimation is that the algorithm allows an analyst to set the value of the initial tuning constant φ and thereby determine the minimum size of the weighted regression residuals that will be considered as potential influential values. This facilitates the efficient use of staff time in examining proposed adjustments. However, setting the initial φ is important to the effectiveness of the algorithm and needs to be a data-driven decision based on exploratory analysis. Some further refining may occur as the procedure is used in practice. In addition, there is a need to have a back-up strategy for

situations when the algorithm does not converge and for situations when the algorithm converges but does not provide helpful results. In the latter cases, an influential value is present, but the MSE is either a strictly decreasing or a strictly increasing function of the tuning constant φ resulting in adjustments for almost all or none of the observations. If the MSE does have a global minimum but the algorithm does not converge, then changing the initial φ to be close to the value of φ corresponding to the minimum MSE usually results in the algorithm converging.

Research is currently underway on how to set the initial φ in an ongoing monthly survey that may or may not be subject to seasonal effects, but the approaches under study require at least minimal prior knowledge of the population. If one has no prior knowledge of the population, one could take the approach of applying Clark Winsorization. If Clark Winsorization produces no adjustment or merely trimming, then no adjustment is an acceptable choice.

Other research on M-estimation and Winsorization methods have either supported or not contradicted our findings. In a recent study with the U. S. Census Bureau's Annual Survey of Public Employment and Payroll, M-estimation also performed better than Clark Winsorization (Barth et al. 2012). In another study, Lewis (2007) attempted to formulate methodology for Winsorization of estimates of change, but did not find a satisfactory method in spite of making more restrictive assumptions than presented here.

Ultimately, we believe that the trimming of some observations by Clark Winsorization that introduces some bias for a small reduction in MSE is a less than desirable feature and instead choose to focus on M-estimation applications, with the full endorsement of the MRTS program managers. Implementing M-estimation in MRTS requires investigating the remaining issues, such as seasonality, data-driven methods of optimizing the selection of the initial tuning constant φ, and – most important – a changing economy. The flexibility of M-estimation makes the approach particularly appealing given these challenges.

6. References

Barth, J., J. Tillinghast, and M.H. Mulry. 2012. "Treatment of Influential Values in the Annual Survey of Public Employment and Payroll." In Proceedings of the 2012 Research Conference of the Federal Committee on Statistical Methods. Office of Management and Budget. Washington, DC. Available at: https://fcsm.sites.usa.gov/files/2014/05/Barth_2012FCSM_III-D.pdf (accessed October 10, 2014)

Beaumont, J.-F. 2004. *Robust Estimation of a Finite Population Total in the Presence of Influential Units.* Report for the Office for National Statistics, dated July 23, 2004. Office for National Statistics, Newport, U.K.

Beaumont, J.-F., and A. Alavi. 2004. "Robust Generalized Regression Estimation." *Survey Methodology* 30: 195–208.

Black, J. 2001. "Changes in Sampling Units in Surveys of Businesses." In Proceedings of the Federal Committee on Statistical Methods Research Conference. Office of Management and Budget. Washington, DC. Available at: http://www.fcsm.gov/files/2014/05/2001FCSM_Black.pdf (accessed October 20, 2014)

Chambers, R.L. and R. Ren. 2004. "Outlier Robust Imputation of Survey Data." In Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM]. American Statistical Association. Alexandria, VA. 3336–3344. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2004/files/Jsm2004-000559.pdf (accessed October 20, 2014)

Chambers, R.L., P. Kokic, P. Smith, and M. Cruddas. 2000. "Winsorization for Identifying and Treating Outliers in Business Surveys." In Proceedings of the Second International Conference on Establishment Surveys. Statistics Canada. Ottawa, Canada. 717–726.

Clark, R. 1995. "Winsorization Methods in Sample Surveys." Masters Thesis. Department of Statistics. Australia National University. Available at: http://hdl.handle.net/10440/1031 (accessed October 21, 2014)

Farrell, P.J. and M. Salibian-Barrera. 2006. "A Comparison of Several Robust Estimators for a Finite Population Mean." *Journal of Statistical Studies* 26: 29–43.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and S.A. Werner. 1986. *Robust Statistics. An Approach Based on Influence Functions*. New York: John Wiley & Sons.

Huang, E. 1984. "An Imputation Study for the Monthly Retail Trade Survey." In Proceedings Joint Statistical Meeting, Survey Research Methods Section, American Statistical Association. Alexandria, VA. 610–615.

Huber, P.J. 1964. "Robust Estimation of a location parameter." *Annals of Mathematical Statistics. Institute of Mathematical Statistics* 35: 73–101.

Hidiroglou, M.A. and J.-M. Berthelot. 1986. "Statistical Editing and Imputation for Periodic Business Surveys." *Survey Methodology* 12: 73–83.

Hulliger, B. 1995. "Outlier Robust Horvitz-Thompson Estimators." *Survey Methodology* 21: 79–81.

Hunt, J.W., J.S. Johnson, and C.S. King. 1999. "Detecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Berthelot Method." In Proceedings of the Section on Survey Research Methods. American Statistical Association. Alexandria, VA. 539–543. Available at: http://www.amstat.org/sections/SRMS/Proceedings/papers/1999_093.pdf (accessed October 20, 2014)

Kokic, P.N. and P.A. Bell. 1994. "Optimal Winsorising Cut-Offs for a Stratified Finite Population Estimator." *Journal of Official Statistics* 10: 419–435.

Lewis, D. 2007. "Winsorisation for estimates of change." *Proceedings of the Third International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA. 1165–1172.

Mulry, M.H. and B. Oliver. 2009. "A Simulation Study of Treatments of Influential Values in the Monthly Retail Trade Survey." *JSM Proceedings*, Survey Research Methods Section. American Statistical Association. Alexandria, VA. 2979–2993. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2009/Files/304284.pdf (accessed October 20, 2014)

Mulry, M.H. and R. Feldpausch. 2007a. "Investigation of Treatment of Influential Values." *Proceedings of the Third International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA. 1173–1179.

Mulry, M.H. and R. Feldpausch. 2007b. "Treating Influential Values in a Monthly Retail Trade Survey." *Proceedings of the Survey Methods Section, SSC Annual Meeting*.

Statistical Society of Canada. Ottawa, Ontario, Canada. Available at: http://www.ssc.ca/survey/documents/SSC2007_M_Mulry.pdf (accessed October 20, 2014)

Ren, R. and R.L. Chambers. 2003. "Outlier Robust Imputation of Survey Data via Reverse Calibration." S3RI Methodology Working Paper M03/19. Southampton Statistical Sciences Research Institute, University of Southampton, U.K. Available at: http://www.eprints.soton.ac.uk/8169/1/8169-01.pdf (accessed October 20, 2014)

Rousseeuw, P.J. 1984. "Least Median of Squares Regression." *Journal of the American Statistical Association* 79: 871–880.

Rousseeuw, P.J. and A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Thompson, J.R. 2000. *Simulation: A Modeler's Approach*. New York: John Wiley and Sons. 87–110.

Thompson, K.J. and K.T. Washington. 2013. "Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study." *Survey Methods: Insights from the Field*. Available at: http://surveyinsights.org/?p=2991 (accessed October 20, 2014)

Thompson, K.J. and R.S. Sigman. 1999. "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data." *Journal Official Statistics* 15: 517–535.

Received November 2012
Revised October 2014
Accepted October 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 749–771, http://dx.doi.org/10.2478/JOS-2014-0046

The Impact of Sampling Designs on Small Area Estimates for Business Data

Jan Pablo Burgard[1], *Ralf Münnich*[1], *and Thomas Zimmermann*[1]

Evidence-based policy making and economic decision making rely on accurate business information on a national level and increasingly also on smaller regions and business classes. In general, traditional design-based methods suffer from low accuracy in the case of very small sample sizes in certain subgroups, whereas model-based methods, such as small area techniques, heavily rely on strong statistical models.

In small area applications in business statistics, two major issues may occur. First, in many countries business registers do not deliver strong auxiliary information for adequate model building. Second, sampling designs in business surveys are generally nonignorable and contain a large variation of survey weights.

The present study focuses on the performance of small area point and accuracy estimates of business statistics under different sampling designs. Different strategies of including sampling design information in the models are discussed. A design-based Monte Carlo simulation study unveils the impact of the variability of design weights and different levels of aggregation on model- versus design-based estimation methods. This study is based on a close to reality data set generated from Italian business data.

Key words: Nonignorable sampling designs; MSE estimation; confidence interval coverage.

1. Business Surveys and Small Area Estimation

Statistical offices increasingly face the challenge of producing estimates on subgroups in addition to national estimates. In business statistics, these subgroups may consist of regions or NACE classes (Nomenclature statistique des activités économiques dans la Communauté européenne, Eurostat 2008). Generally, in business surveys the sampling designs are optimized to furnish national estimates with a desired level of accuracy which may lead to unsuitably small sample sizes for subgroups of interest. Since the precision of direct estimates, for example measured by the variance of the estimator, is inversely proportional to the sample size, the resulting small sample sizes may lead to unreliable direct estimates for these subgroups. Hence, alternative estimators may have to be considered.

[1]University of Trier – Fachbereich IV, Lehrstuhl für Wirtschafts- und Sozialstatistik, Universitätsring 15 Trier D-54286, Germany. Emails: muennich@uni-trier.de, burgardj@uni-trier.de, and thzimmer@uni-trier.de

Acknowledgments: The authors greatly acknowledge the very useful comments of Natalie Shlomo as well as an associate editor and three referees which helped to improve this paper substantially. The research is embedded within the BLUE-ETS project, which is financially supported by the European Commission within the 7th Framework Programme (cf. http://www.blue-ets.eu). The authors would like to thank the Italian National Institute of Statistics (ISTAT) for kindly providing the data sets on which this study is based.

Over the past decades, small area estimation techniques have gained popularity. The main idea behind these methods is to borrow information from other subgroups via statistical models in order to increase the effective sample size of the subgroups of interest (cf. Rao 2003). One major reservation in official statistics against the use of model-based methods is the possible lack of design unbiasedness. In the presence of small sample sizes, however, the design biasedness may play a minor role in assessing the precision of the estimates because of the variability caused by small sample sizes. A widely used measure to assess the precision of estimates is the mean square error (MSE), which considers both the squared bias and the variance of an estimator. Model-based small area methods typically have lower variances but may suffer from design bias. In contrast to model-based methods, design-based methods are design unbiased at the expense of large variances with small sample sizes. Thus there is a trade-off between bias and variance of the different estimators. Therefore, the selection of an estimator of either kind has to be made carefully in any application. While small area estimation is increasingly used in many fields of social statistics, such as the estimation of poverty measures (cf. Molina and Rao 2010, or Lehtonen et al. 2011), it has not yet been widely used in the area of business statistics. Small area estimation techniques use models for the prediction of the quantity of interest. This approach relies heavily on the availability of strong predictive variables for modeling the dependent variable. This auxiliary information usually comes from business registers. The higher the predictive power of the model, the better estimates are produced.

In this article, we want to raise and discuss two issues arising in the application of small area estimation methods for business statistics. First, in many countries business registers do not include strong auxiliary information, leaving the data producer with little choice regarding model building and variable selection. Nevertheless, the data producer might be obliged to publish information on subgroups under these less suitable conditions and without sufficient sample sizes for applying design-based methods. Options available to a data producer are discussed.

Secondly, sampling designs in business statistics are often nonignorable due to a high market concentration of important variables, such as total turnover. The designs, mainly stratified or probability proportional to size, work well within a design-based framework for estimating national figures. However, most small area estimators operate in a model-based framework ignoring the sampling design. In the case of informative sampling designs, this may lead to erroneous statistical inferences (cf. Pfeffermann and Sverchkov 2009). In this case, one option is to correct for the design bias due to the informativeness directly (cf. Pfeffermann and Sverchkov 2007). Another approach incorporates the design weights into the estimation of the statistical model. In a Bayesian context, this issue has been addressed by You and Rao (2003) and Little (2012). A discussion on weighting and prediction in the context of small area estimation from a frequentist's viewpoint is given in Pfeffermann (1993) and Pfeffermann et al. (1998) and for multilevel modeling in general in Asparouhov (2006). In our article we compare different frequentist strategies for including design weights in small area modeling.

In Section 2, we describe the sampling designs used in business surveys and discuss their usefulness for small area estimation. This is followed by the presentation of the small area estimators considered in our study, including their properties with respect to complex survey designs. In Section 3, we describe our data set and outline our design-based

simulation strategy followed by a discussion of the results of our simulation study. Finally, we summarize our findings in Section 4.

2. Small Area Estimation and Modeling

2.1. Sampling Designs for Business Surveys

In business surveys, stratified sampling designs are typically applied. The strata are often determined as cross-classifications of variables such as industry classifications, geographical information or employee size classes (cf. Hidiroglou and Lavallee 2009). Since the present article focuses on enterprise-level business surveys, we omitted multistage designs which, in general, are not applied in business statistics (cf. Thompson and Oliver 2012). Some ideas in the context of small area applications for household surveys can be drawn from Münnich and Burgard (2012).

Frequently, the survey planner who designs the survey and chooses the estimator faces a conflict between obtaining reliable estimates for small domains and for national figures. Furthermore, the planner has to consider the impact of the design on the estimator as well as decide on the level of aggregation at which the estimates are required. This decision-theoretic problem may be addressed by specifying a loss function, which is to be minimized under certain constraints.

Longford (2006) minimizes the weighted sum of domain-specific variances and the variance of the national estimators subject to the sample size restriction, where the weights specify the relative importance of each domain and the priority for the national estimate. Choudhry et al. (2012) consider the problem of minimizing the total sample size subject to the upper bounds of the coefficients of variation for the strata means and the national mean by using nonlinear programming techniques. Another approach introduced by Costa et al. (2004) does not require an explicit loss function but consists of a convex combination of the equal and proportional allocation with L strata ($h = 1, \ldots, L$):

$$n_{h,Costa} = kn\frac{N_h}{N} + (1-k)\frac{n}{L},\ 0 \leq k \leq 1, h = 1, \ldots, L, \tag{1}$$

where n_h denote the stratum-specific sample sizes with total sample size n, N_h is the number of units in the h-th stratum summing up to the total number of units N, and k is a weighting constant, which yields the equal allocation for $k = 0$ and the proportional allocation for $k = 1$. The idea behind the Costa allocation is that the equal allocation is favorable for domain level estimates but not very efficient for national estimates, whereas the opposite holds for proportional allocation. In addition to reaching a compromise between efficient estimation at different levels of aggregation, allocation (1) is also particularly easy to apply. The optimal allocation due to Neyman (1934) and Tschuprow (1923) minimizes the variance of the national mean estimator $\hat{\mu}$ of the variable of interest Y for stratified random sampling. If we are interested in small domain estimates, however, this will not be sufficient, since the optimal allocation leads to very small domain-specific sample sizes in cases where there is hardly any variation within a stratum. This may yield stratum-specific sample sizes $n_h < 2$ which do not allow unbiased estimation of the variances. We therefore consider the box-constraint optimal allocation proposed by

Gabler et al. (2012), which minimizes the 2−norm of the relative root mean square error (RRMSE) of a set of direct statistics $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_D)$ under constraints regarding the lower and upper bounds of the domain-specific sample sizes n_d $(d = 1, \ldots, D)$ of D domains and an upper bound of the total sample size n. The 2−norm (cf. Harville 2008, 60) can be seen as a compensatory functional penalizing larger RRMSEs more than smaller ones. The box-constraint optimal allocation technique allows for control of the sample sizes or sampling fractions and, hence, the variation of the design weights. The domain-specific sample sizes emerge as a solution of the following optimization problem:

$$\begin{aligned} &\min_{n_d} \ \|\boldsymbol{RRMSE}_{<\cdot>}(\hat{\mu})\|_2 = \sqrt[2]{\sum_{d=1}^{D} \mathrm{RRMSE}(\mu_{<d>})} \\ &\text{s.t. } L_d \leq n_d \leq U_d, \ d = 1, \ldots, D \\ &\sum_{d=1}^{D} n_d \leq n, \end{aligned} \tag{2}$$

where L_d and U_d denote the lower and upper bound for the sample size in the d^{th} domain. The issue of obtaining numerically efficient solutions for the optimization problem (2) for very large numbers of strata is explored in detail by Münnich et al. (2012).

Besides these stratified sampling designs, πps−designs are often used in business surveys as past values of the auxiliary variables are available from the enterprise register (cf. Holmberg et al. 2002). In πps sampling, the inclusion probability of each unit is proportional to the value of some size variable available at the design stage. πps sampling is a very efficient design for design-based estimation strategies in cases where a high correlation exists between the target variable and the size variable and the intercept is close to zero (cf. Tillé 2006). In fact, if the variable of interest is proportional to the size variable, the variance of a Hajék-type estimator on a national level would be zero for fixed size designs (cf. Särndal et al. 2003, 89). One issue with πps sampling is that it tends to lead to highly variable design weights when there is a large variation in the auxiliary variable X. This can negatively influence the statistical modeling. An approach to reduce this variation is to incorporate box constraints to inclusion probabilities π_i $(i = 1, \ldots, N)$ yielding new inclusion probabilities π_i^* according to

$$\begin{aligned} &\min_{\pi_i^*} \ \sum_{i=1}^{N} \frac{(1/2)\left(\pi_i^* - \pi_i\right)^2}{\pi_i} \\ &\text{s.t. } \sum_{i=1}^{N} \pi_i^* = n, \\ &\pi_L \leq \pi_i^* \leq \pi_U, \ i = 1, \ldots, N, \end{aligned} \tag{3}$$

where π_L and π_U denote the lower and the upper bound for the new box constraint inclusion probabilities. The solution to problem (3) gives the box-constraint inclusion probabilities π_i^* which satisfy the box constraints. In the same spirit as the box-constraint

optimal allocation, the box-constraint πps design allows for control of the range of the design weights directly. As an additional benefit, the box-constraint approach towards πps sampling avoids very small inclusion probabilities, which are a concern for the sample selection algorithms. Another method has been proposed by Falorsi and Righi (2008) whose strategy may be described as a balanced sampling multiway stratification. They consider a situation in which constraints regarding a multivariate response $\mathbf{y}$ and several partitions hold whilst at the same time the selected sample is balanced on auxiliary variables. Since current algorithms for drawing balanced samples from large universes are still extremely computer intensive, we omitted the approach of Falorsi and Righi (2008) from our simulation study.

2.2. *Small Area Estimators under Complex Designs*

A common aim in small area estimation is the estimation of the domain mean

$$\mu_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj},\ d = 1, \ldots, D, \tag{4}$$

where y_{dj} is the variable of interest for unit j in domain d and N_d denotes the population size in domain d. A traditional estimator often used in survey sampling is the direct estimator given by

$$\hat{\mu}_{d,Direct} = \frac{\sum_{j=1}^{n_d} w_{dj} y_{dj}}{\sum_{j=1}^{n_d} w_{dj}}, \tag{5}$$

with w_{dj} as the design weight of unit j in domain d. Note that with planned domains and stratified random sampling where the strata are nested within the domains, the sum in the denominator of (5) is equal to N_d. Though Estimator (5) is design unbiased, estimates for domains with small sample sizes are expected to be inaccurate. We refer to (5) as the Direct estimator. The group of GREG estimators are given by

$$\hat{\mu}_{d,GREG} = \frac{1}{N_d}\left[\sum_{j=1}^{N_d} \hat{y}_{dj} + \sum_{j=1}^{n_d} w_{dj}\left(y_{dj} - \hat{y}_{dj}\right)\right]. \tag{6}$$

where $\hat{y}_{dj}$ is the predicted value of the variable of interest for unit j in domain d under a specified regression model. Thus the domain estimate in (6) results as the mean of the predicted values for all population units in domain d plus the mean of the weighted residuals for the sampled units in domain d. There are various choices for the assisting model, such as using linear or possibly nonlinear models, considering mixed models or focusing on fixed effects, and including or omitting design weights when fitting the model. A thorough investigation of model choice for GREG estimators is given in Lehtonen et al. (2003, 2005). In our study, we will focus on a linear fixed effects model and refer to this estimator as the GREG estimator. A detailed account on design-based and model-assisted domain estimation is given by Lehtonen and Veijanen (2009).

The unit-level mixed model, which is also known as the nested error regression model, is given by

$$y_{dj} = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D, j = 1, \ldots, N_d, \tag{7}$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$, $\varepsilon_{dj} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$. The domain-specific effects u_d are independent of the sampling error ε_{dj}. $\boldsymbol{x}_{dj}$ is the vector of auxiliary information for unit j in domain d, and $\boldsymbol{\beta}$ the vector of fixed regression parameters. Under Model (7) the small area mean is given by $\mu_d = \overline{\boldsymbol{X}}_d^T\boldsymbol{\beta} + u_d$ for all domains $d = 1, \ldots, D$. $\overline{\boldsymbol{X}}_d$ is the vector of the population mean of the auxiliary information in domain d and $\bar{\boldsymbol{x}}_d$ refers to the sample equivalent.

Assuming that Model (7) holds for the sample as well, the following EBLUP (empirical best linear-unbiased predictor) under negligible sampling fractions for the unknown domain mean μ_d can be derived as (cf. Battese et al. 1988)

$$\begin{aligned} &\hat{\mu}_{d,BHF} = \bar{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad \hat{u}_d = \hat{\gamma}_d\left(\bar{y}_d - \bar{\boldsymbol{x}}_d^T\hat{\boldsymbol{\beta}}\right) \qquad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \left(\hat{\sigma}_\varepsilon^2/n_d\right)} \\ &\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\boldsymbol{x}_d^T\hat{\boldsymbol{V}}_d^{-1}\boldsymbol{x}_d\right)^{-1}\left(\sum_{d=1}^{D}\boldsymbol{x}_d^T\hat{\boldsymbol{V}}_d^{-1}\boldsymbol{y}_d\right) \end{aligned} \tag{8}$$

In Equation (8), $\hat{u}_d$ is the EBLUP of the random effect u_d, $\hat{\gamma}_d$ is the shrinkage factor depending on the estimated variance components ($\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$), and $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$, $\bar{y}_d = n_d^{-1}\sum_{j=1}^{n_d} y_{dj}$, $\hat{\boldsymbol{V}}^{-1}$ refers to the inverse of the variance-covariance matrix. $\hat{\boldsymbol{V}}_d = \hat{\boldsymbol{V}}_d(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$ in domain d (cf. Rao 2003, Sec. 7.2). While Estimator (8) is model unbiased and efficient for self-weighting sampling designs, this is unlikely to hold for general sampling designs. In the following, we will denote the Estimator (8) as BHF for notational convenience since it dates back to Battese et al. (1988).

In typical applications in official statistics, ignoring the design weights may have severe consequences for the quality of model-based estimators (cf. Münnich and Burgard 2012). Several extensions of mixed models to cope with nonignorable sampling designs have been proposed, for example in Pfeffermann et al. (1998), Asparouhov (2006), Rabe-Hesketh and Skrondal (2006) and Lehtonen et al. (2006). Here, we focus on selected approaches which are suitable and easily applicable in official statistics.

A second way of extending the unweighted EBLUP under the unit-level mixed model is by augmenting the design matrix by the design weights. The following model is fitted to the survey data:

$$y_{dj} = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \kappa w_{dj} + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D, \; j = 1, \ldots, n_d, \tag{9}$$

where κ is the additional regression coefficient for the impact of the weights on the variable of interest, estimated by $\hat{\kappa}$. Alternatively, the size variable might also be used instead of w_{dj} under unequal probability sampling. The EBLUP under Model (9) is obtained as

$$\hat{\mu}_{d,augBHF} = \bar{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}} + \hat{\kappa}\bar{W}_d + \hat{u}_d, \tag{10}$$

with $\bar{W}_d = N_d^{-1}\sum_{j=1}^{N_d} w_{dj}$ as the population mean of the design weights in domain d. This estimator was introduced by Verret et al. (2010) in the context of informative

sampling and will be referred to as the augBHF estimator. Note that we could alternatively estimate $\boldsymbol{\beta}$ and κ using design weights.

You and Rao (2002) propose to transform the unit-level model (7) to a survey-weighted domain-level model with normalized weights within the domains. This model is given by

$$\bar{y}_{dw} = \bar{\boldsymbol{x}}_{dw}^T \beta_w + u_d + \bar{\varepsilon}_{dw},\ d = 1, \ldots, D, \tag{11}$$

with

$$\bar{y}_{dw} = \sum_{j=1}^{n_d} \tilde{w}_{dj} y_{dj}, \bar{\boldsymbol{x}}_{dw} = \sum_{j=1}^{n_d} \tilde{w}_{dj} \boldsymbol{x}_{dw}, \bar{\varepsilon}_{dw} = \sum_{j=1}^{n_d} \tilde{w}_{dj} \varepsilon_{dj} \text{ and } \tilde{w}_{dj} = \frac{w_{dj}}{\sum_{j=1}^{n_d} w_{dj}}.$$

The pseudo-EBLUP under Model (11) follows as (cf. You and Rao 2002):

$$\hat{\mu}_{d,YR} = \hat{\gamma}_{dw} \bar{y}_{dw} + \left(\bar{\boldsymbol{X}}_d - \hat{\gamma}_{dw} \bar{\boldsymbol{x}}_{dw}\right)^T \hat{\boldsymbol{\beta}}_w, \text{ with}$$

$$\hat{\gamma}_{dw} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \delta_d^2 \hat{\sigma}_\varepsilon^2}, \tag{12}$$

$$\delta_d^2 = \sum_{j=1}^{n_d} \tilde{w}_{dj}^2, \text{ and}$$

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_d^D \sum_{j=1}^{n_d} w_{dj} \boldsymbol{x}_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dw} \bar{\boldsymbol{x}}_{dw})^T \right)^{-1} \left(\sum_d^D \sum_{j=1}^{n_d} w_{dj} y_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dw} \bar{\boldsymbol{x}}_{dw}) \right)$$

In addition to achieving design consistency, the estimator given by (12) also fulfils the benchmarking property with respect to the national estimate. The Estimator (12) is denoted by YouRao.

In earlier simulation studies, the approach employed by Lehtonen et al. (2011) gave good results for various sampling designs. It is based on incorporating the vector of design weights in the lmer function in the R-package lme4 (cf. Bates et al. 2011; lmer provides a fast mixed-effects model implementation). Despite the fact that it is not meant for including design weights specifically, it has been shown to reduce the bias of the unweighted estimator (8) in many cases. For details regarding the estimation of the model parameters we refer to Bates (2011). This estimator is denoted by wBHF.

In some cases where unit-level data may not be available or the computation of unit-level models may not be feasible, area-level models can be a remedy. An area-level model may be described as follows:

$$\widehat{\bar{y}}_d = \bar{\boldsymbol{X}}_d^T \boldsymbol{\beta} + u_d + \varepsilon_d,\ d = 1, \ldots, D, \tag{13}$$

where $u_d \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_d \overset{\text{ind}}{\sim} N(0, \sigma_{\varepsilon_d}^2)$, which is independent of u_d (cf. Jiang and Lahiri 2006). Note that in the area-level model (13) the small area means of the direct estimator $\hat{\bar{y}}_d$ are modeled but not the observations themselves. This is due to the fact that auxiliary information is available at the domain level only. In the area-level

literature, σ_u^2 is also referred to as the model variance and $\sigma_{\varepsilon_d}^2$ as the sampling variance of the direct estimator, which depends on the domain-specific sample sizes and is therefore not identically distributed between the domains. The EBLUP under the area-level model (13) is given by

$$\hat{\mu}_{d,FH} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_{FH} + \hat{u}_d \tag{14}$$

and we will refer to the estimator as FH, since it was introduced by Fay and Herriot (1979). $\hat{\boldsymbol{\beta}}_{FH}$ refers to the estimator of the regression parameters under Model (13) and is given in Rao (2003, 116).

To estimate the prediction mean square error (PMSE) of the aforementioned EBLUP-type estimators BHF, wBHF, augBHF, YouRao and FH we consider two different strategies: one based on Taylor series expansions and the other based on the parametric bootstrap method. A good reference on these methods is Datta (2009). Prasad and Rao (1990) derived the following PMSE decomposition for EBLUP estimators based on results from Kackar and Harville (1984):

$$\mathrm{PMSE}(\hat{\mu}_d(\hat{\theta})) = g_{1d}(\hat{\theta}) + g_{2d}(\hat{\theta}) + 2g_{3d}(\hat{\theta}), \tag{15}$$

where the terms g_{1d} to g_{3d} depend on the estimated variance components $\hat{\theta}$. Additionally, in the case that the variance components are estimated by Restricted Maximum Likelihood (REML) or Maximum Likelihood (ML), explicit formulae for Estimators (8) and (14) based on decomposition (15) are given in Datta and Lahiri (2000). A second-order correct PMSE estimator for (12) has been derived by Torabi and Rao (2010).

Butar and Lahiri (2003) proposed using parametric bootstrap methods to estimate the PMSE of small area estimators. To account for the finite population, we consider a simplification of the bootstrap proposed by González-Manteiga et al. (2008) to produce PMSE estimates. Their algorithm for computing PMSE estimates for Estimator (8) is as follows:

1. Fit the statistical model to the sample data to obtain the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$.
2. Construct replicates $y_{dj}^* = \boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + u_d^* + \varepsilon_{dj}^*$, where $u_d^* \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ and $\varepsilon_{dj}^* \overset{iid}{\sim} N(0, \hat{\sigma}_\varepsilon^2)$.
3. Calculate the domain means $\mu_d^* = (1/N_d)\sum_{j=1}^{N_d} y_{dj}^*$.
4. Fit the statistical model to the sampled elements of y_{dj}^* to obtain estimates $\hat{\boldsymbol{\beta}}^*$ and $\hat{u}_d^*$.
5. Compute the estimated domain means $\hat{\mu}_d^* = \overline{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}^* + \hat{u}_d^*$.
6. Repeat the Steps 2 to 5 B times.
7. The estimated PMSE is computed by $\widehat{\mathrm{PMSE}}(\hat{\mu}_{d,PB}) = \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\mu}_d^{*(b)} - \mu_d^{*(b)}\right)^2$.

We also used the parametric bootstrap to obtain PMSE estimates for estimators wBHF, augBHF and FH, using the above mentioned models and formulae for estimating the model parameters and computing the estimated domain means.

3. Simulation Study

3.1. Data Set and Sampling Design

Our design-based simulation study extends the work of Burgard et al. (2012) to cover the issues of PMSE estimation and prediction intervals for small domains. The study is based

on synthetic business data resembling the small and medium enterprises from the Italian business register. This data set is a precursor of the fully synthetic data set TRItalia, which is being produced within the BLUE-ETS project (see Kolb et al. 2013). The parameter of interest is the mean of value added. As auxiliary variables we use turnover and the number of employees. Both variables are available in the Italian business register. We use these auxiliary variables since they are the only noncategorical register variables available at the design stage. From a subject-matter viewpoint, both auxiliary variables may influence the value added. A linear regression model without random effects confirmed that both explanatory variables are highly significant. However, the model only yielded $R^2 = 0.0045$ for our population, indicating that the explanatory power of the model is poor. Even if this case is pessimistic, similar situations may occur in many countries where registers often lack strong covariates. Even if the variable of interest is skewed, the application of transformation methods requires further research on the inclusion of weights, which is beyond the scope of this article.

As a stratification variable we used the first digit of the industry classification within each province (103 Italian provinces), resulting in 927 strata. The stratum-specific population sizes vary from 98 enterprises in the smallest stratum to 114,844 enterprises in the largest stratum. Since our data set is restricted to small and medium enterprises with 1 to 99 employees, our stratification does not contain a census-like stratum where all units within the stratum are sampled with certainty. We account for the problem that statistical agencies have to disseminate information at different levels of aggregation by considering two kinds of domains as scenarios. In the first scenario, the 103 Italian provinces are also the domains of interest, whereas in the second scenario the 927 strata are considered as domains. It is important to note that both scenarios reflect the problem of prediction with planned domains, thus avoiding problems of nonsampled areas.

The expected total sample size is set to $n = 60,000$. For the (box-constraint) optimal allocations the auxiliary variable turnover was used to compute the stratum-specific sample sizes. Besides these stratified sampling designs we also consider unequal probability within the strata, where the expected sample size within each stratum is set to the sample size allocated by proportional allocation. As in the case of optimal allocation, we use turnover to compute the inclusion probabilities. We use turnover as a size measure because it is the variable in our data set which has the highest correlation with our dependent variable. Since turnover does not have zero values in our data set, its use as a size measure is straightforward. A major difference between optimal allocation and unequal probability designs is that the former leads to design weights which vary between the strata, whereas for the latter the weights also vary within the strata. Even though the computation of inclusion probabilities is straightforward when using a strictly positive auxiliary variable, the sample selection is very computer intensive. For the case of unconstrained inclusion probabilities, Midzuno's method as described in Tillé (2006, Algorithm 6.13) programmed in C failed to produce the desired samples in due time. This problem was resolved by means of the box-constraint inclusion probabilities given in (3). In accordance with Münnich and Burgard (2012) the Gelman factor (GF) is defined as the ratio of the largest to the smallest design weight. This definition of a Gelman factor should not be confused with the Gelman-Rubin factor which is related to

Table 1. Sampling designs

Abbreviation	Design	Gelman factor
COSTA_{50}	Costa-type allocation with $k = 0.5$	47.66
EQUAL	equal allocation	1,153.85
BCOpt_{25}	box-constraint optimal allocation with $GF = 25$	30.88
BCOpt_{50}	box-constraint optimal allocation with $GF = 50$	60.83
OPT	optimal allocation	554.92
PROP	proportional allocation	1.78
UPS	unequal probability sampling	44,085, 380.58
UPS_{10}	unequal probability sampling under constraint $\max \frac{\pi_L^{-1}}{\pi_U^{-1}} \leq 10$	10
UPS_{100}	unequal probability sampling under constraint $\max \frac{\pi_L^{-1}}{\pi_U^{-1}} \leq 100$	100

MCMC convergence diagnostics. The GF is given by

$$\text{GF} = \frac{\max\limits_{i=1,\ldots,N} \frac{1}{\pi_i}}{\min\limits_{i=1,\ldots,N} \frac{1}{\pi_i}}. \tag{16}$$

The GF for equal allocation thus varies with the stratum sizes. If all the strata are of roughly the same size, then the equal allocation is almost equivalent to the proportional

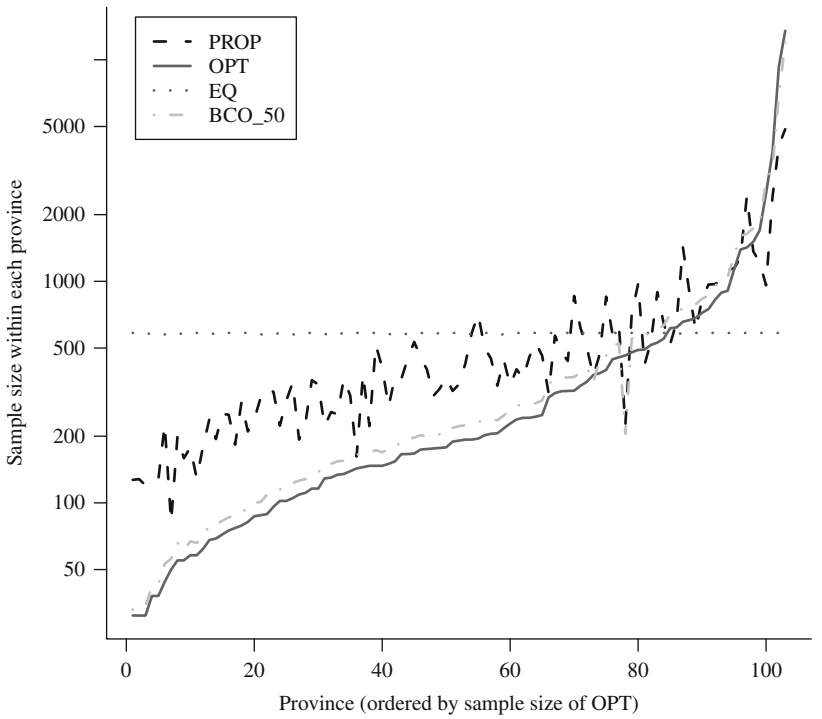


Fig. 1. Domain-specific sample sizes – 1st Scenario (103 domains)

Table 2. Domain-specific sample sizes – 1st Scenario (103 domains)

	PROP	EQ	COSTA_{50}	BCOpt_{25}	BCOpt_{50}	OPT
Min	82	576	330	35	33	31
Max	4,862	585	2,727	11,765	11,978	13,522

allocation and thus the GF will approach 1. If the N_h are highly variable, the equal allocation leads to highly dispersed design weights. Additionally, in the case of πps designs we get $\text{GF} \doteq \max z_i / \min z_i$ with z being the size variable used for the calculation of the πps inclusion probabilities. Typically, the variation of the auxiliary variable in business surveys is very large and thus the GF is very large as well. Table 1 lists our sampling designs with the abbreviations used in the following and the GF. Turning our attention to the box-constraint optimal allocations, we recognize that these allocations do not satisfy the box constraints exactly. This is due to the fact that our approach produces non-integer-valued numbers, which have to be rounded, rather than integer-valued constraints.

The variation of the domain-specific sample sizes under the first scenario is illustrated in Figure 1. We see that the optimal allocation on the one hand and the equal allocation on the other hand are the two extreme cases. The minimum and the maximum for the domain-specific sample sizes under Scenarios 1 and 2 are given in Tables 2 and 3. These tables further illustrate that under Scenario 2 the minimum domain-specific sample sizes are very small except for the equal allocation.

To evaluate the results of our simulation study, we consider several different quality measures related to the accuracy of point estimates and the reliability of confidence intervals. A common measure to estimate the bias of a point estimator is the relative bias. It is given by

$$\text{RB}(\hat{\mu}_d) = \frac{\left(\frac{1}{R}\right)\sum_{l=1}^{R}\hat{\mu}_{l,d} - \mu_d}{\mu_d}, \quad d = 1, \ldots, D, \tag{17}$$

where R denotes the number of Monte Carlo replicates. The relative bias takes values from $-\infty$ to ∞, whilst a relative bias close to 0 is desirable, indicating that the point estimates are on average identical to the true values. Another quality measure is the relative root mean square error (RRMSE), which is computed as

$$\text{RRMSE}(\hat{\mu}_d) = \frac{\sqrt{\frac{1}{R}\sum_{l=1}^{R}(\hat{\mu}_{l,d} - \mu_d)^2}}{\mu_d}, \quad d = 1, \ldots, D. \tag{18}$$

Table 3. Domain-specific sample sizes – 2nd Scenario (927 domains)

	PROP	EQ	COSTA_{50}	BCOpt_{25}	BCOpt_{50}	OPT
Min	2	64	34	2	2	2
Max	1,605	65	836	3,998	3,935	4,766

Table 4. Computing times in seconds

Estimator	BHF	wBHF	augBHF	YouRao	GREG	Direct
Seconds	2732.17	2870.49	2789.30	3.83	0.40	0.21

The values of the RRMSE are in the range between 0 and ∞, where a value close to 0 indicates good results. Moreover, we consider summary statistics of the quality measures over all domains. With respect to the relative bias, we compute the mean absolute relative bias (MARB)

$$\text{MARB}(\hat{\mu}_d) = \frac{1}{D}\sum_{d=1}^{D} |\text{RB}(\hat{\mu}_d)| \tag{19}$$

and for the RRMSE we consider the average relative root mean square error (ARRMSE)

$$\text{ARRMSE}(\hat{\mu}_d) = \frac{1}{D}\sum_{d=1}^{D} \text{RRMSE}(\hat{\mu}_d). \tag{20}$$

We construct confidence intervals based on MSE or PMSE estimators as described in Subsection 2.2. The traditional approach is to compute the confidence interval (CI) as follows (cf. Chatterjee et al. 2008):

$$\text{CI}(\hat{\mu}_d)_{1-\alpha} = \left[\hat{\mu}_d - \sqrt{\widehat{\text{PMSE}}(\hat{\mu}_d)} \cdot z_{1-\alpha/2}; \hat{\mu}_d + \sqrt{\widehat{\text{PMSE}}(\hat{\mu}_d)} \cdot z_{1-\alpha/2}\right] \tag{21}$$

with $z_{1-\alpha/2}$ as the $(1-\alpha/2)$-quantile of the standard normal distribution. Since $\sqrt{\widehat{\text{PMSE}}(\hat{\mu}_d)}$ is estimated, confidence intervals based on quantiles of the t-distribution with $(n_d - 1)$ degrees of freedom could be considered. Note that differences between these two approaches to computing confidence intervals vanish as the domain-specific sample size n_d increases. Additionally, we also considered using bootstrap confidence intervals as proposed by Chatterjee et al. (2008). The reliability of confidence intervals is measured by the coverage rate, computed as the percentage of confidence intervals covering the true value μ_d.

In the following section we will report results based on 1,000 Monte Carlo replications. For the parametric bootstrap methods we use 499 bootstrap replications. Due to the small number of bootstrap replications, the bootstrap confidence intervals are outperformed

Table 5. Types of estimators

Abbreviation	Estimator
Direct	Hajek-type estimator (5)
GREG	linear fixed effects generalized regression estimator (6)
YouRao	pseudo-EBLUP (12)
wBHF	weighted EBLUP using weights option in lmer
augBHF	augmented EBLUP (10)
BHF	EBLUP (8) under unit-level mixed model
FH	EBLUP (14) under area-level mixed model

Table 6. MARB – 1st Scenario (103 domains)

	PROP	EQ	UPS_{10}	UPS_{100}	$COSTA_{50}$	$BCOpt_{25}$	$BCOpt_{50}$	OPT
Direct	0.004	0.004	0.005	0.006	0.003	0.005	0.005	0.007
GREG	0.004	0.004	0.027	0.045	0.004	0.005	0.006	0.007
YouRao	0.023	0.026	0.034	0.069	0.020	0.059	0.066	0.089
wBHF	0.022	0.014	0.019	0.014	0.021	0.012	0.010	0.009
augBHF	0.022	0.023	0.226	0.532	0.021	0.019	0.018	0.019
BHF	0.022	0.023	0.226	0.533	0.020	0.040	0.041	0.040
FH	0.051	0.059	0.069	0.090	0.046	0.100	0.106	0.115

by the other methods. Therefore, the bootstrap confidence intervals are not presented in Subsection 3.3. The average CPU time (AMD Opteron 6164 HE with 1.7 GHz and 4GB RAM for each kernel) for the estimators is given in Table 4, where in case of the BHF, wBHF, and augBHF 499 bootstrap resamples are performed.

3.2. *Results of Point Estimates*

In this section we summarize the most important aspects regarding the simulation results on the point estimates. For convenience, our estimators are listed in Table 5. We also considered a GREG based on mixed models, but we did not observe major differences between a GREG with or without random effects. To keep the presentation of the results as short as possible, the focus subsequently lies on the GREG without random effects.

For our first scenario (103 domains) the mean absolute relative bias over all domains is given in Table 6. The analysis of MARB in Table 6 indicates that the Direct estimator is indeed unbiased under all sampling designs. The model-assisted GREG has some problems under unequal probability designs. This can be traced back to the fact that we did not include design weights when estimating $\boldsymbol{\beta}$. With respect to the estimators based on unit-level models, we see that there are only minor differences under proportional allocation, equal allocation and convex combinations thereof. As soon as we consider (box-constraint) optimal allocations, the bias of the unweighted BHF estimator is more pronounced compared to the wBHF and augBHF estimator. With respect to unequal probability designs, we observe severe biases for the augBHF and BHF estimators, whereas the wBHF estimator is still accurate. The YouRao estimator performs similarly to the wBHF under proportional and equal allocation, but its bias increases for other designs and is higher than the bias of the unweighted BHF estimator for optimal

Table 7. ARRMSE – 1st Scenario (103 domains)

	PROP	EQ	UPS_{10}	UPS_{100}	$COSTA_{50}$	$BCOpt_{25}$	$BCOpt_{50}$	OPT
Direct	0.142	0.160	0.162	0.193	0.135	0.220	0.234	0.256
GREG	0.140	0.158	0.167	0.205	0.134	0.219	0.233	0.255
YouRao	0.139	0.156	0.160	0.191	0.133	0.211	0.224	0.245
wBHF	0.029	0.071	0.042	0.074	0.033	0.103	0.123	0.155
augBHF	0.030	0.033	0.229	0.539	0.029	0.031	0.032	0.034
BHF	0.029	0.032	0.229	0.539	0.030	0.044	0.046	0.046
FH	0.122	0.136	0.136	0.153	0.119	0.166	0.172	0.182

Table 8. MARB – 2nd Scenario (927 domains)

	PROP	EQ	UPS_{10}	UPS_{100}	$COSTA_{50}$	$BCOpt_{25}$	$BCOpt_{50}$	OPT
Direct	0.015	0.008	0.017	0.024	0.010	0.022	0.023	0.026
GREG	0.015	0.009	0.016	0.020	0.010	0.022	0.023	0.026
YouRao	0.247	0.107	0.294	0.439	0.131	0.391	0.409	0.457
wBHF	0.081	0.068	0.073	0.060	0.078	0.047	0.042	0.037
augBHF	0.080	0.075	0.251	0.553	0.074	0.068	0.071	0.074
BHF	0.081	0.076	0.251	0.553	0.078	0.094	0.096	0.096
FH	0.221	0.127	0.257	0.304	0.152	0.314	0.318	0.322

allocations. The FH estimator exhibits bias under all designs considered due to full shrinkage towards the synthetic component.

As soon as we consider the ARRMSE given in Table 7, the picture is completely different. We observe that there is no single best estimator under all designs and hence sampling design plays an important role. The results under proportional and Costa-type allocations are almost the same. Under these designs, all model-based unit-level estimators work well. Even though the area-level FH estimator is biased, it has the lower ARRMSE compared to design-based estimators and the YouRao estimator. Under equal allocation, the results are similar to the proportional and Costa-type allocations except for the weighted wBHF estimator, which has a considerably higher ARRMSE than the other unit-level estimators. For designs based on (box-constraint) optimal allocations, the augmented augBHF estimator performs best, with the unweighted BHF estimator as the only other estimator with an ARRMSE under ten percent. The wBHF estimator suffers from a much higher ARRMSE despite a lower bias, a result that can be attributed to the increase of the variability of the model parameter estimates. Moreover, under unequal probability designs, the wBHF estimator is the only reasonable estimator in terms of an ARRMSE under ten percent. The performance of the augBHF and BHF is identical up to the third decimal number, indicating that augmenting the design matrix does not increase the precision under unequal probability sampling in this setting. A closer look at Tables 6 and 7 reveals that due to the shrinkage to the synthetic component the FH estimator is biased under all designs but does not perform badly with respect to RRMSE. The comparison between the Direct estimator and the GREG shows that the working model does not have much predictive power. Concentrating on the Direct estimator, we note that designs optimized for national-level estimation are not the best choice for domain estimation.

Table 9. ARRMSE – 2nd Scenario (927 domains)

	PROP	EQ	UPS_{10}	UPS_{100}	$COSTA_{50}$	$BCOpt_{25}$	$BCOpt_{50}$	OPT
Direct	0.575	0.333	0.650	0.771	0.374	0.889	0.940	1.006
GREG	0.569	0.329	0.649	0.769	0.370	0.885	0.935	1.002
YouRao	0.451	0.312	0.498	0.573	0.344	0.570	0.581	0.608
wBHF	0.085	0.114	0.103	0.196	0.088	0.344	0.441	0.556
augBHF	0.091	0.082	0.256	0.561	0.080	0.080	0.085	0.089
BHF	0.085	0.082	0.256	0.561	0.083	0.099	0.101	0.102
FH	0.350	0.281	0.385	0.412	0.302	0.420	0.425	0.433

The MARB in the presence of smaller domains (Scenario 2, 927 domains) are given in Table 8. Compared to Scenario 1 (Table 6), the biases increase (almost) uniformly which is due to the smaller sample sizes. We see that for all designs the GREG and Direct estimators have the lowest relative bias. With respect to the unit-level estimators the impact of designs is similar to the first scenario, but with generally higher absolute relative biases. The FH suffers most from severe bias, especially under designs with largely varying sample sizes such as (box-constraint) optimal allocations or proportional allocation and the unequal probability designs. The most striking aspect about these results is the large bias of the YouRao estimator.

With regards to the ARRMSE, which is shown in Table 9, we see that the ARRMSE increases drastically compared to Scenario 1. Unlike Scenario 1, we now observe significant differences between Costa-type allocation and proportional allocation, which is due to the fact that the proportional allocation leads to very small domain-specific sample sizes in Scenario 2 (see Table 3). This causes a severe loss in estimation quality in comparison to the Costa-type allocation for design-based estimators. This does not apply for model-based estimators, which manage to borrow strength from other domains to

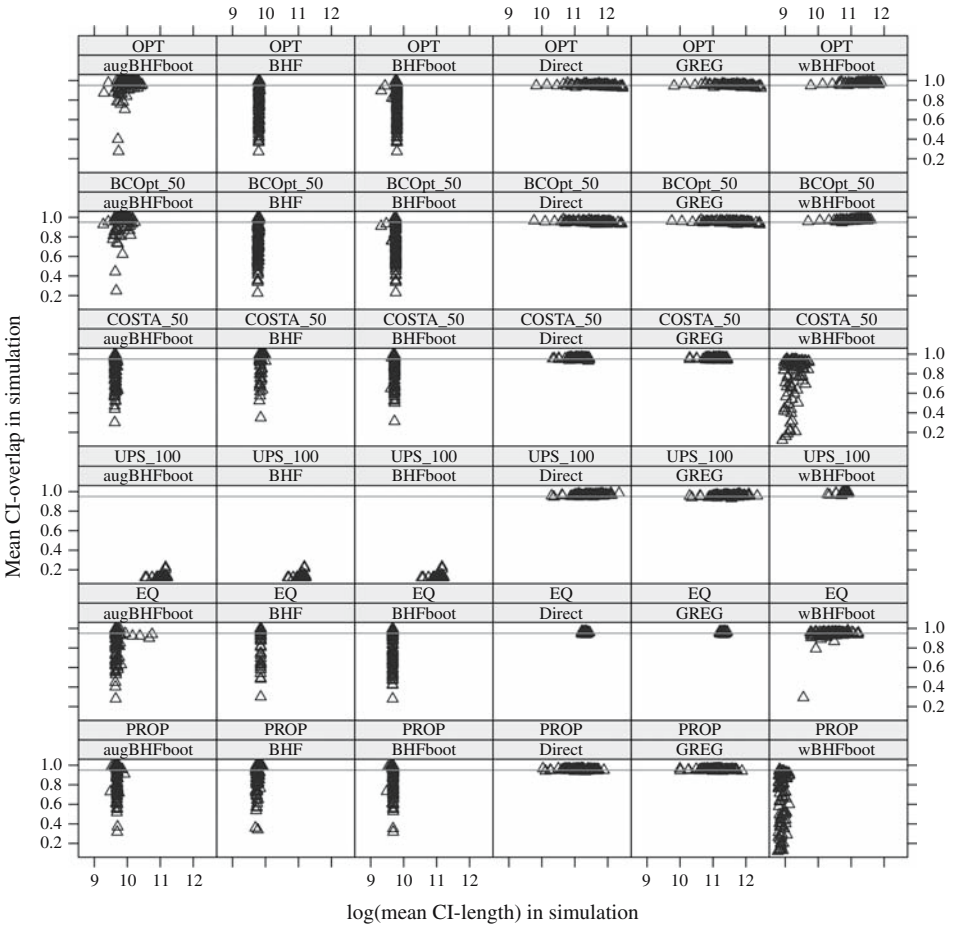


Fig. 2. Coverage Rates – 1st Scenario (103 domains)

Table 10. Mean of coverage rates – 1st Scenario (103 domains)

	PROP	EQ	UPS_{100}	$COSTA_{50}$	$BCOpt_{50}$	OPT
augBHFboot	0.867	0.845	0.127	0.850	0.937	0.944
BHF	0.898	0.927	0.128	0.933	0.748	0.765
BHFboot	0.861	0.831	0.127	0.868	0.738	0.762
Direct	0.955	0.954	0.965	0.956	0.951	0.949
GREG	0.955	0.955	0.956	0.956	0.951	0.949
wBHFboot	0.670	0.939	0.987	0.781	0.973	0.973

compensate for small domain-specific sample sizes and perform best under both designs. With respect to the equal allocation the ranking of the estimators is similar, except that the weighted wBHF estimator performs worse than the other two unit-level estimators. Under (box-constraint) optimal allocations, the augmented augBHF estimator performs slightly better than the BHF estimator due to the lower bias. Other estimators cannot be recommended in this case, since their ARRMSE is 30 percent and more. Under unequal probability designs, the weighted wBHF estimator seems the only reasonable choice, even though its ARRMSE is close to 20 percent when the Gelman factors are constrained to 100.

3.3. Results of Precision Estimates

In this section, we report results of precision estimates for the most interesting designs and estimators only. The coverage rates and mean confidence interval lengths for the first scenario are illustrated in Figure 2 and means of the coverage rates are given in Table 10. Figure 2 depicts both the confidence interval coverage rates and mean confidence interval length for each domain. Ideally, these points would lie on the horizontal line, indicating a 95 % coverage rate, and at the left side in each panel, demonstrating high accuracy of the point estimates by a shorter average length of confidence intervals. It is obvious that some small area estimators yield lower coverages, which is mainly caused by a worse fit of the statistical model in several areas. This reflects the situation of many registers only containing a limited set of potentially predictive covariates.

Focusing on the length of the confidence intervals, we see that equal allocation is the best choice if one wishes to use design-based estimators. Whereas the coverage rates of the Direct and GREG estimators are reasonable under all the sampling designs considered, this does not apply for the other estimators. Under the UPS_{100} design, most model-based estimation methods suffer from severe undercoverage. With respect to the two approaches to PMSE estimation, either by Taylor approximation or by parametric bootstrap, we hardly observe any differences in the case of the unweighted BHF estimator. Interestingly, the augBHF does not perform badly in the case of (box-constraint) optimal allocations, even though it does not achieve the nominal coverage rate on average in any design. The mean coverage rates indicate overcoverage for the wBHF using parametric bootstrap under (box-constraint) optimal allocations, which clearly shows that the confidence intervals are not efficient. Altogether, it is indisputable that the coverage rates of the model-based estimators are not satisfactory.

The coverage rates for the second scenario are depicted in Figure 3. In addition to the dark "+" signs related to the computation of the confidence intervals based on quantiles of

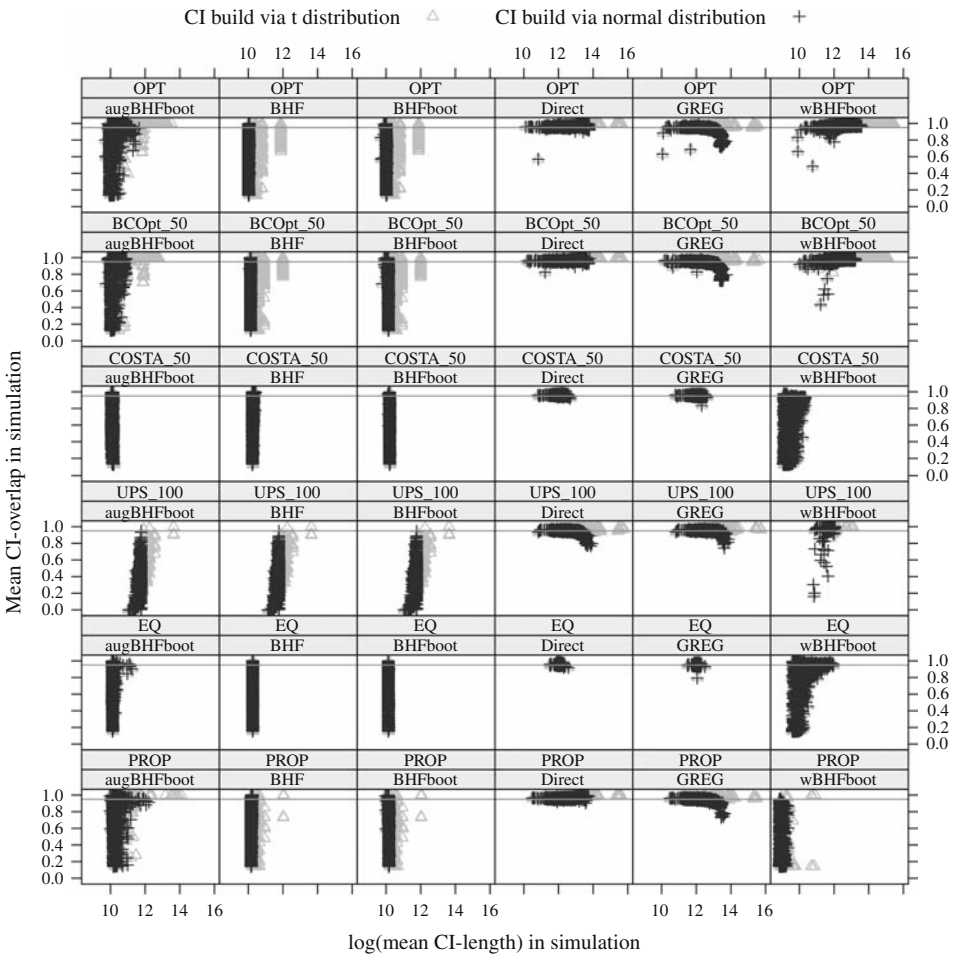


Fig. 3. Coverage Rates – 2nd Scenario (927 domains)

the normal distribution, lighter triangles indicate confidence intervals based on quantiles of t-distributions with $(n_d - 1)$ degrees of freedom as explained at the end of Subsection 3.1. These two methods differ only in the presence of very small domain-specific sample sizes n_d, which was not a concern in Scenario 1. Looking at the x-scale, we observe that the CI-length increases dramatically compared to Scenario 1. Furthermore, in the case of the Direct estimator, we observe some problems under UPS_{100} for the CIs built via the normal distribution. These problems vanish as soon as we use the t-distribution, which seems to be the better choice for very small domains. For the GREG estimator the use of normal-quantiles is critical except under equal and Costa-type allocations. With respect to the model-based estimators the poor performance of all strategies is striking, as can also be seen from Tables 11 and 12.

4. Summary

This article explores two major issues official statistics face when implementing small area estimation techniques in business surveys. First, business registers of many countries

Table 11. Mean of coverage rates – 2nd Scenario (927 domains) – Normal Quantiles

	PROP	EQ	UPS_{100}	$COSTA_{50}$	$BCOpt_{50}$	OPT
augBHFboot	0.717	0.619	0.386	0.642	0.745	0.726
BHF	0.686	0.687	0.389	0.699	0.638	0.588
BHFboot	0.672	0.612	0.386	0.663	0.634	0.585
Direct	0.962	0.960	0.941	0.959	0.966	0.967
GREG	0.943	0.954	0.946	0.952	0.904	0.895
wBHFboot	0.349	0.735	0.984	0.476	0.968	0.965

do not yield many variables with strong predictive power. Second, the sampling designs applied, in general, are nonignorable and may have a major impact on model-based estimates. In this context, several strategies for incorporating design weights into statistical models are discussed. The application focuses on registers where only a few variables with limited predictive power are available. This reflects the situation in many countries and several branches of official statistics and shows the usefulness of the estimators under less favorable circumstances.

Our results suggest that model-based estimators should be considered in addition to purely design-based estimators due to lower RRMSEs in many settings. Furthermore, estimators ignoring the sampling design cannot be recommended since they may yield considerably biased estimates. Besides the influence of the range of design weights, our results stress the relevance of the source of design weight variation – between or within areas and strata. Altogether, our study illustrates the efficiency gains made possible by using model-based small area estimators even under less favorable circumstances.

A comparison of the augBHF and the wBHF estimator illustrates that the origin of the variation of the design weights is an essential basis for selecting the appropriate estimator. Under purely stratified designs with large Gelman factors the augBHF estimator gives reasonable results and should be the estimator of choice with respect to minimal ARRMSE, whilst the wBHF estimator suffers from the variability of $\boldsymbol{\beta}$ estimates. In contrast, under unequal probability designs the wBHF estimator is clearly the best estimator in both scenarios if one wishes to minimize the ARRMSE of the estimates. The poor performance of the augBHF estimators in this case is partly explained by the huge discrepancy between $\overline{W}_d$ and the expected mean of the sampling weights in domain d under unequal probability sampling. This causes a bias due to informative sampling where the model which holds for the population does not hold for the sample as well (cf. Pfeffermann and Sverchkov 2009). This problem of the augBHF estimator under unequal

Table 12. Mean of coverage rates – 2nd Scenario (927 domains) – t Quantiles

	PROP	EQ	UPS_{100}	$COSTA_{50}$	$BCOpt_{50}$	OPT
augBHFboot	0.744	0.623	0.443	0.648	0.834	0.825
BHF	0.714	0.692	0.446	0.706	0.766	0.738
BHFboot	0.699	0.616	0.442	0.669	0.763	0.735
Direct	0.973	0.963	0.958	0.964	0.982	0.983
GREG	0.959	0.958	0.963	0.958	0.960	0.960
wBHFboot	0.366	0.741	0.993	0.483	0.981	0.980

probability sampling could be resolved by estimating the model parameters using design weights. In our simulations, the YouRao estimator especially suffers from a poor model. In other simulations, the YouRao performed much better when auxiliary information with better predictive power was available. Similar results hold for the area-level FH estimator.

In addition to the Gelman factors and their sources of variation, the domain-specific sample size plays a crucial role for domain estimation. This can be seen from the comparatively good results of most estimators under equal and Costa-type allocation achieved at the expense of less efficient estimation at the national level. Furthermore, we note that under Scenario 2 with many small sample sizes the precision of domain estimates generally decreases compared to the first scenario with larger domains. This decrease is most pronounced for design-based estimators which cannot compensate for the small sample sizes by borrowing strength from other domains.

Focusing on the precision estimates, we observe that the confidence interval coverage rates of the design-based estimators are as expected. The shortest CI lengths result under equal allocation designs. Minor problems of the design-based estimators with very small domain-specific sample sizes are corrected by plugging-in quantiles from a t_{n_d-1} distribution. The coverage rates for the BHF were not satisfactory under either Taylor linearization of the PMSE or PMSE estimation by parametric bootstrap due to high biases. We have seen that very small domains may be problematic for precision estimates, as the severe cases of under coverage in Scenario 2 point out. Moreover, our results indicate that under (box constraint) optimal allocations in Scenario 1, the reliability of the confidence intervals of the augBHF estimator is better than the reliability of the unweighted BHF estimator. With respect to the parametric bootstrap method for the wBHF estimator, mainly in Scenario 1, we have seen overcoverage for the (box constraint) optimal allocations and unequal probability sampling, implying that the PMSE estimates are too conservative.

The present application used small and medium enterprises. When dealing with large enterprises one could expect extremely skewed distributions with outliers. Under these settings, either transformation methods (Berg and Chandra 2012 or Shlomo and Priam 2013) or robust models should be considered (Sinha and Rao 2009 or Chambers and Tzavidis 2006). A comparison of robust small area methods including computational issues can be drawn from Schmid (2012). When using nonignorable sampling designs in business surveys, the robustification of design weights should be investigated in addition to the robust modeling.

6. References

Asparouhov, T. 2006. "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics – Theory and Methods* 35: 439–460. DOI: http://dx.doi.org/10.1080/03610920500476598.

Bates, D. 2011. *Computational Methods for Mixed Models*. Available at: http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf (accessed October 16, 2014).

Bates, D., M. Maechler, and B. Bolker. 2011. *Linear Mixed-Effects Models Using S4 Classes. R package Version: 0.999375-42*. Available at: http://www.r-project.org/conferences/useR-2012/TutorialBates.pdf (accessed October 14, 2014).

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the*

American Statistical Association 83: 28–36. DOI: http://dx.doi.org/10.1080/01621459.1988.10478561.

Berg, E. and H. Chandra. 2012. "Small Area Prediction for a Unit Level Lognormal Model." Federal Committee on Statistical Methodology Research Conference. DOI: http://dx.doi.org/10.1016/j.csda.2014.03.007.

Burgard, J.P., R. Münnich, and T. Zimmermann. 2012. "Small Area Modelling Under Complex Survey Designs for Business Data." In Proceedings of the Fourth International Conference of Establishment Surveys, June 11–14, 2012. Montreal. Available at: http://www.amstat.org/meetings/ices/2012/papers/301906.pdf (accessed Oct 16, 2014).

Butar, F.B. and P. Lahiri. 2003. "On Measures of Uncertainty of Empirical Bayes Small-Area Estimators." *Journal of Statistical Planning and Inference* 112: 63–76. DOI: http://dx.doi.org/10.1016/S0378-3758(02)00323-3.

Chambers, R., and N. Tzavidis. 2006. "M-Quantile Models for Small Area Estimation." *Biometrika* 93: 255–268. DOI: http://dx.doi.org/10.1093/biomet/93.2.255.

Chatterjee, S., P. Lahiri, and H. Li. 2008. "Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models." *The Annals of Statistics* 36: 1221–1245.

Choudhry, G.H., J.N.K. Rao, and M.A. Hidiroglou. 2012. "On Sample Allocation for Efficient Domain Estimation." *Survey Methodology* 38: 23–29.

Costa, A., A. Satorra, and E. Ventura. 2004. "Improving Both Domain and Total Area Estimation by Composition." *Statistics and Operations Research Transactions* 28: 69–86.

Datta, G.S. 2009. "Model-Based Approach to Small Area Estimation." In *Handbook of Statistics*, Vol. 29B, 251–288. New York: Elsevier.

Datta, G.S., and P. Lahiri. 2000. "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems." *Statistica Sinica* 10: 613–627.

Eurostat. 2008. "NACE Rev. 2: Statistical Classification of Economic Activities in the European Community." Eurostat methodologies and working papers, European Communities, cat. No. KS-RA-07-015-EN-N.

Falorsi, P.D. and P. Righi. 2008. "A Balanced Sampling Approach for Multi-Way Stratification Designs for Small Area Estimation." *Survey Methodology* 34: 223–234.

Fay, R.E. and R.A. Herriot. 1979. "Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–277. DOI: http://dx.doi.org/10.1080/01621459.1979.10482505.

Gabler, S., M. Ganninger, and R. Münnich. 2012. "Optimal Allocation of the Sample Size to Strata Under Box Constraints." *Metrika* 75: 15–161. DOI: http://dx.doi.org/10.1007/s00184-010-0319-3.

Gonzalez-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2008. "Bootstrap Mean Squared Error of a Small-Area EBLUP." *Journal of Statistical Computation and Simulation* 78: 443–462.

Harville, D.A. 2008. *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hidiroglou, M.A. and P. Lavallee. 2009. "Sampling and Estimation in Business Surveys." In *Handbook of Statistics*, Vol. 29A, 441–470. New York: Elsevier.

Holmberg, A., P. Flisberg, and M. Rönnqvist. 2002. "On the Choice of Sampling Design in Business Surveys with Several Important Study Variables." R&D Report 2002:3, Statistics Sweden.

Jiang, J. and P. Lahiri. 2006. "Mixed Model Prediction and Small Area Estimation." *Test* 15: 1–96. DOI: http://dx.doi.org/10.1007/BF02595419.

Kackar, R.N., and D.A. Harville. 1984. "Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models." *Journal of the American Statistical Association* 79: 853–862. DOI: http://dx.doi.org/10.1080/01621459.1984.10477102.

Kolb, J.-P., R. Münnich, F. Volk, and T. Zimmermann. 2013. "TRItalia dataset." In *BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys*, edited by R. Bernardini Papalia, C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Fernandez-Vazquez, M. Ferrante, J.P. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk, and T. Zimmermann, 168–188. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf (accessed October 16, 2014).

Lehtonen, R., C.-E. Särndal, and A. Veijanen. 2003. "The Effect of Model Choice in Estimation for Domains, Including Small Domains." *Survey Methodology* 29: 33–44.

Lehtonen, R., C.-E. Särndal, and A. Veijanen. 2005. "Does the Model Matter? Comparing Model-Assisted and Model-Dependent Estimators of Class Frequencies for Domains." *Statistics in Transition* 7: 649–673.

Lehtonen, R., C.-E. Särndal, A. Veijanen, and M. Myrskylä. 2006. "The Role of Models in Model-Assisted and Model-Dependent Estimation for Domains and Small Areas." Workshop on survey sampling and methodology, Ventspils. Available at: http://home.lu.lv/~pm90015/workshop2006/papers/Workshop2006_04_Lehtonen.pdf (accessed October 16, 2014).

Lehtonen, R. and A. Veijanen. 2009. "Design-Based Methods of Estimation for Domains and Small Areas." In *Handbook of Statistics*, Vol. 29B, 219–249. New York: Elsevier.

Lehtonen, R., A. Veijanen, M. Myrskylä, and M. Valaste. 2011. "*Small Area Estimation of Indicators on Poverty and Social Exclusion*." Technical report, AMELI deliverable D2.2. Available at: http://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Deliverables/AMELI-WP2-D2.2.20110402.pdf (accessed October 16, 2014).

Little, R.J. 2012. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 309–334.

Longford, N.T. 2006. "Sample Size Calculation for Small Area Estimation." *Survey Methodology* 32: 87–96.

Molina, I., and J.N.K. Rao. 2010. "Small Area Estimation of Poverty Indicators." *Canadian Journal of Statistics* 38: 369–385. DOI: http://dx.doi.org/10.1002/cjs.10051.

Münnich, R. and J.P. Burgard. 2012. "On the Influence of Sampling Design on Small Area Estimates." *Journal of the Indian Society of Agricultural Statistics* 66: 145–156.

Münnich, R., E. Sachs, and M. Wagner. 2012. "Numerical Solution of Optimal Allocation Problems in Stratified Sampling Under Box Constraints." *Advances in Statistical Analysis* 96: 435–450. DOI: http://dx.doi.org/10.1007/s10182-011-0176-z.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–625. DOI: http://dx.doi.org/10.2307/2342192.

Pfeffermann, D. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61: 317–337. DOI: http://dx.doi.org/10.2307/1403631.

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society Series B* 60: 23–40. DOI: http://dx.doi.org/10.1111/1467-9868.00106.

Pfeffermann, D. and M. Sverchkov. 2007. "Small-Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas." *Journal of the American Statistical Association* 102: 1427–1439. DOI: http://dx.doi.org/10.1198/016214507000001094.

Pfeffermann, D. and M. Sverchkov. 2009. "Inference Under Informative Sampling." In *Handbook of Statistics*, Vol. 29B, 455–487. New York: Elsevier.

Prasad, N.G.N., and J.N.K. Rao. 1990. "The Estimation of the Mean Squared Error of Small Area Estimators." *Journal of the American Statistical Association* 85: 163–171. DOI: http://dx.doi.org/10.1080/01621459.1990.10475320.

Rabe-Hesketh, S. and A. Skrondal. 2006. "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169: 805–827. DOI: http://dx.doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: John Wiley and Sons.

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

Schmid, T. 2012. "*Spatial Robust Small Area Estimation applied on Business Data*." Ph.D. thesis, University of Trier.

Shlomo, N. and R. Priam. 2013. "Improving Estimation in Business Surveys." In *BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys*, edited by R. Bernardini Papalia, C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Fernandez-Vazquez, M. Ferrante, J.P. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk, and T. Zimmermann, 52–70. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf (accessed October 16, 2014).

Sinha, S.K. and J.N.K. Rao. 2009. "Robust Small Area Estimation." *Canadian Journal of Statistics* 37(3): 381–399. ISSN 1708-945X.

Thompson, K.J. and B.E. Oliver. 2012. "Response Rates in Business Surveys: Going Beyond the Usual Performance Measure." *Journal of Official Statistics* 28: 221–237.

Tillé, Y. 2006. *Sampling Algorithms. Springer Series in Statistics*. New York: Springer.

Torabi, M. and J.N.K. Rao. 2010. "Mean Squared Error Estimators of Small Area Means Using Survey Weights." *Canadian Journal of Statistics* 38: 598–608.

Tschuprow, A. 1923. "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations." *Metron* 2: 461–493.

Verret, F., M.A. Hidiroglou, and J.N.K. Rao. 2010. "Small Area Estimation Under Informative Sampling." In Proceedings of the Survey Methods Section SSC Annual Meeting, May 2010.

You, Y. and J.N.K. Rao. 2002. "A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights." *The Canadian Journal of Statistics* 30: 431–439.

You, Y. and J.N.K. Rao. 2003. "Pseudo Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights." *Journal of Statistical Planning and Inference* 111: 197–208. DOI: http://dx.doi.org/10.1016/S0378-3758(02)00301-4.

Received January 2013
Revised May 2014
Accepted June 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 773–785, http://dx.doi.org/10.2478/JOS-2014-0047

On Precision in Estimates of Change over Time where Samples are Positively Coordinated by Permanent Random Numbers

Annika Lindblom[1]

Measures of period-to-period change are key statistics for many economy surveys. To improve the precision of these estimates of change, the majority of the business surveys at Statistics Sweden select stratified simple random samples (STSI) at different points in time, ensuring positive correlation between samples (overlapping samples) by using permanent random numbers (PRN). Statistics Sweden normally selects positively coordinated STSIs drawn from an updated Business Register (BR). In these samples, the industry strata are usually stratified further within industry into size strata. When the most recent sampling frame contains updated classification variables for all units, enterprises can change stratum between two sampling occasions. A drawback of the coordinated sample selection procedure is that the desired correlation between the two samples decreases if the proportion of enterprises that change strata is substantial. Consequently, the sample design must anticipate the potential effect of stratum changes between samples. This article presents a study that examines how the design of a repeated business survey affects the precision in estimates of change over time using the Turnover in the Service Sector survey conducted by Statistics Sweden as an example.

Key words: Measures of change; sample coordination; survey design; variance estimation.

1. Introduction

An important issue in many repeated business surveys is to determine whether the period-to-period change in an estimated total is statistically significant. To improve the precision of estimates of change, the majority of the business surveys at Statistics Sweden use samples from separate points in time ("sample occasions") that are positively coordinated (overlapping) by permanent random numbers (PRN). This positive coordination over time introduces dependence between the obtained samples, inducing positive correlation between the two level estimates, which in turn increases the precision in estimates of change over that obtained from independent samples.

Statistics Sweden normally uses positively coordinated stratified simple random samples (STSI) drawn from an updated Business Register (BR). The stratification is usually performed by industry, further grouping units within an industry into size strata. One drawback of this coordinated sampling procedure is that the desired correlation decreases between the two samples if the proportion of enterprises that change strata

[1]Senior Methodologist, Statistics Sweden, 701 89 Örebro, Sweden. Email: annika.lindblom@scb.se
Acknowledgments: I would like to thank Dr. Lennart Nordberg for helpful discussions during the course of this work which have led to substantial improvements. I would also like to thank the Editor, the Associate Editor and three referees for their useful review of earlier versions of this article.

is substantial. The sample designer must therefore anticipate the potential effect of stratum changes between samples. A detailed size stratification procedure (creating numerous small strata within a given industry) promotes high precision in each level estimate but often results in a smaller overlap (less correlation) between samples. On the other hand, a less detailed stratification (allowing wider ranges within size strata) yields less precise level estimates but a larger overlap (higher correlation) between samples.

Despite the fact that coordinated samples are commonly used at Statistics Sweden, there has been little research conducted exploring the "tradeoff" between the usage of a detailed stratification and size of overlap on the precision of estimates of change. Such knowledge would be very useful for future sample designs. This article presents the results of a simulation study conducted at Statistics Sweden that compares the precision of the same change estimates obtained by using three different STSI sampling designs selected from the frame data of the Turnover in the Service Sector survey (hereafter referred to as the TSSS). This study is based on the actual frame populations established in March 2009 and in March 2010. The study variable in the survey TSSS is Monthly Turnover and exactly the same variable can be found (in retrospect) in the monthly Value Added Tax (VAT) returns. This means that we have values on the study variable for all enterprises in both frame populations.

Although this study employs the specific sample coordination PRN technique used at Statistics Sweden, similar PRN techniques are used for sample coordination in several countries. It is not unlikely that the significance and properties of the correlation obtained by these other methods would be quite similar to the correlation obtained by the method presented here. In Section 2, we describe the system used at Statistics Sweden for coordinated frame development and sample selection (the SAMU System). Section 3 presents background information on the TSSS. We present the formulae used for variance and correlation estimation in Section 4. Section 5 presents the simulation study. We conclude in Section 6 with general comments and ideas for future research.

2. The SAMU System

Statistics Sweden uses the SAMU system (Ohlsson 1995, Lindblom 2003) for the coordination of frame populations and sample selection from the BR. The SAMU system has three main objectives: (1) to obtain statistics comparable both in time and between surveys; (2) to ensure high precision in estimates of change over time; and (3) to spread the response burden between the businesses.

The SAMU utilizes a very clever and simple method of drawing coordinated samples. A random number, independently selected from a set of random numbers uniformly distributed over the interval (0, 1), is assigned to every new unit as it enters the BR, and the unit retains this value as long as it remains in the BR. "Closed-down" units (deaths) are deleted from the BR along with their random numbers. After the random number assignment is complete, the entire frame population is ordered by strata, with units in each stratum sorted in ascending random number sequence, and the first n_h units in the strata are selected. Ohlsson (1992) formally proves that the sampling technique used in SAMU is equivalent to STSI without replacement.

The sample coordination in SAMU introduces a dependence between the realized samples that would not be present if new independent random numbers were assigned to each unit on the updated frame prior to sample selection. Since the random number assignment by SAMU is permanent, the same random number is used in each subsequent sample selection after its initial assignment. Each new STSI is drawn using these permanent random numbers. In this way, the STSI incorporates the most recent changes from the updated BR. Furthermore, a large overlap with the most recent sample can be expected since a persistent unit has the same random number on both occasions. All current surveys benefit from frame populations stemming from the same updated version of the BR. However, a drawback of this is that the precision in estimates of change will be sensitive to the proportion of units that change stratum between sample occasions. On the other hand, the use of the latest updated version of the BR is also important, especially for the level estimates.

3. Background on the Turnover in the Service Sector Survey (TSSS)

The "Turnover in the Service Sector" survey (TSSS) conducted by Statistics Sweden produces detailed monthly and quarterly estimates of turnover changes in 138 domains according to the Statistical Classification of Economic Activities in the European Community (NACE Rev. 2). Monthly Turnover is the only variable collected in this survey, and change estimates – not levels – are published.

The year-to-year change in turnover, $\hat{t}_{m1}/\hat{t}_{m0}$, is an important published statistic, where $\hat{t}_{mt}$ is the combined (size strata within industry) ratio estimate (Särndal et al. 1992) of the turnover level for month (m) year ($t = 0$ or $t = 1$, with 1 as the most recent year). The auxiliary information used is annual turnover, the same information used for cut-off and in the stratification (see below). Large enterprises (selected with certainty) are excluded from the combined ratio estimator due to their large impact on the estimates. Their turnover sum is added to the combined ratio estimates (each of the minimal number of nonresponding large enterprises are individually imputed).

The survey covers the following industries, classified into the service sector according to NACE Rev. 2: Motor, wholesale and retail trade (45–47); Transportation and storage (49–53); Accommodation and food service activities (55–56); IT and real estate businesses (58–75); Administrative and support service (77–82); Education, human health and social work (85–88); Art, entertainment and recreation (90–96). NACE is derived from the International Standard Industrial Classification of all Economic Activities (ISIC).

The frame population for the TSSS consists of all active enterprises in the BR classified into the service sector according to the above definition. Annual turnover is used as a unit size measure in TSSS and enterprise level information on monthly turnover is collected from monthly VAT returns. The variable Annual Turnover is defined in TSSS as the sum of monthly turnover for the most recent 12-month period available at the sampling occasion. A cut-off limit is used in the survey, so that enterprises with an Annual Turnover less than 200,000 SEK (about $ 30,000) are excluded from the frame population and the samples. The final frame population consists of about 300,000 enterprises.

The stratification divides the frame population into 138 industrial strata based on economic activity. This stratification accommodates specialized domains of study as much as possible. Each industrial stratum is further subdivided into five size strata, with Annual Turnover as the unit size measure. Within each industry, one size stratum includes the largest enterprises, which are completely enumerated (a certainty or "take-all" stratum). The remaining units are grouped into four strata using the $cum\sqrt{f}$ method to determine stratum boundaries (Dalenius and Hodges 1959). Sample sizes in each stratum are obtained via optimum allocation (Neyman 1934), with Annual Turnover as the allocation variable. The total sample consists of about 12,000 enterprises. Approximately 2,500 enterprises are completely enumerated. These completely enumerated enterprises account for approximately 50 percent of the total turnover in the frame population.

Once a year, in March, a new frame population is established, and a new STSI is drawn using the SAMU. The frame population established and the sample drawn in March of a given year (t) are used for the period April year (t) to March year ($t+1$).

4. Variance and Correlation for Estimates of Change

4.1. Variance Estimation

As mentioned in Section 1, the complete frame population data is available for our study. Therefore, we can directly obtain the variances of the Monthly Turnover estimates at times $m0$ and $m1$ ($V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$, respectively) using the sampling formula variances for a STSI sample. The theoretical variance for the change estimate of Monthly Turnover is approximated by the Taylor Linearization formula:

$$V\left(\frac{\hat{t}_{m1}}{\hat{t}_{m0}}\right) \approx \left(\frac{t_{m1}}{t_{m0}}\right)^2 \left[\frac{V(\hat{t}_{m1})}{t_{m1}^2} + \frac{V(\hat{t}_{m0})}{t_{m0}^2} - 2\frac{C(\hat{t}_{m0},\hat{t}_{m1})}{t_{m0}t_{m1}}\right] \tag{1}$$

This is equivalent to:

$$V\left(\frac{\hat{t}_{m1}}{\hat{t}_{m0}}\right) \approx \left(\frac{t_{m1}}{t_{m0}}\right)^2 \left[\frac{V(\hat{t}_{m1})}{t_{m1}^2} + \frac{V(\hat{t}_{m0})}{t_{m0}^2} - 2\frac{\rho(\hat{t}_{m0},\hat{t}_{m1})\sqrt{V(\hat{t}_{m0})V(\hat{t}_{m1})}}{t_{m0}t_{m1}}\right] \tag{2}$$

However, a theoretical expression of $\rho(\hat{t}_{m0},\hat{t}_{m1})$ in Formula 2 would require the generation of all possible outcomes of pairwise coordinated samples from the two frame populations, and would require prohibitive computational resources.

4.2. Covariance/Correlation Estimation

The sample coordination method employed by SAMU makes estimating the correlation between the level estimates quite complicated because the size of the overlap between two samples is stochastic. Nordberg (2000) presents a complete and workable method for estimating this correlation under the SAMU sampling scheme. Related approaches can be found in Tam (1984), Laniel (1988), Hidiroglou et al. (1995), Berger (2004) and Wood (2008); Garås (1989) summarizes the preceding work on this approach conducted at Statistics Sweden.

Nordberg's method works when only sample data from each time period are available, as well as when values on the study variables are available for the whole frame population. For our study, we estimated the correlation by straightforward simulation and used those estimates as input to the analysis. In addition, we obtained correlation estimates using the method proposed by Nordberg (when values on the study variable are available for the whole frame populations). It is very useful for Statistics Sweden to compare the empirical measures to those obtained using Nordberg's method. This comparison (evaluation) confirms that Nordberg's method gives unbiased estimates. Comparison statistics for these empirical measures to those obtained using Nordberg's method are available upon demand.

Note that in this study the year-to-year change in turnover is based on the Horvitz-Thompson (HT) estimator of the turnover level for month (m) year ($t = 0$ or $t = 1$) instead of the ratio estimator used in the actual TSSS. This study aims to analyze how different choices of stratification variable, number of strata, and study variable are related to the overlap correlation (i.e., $\rho(\hat{t}_{m0}, \hat{t}_{m1})$ in Formula (2)) and the variances of the estimates of change. Use of the HT estimator, instead of the ratio estimator, makes the analysis presented in Section 5 more transparent and avoids confounding. The ratio estimator would add another factor to consider in the analysis, namely the correlation between the study variable and the auxiliary variable, which is very high in the TSSS but could possibly be lower for another choice of study variable.

To obtain empirical estimates of the correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ (for each domain) in our simulation, we independently selected 10,000 coordinated samples from the frame population. Recall that the true variances of $\hat{t}_{m0}$ and $\hat{t}_{m1}$($V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$) are known. Let K = the number of generated pairwise coordinated samples (i.e., the number of replicates) selected in the simulation study ($k = 1, 2, \ldots,$ K). We obtained empirical sample-based estimates of variance and covariance as

$$\hat{V}(\hat{t}_{mt}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{mtk} - \hat{\bar{t}}_{mt})^2, \tag{3}$$

$$\hat{C}(\hat{t}_{m0}, \hat{t}_{m1}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{m0k} - \hat{\bar{t}}_{m0})(\hat{t}_{m1k} - \hat{\bar{t}}_{m1}) \tag{4}$$

where $t = 0$ or 1 and $\hat{\bar{t}}_{mt}$ is the average level estimate over the K samples.

We verified that 10,000 was a sufficient number of replicates by comparing the sample-based values of $\hat{V}(\hat{t}_{m0})$ and $\hat{V}(\hat{t}_{m1})$ to $V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$, respectively. The large number of replicates yielded variance estimates that were essentially unbiased over repeated samples, implying that the estimated correlation ($\hat{\rho}$) was likewise unbiased for ρ. Table 1 compares the abovementioned $\hat{V}(\hat{t}_{m0})$ and $\hat{V}(\hat{t}_{m1})$ to $V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$ obtained by samples using the STSI sampling design selected from the frame population data of the TSSS.

In addition, the number of sufficient replicates was confirmed by comparing the difference in obtained correlation estimates after 100, 500, 1,000, and up to 10,000 replicates to validate that 10,000 replicates were sufficient to ensure convergence.

Using Formulas (3) and (4), we estimated $\rho(\hat{t}_{m0}, \hat{t}_{m1})$ in Formula (2) in each domain.

Table 1. Comparison between sample-based and theoretical variances obtained by the sampling design used in TSSS

NACE Industry	$V(\hat{t}_{m0})$	$\hat{V}(\hat{t}_{m0})$	$V(\hat{t}_{m1})$	$\hat{V}(\hat{t}_{m1})$
45	117,419	117,901	147,626	148,695
46	1,052,889	1,031,004	1,304,284	1,294,880
47	518,376	520,264	583,215	590,050

Note that unlike the variance estimates, whose higher domain level estimates can be obtained by aggregating the independent lower level domain estimates, the covariance estimates must be computed separately for the aggregate domain and for the separate lower level subdomains. The covariance estimates are based on information collected at two time points and are therefore affected by enterprises changing lower level subdomain between the two sample occasions.

5. The Simulation Study

5.1. Simulation Study Design

The actual frame populations established in March 2009 and in March 2010 for the TSSS provide the study data. The study variable is Monthly Turnover, obtained retrospectively for all enterprises from the monthly VAT returns (for the majority of the enterprises) or from the Annual Income Tax returns (for a minor portion of the enterprises). In the latter case, an estimated Monthly Turnover was produced by dividing Annual Turnover by twelve. Due to the timing of the VAT returns, it is not possible to use the turnover values from monthly VAT returns in the production of the survey statistics.

We compare the three different STSI sampling designs, ranging from highly detailed (numerous size strata) to a single noncertainty stratum with a very heterogeneous population:

1. Each industry stratum has four sampled size groups and one take-all stratum (4-size gr.). This is the current design of the TSSS.
2. Each industry stratum has three sampled size groups and one take-all stratum (3-size gr.)
3. Each industry stratum has one sampled size group and one take-all stratum (1-size gr.)

Each design was applied to the same frame populations, with industry as the first level stratification variable. After determining the take-all (certainty) units, the remaining units were stratified into four, three and one noncertainty strata by unit size strata within industry (depending on design) using the $cum\sqrt{f}$ rule. In the tables below, we label four, three, and one noncertainty strata designs as "4-size gr.," "3-size-gr.," and "1-size gr."

Besides varying the number of strata, we considered the effects of alternative second level stratification variables (unit size variables) on the estimated precision. With the TSSS, the correlation between the stratification variable (Annual Turnover) and the study variable (Monthly Turnover) is very high. To extend the results to a less "ideal" situation – that is, reducing the correlation between the size measure and the study variable/s – we

restratified the frame populations using Number of Employees (collected from the BR) as a size measure and repeated the experiment (Note: although less correlated with Monthly Turnover, Number of Employees is a much more stable variable compared to Annual Turnover). In addition, we consider two study variables: Monthly Turnover and Annual Value Added. The study variable Annual Value Added is obtained retrospectively for all enterprises from the Annual Income Tax return. For both stratifications, optimum allocation based on Annual Turnover was used to determine the sample sizes in each stratum under the constraints that total sample size on the three-digit NACE level should be almost the same in all designs.

The estimates were produced at the two- and three-digit NACE Rev. 2 levels. To ensure comparability between the three different sampling designs, all designs have approximately the same sample size for each year in each three-digit NACE domain. Unfortunately, it was too time consuming to include all industries covered by the survey. Consequently, we restricted the analysis to a subset of the TSSS industries: Motor Trade (45), Wholesale Trade (46) and Retail Trade (47). These industries comprise about 75,000 enterprises and were chosen for their importance in the TSSS.

We selected independent pairwise samples per design from the 2009 and 2010 frames, replicating the SAMU PRN-coordination sampling procedure 10,000 times. For each replicate k, we generated a unique seed as the integer part of a random number uniformly distributed over the interval (0, 1) using the SAS RANUNI function (Fishman and Moore 1982), multiplied by a million. The replicate seeds were used to generate the permanent random numbers assigned to all enterprises in the frame population at time 0 (2009) and to the new enterprises in the frame population at time 1 (2010).

Tables 2a and 2b present aggregated information, from each stratification, on the number of enterprises in the frame populations, the number of enterprises in the samples (take-all and sampled), along with aggregated information on frame population overlap and sample overlap (averaging over repeated samples). The counts in the Overlap columns exclude take-all units as well as strata whose frame populations contain one common enterprise in the two years.

Since different variables are used for the two stratifications, the sets of take-all enterprises presented in Tables 2a and 2b do not coincide entirely. However, the difference between the two sets is very slight because an enterprise with large turnover usually has a large number of employees.

5.2. *Results*

We conducted all analyses on both the two- and three-digit NACE Rev. 2 levels. To save space, only the two-digit level results are included; however, the results on the three-digit level support the results on the two-digit level. Tables 3a and 3b show the gain in efficiency in terms of variance reduction (in percent) for the two-digit level change estimates, comparing the variance estimates obtained by using dependent SAMU samples ($VDep$) to the corresponding variance estimates obtained by using independent samples ($VInd$) with gain measured by $100 \bullet \left(1 - \frac{VDep}{VInd}\right)$.

The efficiency gained by using dependent SAMU samples rather than independent samples is quite substantial. At a minimum, a variance reduction of at least about

Table 2a. Sample Design Characteristics with Annual Turnover as Stratification Variable

				Realized Sample Sizes					
		Population		2009		2010			
Design	NACE industry	2009	2010	Take-all	Sampled	Take-all	Sampled	Overlap in population	Overlap in sample
4-size gr.	45	12,868	12,710	207	405	197	403	11,051	326
	46	26,855	26,366	717	783	664	790	22,481	584
	47	34,945	34,132	515	1,133	526	1,139	29,116	818
3-size gr.	45	12,868	12,710	207	402	197	401	11,064	331
	46	26,855	26,366	717	797	664	790	22,490	581
	47	34,945	34,132	515	1,103	526	1,102	29,129	816
1-size gr.	45	12,868	12,710	207	400	197	400	11,076	342
	46	26,855	26,366	717	800	664	802	22,593	673
	47	34,945	34,132	515	1,087	526	1,090	29,227	868

Table 2b. Sample Design Characteristics with Number of Employees as Stratification Variable

				Realized Sample Sizes					
		Population		2009		2010			
Design	NACE industry	2009	2010	Take-all	Sampled	Take-all	Sampled	Overlap in population	Overlap in sample
4-size gr.	45	12,868	12,710	192	418	183	418	11,078	332
	46	26,855	26,366	535	947	520	947	22,664	745
	47	34,945	34,132	477	1,149	494	1,128	29,190	818
3-size gr.	45	12,868	12,710	192	419	183	419	11,094	342
	46	26,855	26,366	535	945	520	945	22,692	762
	47	34,945	34,132	477	1,150	494	1,132	29,208	856
1-size gr.	45	12,868	12,710	192	418	183	427	11,090	359
	46	26,855	26,366	535	945	520	960	22,718	799
	47	34,945	34,132	477	1,147	494	1,130	29,265	891

Table 3a. Stratification by Annual Turnover

NACE industry	Measure Monthly Turnover			Measure Annual Value Added		
	4-size gr. Gain	*3-size gr. Gain*	*1-size gr. Gain*	*4-size gr. Gain*	*3-size gr. Gain*	*1-size gr. Gain*
45	22.3%	30.7%	74.1%	38.0%	42.7%	80.2%
46	24.6%	32.5%	66.6%	41.9%	48.7%	71.8%
47	36.3%	47.0%	78.1%	52.1%	57.1%	75.0%

Table 3b. Stratification by Number of Employees

NACE industry	Measure Monthly Turnover			Measure Annual Value Added		
	4-size gr. Gain	*3-size gr. Gain*	*1-size gr. Gain*	*4-size gr. Gain*	*3-size gr. Gain*	*1-size gr. Gain*
45	54.4%	63.1%	81.8%	63.8%	71.2%	82.0%
46	69.7%	74.3%	73.0%	56.1%	59.4%	71.7%
47	62.6%	67.9%	80.2%	55.9%	65.3%	78.3%

20 percent is attained with the highly stratified design (4-size gr). As the number of strata decreases, the efficiency gains from the dependent SAMU samples are more evident. The gains in efficiency are especially noticeable when the stratification and study variables are less strongly correlated (Table 3b), although the gain is not negligible when the stratification and study variables are highly correlated (Table 3a).

Tables 4a through 4d present the standard errors of the change estimates in percentage points for each sampling design. *SEDep* is the standard error obtained by using overlapping SAMU samples, *SEInd* is the standard error obtained by using independent samples and *Corr* $\left(\hat{\rho}(\hat{t}_{m0}, \hat{t}_{m1})\right)$ is the estimated overlap correlation obtained using SAMU samples.

For the majority, the most detailed stratification (4-size gr.) yields the smallest *SEDep*. In general, the improvements in precision for the input level (total estimates) offset the smaller sample overlap compared to the other design. The magnitude of the overlap correlation increases as the number of size groups (strata) decreases. The difference in precision with four and three size groups (noncertainty strata) is small for *SEDep*, compared to the difference in precision with three and one size groups in many cases. Often, the increase in *SEInd* caused by reducing the number of size groups from four to three is offset by the increased overlap correlation, and there is no detrimental effect on the precision of the estimate of change. However, when only one size group is employed, both the *Corr* and *SEInd* increase substantially, and the increased overlap correlation cannot compensate for the increased *SEInd*.

By comparing corresponding cells in Tables 4a and 4b and in Tables 4c and 4d, we can examine the relationship between the stratification variable and the study variable on the overlap correlation. The results presented in Tables 4a and 4b show that the overlap correlation of Monthly Turnover increases substantially when Number of Employees is the stratification variable. This increase is probably a function of the stability of Number of Employees in contrast to the more volatile Annual Turnover. Because the Number of

Table 4a. Stratification by Annual Turnover, Monthly Turnover Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	2.2%	2.5%	0.22	2.5%	3.0%	0.31	4.1%	8.0%	0.74
46	1.4%	1.6%	0.25	1.5%	1.8%	0.32	3.1%	5.3%	0.67
47	1.7%	2.1%	0.36	1.7%	2.3%	0.47	2.7%	5.8%	0.78

Table 4b. Stratification by Number of Employees, Monthly Turnover Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	6.8%	10.1%	0.55	6.7%	11.1%	0.63	6.3%	14.7%	0.83
46	4.8%	8.7%	0.71	4.8%	9.5%	0.75	6.2%	11.9%	0.75
47	1.9%	3.1%	0.63	1.8%	3.2%	0.68	2.5%	5.6%	0.80

Table 4c. Stratification by Annual Turnover, Annual Value Added Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	4.3%	5.4%	0.40	4.3%	5.7%	0.45	5.0%	11.2%	0.80
46	2.8%	3.6%	0.42	2.7%	3.8%	0.49	4.2%	7.8%	0.72
47	1.9%	2.7%	0.52	1.9%	2.9%	0.57	2.9%	5.8%	0.75

Table 4d. Stratification by Number of Employees, Annual Value Added Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	4.0%	6.6%	0.64	4.4%	8.2%	0.71	4.8%	11.4%	0.82
46	3.1%	4.7%	0.57	3.4%	5.4%	0.60	4.4%	8.3%	0.72
47	1.7%	2.6%	0.56	1.7%	3.0%	0.65	2.4%	5.1%	0.78

Employees in an enterprise tends to remain constant, the enterprise is often retained in the same stratum in consecutive sampling occasions, facilitating larger sample overlap. Although the correlation due to overlap is higher when obtained with the more stable stratification variable, this does not imply that the change estimates are likewise more precise. The correlations presented in Table 4a are consistently lower than their Table 4b counterparts, but the *SEDep* estimates are also considerably lower. Recall that the stratification variable and study variable used in Table 4a are very highly correlated, whereas the stratification and study variables used in Table 4b are not. In the former case, the variance estimates of monthly turnover (*SEInd*) are much lower than those obtained

using the other stratification. The increased *Corr* due to a larger overlap does not compensate for the larger variance estimates of the level estimates.

Tables 4c and 4d demonstrate similar patterns with a different study variable (Annual Value Added). Here, the overlap *Corr* increases as the number of strata decreases. As in Tables 4a and 4b, using Number of Employees as a stratification variable again increases the magnitude of the overlap *Corr*. Again, the differences in precision (*SEDep*) obtained between three and four size group stratifications are very small. Finally, the increased *Corr* due to the large overlap in the one sampled size group design largely compensates for the increased variance of the level estimates, although overall precision still tends to be lower than with the more stratified designs. The comparisons of the *Corr* between the designs with different stratification variables may be somewhat confounded by the different size measures. Recall that there are slightly different sets of take-all enterprises for both designs, which in turn affects the sampling variance.

Finally, we compare corresponding *Corr* values in Tables 4a to 4c and Tables 4b to 4d. The results in Tables 4a and 4c are based on exactly the same sampling design; the only difference is that Monthly Turnover is replaced by Annual Value Added as study variable. A comparison between Tables 4a and 4c reveals that the realized values of *Corr* are very close when the sampling design employs one sampling strata (One Sampled Size Group). In this case, the effect of stratification variable and size of sample overlap is eliminated and the only difference is due to different study variables. This indicates that the correlation between two Annual Value Added values, observed on the same unit at two different occasions, have similar patterns as those seen with Monthly Turnover when the stratification is not very detailed. However, the amount of realized *Corr* increases substantially when four (and three) sampled size groups are used (regardless of stratification variable) when Monthly Turnover is replaced by Annual Value Added as the study variable. We suspect that this phenomenon is related to size of sample overlap and the correlation between stratification and study variables.

6. Conclusions and Future Research

In this article, we present a study that examines the effects of degree of stratification, correlation between stratification variables and study variables, and overlap correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ obtained by using the PRN technique utilized at Statistics Sweden on the precision of estimates of change (level estimates produced by the HT-estimator). The studied SAMU method is easy to implement, but the sample designers have to make many decisions. Specifically, they must balance the need for highly stratified designs – which reduce the variance of the level estimates – with the need for a substantive sample overlap to increase the correlation between the adjacent level estimates to increase the precision of the change estimates (the primary statistics of interest).

One conclusion from the study is that the overlap correlation is of less importance for the precision in estimates of change over time when study variable and stratification variable are highly correlated. In this case, the precision in estimates of change benefits most from the high precision in each level estimate. When the correlation between the stratification variable and the study variable decreases or when a more stable stratification variable was used, such as Number of Employees, we found that using a moderately

stratified design (three noncertainty strata instead of four) with the overlapping SAMU samples created a sufficiently high correlation to offset the increase in level estimate variances.

Since the study variable in TSSS (Monthly Turnover) is known in retrospect for the whole frame population, it was possible to estimate the overlap correlation by simulation in this study. In most other surveys the study variable values would be available only from a *single* sample from each time period. The method proposed by Nordberg (2000) yields unbiased estimates of the correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ obtained by overlapping SAMU samples. However, if the proportion of enterprises that change stratum between two sample occasions is substantial the correlation estimates can become quite variable. This is the case in the TSSS, where the stratification variable Annual Turnover is fairly volatile, causing enterprises to change stratum rather frequently. If Monthly Turnover from an earlier time period can be used as a proxy variable for Monthly Turnover for the actual time period, then the overlap correlation could be estimated in practice by the same simulation method as used in the present study. Examining the effect of this procedure will be an issue for further study. Another important question for future study is the effect on the overlap correlation occurring when different survey designs, as well as different estimators, use the SAMU PRN-coordination method.

7. References

Berger, Y. 2004. "Variance Estimation for Measures of Change in Probability Sampling." *The Canadian Journal of Statistics* 32: 451–467. DOI: http://dx.doi.org/10.2307/3316027.

Dalenius, T. and J.L. Hodges. 1959. "Minimum Variance Stratification." *Journal of the American Statistical Association* 54: 88–101. DOI: http://dx.doi.org/10.2307/2282141.

Fishman, G.S. and L.R. Moore. 1982. "A Statistical Evaluation of Multiplicative Congruential Random Number Generators with Modulus $2^{31}-1$." *Journal of the American Statistical Association* 77: 29–136. DOI: http://dx.doi.org/10.1080/01621459.1982.10477775.

Garås, T. 1989. *Estimators of Change in Dynamic Populations*. Memo: Statistics Sweden. (In Swedish).

Hidiroglou, M., C.-E. Särndal, and D. Binder. 1995. "Weighting and Estimation in Business Surveys." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P.S. Scott, 477–502. New York: John Wiley & Sons.

Laniel, N. 1988. "Variances for a Rotating Sample from a Changing Population." In *Proceedings of the Business and Economic Statistics Section: American Statistical Association*. 246–250.

Lindblom, A. 2003. *SAMU – The System for Coordination of Frame Populations and Samples from the Business Register at Statistics Sweden*. Background Facts on Economic Statistics 2003:3, Statistics Sweden. Available at: http://www.scb.se/statistik/OV/AA9999/2003M00/X100ST0303.pdf (accessed September 1, 2014).

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method Stratified Sampling and the Method of Purposive Selection." *Journal of Royal Statistical Society* 97: 558–606. DOI: http://dx.doi.org/10.2307/2342192.

Nordberg, L. 2000. "On Variance Estimation for Measures of Change When Samples are Coordinated by the Use of Permanent Random Numbers." *Journal of Official Statistics* 16: 363–378. Available at: http://www.jos.nu/Articles/abstract.asp?article=164363 (accessed September 1, 2014).

Ohlsson, E. 1992. *SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden*. R&D Report, Statistics Sweden, 1992:18.

Ohlsson, E. 1995. "Coordination of Samples using Permanent Random Numbers." In *Business Survey Methods*, edited by B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott, 153–169. New York: John Wiley.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Tam, S.M. 1984. "On Covariances from Overlapping Samples." *The American Statistician* 38: 288–292. DOI: http://dx.doi.org/10.1080/00031305.1984.10483227.

Wood, J. 2008. "On the Covariance Between Related Horvitz-Thompson Estimators." *Journal of Official Statistics* 24: 53–78. Available at: http://www.jos.nu/Articles/abstract.asp?article=241053 (accessed September 1, 2014).

Received December 2012
Revised August 2014
Accepted August 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 787–810, http://dx.doi.org/10.2478/JOS-2014-0048

Analytic Tools for Evaluating Variability of Standard Errors in Large-Scale Establishment Surveys

MoonJung Cho[1], *John L. Eltinge*[1], *Julie Gershunskaya*[2], *and Larry Huff*[2]

Large-scale establishment surveys often exhibit substantial temporal or cross-sectional variability in their published standard errors. This article uses a framework defined by survey generalized variance functions to develop three sets of analytic tools for the evaluation of these patterns of variability. These tools are for (1) identification of predictor variables that explain some of the observed temporal and cross-sectional variability in published standard errors; (2) evaluation of the proportion of variability attributable to the abovementioned predictors, equation error and estimation error, respectively; and (3) comparison of equation error variances across groups defined by observable predictor variables. The primary ideas are motivated and illustrated by an application to the U.S. Current Employment Statistics program.

Key words: Degrees of freedom; design effect; generalized variance function (GVF); U.S. Current Employment Statistics program.

1. Introduction: Temporal and Cross-Sectional Variability of Published Standard Errors

Large-scale establishment surveys often exhibit substantial temporal or cross-sectional variability in their published standard errors or relative standard errors. To illustrate, consider a set of domains j and periods t, $j = 1, \ldots, J; t = 1, \ldots, T$; let θ_{jt} be a finite population parameter for domain j at time t; let $\hat{\theta}_{jt}$ be the associated design-based point estimator; let $\hat{V}_p(\hat{\theta}_{jt})$ be an estimator of the design variance of $\hat{\theta}_{jt}$; and define the associated estimated standard errors

$$s(\hat{\theta}_{jt}) = \{\hat{V}_p(\hat{\theta}_{jt})\}^{1/2}$$

and relative standard errors

$$r(\hat{\theta}_{jt}) = \frac{s(\hat{\theta}_{jt})}{\hat{\theta}_{jt}}.$$

Throughout this article, the subscript "p" denotes an expectation or variance evaluated with respect to the sample design.

[1] U.S. Bureau of Labor Statistics-Office of Survey Methods Research, PSB 1950 2 Massachusetts Ave. N.E., Washington, DC, 20212, U.S.A. Emails: Cho.Moon@bls.gov, Eltinge.John@bls.gov
[2] U.S. Bureau of Labor Statistics-Office of Employment and Unemployment Statistics, Washington, DC, U.S.A. Emails: Gershunskaya.Julie@bls.gov and Huff.Larry@bls.gov
Acknowledgment: The authors thank Ken Robertson for many helpful discussions of the CES and the Associate Editor for the insightful and constructive suggestions. The views expressed in this article are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

Variability of $s(\hat{\theta}_{jt})$ and $r(\hat{\theta}_{jt})$ can have a substantial practical effect on data users. Consequently, it is important for survey management to have diagnostic tools to assess this variability. For that assessment, four sources are of primary interest:

(A) Temporal or cross-sectional differences in the true design variances that are attributable to changes in factors that can be controlled (to some extent). For example, let n_{jt} equal the realized sample size for domain j at time t. If the variability in $s(\hat{\theta}_{jt})$ or $r(\hat{\theta}_{jt})$ were considered large enough to be problematic, and if it were attributable primarily to variability in n_{jt}, then one could consider a design modification that would reduce variability in n_{jt} values.
(B) Differences in the true design variance $V_p(\hat{\theta}_{jt})$ that are attributable to changes in factors that can be observed (or estimated from available data) but not controlled. For example, the true design variance and relative variance may be functions of estimable parameters of the underlying finite population, for example, functions of the element-level population variance, and of the true population parameter θ_{jt}.
(C) Differences in the true design variance $V_p(\hat{\theta}_{jt})$ that are attributable to factors that are neither controllable, nor observable, nor readily estimable. Examples include changes in $V_p(\hat{\theta}_{jt})$ that arise from short-term local changes in economic conditions.
(D) Sampling variability of the variance estimator $\hat{V}_p(\hat{\theta}_{jt})$. For surveys to which case (D) applies, one may wish to consider using an alternative to the current variance estimator.

Issues (A) through (D) can arise for both household and establishment surveys. For establishment surveys, these issues can be especially interesting due to two factors. First, many survey variables are approximately continuous and have heavily skewed population distributions. For example, in the establishment survey application considered below, individual employment counts range from single digits to tens of thousands, but most population units had counts in the single or double digits. Second, initiation of new sample units can be expensive and time consuming. To address these issues, many establishment surveys use a panel structure, and realized sample sizes may vary due to the effects of slow sample initiation, as well as attrition. This in turn may lead to increased variability in the true design variances.

The remainder of this article develops methods for exploration of sources (A) through (D) outlined above. These methods are based on relatively simple parametric models for the regression of $ln(\hat{V}_p(\hat{\theta}_{jt}))$ on predictor variables associated with sources (A) and (B). Such regression models may be viewed as extensions of generalized variance function models developed previously in the sample survey literature. Specifically, Section 2 provides a brief introduction to a case study based on the U.S. Current Employment Statistics Program. Section 3 develops some notation for the predictors, coefficients and error terms that will be important for these generalized variance function (GVF) extensions, and outlines estimation and inference methods for the applicable GVF models. Section 4 considers sources (A) and (B) through evaluation of the extent to which variability in $s(\hat{\theta}_{jt})$ may be associated with variability in observed predictors. Section 5 applies the main ideas of Sections 3 and 4 to the CES example introduced in Section 2; and also uses estimators of the equation-error variance to evaluate source (C). Section 6 reviews the main ideas of this article; discusses conditions under which source (D) may also be of practical importance; and considers several possible extensions of the methods developed here.

2. An Example: Monthly Estimation from the U.S. Current Employment Statistics Program

This article was motivated by variability in the direct standard errors computed for the U.S. Current Employment Statistics (CES) survey. The CES survey collects data each month on employment, hours, and earnings from a sample of nonagricultural establishments. The sample includes approximately 140,000 businesses and government agencies, covering around 440,000 individual worksites. Approximately 55,000 new sample units are enrolled in the CES survey each year to account for the establishment of new firms and to rotate a portion of the sample. When firms are rotated into the sample, they are retained for two years or more. The active CES sample includes approximately one third of all nonfarm payroll employees.

The CES design uses a stratified simple random sample of unemployment insurance (UI) accounts. A UI account is a cluster that may contain single or multiple establishments. The sample strata or subpopulations are defined by state, industry, and employment size class, yielding a state-based design. For a given sample size per state, sampling rates for each stratum are determined through optimum allocations to minimize the overall sampling error variance of the estimated statewide total private employment. All data on employment, hours, and earnings for the nation and for states and areas are classified in accordance with the 2007 North American Industry Classification System (NAICS). See the BLS Handbook of Methods (U.S. Bureau of Labor Statistics 2011, ch. 2), Butani et al. (1997) and Werking (1997) for further details.

CES uses a "weighted link relative estimator" of the employment in domain j for month t. This estimator is computed as the product

$$\hat{y}_{jt} = x_{j0}\hat{R}_{jt}, \tag{1}$$

where x_{j0} is the known Quarterly Census of Employment and Wages (QCEW) employment total for all establishments in domain j for the benchmark month 0; $\hat{y}_{jt}$ is an estimator of the unknown true employment total for domain j in month t; and $\hat{R}_{jt}$ is an estimator of the relative employment growth that took place from benchmark month 0 to the current month t as detailed in BLS Handbook of Methods (U.S. Bureau of Labor Statistics 2011) and Gershunskaya and Lahiri (2005). For the current article, the domains of interest are 14 large industries described in Table 1.

For a given reference month, the CES publishes estimates labeled "first closing", "second closing" and "third closing"; the second and third closing estimates use additional information from respondents not available for the first-closing estimates at the time of production. All results reported in this article are for sample sizes, point estimates and variance estimates for the third-closing data. For an additional discussion of the first, second and third closing for the CES, see Copeland and Valliant (2007).

The CES publishes many estimates of employment changes over time periods of varying lengths. However, this article will focus attention on only three distinct estimators: total employment, $\hat{y}_{jt}$, one-month change, $\hat{y}_{jt} - \hat{y}_{j,t-1}$, and one-month relative change $(\hat{y}_{j,t-1})^{-1}\hat{y}_{jt}$. In the discussion below, the generic term $\hat{\theta}_{jt}$ may represent any of these three estimators. In addition, the estimates $\hat{y}_{jt}, \hat{y}_{jt} - \hat{y}_{j,t-1}$ and $(\hat{y}_{j,t-1})^{-1}\hat{y}_{jt}$ and their associated variance estimates are computed for each month $t = 1, 2, \ldots, 20$. These months

Table 1. Description of industries

Industry	Description	Classification
1	Mining and logging	Goods-producing
2	Construction	Goods-producing
3	Durable goods manufacturing	Goods-producing
4	Nondurable goods manufacturing	Goods-producing
5	Wholesale trade	Service-providing
6	Retail trade	Service-providing
7	Transportation and warehousing	Service-providing
8	Utilities	Service-providing
9	Information	Service-providing
10	Financial activities	Service-providing
11	Professional and business services	Service-providing
12	Education and health services	Service-providing
13	Leisure and hospitality	Service-providing
14	Other services	Service-providing

correspond to March of a given year through October of the subsequent year. However, for a specified benchmark month 1, only results from the corresponding months 8 through 19 (October through the following September) are included in official third-closing publications. Consequently, all results presented in this article are based on data from these reference months 8 through 19.

Figure 1 presents boxplots of monthly realized sample sizes n_{jt} for the fourteen industries in the years 2005–2010. For CES national-level estimators, variance estimators are computed using balanced half-sample (BHS) methods, with Fay factors (Judkins 1990). These estimators include stratum-level finite population corrections. This article will use the symbols $\hat{V}_{pjt}$ to denote the BHS variance estimator for domain j and time t.

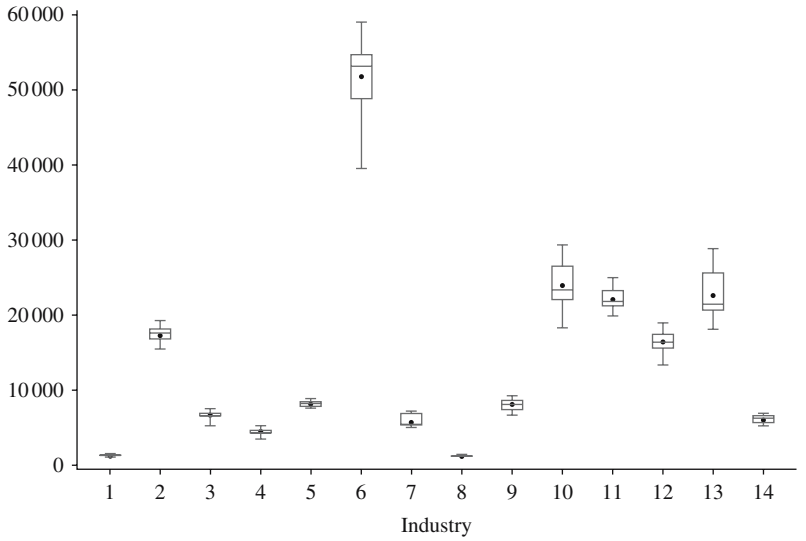


Fig. 1. Boxplots of the monthly numbers of responding sample units (n_{jt}) from years 2005–2010 for each of industries 1 though 14

Figure 2 presents boxplots of the natural logarithms of the BHS variance estimates, $ln(\hat{V}_{pjt})$ for monthly total employment in the specified industries. Note that log-scale differences $ln(\hat{V}_1) - ln(\hat{V}_2) = 1.5, 2.0$ and 3.0 correspond to variance ratios $(\hat{V}_1/\hat{V}_2)$ equal to 4.5, 7.4 and 20.1, respectively, and standard error ratios, $(\hat{V}_1/\hat{V}_2)^{1/2}$ equal to 2.1, 2.7 and 4.5, respectively. Consequently, the log-scale differences displayed in Figure 2 correspond to substantial differences on the standard error and variance scales.

To explore these patterns of variability at an industry level, Figure 3 presents a time plot of n_{jt} for construction; Figures 4 and 5 present the corresponding time plots of $ln(\hat{V}_{pjt})$ for total employment and one-month change respectively. Figures 3 displays "saw-tooth" patterns due to the periodic initiation of new units and continuing attrition of current units. In addition, the numbers of respondents n_{jt} generally show a marked increase between October and November of a given year. Similar plots were produced for other industries such as retail trade but are not shown in the article.

Furthermore, for a given benchmark year, the BHS variance estimator of total employment tends to increase across months, that is, the variance increases as the reference month moves farther away from the benchmark month. However, temporal trends with respect to months are considerably less pronounced in cases of one-month change and one-month relative change estimators.

3. Model Development, Estimation and Inference for Generalized Variance Functions

3.1. General Models for the True Design Variance

Due to the temporal variability in the standard errors computed from the BHS method, $s(\hat{\theta}_{jt}) = \{\hat{V}_p(\hat{\theta}_{jt})\}^{1/2}$, the CES program does not currently publish the values of $s(\hat{\theta}_{jt})$ as such. Instead, it publishes temporal medians of these standard errors. However, the CES

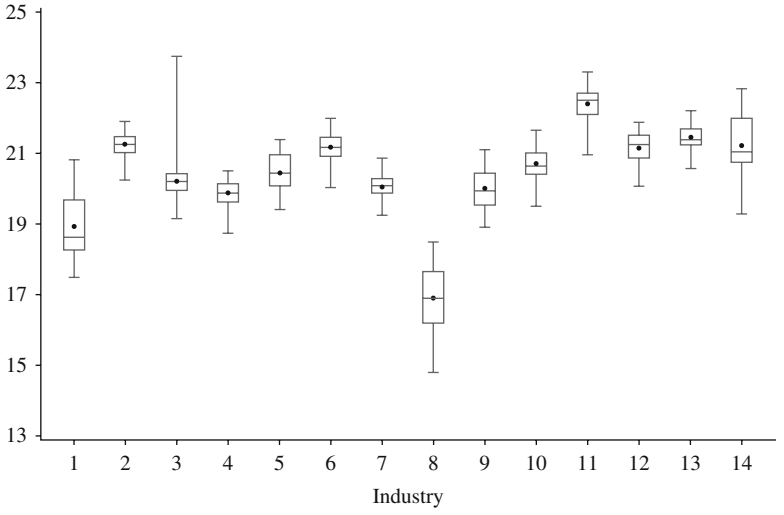


Fig. 2. Boxplots of $ln(\hat{V}_{pjt})$ for monthly estimates of total employment from years 2005–2010, separately for industries 1–14

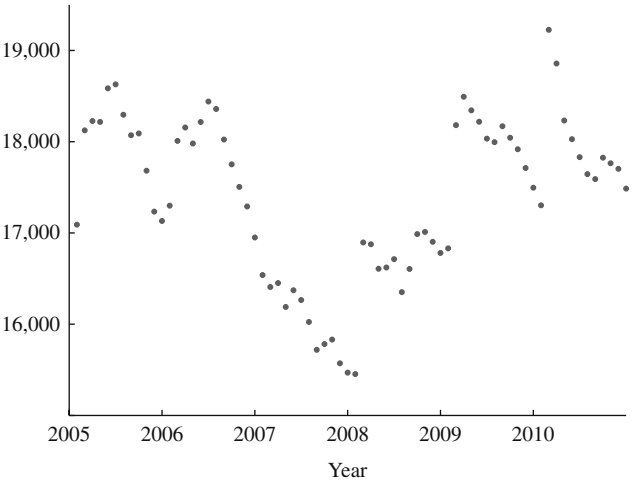


Fig. 3. Number of responding sample units (n_{jt}) across years: construction (monthly realized sample sizes for October 2005 through September 2011)

program is interested in exploring the reasons for variability of $s(\hat{\theta}_{jt})$, and in using the results of that exploration to develop alternative variance estimators.

To begin that exploratory study, let $X_{A_{jt}}$ be a vector of predictors that can be observed and controlled; let $X_{B_{jt}}$ be an additional vector of predictors that can be observed or estimated but not controlled; define $X_{jt} = \left(X_{A_{jt}}, X_{B_{jt}}\right)$; define $V_{pjt} = V_p(\hat{\theta}_{jt})$; and consider a general model

$$ln\left(V_{pjt}\right) = g(X_{jt}, \gamma) + q^*_{jt} \tag{2}$$

where q^*_{jt} is a univariate "equation error" with a mean equal to zero, and γ is a b-dimensional vector of variance function parameters. Note especially that q^*_{jt} represents

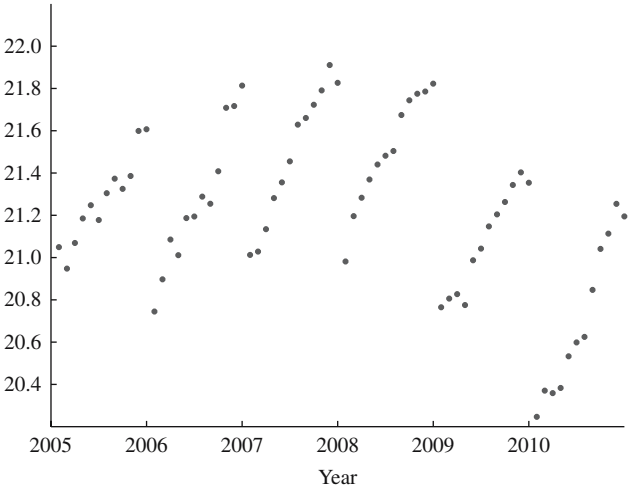


Fig. 4. Plot of $ln(\hat{V}_{pjt})$ of total employment across years: construction industry (estimates for October 2005 through September 2011)

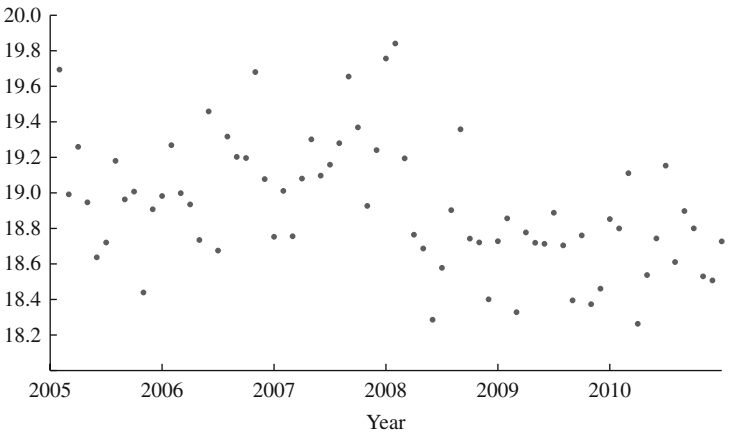


Fig. 5. Plot of $ln(\hat{V}_{pjt})$ of One-month change across years: construction (estimates for October 2005 through September 2011)

the deviation of logarithm of the true design variance V_{pjt} from its modeled value $g(X_{jt}, \gamma)$. Model (2) may be considered a type of generalized variance function, as developed in Johnson and King (1987), Valliant (1987), O'Malley and Zaslavsky (2005), Wolter (2007, sec. 7.2), Cho et al. (2002), Cho et al. (2014) and references cited therein. Some previous authors (e.g., Johnson and King 1987) have also developed generalized variance function models on logarithmic scales. Use of a logarithmic scale converts multiplicative relationships to linear relationships, and reduces the effects of extreme values.

Much of the GVF literature has focused on the variances of point estimators $\hat{\theta}_{jt}$ for population proportions or population totals related to a binary outcome variable; and has tended to emphasize predictors $X_{B_{jt}}$. In addition, much of this literature has used θ_{jt} as one component of the predictor vector $X_{B_{jt}}$. The current article, however, considers the more complex setting in which the point estimator of interest depends primarily on survey variables that are not binary; it will use predictors $X_{A_{jt}}$ and $X_{B_{jt}}$ that are not necessarily related to the value of θ_{jt}, but are related to important features of the sample design or estimation process.

On a logarithmic scale, one example of Model (2) is

$$ln(V_{pjt}) = \gamma_0 + \gamma_A X_{A_{jt}} + \gamma_B X_{B_{jt}} + q^*_{jt} \tag{3}$$

where $\gamma = (\gamma_0, \gamma_A, \gamma_B)$, γ_0 is univariate, γ_A is $1 \times b_A$, γ_B is $1 \times b_B$, $X_{A_{jt}}$ is $b_A \times 1$, $X_{B_{jt}}$ is $b_B \times 1$, $b = 1 + b_A + b_B$ and q^*_{jt} is a random variable with mean equal to zero and variance equal to $\sigma^2_{q^*_{jt}}$.

Before exploring specific forms of the Models (2) and (3), it is useful to add four comments on the conceptual basis for generalized variance functions. First, these functions are intended to approximate the true variances of $\hat{\theta}_{jt}$, considered over the set defined by $j = 1, \ldots, J$ and $t = 1, \ldots, T$, and averaging over all of the sources of random variability considered important for understanding the properties of $\hat{\theta}_{jt}$. In some of the original GVF literature, the only source considered was traditional sampling variability. However, in many cases, practical interest encompasses additional sources of variability, for example, the effects of nonresponse and measurement error. For these latter

applications, one would need to define V_{pjt} to include the relevant sources of both sampling and nonsampling error.

Second, practical fitting of Model (2) involves linear or nonlinear regression of the BHS variance estimators $\hat{V}_{pjt}$ on the corresponding predictors X_{jt}. Thus it is important for the BHS estimators $\hat{V}_{pjt}$ to be approximately unbiased for the variance terms V_{pjt} of interest. For example, if interest centers on variance terms V_{pjt} that include the effects of nonresponse, and of weighting adjustment or imputation used to construct $\hat{\theta}_{jt}$, then one would need to use BHS variance estimators $\hat{V}_{pjt}$ that incorporate these effects, for example, through Rao-Shao adjustments or multiple imputation. Similarly, if one intends to account for the effects of measurement errors on $\hat{\theta}_{jt}$, then it would be important to use initial estimators $\hat{V}_{pjt}$ that account for the combined effects of sampling error and measurement error, per Wolter (2007, app. D).

Third, similar comments apply to variance function models, such as Model (3), that are fit following a nonlinear transformation. For these cases, it is important to account for transformation effects in discussion of unbiased estimation. Valliant (1987) provides a rigorous conceptual basis for generalized variance functions under some specific superpopulation models.

Fourth, the choice of approximate predictors $X_{A_{jt}}$ and $X_{B_{jt}}$ will depend on specific features of a given application. Important criteria include availability of the predictors at the appropriate level of aggregation; potential relevance of the predictors, based on features of the sampling and estimation process; empirical assessment of the statistical significance of the coefficients of the predictors in specific models; and related diagnostics for the goodness-of-fit for the variance function model when specific predictors are included. The remainder of this article explores these ideas in additional detail.

3.2. *Point Estimation and Variance Estimation for Coefficients*

For several versions of a Model (3), we computed estimators $\hat{\gamma}$ of the coefficients γ through ordinary least squares (OLS) regression of $ln(\hat{V}_{pjt})$ on the corresponding vector of predictors. In keeping with Valliant (1987), one could consider alternative estimators of γ based on weighted least squares methods, with weights proportional to the inverses of preliminary estimators of variances of the error terms in Model (3). However, exploratory application of this idea to the CES data encountered issues with numerical stability; see Section 6 for related comments.

In addition, practical work with GVFs can require one to identify groups of estimators $\hat{\theta}_{jt}$ for which a common set of coefficients γ may be used. Some authors have addressed this need through qualitative identification of estimators with similar design or population features; for an example see Wolter (2007, 276). To complement this qualitative approach, it is useful to produce estimators of the variance of the coefficient vector estimator $\hat{\gamma}$, and to carry out significance testing for homogeneity of the coefficients across groups. For example, Subsection 5.2 will present results on comparison of coefficients across years and across industry groups. For this goal, we obtained an estimator $\hat{V}_p(\hat{\gamma})$ of the variance of the approximate distribution of $\hat{\gamma}$ from an extension of standard estimating equation approaches for complex-survey estimators (Binder 1983). Details of the estimating equation formulation for GVF cases and its applications were provided in Cho et al. (2014). This formulation

accounted directly for the features of the sample design and the point estimators $\hat{\theta}_{jt}$. In the CES example, the dependent variables $\hat{V}_{pjt}$ may be strongly correlated across months, due to the form of the weighted link relative estimators as well as the use of a rotation sample design. However, sampling is essentially independent across domains. Thus we decomposed the estimating equation into sums of terms across independent domains. Based on this design-adjusted variance estimator for $\hat{\gamma}$, Cho et al. (2014) showed that standard (unadjusted) variance estimates for $\hat{\gamma}$ may be much smaller than the unbiased estimates. Consequently, it is important to use design-adjusted estimators, $\hat{V}_p(\hat{\gamma})$, in inference for γ.

3.3. Models for Variance Estimation Error

Now consider again the temporal and cross-sectional variability in standard errors discussed in Section 1. Within the framework defined by Model (3), Sources (A) and (B) correspond to the regression terms $\gamma_A X_{A_{jt}}$ and $\gamma_B X_{B_{jt}}$, respectively; and Source (C) corresponds to the equation error term q^*_{jt}. In addition, design features associated with Source (A) and the choice of a specific variance estimator $\hat{V}_{pjt}$ can both have an effect on the sampling errors defined by the differences

$$e_{jt} = \hat{V}_{pjt} - V_{pjt} \tag{4}$$

for Source (D). Note especially that the sampling errors e_{jt} are conceptually distinct from the equation errors q_{jt} in Expression (2). Similar distinctions arise in other work with sampling errors and measurement errors. See, for example, Fuller (1987). In some cases, one may treat the distribution of the e_{jt} terms as a rescaled and centered version of a chi-squared distribution on d_{jt} degrees of freedom, that is,

$$V_{pjt}^{-1}\, d_{jt}\, \hat{V}_{pjt} = V_{pjt}^{-1}\, d_{jt}\, e_{jt} + d_{jt} \sim \chi^2_{d_{jt}}.$$

Some of the sample survey literature approximates d_{jt} as the difference between the number of primary sampling units and the number of strata applicable to domain j at time t. For some discussion of conditions under which this approximation may be appropriate, see Korn and Graubard (1990), Valliant and Rust (2010) and references cited therein. Our CES analyses will consider only estimators of national-level population parameters for relatively large industries. For such cases, the abovementioned computations lead to values of d_{jt} greater than 100. Consequently, the current article will devote relatively limited attention to the sampling error terms e_{jt}.

4. Differences Attributable to Variability in the Predictors X_{jt}; Equation Error; and Estimation Error

In keeping with standard approaches to decomposition of sums of squares in regression (e.g., Draper and Smith 1998, ch. 6), one may decompose the variability of $ln(\hat{V}_{pjt})$ into four terms:

SS_A: The sum of squared differences associated with controllable predictors X_A
$SS_{B|A}$: The sum of squared differences associated with predictors X_B, after accounting for the controllable-predictor terms X_A variability

SS_Q: The variability associated with equation error (sometimes called "lack of fit" error in the regression literature)
SS_{PE}: The variability attributable to the random variability of $ln(\hat{V}_{pjt})$ conditional on $ln(V_{pjt})$ (sometimes called "pure error" in the regression literature)

For the CES national-level work, Subsection 3.3 noted that the $\hat{V}_{pjt}$ estimators are associated with relatively large "degrees of freedom" terms d_{jt}. Consequently, our analysis will use the assumption that the conditional variance $V_p\left[ln(\hat{V}_{pjt})|V_{pjt}\right]$ is approximately equal to zero. Note that $V_p\left[ln(\hat{V}_{pjt})|V_{pjt}\right]$ reflects the sampling variability of $ln(\hat{V}_{pjt})$ after conditioning on the true variance term V_{pjt}, and thus is essentially conditioning on the predictors X_{jt} and the equation errors q^*_{jt}. Thus we will use the corresponding assumption that $SS_{PE} = 0$. With this approximation, we have the decomposition of the "corrected total" sum of squares

$$SSCT = \sum_{j=1}^{J}\sum_{t=1}^{T}\left\{ln(\hat{V}_{pjt}) - L_{..}\right\}^2 \tag{5}$$

$$= SS_A + SS_{B|A} + SS_Q \tag{6}$$

$$= SS_B + SS_{A|B} + SS_Q \tag{7}$$

where $L_{..} = J^{-1}T^{-1}\sum_{j=1}^{J}\sum_{t=1}^{T} ln(\hat{V}_{pjt})$. In addition, for a full-model fit

$$ln(\hat{V}_{pjt}) = \gamma_0 + \gamma_A X_{A_{jt}} + \gamma_B X_{B_{jt}} + q^*_{jt},$$

the customary model R^2 equals the ratio

$$(SSCT)^{-1}\left(SS_A + SS_{B|A}\right) = (SSCT)^{-1}\left(SS_B + SS_{A|B}\right). \tag{8}$$

Furthermore, for the partial model fit $ln(\hat{V}_{pjt}) = \gamma_0 + \gamma_A X_{A_{jt}} + q^*_{jt}$, the resulting model R^2 equals the ratio $(SSCT)^{-1}SS_A$. Similar comments apply to the partial model fit

$$ln(\hat{V}_{pjt}) = \gamma_0 + \gamma_B X_{B_{jt}} + q^*_{jt}$$

with model R^2 equal to $(SSCT)^{-1}SS_B$.

5. Application to the U.S. Current Employment Statistics Program

5.1. Models from the Decomposition of the Design Variance

To identify some potential predictors X_A and X_B for the CES application, recall that our sample consists of unemployment insurance accounts, which report nonzero employment for previous and current months. Let n_{jt} be a number of responding UI accounts, N_{jt} be a number of total UI accounts, and S^2_{jt} be a finite population variance within the domain j at time t. Then, we can express the variance of $\hat{y}_{jt}$ as a function of a design effect, Δ_{jt}, for $\hat{y}_{jt}$. Ignoring the finite-population correction term, we write the variance of $\hat{y}_{jt}$, in terms of Δ_{jt}

on the original variance scale:

$$V_p(\hat{y}_{jt}) = \Delta_{jt}\left(n_{jt}^{-1} S_{jt}^2 N_{jt}^2\right). \tag{9}$$

For some general background on design effects and their use in variance approximations, see, for example, Kish (1995), Park and Lee (2004) and references cited therein.

Note that Expression (9) uses the variance term S_{jt}^2, the finite population variance of the original employment counts y_{jt}. The design effect term Δ_{jt} incorporates all of the ratio estimator effects. In addition, for point estimators such as (1) that are based on estimators of cumulative growth from a benchmark month, Δ_{jt} may be an increasing function of t (i.e., the design variance increases as the reference month moves further away from the benchmark month). For example, one could consider the approximation

$$\Delta_{jt} \doteq \Delta\, t^{\alpha_0}, \tag{10}$$

where Δ is a common design effect term shared across domains j. Moreover, several authors have considered cases in which (sub)population variances are functions of associated (sub)population means or totals. For example, Cochran (1977, 243) discusses approximation of a finite population variance of an area unit as proportional to a positive power of the size of that unit. Similarly, Box-Cox transformations are often based on the assumption of a power relationship between the means and variances of sets of observations. Application of this idea to the CES leads to the approximation

$$S_{jt}^2 N_{jt}^2 \doteq \alpha_1 (x_{j0})^{\alpha_2}, \tag{11}$$

where α_1 and α_2 are constants, and x_{j0} is the QCEW employment total for all establishments in domain j for the month $t = 0$. Taken together, Expressions (9) through (11) suggest that on a logarithmic scale, one may consider the variance model

$$ln\{V(\hat{y}_{jt})\} = ln(\Delta) + \alpha_0 ln(t) - ln(n_{jt}) + ln(\alpha_1) + \alpha_2 ln(x_{j0}) + q_{jt}^*$$

or in a slightly more general form,

$$ln\{V(\hat{y}_{jt})\} = \gamma_0 + \gamma_1 ln(n_{jt}) + \gamma_2 ln(t) + \gamma_3 ln(x_{j0}) + q_{jt}^* \tag{12}$$

where, for example, $\gamma_0 = ln(\Delta) + ln(\alpha_1)$, $\gamma_1 = -1$, $\gamma_2 = \alpha_0$ and $\gamma_3 = \alpha_2$.

In addition, under some standard designs, the selected sample size n_{jt} may be a function of variables related to x_{jt}. For example, under Neyman allocation (e.g., Cochran 1977, 99) n_{jt} is proportional to $S_{jt}N_{jt}$ provided the domains were equal to individual strata, and so the log transformed Model (12) reduces to

$$ln(V_{pjt}) = \gamma_0 + \gamma_1 ln(x_{j0}) + \gamma_2 ln(t) + q_{jt}^* \tag{13}$$

with appropriate redefinitions of the coefficients γ_0, γ_1 and γ_2. For Model (13), $X_{B_{jt}} = \left[ln(x_{j0}), ln(t)\right]$. This model does not include any variables under the direct control of the designer, so $X_{A_{jt}}$ is empty. For the CES application, the domains were unions of several strata. Consequently, in preliminary work, we considered versions of Model (13) that included $X_{A_{jt}} = ln(n_{jt})$. However, our empirical results indicated that after inclusion of the predictor $ln(x_{j0})$, the additional predictor $ln(n_{jt})$ provided very limited additional value. Consequently, our modeling work for this article centered on versions of Model (13).

Finally, recall from Subsection 3.1 that GVF models often include the point estimator $\hat{\theta}_{jt}$ as a predictor. For the CES application, this would suggest inclusion of the population total estimators $\hat{y}_{jt}$. However, these estimators are strongly associated with the benchmark values x_{j0} which are already included in Model (13). Consequently, we did not include $\hat{y}_{jt}$ as an additional predictor in Model (13) for the CES data.

5.2. *Differences in Model Coefficients γ*

Model (13) was based on the assumption that the coefficient vector γ was constant over all years and all domains. However, this assumption may not hold, for example, if the underlying terms $\Delta, \alpha_0, \alpha_1$, or α_2 are not constant over years and domains. Consequently, we explored the possible heterogeneity of γ over years and domains, respectively.

5.2.1. Temporal Homogeneity

To explore the temporal heterogeneity of $\hat{V}_{pjt}$, we divided years into two groups. National Bureau of Economic Research (NBER) declared the current recession starting December 2007. Moreover, the data from the BLS payroll employment site (http://data.bls.gov/timeseries/CES0000000001?output_view=net_1mth) are generally consistent with the NBER recession timing. Consequently, we fit Model (13) separately for the years 2005–2007 and 2008–2010, respectively.

$$ln(V_{pjt}) = \begin{cases} \gamma_{10} + \gamma_{11} ln(x_{j0}) + \gamma_{12} ln(t) + q^*_{jt} & \text{if } Year = 2005\text{-}2007 \\ \gamma_{20} + \gamma_{21} ln(x_{j0}) + \gamma_{22} ln(t) + q^*_{jt} & \text{if } Year = 2008\text{-}2010 \end{cases} \tag{14}$$

In addition, we tested for the homogeneity of coefficients across year groups, based on the null hypothesis

$$H_0 : (\gamma_{10}, \gamma_{11}, \gamma_{12}) = (\gamma_{20}, \gamma_{21}, \gamma_{22}).$$

For this test, the Wald test statistic is:

$$W = (A\,\hat{\gamma})' \left\{ A\,\hat{V}(\hat{\gamma}) A' \right\}^{-1} (A\,\hat{\gamma}) \tag{15}$$

where $\gamma = \left(\gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_{20}, \gamma_{21}, \gamma_{22}\right)$, $\hat{V}(\hat{\gamma})$ is a 6×6 design-based estimator of the covariance matrix of $\hat{\gamma}$ as described in Subsection 3.2 and

$$A = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}.$$

Standard arguments adapted to the current case (e.g., Korn and Graubard 1990) indicate that $(W/d)\{(d-p+1)/p\}$ has approximately a noncentral F distribution with p and $(d-p+1)$ degrees of freedom and with noncentrality parameter $(A\gamma)'\{AV(\gamma)A'\}^{-1}(A\gamma)$ where $l = 28$ is number of clusters (due to the presence of two groups of years intersected with 14 industries); $d = l - 1 = 27$; and $p = 3$ is number of rows in the contrast Matrix A.

Table 2 presents the resulting coefficient estimates, standard errors and test statistics. The separate blocks of rows in Table 2 correspond to separate model fits for variance

Table 2. Coefficient point estimates, standard errors and tests for homogeneity between early years (2005–7) and late years (2008–10); critical values for Wald test statistics W: 9.69 for $\alpha = 0.05$, 7.51 for $\alpha = 0.10$

	Early years (2005–2007)			Late years (2008–2010)					
Estimator	γ_{10}	γ_{11}	γ_{12}	γ_{20}	γ_{21}	γ_{22}	R^2_γ	$\hat{\sigma}^2_e$	W
Total employment	0.26	1.08	1.33	0.38	1.15	0.87	0.71	0.57	11.14
(s.e.)	(4.46)	(0.27)	(0.12)	(2.50)	(0.16)	(0.09)			
1-Month change	−2.18	1.27	0.32	−1.45	1.29	−0.12	0.72	0.67	6.95
(s.e.)	(2.65)	(0.17)	(0.13)	(2.84)	(0.18)	(0.11)			
1-Month relative change	−2.27	−0.72	0.30	−1.60	−0.71	−0.09	0.45	0.67	5.73
(s.e.)	(2.59)	(0.17)	(0.13)	(2.75)	(0.17)	(0.10)			

estimates $\hat{V}_{pjt}$ associated with total employment, one-month change and one-month relative change, respectively. Note especially the strong indications of statistically significant coefficients for $ln(x_{j0})$ for each of the model fits. The coefficient estimates of $ln(x_{j0})$ are positive for total employment and one-month change, reflecting the fact that larger values of $ln(V_{pjt})$ were generally associated with domains that had larger levels of employment and employment change.

For total employment, one-month change and one-month relative change estimators, the W values are 11.14, 6.95 and 5.73, respectively. The cutoff points $\{d\,p/(d-p+1)\}$ $F_{0.05}\{p,(d-p+1)\}$ were 9.69 for $\alpha = 0.05$ and 7.51 for $\alpha = 0.10$. Note that the test statistic for total employment is much larger than both cutoff points. Thus, at conventional levels of significance, for the case of total employment, we reject the null hypothesis of equality of the GVF coefficients across the two groups of years. In addition, note that for total employment, the coefficient for the predictor $ln(t)$ changes substantially between 2005–2007 ($\hat{\gamma}_{12} = 1.33$) and 2008–2010 ($\hat{\gamma}_{22} = 0.87$), relative to the magnitude of $se(\hat{\gamma}_{12}) = 0.12$. This illustrates the importance of carrying out empirical checks on the homogeneity of variance function models across years, rather than just assuming that the coefficients are constant.

5.2.2. Cross-Sectional Homogeneity

To explore the cross-sectional variability of $\hat{V}_{pjt}$, we fit Model (13) separately for domains in goods-producing and service-providing industries, respectively, which led to the model

$$ln(V_{pjt}) = \begin{cases} \gamma_{10} + \gamma_{11} ln(x_{j0}) + \gamma_{12} ln(t) + q^*_{jt} & \text{if Goods (four industries)} \\ \gamma_{20} + \gamma_{21} ln(x_{j0}) + \gamma_{22} ln(t) + q^*_{jt} & \text{if Services (ten industries)} \end{cases} \quad (16)$$

In addition, we tested the null hypothesis $H_0 : (\gamma_{10}, \gamma_{11}, \gamma_{12}) = (\gamma_{20}, \gamma_{21}, \gamma_{22})$ using the Wald test statistic (15) where $l = 14$ is number of clusters because there are two industry groups: one with four industries and the other with ten industries; $d = l - 1 = 13$; and $p = 3$ is number of rows in the contrast Matrix A. Table 3 presents the results of these analyses. As with Table 2, we have three sets of results for total employment, one-month change and one-month relative change, respectively. For estimators of total, one-month change and one-month relative change, the W values were 15.94, 65.33 and 53.52, respectively. The cutoff points were 12.72 for $\alpha = 0.05$, and 9.43 for $\alpha = 0.10$.

Thus we have strong indication of differences in the Goods and Services coefficients for all three sets of estimators.

Finally, note that in both Table 2 and Table 3, the R^2 values for the total employment and one-month change were relatively strong (greater than 0.7 in each case). For the two GVF model fits for one-month relative change, the R^2 values were somewhat lower (0.45 and 0.49, respectively).

5.3. *Evaluation of Sources of Variability in the CES Variance Estimators*

After evaluating the coefficient estimators $\hat{\gamma}$ for the CES data, we applied the diagnostic ideas outlined in Section 4.

Table 3. Coefficient point estimates, standard errors and tests for homogeneity between two groups of industries by product types; critical values for Wald test statistics W: 12.72 for $\alpha = 0.05$, 9.43 for $\alpha = 0.10$

	Goods			Services					
Estimator	γ_{10}	γ_{11}	γ_{12}	γ_{20}	γ_{21}	γ_{22}	R^2_γ	$\hat{\sigma}^2_e$	W
Total employment	6.69	0.69	1.15	−2.35	1.29	1.08	0.75	0.50	15.94
(s.e.)	(2.17)	(0.16)	(0.10)	(3.02)	(0.18)	(0.12)			
1-Month change	4.87	0.90	−0.25	−4.73	1.44	0.23	0.75	0.62	65.33
(s.e).	(1.61)	(0.12)	(0.07)	(1.86)	(0.13)	(0.12)			
1-Month relative change	4.27	−1.07	−0.21	−4.68	−0.56	0.23	0.49	0.62	53.52
(s.e.)	(1.64)	(0.12)	(0.08)	(1.90)	(0.13)	(0.12)			

Table 4 presents results for full and partial model fits for the variance estimators $\hat{V}_{pjt}$ for total employment. In keeping with the results of Table 2, we allowed separate coefficients for the early years (2005–2007) and the late years (2008–2010), respectively. Note that in Table 4, in the full model fit for both early and late years, all coefficients (except for the intercept) are statistically significant at conventional α levels. In addition, $R^2 = 0.71, 0.67$ and 0.05 for the full model fit, the fit with $ln(t)$ omitted, and the fit with $ln(x_{j0})$ omitted, respectively. In that sense, most of the explanatory power of Model (14) is attributable to the predictors $ln(x_{j0})$. This also indicates that although the coefficient of $ln(t)$ satisfies significance testing criteria at customary α levels, it does not contribute much power for prediction of $ln(V_{jt})$ as reflected in R^2 and σ_e^2 values. This illustrates the importance of using the diagnostics of Section 4 as complements to the coefficient testing idea from Section 3.

Tables 5 and 6 present related results for the variance estimators $\hat{V}_{pjt}$, associated with one-month change and one-month relative change, respectively. Table 5 displays patterns of statistical significance and R^2 results that are similar to those observed in Table 4, except that for the late years, the full model fit does not lead to statistically significant coefficients for the predictor $ln(t)$. The results in Table 6 differ from those in Tables 4 and 5 in two notable ways. First, in the full-model fit, the estimates for γ_0 and γ_1 are negative in Table 6, but positive in Table 4. Second, the R^2 values in Table 6 are notably smaller than those in Tables 4 and 5 for the full model fit and the $ln(t)$-omitted fits. Because the underlying point estimator for Table 6 is a ratio, one would not necessarily expect Table 6 to display the same pattern as observed for point estimators for totals and differences of totals as in Tables 4 and 5, respectively.

5.4. *Magnitude of Equation Error Variances*

To address issues (A) and (B) of Section 1, Subsection 5.2 developed methods for the identification of predictors X_{jt} that account for some of the observed variability in the

Table 4. Total employment: coefficient estimates, inferential statistics and R^2 values for full-model and reduced-model fits

	Early years (2005–2007)			Late years (2008–2010)				
Model	Intercept γ_0 (s.e.) (t_{γ_0})	$ln(x_{j0})$ γ_1 (s.e.) (t_{γ_1})	$ln(t)$ γ_2 (s.e.) (t_{γ_2})	Intercept γ_0 (s.e.) (t_{γ_0})	$ln(x_{j0})$ γ_1 (s.e.) (t_{γ_1})	$ln(t)$ γ_2 (s.e.) (t_{γ_2})	R^2_γ	$\hat{\sigma}^2_e$
Full	0.26 (4.46) (0.06)	1.08 (0.27) (3.93)	1.33 (0.12) (11.31)	0.38 (2.50) (0.15)	1.15 (0.16) (7.42)	0.87 (0.09) (9.46)	0.71	0.57
$ln(t)$ omitted	3.66 (4.35) (0.84)	1.08 (0.27) (3.93)	– – –	2.62 (2.41) (1.09)	1.15 (0.16) (7.42)	– – –	0.67	0.66
$ln(x_{j0})$ omitted	16.97 (0.62) (27.41)	– – –	1.33 (0.12) (11.31)	18.21 (0.48) (38.05)	– – –	0.87 (0.09) (9.46)	0.05	1.89

Table 5. One-month change: coefficient estimates, inferential statistics and R^2 values for full-model and reduced-model fits

	Early years			Late years				
Model	Intercept γ_0 (*s.e.*) (t_{γ_0})	$ln(x_{j0})$ γ_1 (*s.e.*) (t_{γ_1})	$ln(t)$ γ_2 (*s.e.*) (t_{γ_2})	Intercept γ_0 (*s.e.*) (t_{γ_0})	$ln(x_{j0})$ γ_1 (*s.e.*) (t_{γ_1})	$ln(t)$ γ_2 (*s.e.*) (t_{γ_2})	R^2_γ	$\hat{\sigma}^2_e$
Full	−2.18 (2.65) (−0.82)	1.27 (0.17) (7.32)	0.32 (0.13) (2.44)	−1.45 (2.84) (−0.51)	1.29 (0.18) (7.20)	−0.12 (0.11) (−1.16)	0.72	0.67
$ln(t)$ omitted	−1.37 (2.74) (−0.50)	1.27 (0.17) (7.32)	– – –	−1.76 (2.80) (−0.63)	1.29 (0.18) (7.20)	– – –	0.72	0.68
$ln(x_{j0})$ omitted	17.53 (0.39) (44.96)	– – –	0.32 (0.13) (2.44)	18.43 (0.36) (50.98)	– – –	−0.12 (0.11) (−1.16)	0.01	2.41

estimators $\hat{V}_{pjt}$. Furthermore, Sections 4 and 5.3 used standard regression diagnostics to evaluate the properties of variability in $ln(\hat{V}_{pjt})$ that is attributable to specific predictors, X.

To address issue (C), this section will consider the variability of the residual terms $\hat{q}^*_{jt} = ln(\hat{V}_{pjt}) - X_{jt}\hat{\gamma}$. In particular, we address issue (C) by exploring the extent to which the variances of the residuals $\hat{q}^*_{jt}$ may vary across industry, employment size at benchmark month, or month.

Figure 6 presents a scatter plot of these monthly residuals for total employment against the predicted values with separate plotting symbols for industries that are goods producing (1–4) and service providing (5–14), respectively. To explore this further, Table 7 presents

Table 6. One-month relative change: coefficient estimates, inferential statistics and R^2 values for full-model and reduced-model fits

	Early years			Late years				
Model	Intercept γ_0 (*s.e.*) (t_{γ_0})	$ln(x_{j0})$ γ_1 (*s.e.*) (t_{γ_1})	$ln(t)$ γ_2 (*s.e.*) (t_{γ_2})	Intercept γ_0 (*s.e.*) (t_{γ_0})	$ln(x_{j0})$ γ_1 (*s.e.*) (t_{γ_1})	$ln(t)$ γ_2 (*s.e.*) (t_{γ_2})	R^2_γ	$\hat{\sigma}^2_e$
Full	−2.27 (2.59) (−0.88)	−0.72 (0.17) (−4.27)	0.30 (0.13) (2.38)	−1.60 (2.75) (−0.58)	−0.71 (0.17) (−4.08)	−0.09 (0.10) (−0.84)	0.45	0.67
$ln(t)$ omitted	−1.50 (2.68) (−0.56)	−0.72 (0.17) (−4.27)	– – –	−1.82 (2.72) (−0.67)	−0.71 (0.17) (−4.08)	– – –	0.45	0.67
$ln(x_{j0})$ omitted	−13.52 (0.42) (−32.19)	– – –	0.30 (0.13) (2.38)	−12.57 (0.42) (−30.00)	– – –	−0.09 (0.10) (−0.84)	0.00	1.21

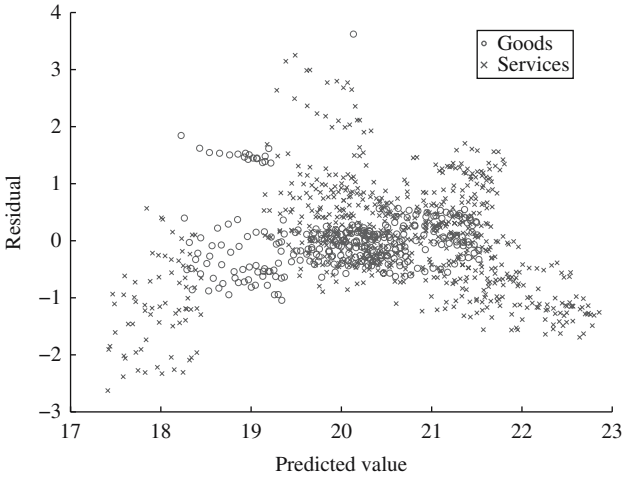


Fig. 6. Scatter plot of log-scale residuals $\hat{q}^{*}_{jt}$ against predicted values $X_{jt}\hat{\gamma}$ for the variance of total employment

selected sample quantiles of these residuals for goods-producing and service-providing industries, respectively, based on data from 2005–2010. Note especially that for each of total employment, one-month change, and one-month relative change, the interquartile range (IQR) for goods is somewhat larger than the IQR for services. However, the difference between the 99th percentile and the first percentile is larger for services than for goods with total employment, and are approximately equal with one-month change and one-month relative change.

In addition, we fit the models,

$$\left(\hat{q}^{*}_{jt}\right)^2 = \begin{cases} \omega_{G0} + \omega_{G1} ln(x_{j0}) + \omega_{G2} ln(t) & \text{if } j \in \text{ Goods} \\ \omega_{S0} + \omega_{S1} ln(x_{j0}) + \omega_{S2} ln(t) & \text{if } j \in \text{ Services} \end{cases} \tag{17}$$

and tested $H_0 : (\omega_{G0}, \omega_{G1}, \omega_{G2}) = (\omega_{S0}, \omega_{S1}, \omega_{S2})$ using estimators and test statistics similar to those developed in Subsection 5.2.

Table 8 presents the resulting coefficient estimates, standard errors and test statistics. Note that the Wald tests do not reject the null hypothesis of no differences for the total employment, one-month change and one-month relative change analyses. However, for one-month change and one-month relative change, *t*-tests on individual coefficients are fairly distinct for the "Goods" and "Services" models, respectively. In particular, for the "Goods" analyses, the coefficients for $ln(x_{j0})$ are not significant for these two cases; and for the "Services" analyses, the coefficients for $ln(x_{j0})$ are significant for the corresponding two cases.

Reviewers of an earlier form of this article noted that a version of Figure 6 displays curvature for the service-providing industries. To address this, we fit an alternative form of Model (13) that included the predictor $\left\{ln(x_{j0})\right\}^2$. Table 9 presents results for the model

$$ln(\hat{V}_{pjt}) = \begin{cases} \gamma_{10} + \gamma_{11} ln(x_{j0}) + \gamma_{111} \left\{ln(x_{j0})\right\}^2 + \gamma_{12} ln(t) + q^{*}_{jt} & \text{if Year} = 2005\text{-}2007 \\ \gamma_{20} + \gamma_{21} ln(x_{j0}) + \gamma_{211} \left\{ln(x_{j0})\right\}^2 + \gamma_{22} ln(t) + q^{*}_{jt} & \text{if Year} = 2008\text{-}2010 \end{cases} \tag{18}$$

Table 7. Quantiles of residuals from Model (16) for two groups of industries

Estimator	Group	0.01	0.10	0.25	0.50	0.75	0.90	0.99	IQR
Total employment	Goods	−0.98	−0.69	−0.52	−0.23	0.65	1.07	1.52	1.17
	Services	−1.63	−0.79	−0.42	−0.02	0.38	0.84	2.26	0.80
1-Month change	Goods	−1.23	−0.81	−0.56	−0.16	0.49	0.85	2.91	1.05
	Services	−1.48	−0.87	−0.48	−0.08	0.37	0.90	2.60	0.86
1-Month relative change	Goods	−1.25	−0.82	−0.54	−0.13	0.48	0.86	2.88	1.02
	Services	−1.47	−0.87	−0.49	−0.07	0.36	0.92	2.57	0.85

Table 8. Coefficient point estimates, standard errors and test for homogeneity between two groups of industries for the equation-error variance Model (20): critical value for Wald test statistics W: 12.72 for $\alpha = 0.05$, 9.43 for $\alpha = 0.10$

	Goods			Services			
Estimator	ω_{G0}	ω_{G1}	ω_{G2}	ω_{S0}	ω_{S1}	ω_{S2}	W
Total employment	2.00	−0.08	−0.09	3.17	−0.12	−0.30	1.08
(*s.e.*)	(0.46)	(0.05)	(0.12)	(1.32)	(0.06)	(0.19)	
(t)	(4.35)	(−1.67)	(−0.74)	(2.41)	(−1.96)	(−1.59)	
1-Month change	−0.57	0.14	−0.35	2.87	−0.15	0.05	7.88
(*s.e.*)	(1.29)	(0.13)	(0.30)	(0.95)	(0.05)	(0.11)	
(t)	(−0.44)	(1.06)	(−1.16)	(3.03)	(−3.20)	(0.44)	
1-Month relative change	−0.84	0.15	−0.28	3.04	−0.16	0.03	6.76
(*s.e.*)	(1.15)	(0.13)	(0.30)	(0.96)	(0.05)	(0.12)	
(t)	(−0.73)	(1.16)	(−0.94)	(3.16)	(−3.31)	(0.29)	

Table 9. Coefficient point estimates, standard errors and tests for homogeneity between early years (2005–7) and late years (2008–10); critical values for Wald test statistics W: 18.09 for $\alpha = 0.05$, 13.55 for $\alpha = 0.10$

	Goods				Services						
Estimator	γ_{10}	γ_{11}	γ_{111}	γ_{12}	γ_{20}	γ_{21}	γ_{211}	γ_{22}	R^2_γ	$\hat{\sigma}^2_e$	W
Total employment	−17.30	3.99	−0.11	1.15	−56.54	8.51	−0.24	1.08	0.78	0.44	777
(*s.e.*)	(88.67)	(12.23)	(0.42)	(0.10)	(21.50)	(2.96)	(0.10)	(0.12)			
(t)	(−0.20)	(0.33)	(−0.27)	(11.31)	(−2.63)	(2.87)	(−2.37)	(9.05)			
1-Month change	33.78	−3.08	0.14	−0.25	−38.36	5.92	−0.15	0.23	0.76	0.59	1117
(*s.e.*)	(57.30)	(7.86)	(0.27)	(0.07)	(17.98)	(2.46)	(0.08)	(0.12)			
(t)	(0.59)	(−0.39)	(0.51)	(−3.37)	(−2.13)	(2.41)	(−1.77)	(1.94)			
1-Month relative change	33.90	−5.14	0.14	−0.21	−39.75	4.11	−0.16	0.23	0.52	0.59	932
(*s.e.*)	(58.87)	(8.07)	(0.28)	(0.08)	(18.07)	(2.47)	(0.08)	(0.12)			
(t)	(0.58)	(−0.64)	(0.51)	(−2.54)	(−2.20)	(1.66)	(−1.84)	(1.98)			

A plot of the residuals from Model (18) against the predicted values produced by Model (18) did not display the curvature pattern observed in Figure 6. However, in comparison with the results in Table 3 for Model (16), Table 9 indicates that for Model (18), inclusion of the squared predictor to modest changes in the values of σ_e^2 (0.50 vs 0.44) and R^2 (0.75 vs 0.78). In addition, for goods-producing industries, the t-statistics reported for the coefficient γ_{111} of $\{ln(x_{j0})\}^2$ are not significant at conventional levels of significance; however, for service-providing industries, the t-statistics reported for the coefficients γ_{211} equal -2.37, -1.77 and -1.84 for total employment, one-month change and one-month relative change respectively. Use of an additional regressor $ln(n_{jt})$ was explored through analyses that are not detailed in the current text; inclusion of this regressor did not produce notable changes in the analyses.

In summary, Subsections 5.2 and 5.3 indicated that much of the observed variability in the $ln(\hat{V}_{pjt})$ values may be attributed to variability in the conditional-expectation structure described by the regression Model (13). In addition, those sections indicated importance of testing for homogeneity of model fit across different temporal and cross-sectional groups. The current section indicates that the patterns of residual variability (reflected in the variances of the equation error terms q_{jt}^*) differ substantially between goods-producing and service-providing industries, and in some cases may be associated with the predictors $ln(x_{j0})$.

6. Discussion

Historically, survey organizations have developed generalized variance function models based on relatively broad concepts like commonality in design and population features. This article complements these previous approaches by using formal significance procedures to test for homogeneity of GVF coefficients across groups of estimators; by using regression diagnostics to evaluate the impact of adding particular predictors; and by using models for the variances of the equation errors in GVF models. The CES application presented in Sections 2 and 5 illustrated the main ideas of this article, with special emphasis on comparison of models across years (2005–2007 vs 2008–2010) and across industry groups (goods-producing vs service-providing).

One could consider several extensions of the ideas developed here. First, this article used the assumption that the rescaled variance estimators $V_{pjt}^{-1} d_{jt} \hat{V}_{pjt}$ followed a chi-square distribution on d_{jt} degrees of freedom with values of d_{jt} that were large (over 100). This was appropriate for the national-level analyses considered here. It would be of interest to extend the current work to state and local area analyses; for some of those analyses, the effective degrees of freedom d_{jt} may be relatively small. In addition, one could consider alternative approaches under which scaled forms of $V_{pjt}^{-1} \hat{V}_{pjt}$ followed a heavy-tailed distribution, for example, a contaminated chi-square or contaminated lognormal. These alternatives may be of special interest for cases in which the underlying data may be subject to outliers.

Second, one could consider versions of Models (2) and (3) that directly incorporate finite population corrections (fpc). This would be of interest primarily in applications for which some strata have substantial sampling fractions, and for which explicit inclusion of fpc terms may lead to substantial improvements in the GVF model fit. For cases in which

an estimator $\hat{\theta}_{jt}$ is based on data from a single stratum (as is the case for some types of domain estimation), explicit inclusion of a finite population correction term leads to an adjustment of the intercept terms in a logarithm scale fit of Model (3). For other cases, inclusion of a finite population correction leads to more complex adjustments that are beyond the scope of the current work.

Third, in keeping with the comments in Subsection 3.2, one could consider weighted least squares (WLS) or generalized least squares (GLS) point estimators of the coefficient vectors γ. These alternatives would be of special interest in cases for which ordinary least squares residual plots displayed patterns of heteroscedasticity that were more severe than the pattern in Figure 6 for the log-transformed fit. Under the alternative models described in the previous paragraph, it would be of special interest to explore conditions under which GLS estimators of γ are more efficient than the ordinary least squares estimators used in this article, to develop variance estimators for these GLS point estimators, and to evaluate properties of the GLS estimators under violation of the abovementioned model assumptions.

Fourth, the proposed parametric GVF model in this article assumes that the model is fully described by a very small set of parameters. However, for other applications, the relationships between sampling variances and predictor variables may follow patterns that require more complex models with a larger number of parameters. Semiparametric analysis may provide a flexible tool for studying the dependence of a variable of interest on auxiliary information, without constraining the dependence to a fixed form with few parameters. It would be of interest to extend our model to the semiparametric setting.

7. References

Binder, D.A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–292.

Butani, S., G. Stamas, and M. Brick. 1997. "Sample Redesign for the Current Employment Statistics Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 10–14, 1997. 517–522. Available at: http://www.amstat.org/sections/SRMS/Proceedings/papers/1997088.pdf (accessed May 2014).

Cho, M.J., J.L. Eltinge, J. Gershunskaya, and L. Huff. 2002. "Evaluation of the Predictive Precision of Generalized Variance Functions in the Analysis of Complex Survey Data." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 11–15, 2002. 534–539. Available at http://www.amstat.org/sections/SRMS/Proceedings/y2002/Files/JSM2002-000845.pdf (accessed May 2014).

Cho, M., J.L. Eltinge, J. Gershunskaya, and L. Huff. 2014. "Evaluation of Generalized Variance Functions in the Analysis of Complex Survey Data." *Journal of Official Statistics* 30: 63–90.

Cochran, W.G. 1977. *Sampling Techniques*, (Third Edition). New York: John Wiley and Sons.

Copeland, K.R., and R. Valliant. 2007. "Imputing for Late Reporting in the U.S. Current Employment Statistics Survey." *Journal of Official Statistics* 23: 69–90.

Draper, N.R., and H. Smith. 1998. *Applied Regression Analysis*, (Third Edition). New York: John Wiley and Sons.

Fuller, W.A. 1987. *Measurement Error Models*. New York: John Wiley and Sons.

Gershunskaya, J., and P. Lahiri. 2005. "Variance Estimation for Domains in the U.S. Current Employment Statistics Program." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 7–11, 2005. 3044–3051. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSM2005-000411.pdf (accessed May 2014).

Johnson, E.G., and B.F. King. 1987. "Generalized Variance Functions for a Complex Sample Survey." *Journal of Official Statistics* 3: 235–250.

Judkins, D.R. 1990. "Fay's Method for Variance Estimation." *Journal of Official Statistics* 6: 223–239.

Kish, L. 1995. "Methods for Design Effects." *Journal of Official Statistics* 11: 55–77.

Korn, E.L., and B.I. Graubard. 1990. "Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni *t* statistics." *The American Statistician* 44: 270–276.

O'Malley, A.J., and A.M. Zaslavsky. 2005. "Variance-Covariance Functions for Domain Means of Ordinal Survey Items." *Survey Methodology* 31: 169–182.

Park, I., and H. Lee. 2004. "Design Effects for Weighted Mean and Total Estimators Under Complex Survey Sampling." *Survey Methodology* 30: 183–193.

U.S. Bureau Of Labor Statistics. 2011. *BLS Handbook of Methods*. Washington, DC: U.S. Department of Labor. Available at: http://www.bls.gov/opub/hom/pdf/homch2.pdf (accessed May 2014).

Valliant, R. 1987. "Generalized Variance Functions in Stratified Two-Stage Sampling." *Journal of American Statistical Association* 82: 499–508.

Valliant, R., and K. Rust. 2010. "Degrees of Freedom Approximation and Rules-of-Thumb." *Journal of Official Statistics* 26: 585–602.

Werking, G. 1997. "Overview of the CES Redesign." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 10–14, 1997. 512–516. Available at: http://www.amstat.org/sections/SRMS/Proceedings/papers/1997 087.pdf (accessed May 2014).

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

Received December 2012
Revised May 2014
Accepted July 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 811–838, http://dx.doi.org/10.2478/JOS-2014-0049

Estimation of Mean Squared Error of X-11-ARIMA and Other Estimators of Time Series Components

Danny Pfeffermann[1] *and Michail Sverchkov*[2]

This article considers the familiar but very important problem of how to estimate the mean squared error (MSE) of seasonally adjusted and trend estimators produced by X-11-ARIMA or other decomposition methods. The MSE estimators are obtained by defining the unknown target components such as the trend and seasonal effects to be the hypothetical X-11 estimates of them that would be obtained if there were no sampling errors and the series were sufficiently long to allow the use of the symmetric filters embedded in the programme, which are time invariant. This definition of the component series conforms to the classical definition of the target parameters in design-based survey sampling theory, so that users should find it comfortable to adjust to this definition. The performance of the MSE estimators is assessed by a simulation study and by application to real series obtained from an establishment survey carried out by the Bureau of Labor Statistics in the U.S.A.

Key words: Bias correction; canonical decomposition; seasonal adjustment; state-space model; survey sampling; trend; X-13ARIMA-SEATS.

1. Introduction

In this article, we consider estimation of the mean squared error (MSE) of seasonally adjusted and trend estimators produced by X-11-ARIMA or other decomposition methods. In particular, we compare the MSE of estimators obtained by application of X-11-ARIMA with the MSE of estimators obtained by fitting state-space models that account for correlated sampling errors. We define the target seasonally adjusted and trend components to be the hypothetical X-11 estimates of those that would be obtained in the absence of sampling errors and if the time series under consideration was sufficiently long for application of the symmetric filters embedded in the original X-11 procedure, which are time invariant. This definition of the component series conforms to the classical definition of target finite population parameters in design-based survey sampling theory. In fact, in one variant of the proposed definition, the target components are shown to be linear combinations of finite

[1] Central Bureau of Statistics, Israel, Hebrew University of Jerusalem, Israel and University of Southampton, Southampton SO17 1BJ, UK. Email: msdanny@soton.ac.uk
[2] Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Suite 1950, Washington DC 20212, U.S.A. Email: Sverchkov.Michael@bls.gov
Acknowledgments: We are grateful to the Associate Editor and two reviewers for many excellent and constructive comments. We also thank Brian Monsell from the Census Bureau in the U.S.A. for writing a special module within X-13ARIMA-SEATS, which permits forecasting any number of signal components' values. The opinions expressed in this article are those of the authors and do not necessarily represent the policies of the Central Bureau of Statistics in Israel or the Bureau of Labor Statistics in the U.S.A.

population means or totals. The MSE of X-11-ARIMA and state-space model estimators are defined with respect to this definition.

We estimate the MSE by conditioning on the target components, thereby accounting for possible conditional bias in estimating them. The results are illustrated by use of simulated series and by application to real series obtained from an establishment survey carried out by the Bureau of Labor Statistics (BLS) in the U.S.A. The latter results also contrast the performance of our proposed MSE estimators with estimators proposed by Bell and Kramer (1999).

2. Target Components, Bias and MSE of X-11-ARIMA Estimators

2.1. Target Components

We begin with the usual notion that an economic time series, Y_t; $t = 1, 2, \ldots$ can be decomposed into a trend or trend-cycle component T_t, a seasonal component S_t, and an irregular component I_t; $Y_t = T_t + S_t + I_t$. Here we consider the additive decomposition, but the results of this article can be generalized to the multiplicative decomposition $Y_t = T_t \times S_t \times I_t$ by applying the log transformation and employing similar considerations to those in Pfeffermann et al. (1995). In practice, it is often the case that the series Y_t is unobserved and the available series consists of sample estimates, y_t, obtained from repeated sample surveys. Consequently, the series y_t can be expressed as the sum of the true population value, Y_t, and a sampling error, ε_t. More generally, the observed series can be viewed as the sum of a signal, G_t, and an error, e_t; $y_t = G_t + e_t$, where the signal, and hence the error, may be defined in two alternative ways:

*GE*1. $G_t = T_t + S_t$, $e_t = I_t + \varepsilon_t$. In this case e_t is the combined error of the time series irregular term and the sampling error (Pfeffermann 1994);

*GE*2. $G_t = T_t + S_t + I_t$, $e_t = \varepsilon_t$. In this case the irregular term is part of the signal, and e_t is the sampling error (Bell and Kramer 1999)

We assume without loss of generality that the series started at time $-\infty < t_{start} < 1$, but y_t is only observed for the time points $t = 1, \ldots, N$, such that

$$y_t = G_t + e_t, \quad t = \underbrace{t_{start}, \ldots, 0}_{unobserved}, \underbrace{1, \ldots, N}_{y_t\ observed}, \underbrace{N+1, \ldots, \infty}_{unobserved} \tag{1}$$

It is assumed also that under both definitions of the signal, e_t is independent of $\mathbf{G} = \{G_t,\ t = t_{start}, \ldots, \infty\}$ for all t, with $E(e_t) = 0$, $Var(e_t) < \infty$, although in practice the sampling error, and in particular the variance of the sampling error, sometimes depends on the magnitude of the signal.

The X-11-ARIMA program first forecasts and backcasts the time series under consideration based on an ARIMA model fitted to the observed series, and then applies a sequence of moving averages (linear filters) to the series augmented by the forecasts and backcasts. It follows that the X-11-ARIMA estimators of the trend and the seasonal components can be approximated as,

$$\hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k}, \quad \hat{S}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^S y_{t+k}, \tag{2}$$

where the coefficients $\{w_{kt}^T\}$, $\{w_{kt}^S\}$ are defined in general by the program options for the observed time interval $t = 1, \ldots, N$, by the ARIMA model used to forecast and backcast the series and by the number of backcasts and forecasts. However, at the central part of the series, the filters in (2) are time-invariant and symmetric; $w_{kt}^T = w_k^T$, $w_{-k}^T = w_k^T$ for $a_T < t \leq N - a_T$; $w_{kt}^S = w_k^S$, $w_{-k}^S = w_k^S$ for $a_S < t \leq N - a_S$, where a_T, a_S are defined by the X-11-ARIMA program options. The length of the symmetric filters is thus $2a_T + 1$ $(2a_S + 1)$. For example, for the default X-11-ARIMA options, $a_S = 84$, $a_T = 90$, but a_S, for example, may be as low as 70 or as high as 149 when using other options. Note that in the central part of the series the X-11-ARIMA estimators are the same as the X-11 estimators with no ARIMA extrapolations, such that the symmetric filters only depend on the X-11 program options and not on the ARIMA extrapolations.

Remark 1. The use of X-11-ARIMA also involves 'non-linear' operations such as the identification and estimation of ARIMA models used for forecasting and backcasting the original series, and the identification and gradual replacement of extreme observations. We assume that the time series under consideration is already modified for extreme values, thus robustifying the variance estimates described in Subsection 2.3. As illustrated in Pfeffermann et al. (1995) and Pfeffermann et al. (2000), the effects of the identification and nonlinear estimation of ARIMA models are generally minor.

Definition 1. *Assuming* $t_{start} < \min(-a_T, -a_S)$ *and following Bell and Kramer (1999), we define the trend component at time t to be* $T_t^{X11} = \sum_{k=-a_T}^{a_T} w_k^T G_{t+k}$. *Analogously, the seasonal component is defined as* $S_t^{X11} = \sum_{k=-a_S}^{a_S} w_k^S G_{t+k}$. *The target components* T_t^{X11} *and* S_t^{X11} *are thus the hypothetical components that would be obtained by application of the X-11 symmetric filters to the signal* $\boldsymbol{G}$ *at time point t,* $t = 1, \ldots, N$. *It follows therefore that the observed series may be decomposed as the sum of the 'X-11-trend',* T_t^{X11}, *the 'X-11-seasonal component',* S_t^{X11}, *and the 'X-11 error',* $e_t^{X11} = y_t - T_t^{X11} - S_t^{X11}$:

$$y_t = T_t^{X11} + S_t^{X11} + e_t^{X11}. \tag{3}$$

Result 1. For $a_T < t \leq N - a_T$, $T_t^{X11} = E(\hat{T}_t|\mathbf{G})$ and for $a_S < t \leq N - a_S$, $S_t^{X11} = E(\hat{S}_t|\mathbf{G})$, where $\hat{T}_t, \hat{S}_t$ are the X-11-ARIMA estimators defined in (2) and the expectation is taken over the distribution of the errors $\{e_t,\ t = 1, \ldots, N\}$, with the signal $\boldsymbol{G}$ held fixed. It follows therefore from our definition that in the central part of the series, the X-11-ARIMA estimators $\hat{T}_t, \hat{S}_t$ of the trend and the seasonal component are unbiased. (As noted before, we assume that the observed series is already modified for extreme values. The identification and estimation of ARIMA models are irrelevant at the center of the series.)

Remark 2. For X-11 filters $a_T > a_S$ because the final trend filter is applied after the final seasonal and seasonally adjusted components are computed. Thus, $\max(a_T, a_S) = a_T$.

Remark 3 We define the trend and seasonal components to be the (hypothetical) outputs that would be obtained when applying the symmetric filters to the signal, since the filters at the non-central parts of the series are asymmetric and depend on the time points with data.

In particular, the filters applied for a time point $t > N - a_T$ change every time that a new observation is added to the series until $t \leq N - a_T$, when the symmetric filter is applied. As mentioned before, the decomposition (3) has been used by Bell and Kramer (1999) with the error defined by the sampling error, such that the irregular term is part of the signal; $G_t = T_t + S_t + I_t$ (Definition GE2). See Subsection 2.5 for details of their approach. Note that with this definition, the target values are just linear combinations of the unadjusted population values of the series, which in most cases are finite population means or totals, in line with classical survey sampling theory.

2.2. Conditional Bias and MSE of X-11-ARIMA Estimators

The conditional bias, variance and MSE of the X-11-ARIMA estimator of the trend with respect to the decomposition (3), given the signal, are as follows:

$$Bias(\hat{T}_t|\mathbf{G}) = E\left[\left(\hat{T}_t - T_t^{X11}\right)|\mathbf{G}\right] = \sum_{k=-(t-1)}^{N-t} w_{kt}^T G_{t+k} - \sum_{k=-a_T}^{a_T} w_k^T G_{t+k}. \tag{4}$$

$$\begin{aligned} Var[\hat{T}_t|\mathbf{G}] &= E\left\{\left[\sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k} - E\left(\sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k}\middle|\mathbf{G}\right)\right]^2\middle|\mathbf{G}\right\} \\ &= E\left\{\left[\sum_{k=-(t-1)}^{N-t} w_{kt}^T (y_{t+k} - G_{t+k})\right]^2\middle|\mathbf{G}\right\} = E\left(\sum_{k=-(t-1)}^{N-t} w_{kt}^T e_{t+k}\right)^2 \end{aligned} \tag{5}$$

$$MSE(\hat{T}_t|\mathbf{G}) = E\left[\left(\hat{T}_t - T_t^{X11}\right)^2\middle|\mathbf{G}\right] = Var(\hat{T}_t|\mathbf{G}) + Bias^2(\hat{T}_t|\mathbf{G}). \tag{6}$$

Similar expressions hold for the seasonal and seasonally adjusted estimators.

Expressions (4)–(6) are general and apply to any linear estimator with arbitrary coefficients $\{w_{kt}^T\}$, as defined by the X-11-ARIMA options, the ARIMA model used for extrapolations and the number of forecasts and backcasts. In fact, as will be shown in Section 3, the Expressions (4)–(6) hold equally for other linear filters, not necessarily embedded in the X-11-ARIMA program. In the following sections we discuss ways of estimating the MSE in (6).

2.3. Variance Estimation

Under Definition GE2 of the signal and error in Subsection 2.1, $e_t = \varepsilon_t$ is the sampling error, and by (5),

$$Var(\hat{T}_t|\mathbf{G}) = E\left(\sum_{k=-(t-1)}^{N-t} w_{kt}^T \varepsilon_{t+k}\right)^2 = \sum_k \sum_l w_{kt}^T w_{lt}^T Cov(\varepsilon_{t+k}, \varepsilon_{t+l}).$$

Similar expressions apply when estimating the seasonal or the seasonally adjusted component. We assume the availability of estimates of the variances and covariances of the sampling errors, which enables estimation of the variance $Var(\hat{T}_t|\mathbf{G})$ and the variance of any other component estimator.

Next, consider the estimation of the variance under Definition GE1 of the signal and error, by which $e_t = I_t + \varepsilon_t$. By (5), the variance of the X-11-ARIMA trend estimator is in this case a linear combination of the covariances $v_{tm} = Cov(e_t, e_m),\ \ t, m = 1, \ldots, N$. Following Pfeffermann (1994) and Pfeffermann and Scott (1997), let $R_t = y_t - \hat{S}_t - \hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^R y_{t+k}$ define the linear approximation of the X-11-ARIMA residual term at time t, where $w_{0t}^R = 1 - w_{0t}^S - w_{0t}^T$ and $w_{kt}^R = -w_{kt}^S - w_{kt}^T$ for $k \neq 0$. Then,

$$Var(R_t|\mathbf{G}) = E\left\{\left[\sum_{k=-(t-1)}^{N-t} w_{kt}^R\left(y_{t+k} - E(y_{t+k}|\mathbf{G})\right)\right]^2 \Big| \mathbf{G}\right\} = Var\left(\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k}\right),$$

$$Cov(R_t, R_m|\mathbf{G}) = Cov\left[\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k}, \sum_{l=-(m-1)}^{N-m} w_{lm}^R e_{m+l}\right] = \sum_k \sum_l w_{kt}^R w_{lm}^R Cov(e_{t+k}, e_{m+l}). \tag{7}$$

The residuals R_t are not stationary because of the use of asymmetric filters towards the two ends of the series. However, let $U(m) = \frac{1}{N-m}\sum_{t=1}^{N-m} Cov(R_t R_{t-m}),\ m = 0, \ldots, N-1$, and suppose that the errors $e_t = I_t + \varepsilon_t$ are stationary (see Remark 4 below). Then, by (7), the vector $\mathbf{U}$ of the means $U(m)$ and the vector $\mathbf{V}$ of the covariances $V_k = Cov(e_t, e_{t+k}) = Cov(I_t + \varepsilon_t, I_{t+k} + \varepsilon_{t+k}),\ k = 0, \ldots, N-1$ are related by the system of linear equations,

$$\mathbf{U} = D\mathbf{V}, \tag{8}$$

where the matrix D is defined by the known weights $\{w_{kt}^R\}$. Since the X-11-ARIMA residuals are known for every $t = 1, \ldots, N$, one may estimate $U(m)$ by $\tilde{U}(m) = \frac{1}{N-m}\sum_{t=1}^{N-m} R_t R_{t-m}$. Substituting $\tilde{U}(m)$ for $U(m)$ in (8) enables estimation of $\mathbf{V}$ by solving the resulting equations; see Pfeffermann (1994) and Pfeffermann and Scott (1997). Note that the use of (8) does not require the availability of estimates of the variances and covariances of the sampling errors. However, the estimators obtained in this way can be very unstable since the number of unknown variances and covariances generally equals the number of equations. A possible solution to this problem is to assume that the covariances V_k are negligible beyond some lag C and set them to zero, and then solve the reduced set of equations for $V_0, \ldots, V_C$. This is a mild ergodicity condition assumed for the series e_t. Note that with this assumption it is no longer necessary to consider the estimates $\tilde{U}(m)$ for large m. Additionally, when estimates for the autocovariances of the sampling errors are available, they can be substituted into the vector $\mathbf{V}$ and taken as known, in which case one only needs to estimate the unknown variance and covariances of the time series irregular terms, I_t. This reduces the number of unknown covariances and hence the number of equations very drastically. Note that all these procedures are basically 'model free'. See Chen et al. (2003) for a different approach to estimating $\mathbf{U}$ and $\mathbf{V}$. Bell and Kramer (1999) consider model-based estimation of the variance and covariances of the sampling errors.

Remark 4. The linear equations in (8) can easily be extended to the case of heteroscedastic sampling errors for which $V_{tk} = Cov(e_t, e_{t+k}) = L_{tk} V_k$ with known

coefficients L_{tk}. Another potential modification consists of utilizing all (or most of) the equations in (8), and estimating $V_0, \ldots, V_C$ by a discounted least-squares procedure.

2.4. Bias and MSE Estimation

Estimation of the conditional bias of the estimator $\hat{T}_t$ (or any other linear estimator) given the signal, and hence the conditional MSE is more involved. We propose to estimate the bias by estimating the signal and then substituting the estimate in the right hand side of the bias expression (4). A possible way of estimating the signal is by application of the programme X-13ARIMA-SEATS (X-13A-S Reference Manual, Version 0.1 2013). This program is now in common use in many statistical bureaus around the world (replacing X-12-ARIMA). The programme enables to extract the models holding for the trend and the seasonal effects from the ARIMA model fitted to the observed series, and use these models in order to estimate the signal within the observation period, and forecast and backcast the signal for a_T time points with no observations. Denote by $\hat{G}_t$ the estimated signal for time t, including before or after times $1, \ldots, N$. The bias is estimated as,

$$\hat{Bias}[\hat{T}_t|\mathbf{G}] = \hat{E}\left[\left(\hat{T}_t - T_t^{X11}\right)|\mathbf{G}\right] = \sum_{k=-(t-1)}^{N-t} w_{kt}^T \hat{G}_{t+k} - \sum_{k=-a_T}^{a_T} w_k^T \hat{G}_{t+k}, \tag{9}$$

$$t = 1, \ldots, N.$$

Use a similar expression for estimating the bias of the seasonally adjusted estimator.

The SEATS models are obtained by application of canonical signal extraction and under correct model specification, the estimators have Minimum MSE (MMSE) (Hilmer and Tiao, 1982).

Remark 5. Wecker (1979) noted that the MMSE signal estimator has a different spectrum from the true signal and proposed another estimator which preserves the spectrum of the true signal. Application of this proposal requires external information and the loss in MSE compared to the use of the MMSE estimator can be large. We do not consider this estimator in the present article.

A simpler way of estimating the signal, which can be implemented by application of the X-11-ARIMA programme (or within X-13ARIMA-SEATS), consists of the following two steps:

(*a*) Use the ARIMA model fitted by X-11-ARIMA to the original series to augment the series with a_T forecasts and backcasts;
(*b*) Estimate the signal of the augmented series as,

$$\hat{G}_t = \sum_{k=-(t^1 a_T-1)}^{N+a_T-t} w_{kt}^{G,aug} y_{t+k}^{aug}, \quad t = -a_T + 1, \ldots, N + a_T. \tag{10}$$

where $y_t^{aug} = y_t$ if y_t is observed, y_t^{aug} is the forecasted (backcasted) value if y_t is not observed and $w_{kt}^{G,aug} = w_{kt}^{T,aug} + w_{kt}^{S,aug}$, with $w_{kt}^{T,aug}, w_{kt}^{S,aug}$ defining the X-11 weights for

the longer (augmented) series. Substituting the estimated signal (10) into (4) yields the trend bias estimator, similarly to (9). Similar expressions hold for the seasonal and seasonally adjusted estimators.

Remark 6. The difference between the two methods of estimating and predicting the signal described above lies in the linear filters applied to the original series. The first method uses the optimal filters for extracting the trend and seasonal component under the ARIMA model fitted to the series. The second method uses the ARIMA model for backcasting and forecasting the original series (for given model coefficients, the backcasts and forecasts are linear combinations of the original series), and then uses the original X-11 filters for estimating the trend and seasonals of the augmented series. The resulting filters generally differ from the optimal filters.

Having estimated the conditional variance and bias, a conservative estimator of the conditional MSE defined by (6) is obtained by adding the variance estimator to the square of the bias, i.e.,

$$\hat{MSE}(\hat{T}_t|\mathbf{G}) = \hat{Var}(\hat{T}_t|\mathbf{G}) + \hat{Bias}^2(\hat{T}_t|\mathbf{G}). \tag{11}$$

The estimator in (11) is conservative since $E[\hat{Bias}^2(\hat{T}_t|\mathbf{G})|\mathbf{G}] = \{E[\hat{Bias}(\hat{T}_t|\mathbf{G})]|\mathbf{G}\}^2 + Var[\hat{Bias}(\hat{T}_t|\mathbf{G})|\mathbf{G}] > \{E[\hat{Bias}(\hat{T}_t|\mathbf{G})]|\mathbf{G}\}^2$. The overestimation of the MSE can be corrected by subtracting an estimate of $Var[\hat{Bias}(\hat{T}_t|\mathbf{G})|\mathbf{G}]$. Note that $\hat{Bias}(\hat{T}_t|\mathbf{G})$ is a linear combination of the signal estimates, $\hat{G}_t$, which in turn are linear combinations of the observed series, y_t. Thus, $\hat{Bias}(\hat{T}_t|\mathbf{G})$ is a linear combination of the y_t's and hence $Var[\hat{Bias}(\hat{T}_t|\mathbf{G})|\mathbf{G}]$ can be estimated similarly to the estimation of $Var[\hat{T}_t|\mathbf{G}]$ discussed in Subsection 2.3. The weights defining $\hat{Bias}(\hat{T}_t|\mathbf{G})$ can be obtained similarly to Burck and Sverchkov (2001) and Findley and Martin (2006) (See Section 3).

Remark 7. The procedure proposed for estimating the bias and MSE of the X-11-ARIMA estimators raises two valid questions:

i) The predictors of the signal many years ahead, required for estimating the MSE of the estimators of the component series may be severely biased for time points far away from the last time point N with an observation, because of possible changes in the behavior of the signal over time. So how can one rely on these predictors? To answer this question, note first that even if the signal predictors are biased (given the true signal), the trend bias estimator in (9) may still be unbiased or only have a small bias. For example, if $E[(\hat{G}_t - G_t)|\mathbf{G}] = constant$ for all t, the bias estimator (9) is unbiased for the true bias since $\sum_{k=-(t-1)}^{N-t} w_{kt}^T = \sum_{k=-a_T}^{a_T} w_k^T = 1$. The same holds when estimating the bias of the seasonally adjusted estimators (SAE). While this may not be a realistic scenario, what is more important is that the weights of the symmetric filters, used to predict the trend and the SAE decay to zero very fast when moving away from the time point of interest, so that even large biases of the predictors of the signal for time points far away from the last time with an observation may have little effect on the bias of the estimator of the bias of the trend or the SAE. Figure 1 shows the central weights of the trend filters used in our simulation study described in Section 4. The plot of the Basic Structural Model (BSM) filter weights looks like

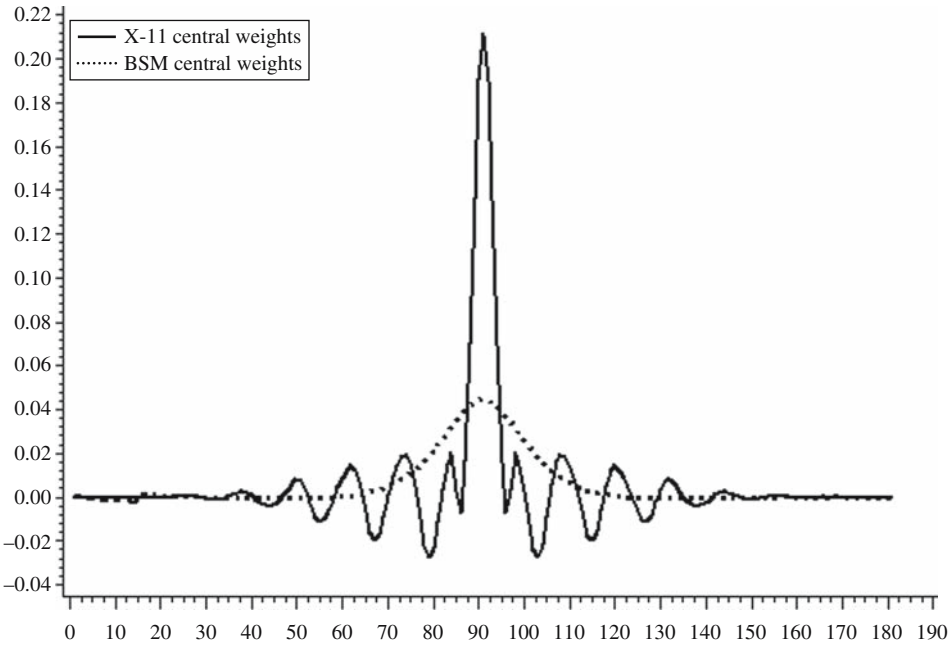


Fig. 1. Central weights applied to the signal for predicting the trend under X-11 and under the basic structural model (BSM, Section 3)

a trend filter for a nonseasonal series because there seems to be no seasonal pattern to the weights, which is counterintuitive. We have no explanation to this behavior of the weights, but all our checks show that they are correct.

ii) If we believe that we have good predictors of the signal and hence good estimators of our target trend, why not use these estimates in the first place instead of using the X-11-ARIMA estimates of the trend? The answer to this question is simple. Our aim in this article is not to propose new trend or seasonally adjusted estimators. In fact, the model-based predictors of the trend and seasonal component that we use to estimate the bias are produced by one of the modules of X-13ARIMA-SEATS, following the pioneering work of Gómez and Maravall (1996). Rather, our aim is to develop a method of estimating the conditional MSE of linear estimators such as the X-11-ARIMA estimators, which are in common use. We may refer to our method of bias estimation as 'model-based'.

Remark 8. When the signal is estimated by the MMSE estimator under the models extracted for the trend and the seasonal component, the estimator of the signal coincides with the conditional expectation of the signal, given the observed series. In this case the bias estimator (9) is the conditional expectation of the bias over all possible realizations of the signal given the observed series.

2.5. *Comparison With the Method of Bell and Kramer*

As noted before, Bell and Kramer (1999) use a similar definition of the target components. The authors estimate these components by augmenting the observed series

with a_T MMSE forecasts and backcasts under an appropriate ARIMA model, such that the symmetric X-11 filters can be applied to the augmented series at every time point t with an observation. The trend estimator, for example, can be written then as $\hat{T}_t^{BK} = \sum_{k=-a_T}^{a_T} w_k^T y_{t+k}^*$, where $y_{t+k}^* = y_{t+k}$ if y_{t+k} is observed ($1 \leq t+k \leq N$), and y_{t+k}^* is the forecasted or backcasted value otherwise. The authors focus on $Var(\hat{T}_t^{BK} - T_t^{X11})$ under the GE2 definition of the signal by which the irregular term is part of the signal, so that the variance is taken over the distributions of the sampling errors and the forecast and backcast prediction errors. Notice that since $E(y_{t+k}^* - y_{t+k}) = 0$ under the model (unconditional on $\boldsymbol{G}$),

$$E\left(\hat{T}^{BK} - T_t^{X11}\right) = E\left[\sum_{k=-a_T}^{a_T} w_k^T y_{t+k}^* - \sum_{k=-a_T}^{a_T} w_k^T G_{t+k}\right]$$

$$= E\left[\sum_{k=-a_T}^{a_T} w_k^T y_{t+k}^* - \sum_{k=-a_T}^{a_T} w_k^T y_{t+k}\right] = 0, \tag{12}$$

such that the estimators of the trend are likewise unbiased unconditionally. However, when conditioning on the signal, in general $E\left[\left(\hat{T}_t^{BK} - T_t^{X11}\right)\middle|\mathbf{G}\right] \neq 0$. As is evident from (4), a bias may also exist even unconditionally when forecasting and backcasting less than a_T observations, depending on the distribution of the signal.

Our approach differs from Bell and Kramer (1999) in three main aspects.

I- Our definition of the MSE and its estimation is not restricted to the case of full forecasts and backcasts, and it can be applied for any linear estimator of the form $\tilde{H}_t = \sum_{k=-(t-1)}^{N-t} h_{kt} y_{t+k}$. In particular, it applies to the case where the seasonally adjusted and trend components are estimated by use of X-11-ARIMA with only one or two years of forecasts and backcasts, the common case in practice, or even without ARIMA extrapolations. It also applies when estimating the components by signal extraction under appropriate ARIMA models as in X-13ARIMA-SEATS, or by fitting a state-space model to the series as in Sections 3 and 4.

II- We attempt to estimate the conditional MSE given the signal, even though the signal is not observed. We believe that many users of seasonally adjusted and trend estimators would feel most comfortable with the notion that the corresponding target components are fixed values, which conforms to classical sampling theory under which the target parameters are functions of the population values, which are viewed as fixed, nonstochastic quantities. In fact, under definition *GE*2 of the signal, the target component values are just linear combinations of the unadjusted population values defining the series, which in most cases are finite population means or totals. On the other hand, as already stated in Remark 8, our bias estimators may also be viewed as estimating the unconditional bias over all possible realizations of the signal under an appropriate model, given the observed series.

III- Our approach is applicable to the case where the signal consists of only the trend and the seasonal effect, and the time series irregular is part of the error (definition *GE*1 of the signal). We mention also that in its present state, the application of Bell and Kramer's procedure is not straightforward and requires many intermediate steps. See Bell and Kramer (1999) and Scott et al. (2012) for details.

3. Estimation of MSE of Model-Based and Other Estimators of X-11 Components

Consider any other set of component estimators of the form

$$\tilde{T}_t = \sum_{k=-(t-1)}^{N-t} h_{kt}^T y_{t+k}, \quad \tilde{S}_t = \sum_{k=-(t-1)}^{N-t} h_{kt}^S y_{t+k}. \tag{13}$$

Then, similarly to the X-11-ARIMA estimators in Section 2, we can calculate the conditional bias and MSE with respect to the target X-11 components defined in Definition 1, yielding the same expressions as in (4)–(6) but with the weights $w_{kt}^T(w_{kt}^S)$ replaced by the weights $h_{kt}^T(h_{kt}^S)$. Note that unlike the X-11 estimators, the estimators defined by (13) are potentially biased when conditioning on the signal, even at the center of the series.

In the present article, besides X-11-ARIMA estimators, we also consider estimators obtained by fitting a simple state-space model (see Subsection 4.1). The state-space model estimates of the seasonal component and the trend for a given time t are again linear combinations of all the observed values. We calculated the weights defining the corresponding filters by using the impulse response method (Findley and Martin 2006). According to this method, the weight of an observation at time τ when applying the filter at time t is computed by applying the model fitted to the observed series to a series composed of 1 at time τ and 0 elsewhere, and then observing the filter value for time t, $t = 1, \ldots, N$. Calculation of the weights for all time points of a series of length N therefore requires running the model N times, each time with a vector observation defined by a different column of the identity matrix I_N. As in Subsection 2.4, in this case the bias is estimated by estimating the augmented signal $\mathbf{G}^{aug} = (G_{-a_T+1}, \ldots, G_0, \ldots, G_N, \ldots, G_{N+a_T})$ under an appropriate model. The bias and MSE estimators are obtained similarly to Eqs. (9)–(11).

In a recent article, Tiller (2012) suggested another approach to trend estimation which consists of applying time series model-based signal extraction to estimate and remove the sampling errors from the original series, and then applying the X-11-ARIMA trend filter to the adjusted series. Under definition *GE*2 of the signal, the use of this approach reduces to applying the trend filter to the estimated signal under the model. Note that since the estimated signal is a linear filter and the X-11-ARIMA trend filter is linear as well, the trend estimators obtained under this approach are again linear combinations of all the observed values and we may apply our proposed approach to estimate the bias and MSE of the trend estimators obtained this way.

Remark 9. We have not considered Tiller's (2012) approach in the simulation study described below, but Pfeffermann et al. (1998) applied this approach to Labour Force series in Australia and found that the resulting trend estimators were very similar to the trend estimators obtained directly under the model.

4. Simulation Study

In this section we apply the estimators considered in Sections 2 and 3 to simulated series, generated from a state-space model fitted by the Bureau of Labour Statistics (BLS) in the

U.S.A. to the series *Employment to Population Ratio in the District of Columbia*, abbreviated hereafter by EP-DC. The EP series is obtained from the Current Population Survey (the US Labour Force Survey) and it estimates the percentage of employed persons out of the total population aged 15+. This is one of the key economic series in the U.S.A., produced monthly by the BLS for each of the 50 States and DC. The BLS uses a similar model for the production of the major employment and unemployment estimates in all the states of the U.S.A.; see Tiller (1992) for details. In order to assess the performance of the various estimators, we generated a large number of series from the EP-DC model. The model depends on 18 estimated hyperparameters but for the present experiment we consider the hyperparameter estimates as true known parameters.

4.1. Model Fitted to EP-DC Series

The EP-DC series is very erratic: The residual component (calculated by X-11-ARIMA) explains 55% of the month to month changes and 32% of the yearly changes. A large portion of the residual component is sampling errors. The series is plotted in Figure 2, along with the trend estimated under the EP-DC model defined below, and the trend estimated by X-11-ARIMA with twelve months forecasts when fitting the familiar airline model ARIMA(0,1,1)(0,1,1), selected by the program. The two trends behave similarly, but the trend estimated under the EP-DC model is smoother, a phenomenon observed in many other series. The X-11-ARIMA trend is below the EP-DC trend for most of the time points, but the average values of the two trends are very close: Av.(trend EP-DC) = 63.11, Av.(trend X11 ARIMA) = 63.01, Av.(original series) = 63.00.

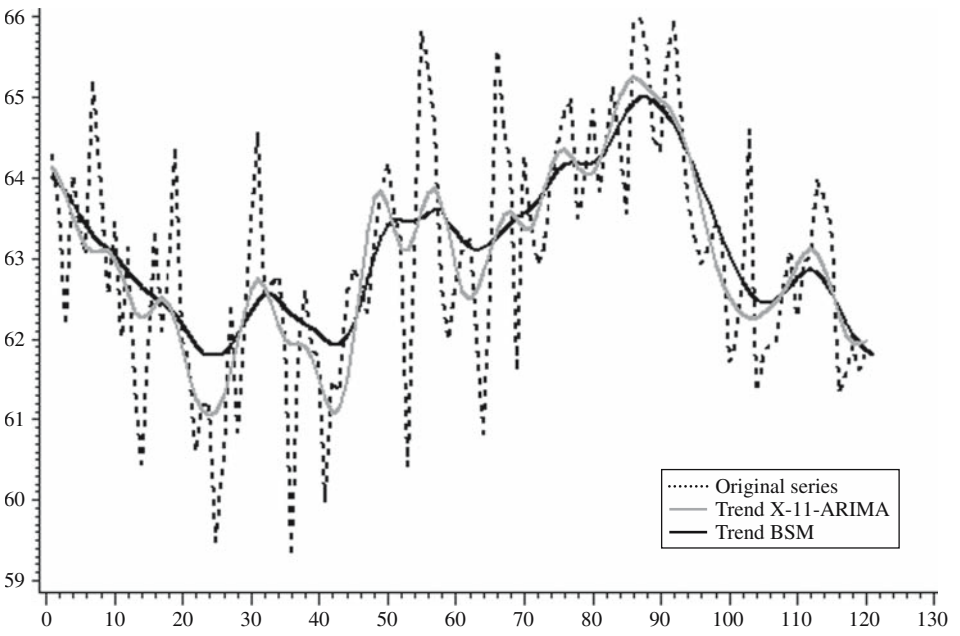


Fig. 2. Employment to Population Ratio in DC (in percentages), Jan2001-Dec2010. Original series and trends estimated by X-11-ARIMA with 12 forecasts, and by the EP-DC model

Let y_t define, the direct sample estimate for time t and Y_t the corresponding true population ratio such that $\varepsilon_t = y_t - Y_t$ is the sampling error. The state-space model fitted to the series y_t combines a model for Y_t with a model for ε_t. The model postulated for Y_t is the basic structural model (BSM, Harvey 1989)

$$\begin{aligned}
&Y_t = T_t + S_t + I_t, \quad I_t \sim N\left(0, \sigma_I^2\right); \quad T_t = T_{t-1} + R_{t-1}, \quad R_t = R_{t-1} + \eta_{Rt}, \\
&\quad \eta_{Rt} \sim N\left(0, \sigma_R^2\right) \\
&S_{j,t} = \cos \omega_j S_{j,t-1} + \sin \omega_j S^*_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim N(0, \sigma_S^2) \\
&S^*_{j,t} = -\sin \omega_j S_{j,t-1} + \cos \omega_j S^*_{j,t-1} + \eta^*_{j,t}, \quad \eta^*_{j,t} \sim N(0, \sigma_S^2), \\
&S_t = \sum\nolimits_{j=1}^{6} S_{j,t}; \quad \omega_j = 2\pi j/12, \quad j = 1. \, . \, .6.
\end{aligned} \tag{14}$$

The error terms $I_t, \eta_{Rt}, \eta_{j,t}, \eta^*_{j,t}$ are mutually independent normal disturbances. In this model, T_t is the trend level, R_t is the slope and S_t is the seasonal effect. The trend model approximates a local linear trend, whereas the model for the seasonal effects uses the traditional decomposition of the seasonal component into eleven cyclical components corresponding to the six seasonal frequencies. The innovations $\eta_{j,t}, \eta^*_{j,t}$ allow the seasonal effects to evolve over time.

The model fitted for the sampling errors is $AR(15)$; see Pfeffermann and Tiller (2005) for the considerations leading to the choice of this model.

The separate models holding for the population ratios and the sampling errors are cast into a single state-space model. In what follows we refer to the combined model holding for the observed series $y_t = Y_t + \varepsilon_t$ as the extended BSM (EBSM). Note that the state vector consists of the trend, slope, seasonal effects and sampling errors. The variances and covariances of the sampling errors are estimated from the survey micro-data using a replication approach with a large number of replications. The $AR(15)$ model coefficients are then estimated by solving the corresponding Yule-Walker equations and they are set to their estimated values when estimating the population model variances by maximum likelihood. The variances and AR coefficients used for the present simulation experiment are the same as in Pfeffermann and Tiller (2005). See that paper for further details on the way we generated series under the model and for the values of the model variances and AR coefficients.

4.2. *Simulation Plan*

We generated three sets of 1,000 monthly series of length 300; $\{y_{t,b},\ t = 1, \ldots, 300,\ b = 1, \ldots, 1,000\}$. The first set was generated by simulating for every month t a trend value, T_t, a seasonal effect, S_t, and an irregular term, I_t, from the model (14), and a sampling error, ε_t, from the AR(15) model, and then adding the separate components; $y_{t,b} = T_{t,b} + S_{t,b} + I_{t,b} + \varepsilon_{t,b},\ b = 1, \ldots, 1,000$. The second set was obtained from the first set by halving the sampling errors, that is, $y_{t,b} = T_{t,b} + S_{t,b} + I_{t,b} + \varepsilon_{t,b}/2$. The third set was obtained from the first set by doubling the sampling errors, i.e., $y_{t,b} = T_{t,b} + S_{t,b} + I_{t,b} + 2\varepsilon_{t,b}$. Considering the three data sets allows an assessment of the effect of the magnitude of the sampling errors on the performance of the estimators.

For the present study we employ the definition *GE*2 of the signal by which $G_{t,b} = T_{t,b} + S_{t,b} + I_{t,b}$. We computed the default X-11 estimator of the trend and seasonal component for each simulated signal of length 300, so as to obtain the target X-11 components defined by (3) for the central 180 months. (For the default X-11 estimator $a_S = 84,\ a_T = 90$, but augmenting the series with only 60 forecasts and backcasts yields almost identical target components.) We defined the target seasonally adjusted component as the difference between the original series without sampling error and the target seasonal component, that is,

$$T_t^{X11} = \sum_{k=-a_T}^{a_T} w_k^T G_{t+k}; \quad S_t^{X11} = \sum_{k=-a_S}^{a_S} w_k^S G_{t+k}; \quad SA_t^{X11} = (y_t - S_t^{X11} - \varepsilon_t). \tag{15}$$

Finally, we removed the first and last 60 monthly observations from the simulated series and applied X-11-ARIMA with twelve and 60 forecasts to the reduced series of length 180, using the default X-11 filters but setting the ARIMA model as the airline model ARIMA(0,1,1),(0,1,1) (Remark 10 below). Thus, the X-11-ARIMA estimators are

$$\hat{T}_{t,b}^{X11} = \sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k,b}, \ \ \hat{SA}_{t,b}^{X11} = y_{t,b} - \sum_{k=-(t-1)}^{N-t} w_{kt}^S y_{t+k,b}, \tag{16}$$

where the weights $\{w_{kt}^T\}, \{w_{kt}^S\}$ are defined by the ARIMA model, the program default options and the number of forecasts (12 or 60). We also computed the EBSM estimators, obtained by replacing the weights $\{w_{kt}^T\}, \{w_{kt}^S\}$in (16) by the weights $\{h_{kt}^T\}, \{h_{kt}^S\}$ (Section 3).

Remark 10. In the simulation study we did not select new ARIMA models or re-estimated the model coefficients for every simulated series. We used the airline model for all the simulated series and estimated the model parameters once for each set of series, based on a randomly selected series from the set. Selecting a model and re-estimating the model coefficients for each simulated series would require new computation of the filter weights for every series and every model, which is not feasible in a simulation study with 3,000 series. See Section 3 for the method used for estimating the filter weights. Notwithstanding, X-13ARIMA-SEATS selected the airline model as the preferred model for most of the series in all three sets. Moreover, for series of length 180 (quite typical for monthly economic series), the estimation of the model coefficients is generally very stable and is not expected to affect the results. (The sampling error variances and hence the AR(15) model coefficients are taken as known. The model variances are estimated by MLE, which are known to be consistent.) We also reiterate the statement made in Remark 3 above that our purpose in this article is to propose a method of estimating the conditional bias and RMSE of linear estimators of the proposed target components, and not to search for the most appropriate model and estimators. In practice, one would let the program select the model and estimate the unknown coefficients, and then compute the required filter weights for the particular choice of model and estimates.

4.3. Computations

Because of space limitations, in subsequent subsections we restrict ourselves to the estimation of the target trend. Estimation of the MSE of seasonally adjusted estimators is considered in Section 5. We computed the following statistics:

4.3.1. Conditional Variance of X-11-ARIMA and EBSM Estimators

$$V_t^{T,X11} = \sum_k \sum_l w_{kt}^T w_{lt}^T Cov(\varepsilon_{t+k}, \varepsilon_{t+l}), \quad V_t^{T,EBSM} = \sum_k \sum_l h_{kt}^T h_{lt}^T Cov(\varepsilon_{t+k}, \varepsilon_{t+l}). \quad (17)$$

The variances and covariances of the sampling errors are taken as known.

4.3.2. Conditional Bias and Root MSE of X-11-ARIMA and EBSM Estimators

The conditional bias and root MSE (RMSE) for a given signal when estimating the target trend in Eq. 3 are:

$$\begin{aligned} B_{t,b}^{T,X11} &= \sum_{k=-(t-1)}^{N-t} w_{kt}^T G_{t+k,b} - \sum_{k=-a_T}^{a_T} w_k^T G_{t+k,b}, \\ RMSE(\hat{T}_{t,b}^{X11}) &= [V_t^{T,X11} + (B_{t,b}^{T,X11})^2]^{1/2}. \end{aligned} \quad (18)$$

The bias and RMSE of the EBSM estimators are obtained in similar manner.

4.3.3. Estimation of Squared Bias and MSE

Denote by $\hat{B}_{t,b}^{T,X11}$ the estimate of the bias, obtained by predicting the unknown signal using the models extracted for the trend and seasonal effects by X-13ARIMA-SEATS (Eq. 9), or by predicting the signal using the X-11-ARIMA forecasts and backcasts (Eq. 10).

The RMSE is estimated as,

$$\begin{aligned} \hat{MSE}_{t,b}^{T,X11} &= V_t^{T,X11} + \left(\hat{B}_{t,b}^{T,X11}\right)^2 - V\left(\hat{B}_{t,b}^{T,X11}\right), \\ \hat{MSE}_{t,b}^{T,EBSM} &= V_t^{T,EBSM} + \left(\hat{B}_{t,b}^{T,EBSM}\right)^2 - V\left(\hat{B}_{t,b}^{T,EBSM}\right). \end{aligned} \quad (19)$$

where $V\left(\hat{B}_{t,b}^{T,X11}\right)$, $V\left(\hat{B}_{t,b}^{T,EBSM}\right)$ are the variances of the bias estimates, computed similarly to the variances of the estimators in Eq. 17. As explained in Subsection 2.4, subtracting the variance of the bias estimator is necessary for unbiased MSE estimation.

4.3.4. Error of Estimators of Squared Bias and RMSE

$$\begin{aligned} EBS_{t,b}^{B,X11} &= (B_{t,b}^{T,X11})^2 - [(\hat{B}_{t,b}^{T,X11})^2 - V(\hat{B}_{t,b}^{T,X11})], \\ EM_{t,b}^{RMSE,X11} &= RMSE(\hat{T}_{t,b}^{X11}) - \sqrt{\hat{MSE}_{t,b}^{T,X11}}. \end{aligned} \quad (20)$$

Similar expressions apply when using the EBSM estimators.

4.4. Results

The results are summarized in Tables 1–3 and Figures 3–8. The tables show average results obtained for the three sets of series, for each of the last six months of the reduced series (time

Table 1. Means of true squared bias and RMSE, simulation means of error when estimating the squared bias and RMSE, and SD of simulation means of error as obtained by application of X-11-ARIMA and by EBSM. First set of 1,000 series, last six months of series

	Month	Jul	Aug	Sep	Oct	Nov	Dec
X-11 ARIMA	ABS	0.020	0.021	0.021	0.035	0.113	0.332
60 forecasts	AEBS(9)	0.002	0.002	0.004	0.017	0.084	0.273
	SDEBS(9)	(0.042)	(0.043)	(0.044)	(0.062)	(0.177)	(0.508)
	AEBS(10)	−0.009	−0.012	0.009	0.000	0.012	0.022
	SDEBS(10)	(0.326)	(0.333)	(0.333)	(0.357)	(0.508)	(0.831)
	ARMSE	1.268	1.267	1.265	1.285	1.347	1.457
	AEM(9)	0.001	0.001	0.001	0.006	0.030	0.089
	SDEM(9)	(0.016)	(0.017)	(0.017)	(0.024)	(0.062)	(0.154)
	AEM(10)	−0.026	−0.026	−0.026	−0.023	−0.008	0.019
	SDEM(10)	(0.046)	(0.048)	(0.049)	(0.055)	(0.090)	(0.187)
X-11 ARIMA	ABS	0.024	0.025	0.026	0.041	0.120	0.343
12 forecasts	AEBS(9)	0.018	0.020	0.021	0.035	0.108	0.036
	SDEBS(9)	(0.035)	(0.034)	(0.036)	(0.060)	(0.182)	(0.514)
	AEBS(10)	−0.003	−0.001	0.004	0.012	0.019	0.022
	SDEBS(10)	(0.254)	(0.247)	(0.241)	(0.276)	(0.445)	(0.774)
	ARMSE	1.268	1.274	1.274	1.291	1.352	1.463
	AEM(9)	0.007	0.008	0.008	0.013	0.039	0.100
	SDEM(9)	(0.014)	(0.013)	(0.014)	(0.023)	(0.063)	(0.154)
	AEM(10)	−0.008	−0.006	−0.004	−0.001	0.014	0.041
	SDEM(10)	(0.027)	(0.025)	(0.023)	(0.032)	(0.075)	(0.176)
EBSM	ABS	0.240	0.221	0.209	0.221	0.287	0.441
	AEBS(9)	0.010	0.022	0.018	0.001	−0.006	0.044
	SDEBS(9)	(0.396)	(0.355)	(0.328)	(0.358)	(0.525)	(0.845)
	AEBS(10)	−0.004	0.038	0.078	0.101	0.097	0.061
	SDEBS(10)	(0.570)	(0.510)	(0.493)	(0.541)	(0.660)	(0.852)
	ARMSE	1.124	1.149	1.185	1.230	1.286	1.367
	AEM(9)	0.005	0.009	0.007	0.001	−0.001	0.016
	SDEM(9)	(0.155)	(0.137)	(0.125)	(0.132)	(0.179)	(0.265)
	AEM(10)	−0.067	−0.047	−0.028	−0.024	−0.036	−0.041
	SDEM(10)	(0.207)	(0.182)	(0.172)	(0.189)	(0.235)	(0.312)

points 175–180). As stated before, we restrict to estimation of the target trend and we show the results of estimating the squared bias and the RMSE (Eq. 20). We use the following abbreviations: ABS is the simulation mean (over 1,000 series) of true $bias^2$ (average of square of Eq. 18), ARMSE is the simulation mean of the true RMSE (Eq. 18); Note that the variance of any given estimator is fixed for all simulated series in a given set, but the signal, and hence the bias, changes from one simulated series to another. AEBS(9) is the simulation mean of the error in estimating $bias^2$ (Eq. 20) when the signal is estimated as in (9), AEBS(10) is the simulation mean of the error in estimating $bias^2$ when the signal is estimated by (10); SDEB(9) and SDEBS(10) are the standard deviations (SD) of the simulation means AEBS(9) and AEBS(10) respectively. AEM(9) is the simulation mean of the error of the RMSE estimates (Eq. 20) when the signal is estimated as in (9), AEM(10) is the simulation mean of the error of the RMSE estimates when the signal is estimated by (10); SDEM(9) and SDEM(10) are the standard deviations of the means AEM(9) and AEM(10) respectively.

Figures 3–8 show the means of the true and estimated squared bias and RMSE, as obtained for the last four years of the series for each of the three sets of series by

Table 2. Means of true squared bias and RMSE, simulation means of error when estimating the squared bias and RMSE, and SD of simulation means of error as obtained by application of X-11-ARIMA and by EBSM. Second set of 1,000 series, last six months of series

	Month	Jul	Aug	Sep	Oct	Nov	Dec
X-11 ARIMA	ABS	0.020	0.021	0.021	0.035	0.113	0.332
60 forecasts	AEBS(9)	0.010	0.011	0.012	0.026	0.098	0.298
	SDEBS(9)	(0.031)	(0.032)	(0.034)	(0.054)	(0.172)	(0.501)
	AEBS(10)	− 0.007	− 0.009	− 0.006	0.002	0.011	0.017
	SDEBS(10)	(0.238)	(0.248)	(0.246)	(0.268)	(0.427)	(0.739)
	ARMSE	0.644	0.644	0.644	0.661	0.725	0.851
	AEM(9)	0.008	0.008	0.009	0.019	0.064	0.161
	SDEM(9)	(0.023)	(0.024)	(0.025)	(0.038)	(0.101)	(0.227)
	AEM(10)	− 0.014	− 0.015	− 0.014	− 0.005	0.033	0.104
	SDEM(10)	(0.042)	(0.046)	(0.048)	(0.057)	(0.113)	(0.241)
X-11 ARIMA	ABS	0.024	0.025	0.026	0.041	0.120	0.343
12 forecasts	AEBS(9)	0.022	0.023	0.024	0.039	0.115	0.323
	SDEBS(9)	(0.034)	(0.033)	(0.035)	(0.059)	(0.181)	(0.512)
	AEBS(10)	− 0.003	− 0.001	0.003	0.011	0.016	0.017
	SDEBS(10)	(0.185)	(0.185)	(0.180)	(0.216)	(0.385)	(0.693)
	ARMSE	0.647	0.651	0.651	0.667	0.731	0.859
	AEM(9)	0.017	0.017	0.018	0.029	0.075	0.175
	SDEM(9)	(0.025)	(0.025)	(0.026)	(0.041)	(0.105)	(0.229)
	AEM(10)	0.007	0.009	0.010	0.019	0.056	0.128
	SDEM(10)	(0.030)	(0.029)	(0.029)	(0.044)	(0.109)	(0.239)
EBSM	ABS	0.240	0.221	0.209	0.221	0.287	0.441
	AEBS(9)	0.160	0.140	0.120	0.120	0.167	0.291
	SDEBS(9)	(0.279)	(0.264)	(0.260)	(0.260)	(0.375)	(0.648)
	AEBS(10)	− 0.002	0.038	0.076	0.096	0.090	0.052
	SDEBS(10)	(0.340)	(0.330)	(0.355)	(0.403)	(0.494)	(0.665)
	ARMSE	0.690	0.694	0.717	0.749	0.794	0.871
	AEM(9)	0.109	0.092	0.078	0.076	0.099	0.152
	SDEM(9)	(0.163)	(0.153)	(0.147)	(0.147)	(0.195)	(0.297
	AEM(10)	0.016	0.021	0.0290	0.033	0.035	0.057
	SDEM(10)	(0.172)	(0.166)	(0.179)	(0.203)	(0.248)	(0.332)

application of X-11-ARIMA and the EBSM. The X-11-ARIMA estimators refer to the case of twelve months forecasts (generally similar results to the case of 60 months forecasts). The signal in all the figures is estimated by use of X-11-ARIMA (Eq. 10), which produces somewhat less biased estimators of the RMSE than the use of Eq. 9, although occasionally with larger SD (see summary below).

The main conclusions from the simulation study can be summarized as follows:

1. The simulation mean of the errors over all realizations of the signal and the sampling errors of the 1,000 series in each set, when estimating the true bias of the estimators are all very close to zero for all the estimators and all three data sets (not shown in the tables). Thus, our proposed estimators of the bias of the estimators of the target trend are unbiased *unconditionally*, although occasionally with large standard errors.
2. The true ARMSEs of the estimators increase by a magnitude of around two when increasing the SD of the sampling errors by two. Thus, the ARMSEs in Table 1 are

Table 3. Means of true squared bias and RMSE, simulation means of error when estimating the squared bias and RMSE, and SD of simulation means of error as obtained by application of X-11-ARIMA and by EBSM. Third set of 1,000 series, last six months of series

	Month	Jul	Aug	Sep	Oct	Nov	Dec
X-11 ARIMA	ABS	0.020	0.021	0.021	0.035	0.113	0.332
60 forecasts	AEBS(9)	−0.042	−0.039	−0.034	−0.025	0.019	0.157
	SDEBS(9)	(0.121)	(0.118)	(0.101)	(0.126)	(0.248)	(0.595)
	AEBS(10)	−0.011	−0.016	−0.015	−0.005	0.012	0.030
	SDEBS(10)	(0.552)	(0.549)	(0.552)	(0.586)	(0.747)	(1.122)
	ARMSE	2.530	2.530	2.520	2.550	2.651	2.802
	AEM(9)	−0.008	−0.008	−0.007	−0.005	0.004	0.027
	SDEM(9)	(0.023)	(0.022)	(0.020)	(0.024)	(0.046)	(0.102)
	AEM(10)	−0.051	−0.049	−0.050	−0.052	−0.058	−0.083
	SDEM(10)	(0.074)	(0.075)	(0.075)	(0.083)	(0.115)	(0.206)
X-11 ARIMA	ABS	0.024	0.025	0.026	0.041	0.120	0.343
12 forecasts	AEBS(9)	−0.009	−0.004	−0.002	0.007	0.060	0.209
	SDEBS(9)	(0.064)	(0.058)	(0.058)	(0.086)	(0.209)	(0.555)
	AEBS(10)	−0.003	−0.001	0.005	0.014	0.023	0.031
	SDEBS(10)	(0.429)	(0.440)	(0.399)	(0.441)	(0.633)	(1.028)
	ARMSE	2.525	2.536	2.532	2.557	2.654	2.807
	AEM(9)	−0.002	−0.001	−0.000	0.001	0.011	0.036
	SDEM(9)	(0.012)	(0.011)	(0.011)	(0.017)	(0.039)	(0.095)
	AEM(10)	−0.027	−0.023	−0.021	−0.023	−0.023	−0.054
	SDEM(10)	(0.044)	(0.038)	(0.036)	(0.044)	(0.079)	(0.173)
EBSM	ABS	0.240	0.221	0.209	0.221	0.287	0.441
	AEBS(9)	−0.173	−0.141	−0.082	−0.043	−0.081	−0.173
	SDEBS(9)	(1.452)	(1.200)	(0.938)	(0.976)	(1.473)	(2.268)
	AEBS(10)	0.006	0.049	0.089	0.107	0.094	0.044
	SDEBS(10)	(1.186)	(0.935)	(0.847)	(0.879)	(1.044)	(1.300)
	ARMSE	2.090	2.150	2.221	2.301	2.391	2.499
	AEM(9)	−0.023	−0.021	−0.011	−0.002	−0.004	−0.010
	SDEM(9)	(0.291)	(0.238)	(0.190)	(0.193)	(0.270)	(0.386)
	AEM(10)	−0.205	−0.156	−0.116	−0.107	−0.137	−0.178
	SDEM(10)	(0.321)	(0.267)	(0.223)	(0.224)	(0.277)	(0.363)

around twice the ARMSEs in Table 2, and the ARMSEs in Table 3 are around twice the ARMSEs in Table 1. The increase in the ARMSEs is somewhat lower for the EBSM estimators. This outcome is explained by the fact that under the present simulation setup, the major component of the RMSE is the variance of the estimator, which of course depends on the variances and covariances of the sampling errors. (The autocorrelations of the sampling errors are the same for all the three sets.)

3. Interestingly enough, the SD of the mean error when estimating the true RMSE does not show a similarly stable pattern. For example, the SDEM(9) values in Table 3 with the largest variance of the sampling errors are, in the case of X-11-ARIMA with 60 forecasts, smaller for the last two months than the corresponding values in the other two tables. This seemingly odd outcome is explained by the fact that the SD of the true RMSE (not shown) actually decreases as the variance of the sampling errors increases. The latter property follows from the fact that for a given estimator and data set, the variance of the estimator is constant and under general conditions $SD\{[bias^2 + \text{var}(est.)]^{1/2}\}$ decreases as $\text{var}(est.)$ increases (can be shown by second-order linearization). Note that unlike the variance, which is fixed in a given

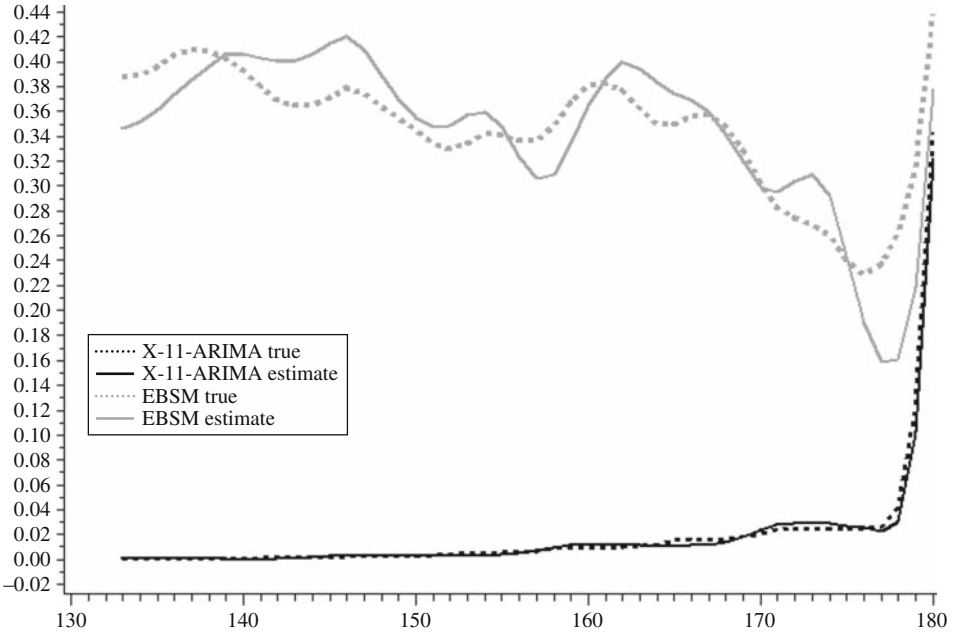


Fig. 3. Means of true and estimated squared bias by application of X-11-ARIMA with twelve months forecasts and EBSM. First set of 1,000 series, last 48 months of data

set, the true $bias^2$ changes from one simulation to another, depending on the random realization of the signal.

The aforementioned phenomenon with the SD of the mean error when estimating the true RMSE does not repeat itself when estimating the true squared bias. Thus, SDEBS(9) and SDEBS(10) are smaller for all estimators and all the months in Table 2 than in Table 1, and smaller in Table 1 than in Table 3.

4. The estimators of the true squared bias when estimating the signal by forecasting the series using X-11-ARIMA (Eq. 10) are generally less biased unconditionally (over all realizations of the signal in a given set) than when estimating the signal by the model identified by X13ARIMA-SEATS (Eq. 9), particularly in the last two months (November, December). This outcome may look odd but note that the models used to generate the series for our simulation study (Eq. 14 for the population values and AR(15) for the sampling errors) do not combine to the airline model fitted to the data, so that the model extracted for the trend levels by X13ARIMA-SEATS is not the correct model. On the other hand, the SD of the mean errors when estimating the squared bias are smaller, and in most cases much smaller, when estimating the signal by application of Eq. 9 than when estimating the signal by application of Eq. 10. (Compare the rows SDEBS(9) and SDEBS(10).) The only exception is in Table 3 when estimating the trend using the EBSM.
5. The conclusions referring to the estimation of the true squared bias generally also apply to the estimation of the true RMSE, particularly with regard to the SD of the mean of the estimation errors. (Compare the rows SDEM(9) and SDEM(10).) In general, we find that our proposed estimators of the RMSE when using the X-11-ARIMA method are

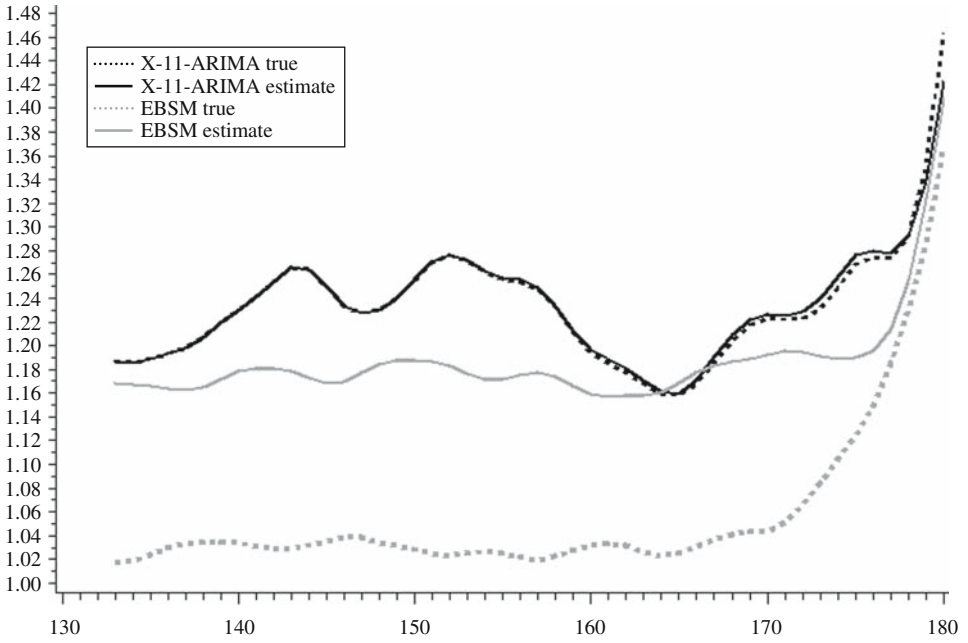


Fig. 4. Means of true and estimated RMSE by application of X-11-ARIMA with twelve months forecasts and EBSM. First set of 1,000 series, last 48 months of data

unbiased in our simulation study when averaging over all possible realizations of the signal, except perhaps for the last two time points, although even there, the bias is never significant using the ordinary *t*-statistic (the ratio of AEM to SDEM is always smaller than 1).

6. Finally, by comparing the results obtained for the three estimators we notice that the ARMSEs are very similar when using the ARIMA estimators with 60 forecasts or with only twelve forecasts. What we find very interesting is that the EBSM produces estimators with lower ARMSEs (lower variances), except in the case of the small sampling errors. As noted before, we used the EBSM for generating the simulated series, but this only partly explains this outcome because the target trend defined by Eq. 3 is not the trend generated under the model and we predicted the signal by use of the airline ARIMA model fitted to the data and not under the EBSM.

Figures 3–8 show the means of the true and estimated squared bias and RMSE as obtained for the last four years of the series for each of the three sets of series, by application of X-11-ARIMA with twelve forecasts (using Eq. 10 for estimating the signal) and the EBSM. The main conclusion from these figures is that the use of X11-ARIMA yields unbiased estimators of the squared bias and the RMSE, except when estimating the RMSE for the last two months in the second set with the small sampling errors. The EBSM estimators seem to be biased, especially in the case of the third set with the large sampling errors, but as can be seen in the tables, the biases are highly insignificant. As already mentioned, the true RMSEs of the EBSM estimators are lower than the true RMSEs of the X11-ARIMA estimators, except for the second set of data with the small sampling errors.

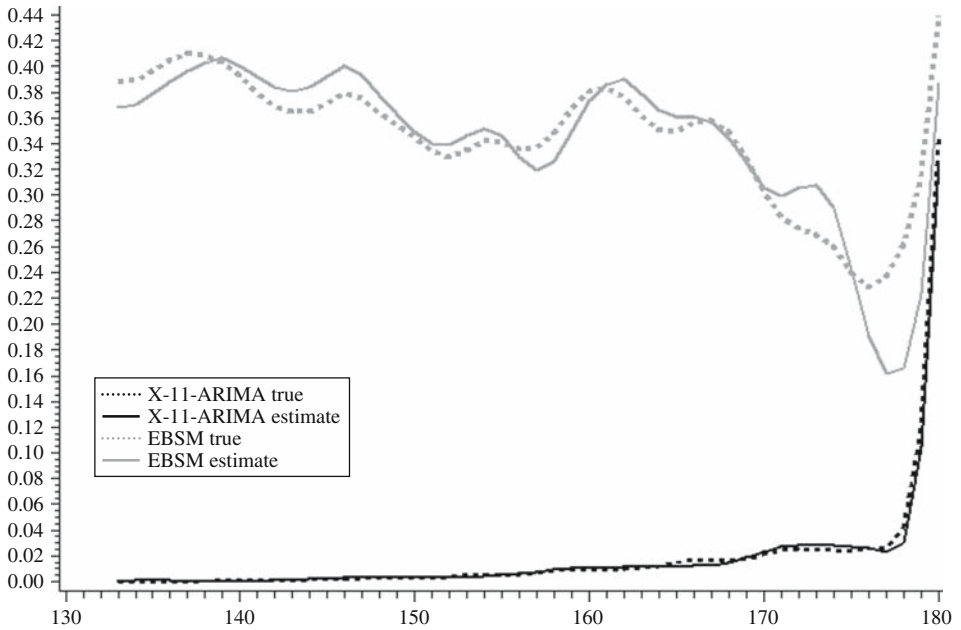


Fig. 5. Means of true and estimated squared bias by application of X-11-ARIMA with twelve months forecasts and EBSM. Second set of 1,000 series, last 48 months of data

Remark 11. We emphasize again that although our proposed estimators condition on a given signal, the results in the tables and figures are unconditional by averaging over the 1,000 realizations of the true signal and the estimators.

5. Application to Current Employment Statistics Series

5.1. Series Considered

In this section we study the performance of the proposed method when applied to real series. We consider four leading monthly employment series, each spreading over 17 years, from February 1990 to January 2007. The series are produced by the BLS based on the Current Employment Statistics (CES) survey, which covers more than 300,000 establishments. The target of interest is the monthly change in employment. The variance and autocovariances of the sampling errors of the unadjusted estimators are estimated each month using the balanced repeated replication (BRR) method, with a modification proposed by Robert Fay, using a factor of 0.5 to reflect the sampling design (see http://www.bls.gov/web/empsit/cestn2.htm#4). The CES survey has the advantage of having time-lagged true population figures from the Unemployment Insurance Program (UIP). Quarterly business tax forms collected by the UIP include monthly employment data which are assembled first at the state level and then at the national level. The true population value for March of each year becomes available by the following January and then all the estimates from March of the previous year up to the current January are

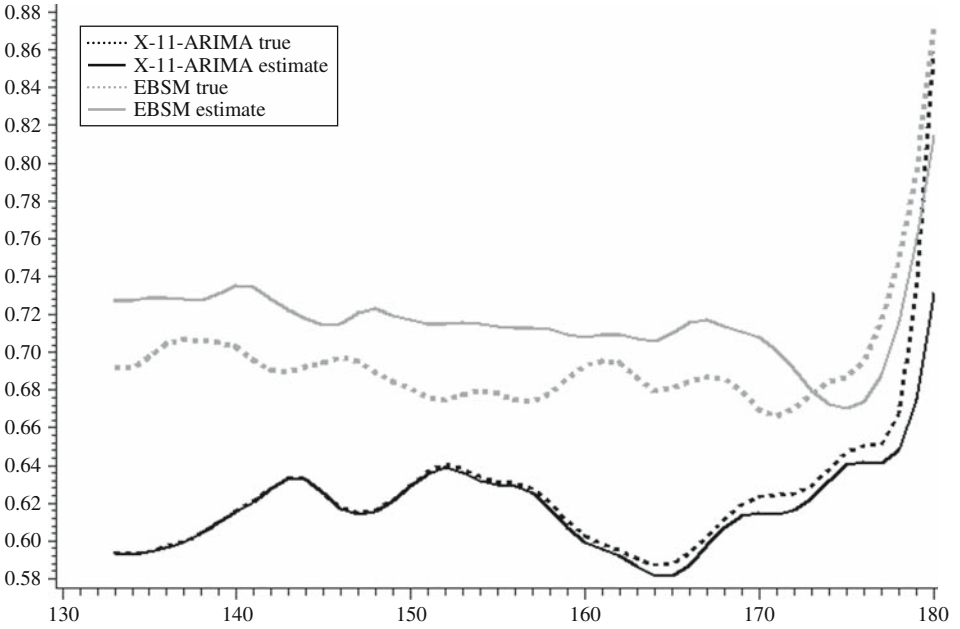


Fig. 6. Means of true and estimated RMSE by application of X-11-ARIMA with twelve months forecasts and EBSM. Second set of 1,000 series, last 48 months of data

benchmarked for the difference between the population and estimated levels in March of last year. No benchmarking is carried out in the months of February to December of a current year. The employment estimate for a current month t is computed as a "link-relative" estimator,

$$\hat{E}_t = E_0 \times r_1 \times r_2 \times \ldots \times r_t, \tag{21}$$

where E_0 is the latest available population value and subsequent subscripts denote months after the benchmark month. The links r_j are ratios between employment estimates in adjacent months,

$$r_j = \frac{\sum_{i \in M_j} d_{ij} x_{ij}}{\sum_{i \in M_j} d_{i,j-1} x_{i,j-1}}, \tag{22}$$

where x_{ij} represents the number of employees in establishment i at month j, d_{ij} is the survey weight and M_j represents the set of establishments for which the number of employees is reported for both months j and $j-1$.

In this study we focus on the estimation of the MSEs of seasonally adjusted estimators (SAE) produced by X-11-ARIMA with 24 months of forecasts, which is the common routine at the BLS for these series. We focus on SAE since it allows us to compare our proposed MSE estimators with estimators produced by application of Bell and Kramer's (1999) approach, reported in Scott et al. (2012). Traditionally, the CES employment series

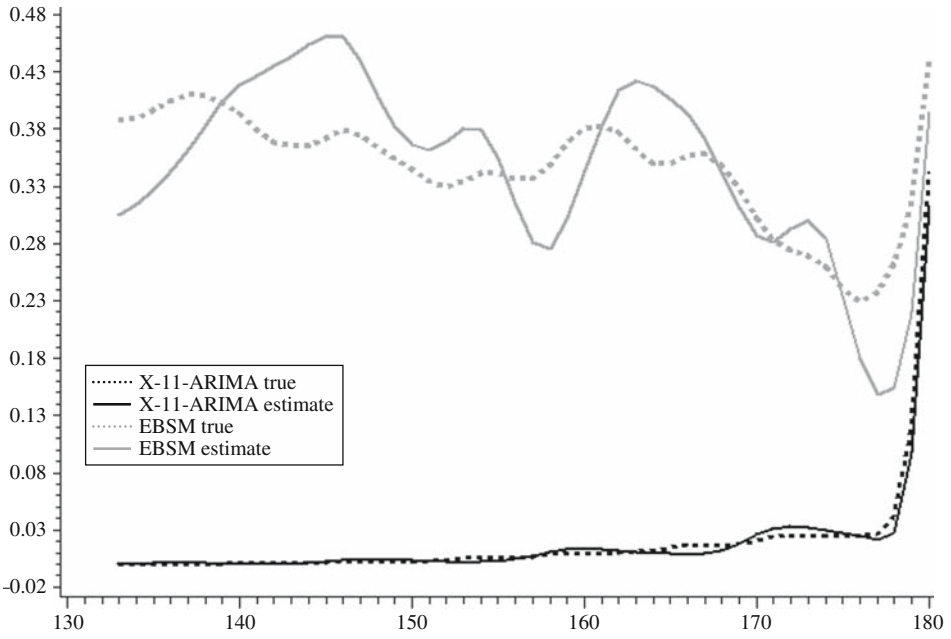


Fig. 7. Means of true and estimated squared bias by application of X-11-ARIMA with twelve months forecasts and EBSM. Third set of 1,000 series, last 48 months of data

have been seasonally adjusted multiplicatively. This suggests considering monthly changes in the log scale,

$$y_t = \log(\hat{E}_t) - \log(\hat{E}_{t-1}) = \log\left(\frac{\hat{E}_t}{\hat{E}_{t-1}}\right). \tag{23}$$

Under the multiplicative decomposition, $y_t = (\log(E_t) - \log(E_{t-1})) + \log(\varepsilon_t/\varepsilon_{t-1})$, decomposing the estimated monthly change as the sum of the population value, $Y_t = \log(E_t) - \log(E_{t-1})$ and a sampling error component, $\tilde{\varepsilon}_t = \log(\varepsilon_t/\varepsilon_{t-1})$.

Remark 12. For the present illustrative study, the input series are the ratios of the benchmarked estimators. Previous studies show that the ratios of the benchmarked estimators are very close to the ratios of the unbenchmarked estimators, and in what follows, we refer to the observed series as the ratios r_t. BRR estimates for the variances and covariances of the sampling errors $\tilde{\varepsilon}_t$ of the log ratios, $\log(r_t)$, have been produced and are used for the computations of the various estimators. As stated above, the benchmarking changes the current estimates, and hence the variances and covariances, very little.

Following the methodology of the previous sections we fit ARIMA models to $\log(\hat{E}_t)$ with one regular difference, such that the observed input series has the general form, $y_t = (1 - B)\log(\hat{E}_t) = \log(r_t)$. Furthermore, assuming that the ratios r_t fluctuate around 1 and using a Taylor expansion, $\log(r_t) \approx r_t - 1 = (\hat{E}_t - \hat{E}_{t-1})/\hat{E}_{t-1}$. Thus, the seasonally adjusted values of the series $y_t = \log(r_t)$ can be interpreted as estimating the seasonally adjusted values of the percentage change in employment, which is the focus of estimation.

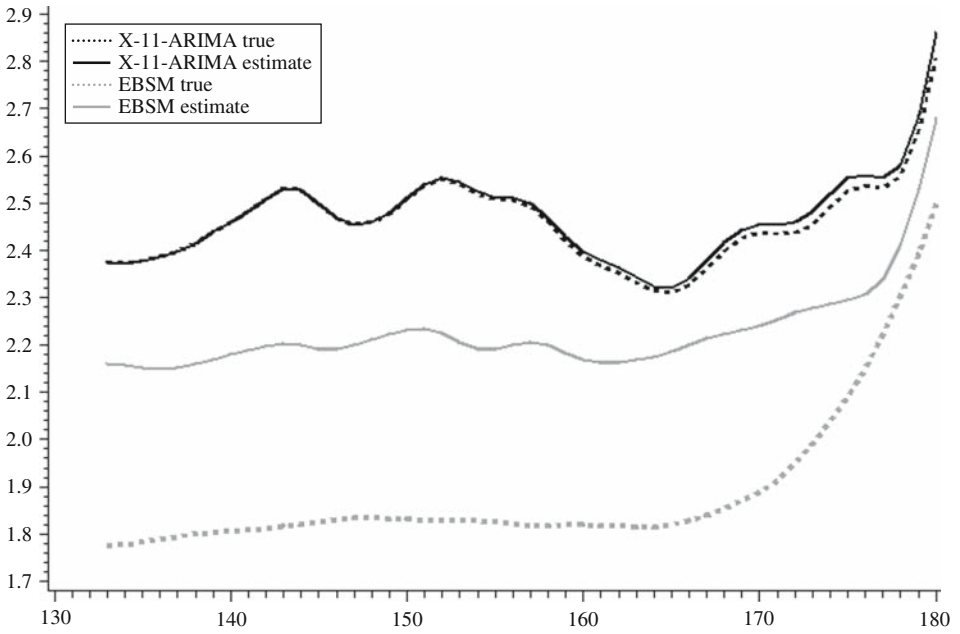


Fig. 8. Means of true and estimated RMSE by application of X-11-ARIMA with twelve months forecasts and EBSM. Third set of 1,000 series, last 48 months of data

5.2. *Results*

We present the results obtained for the last five years of data when applying our proposed method of RMSE estimation and the method proposed by Bell and Kramer (1999, hereafter B-K), to the following four series: "Total Employment in Education and Health Services"; "Total Employment in Manufacturing, Durable Goods"; "Total Employment in Manufacturing, Nondurable Goods" and "Total Employment in Retail Trade". Using standard ARIMA model fitting and diagnostic techniques, we fit the model (1,0,1) (0,1,1) to the first three series and the model (1,0,0) (0,1,1) to the last series. (The input series is in all cases $y_t = \log(I_t^-)$). As mentioned above, we used two years of monthly forecasts when computing the X-11-ARIMA estimator $\hat{SA}_t$ of the seasonally adjusted value SA_t for time t. The MSE estimator under the proposed method is,

$$\hat{MSE}(\hat{SA}_t|\mathbf{G}) = \hat{Var}(\hat{SA}_t|\mathbf{G}) + \hat{Bias}^2(\hat{SA}_t|\mathbf{G}) - \hat{Var}[\hat{Bias}(\hat{SA}_t|\mathbf{G})|\mathbf{G}], \qquad (24)$$

with the signal estimated by X-11 ARIMA (Eq. 10, the same as in the Figures 3–8). See Eq. 15 for the definition of the SA estimator, and Subsection 2.5 for discussion of the difference between the proposed MSE estimator and the B-K method. We also show the RMSE estimators obtained under the proposed method when the irregular term is part of the error (definition *GE*1 of the signal in Subsection 2.1). As noted before, no B-K estimators are available for this definition of the signal. In addition, we show the conditional standard deviations (SD) of the estimators $\hat{SA}_t$ given the signal *GE*2 (SQRT of Eq. 5 with respect to the SAE), and the SD of the original, unadjusted estimators. The last two SDs only account for the variance of the sampling errors. All the values in the figures are multiplied by 10,000.

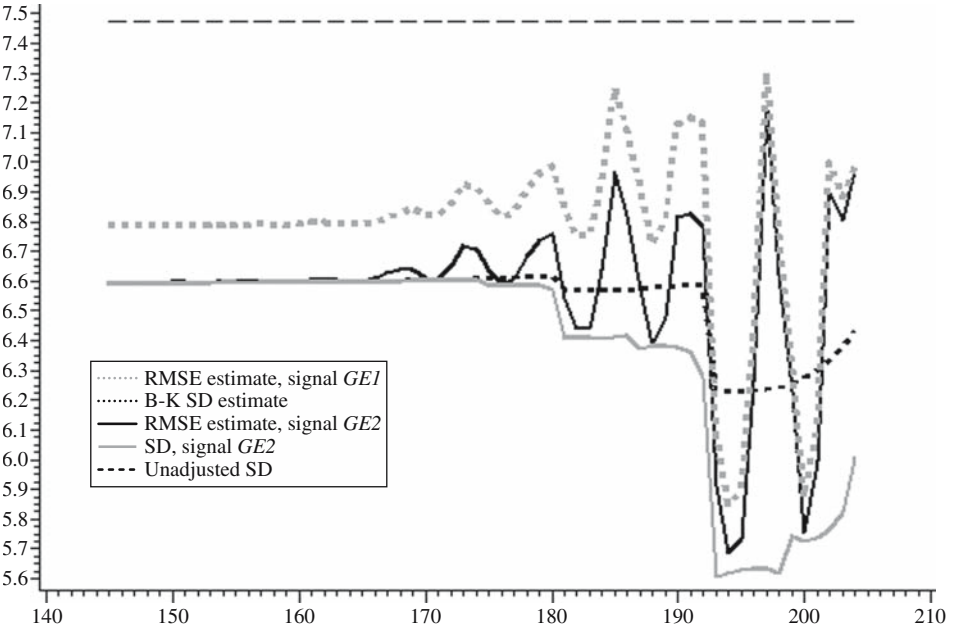


Fig. 9. Results for Total Employment in Education and Health Services

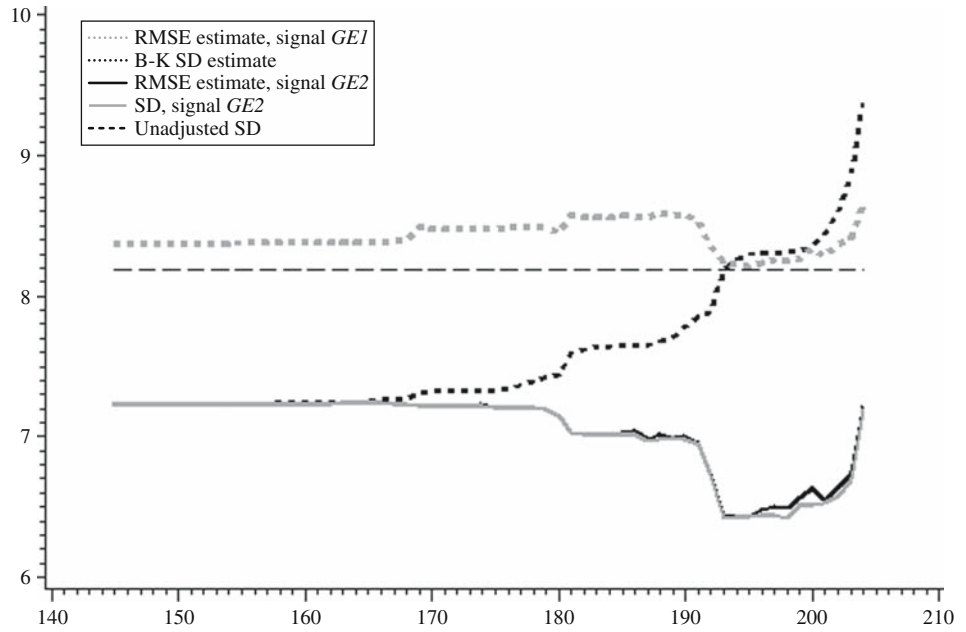


Fig. 10. Results for Total Employment in Manufacturing, Durable Goods

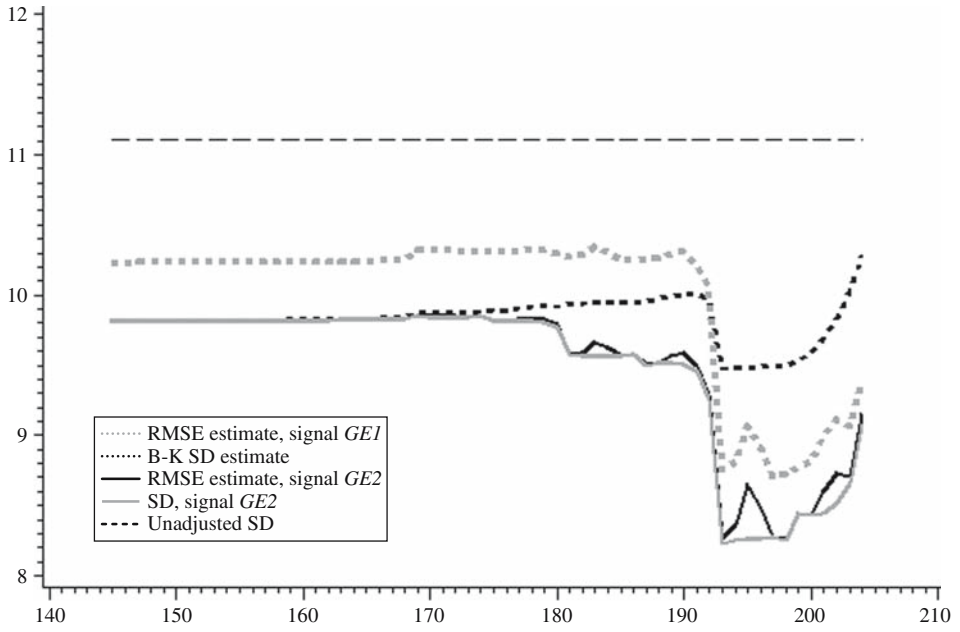


Fig. 11. Results for Total Employment in Manufacturing, Nondurable Goods

It is hard to assess the performance of the various estimators because the true MSEs are unknown when analyzing real series, but the following points are worth mentioning.

1. The SD of the SAE given the signal *GE*2, by which the error consists only of the sampling error without the irregular term, is always smaller than the SD of the unadjusted estimator. This result is explained by the fact that the SAEs are weighted averages of the unadjusted estimators with weights that sum to 1.
2. The RMSE estimates given the signal *GE*1 are always higher than the RMSE estimates given the signal *GE*2, which is obvious since under definition *GE*1 the signal consists only of the trend and seasonal effect and the irregular term is part of the error.
3. The RMSE estimates given the signal *GE*2 are literally the same as the B-K SD estimates in the center of the series (until around time point 168). However, except for Figure 9 (Total Employment in Education and Health Services), for the last three years of data the B-K SD estimates are higher than the conditional RMSE estimates given the signal *GE*2. As discussed in Subsection 2.5, the two estimators differ in the definition of the estimators of the SAE (B-K assume that the X-11 ARIMA SAE use seven years of forecasts whereas in our present application the SAE use only two years of forecasts), and in the definition of the target MSE (we condition on the actual signal, whereas the B-K variance is over all possible realizations of the signal under the ARIMA model fitted to the series, thus accounting for the forecast and backast prediction errors).
4. Except for Figure 9, The RMSE estimates given the signal *GE*2 are very close to the SD of the SAE given the signal *GE*2, and the bias corrections contribute only marginally. On the other hand, in Figure 9, the SDs are much smaller than the

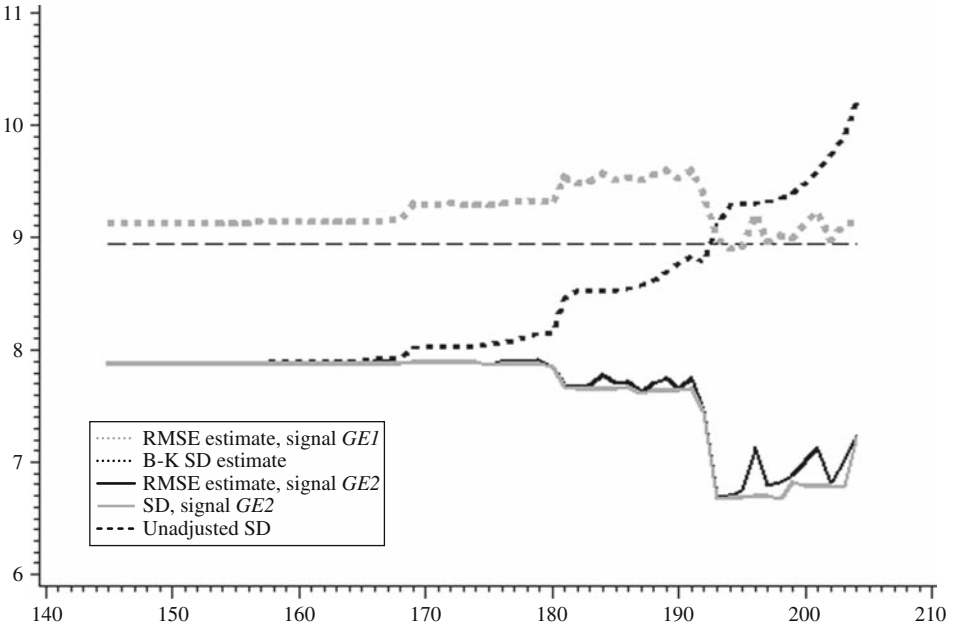


Fig. 12. Results for Total Employment in Retail Trade

RMSEs in the last two years where the SAE use asymmetric weights, indicating a significant contribution of the bias corrections.

5. The RMSE estimates given the signal *GE*1 are slightly higher than the SD of the unadjusted series in two of the series, but are appreciably lower in the two other series.

6. Summary

In this article we propose a new method for the estimation of the MSE of X-11-ARIMA estimators or other linear estimators of the underlying components of a time series. Our approach has some important advantages over other approaches proposed in the literature. First, we follow Bell and Kramer (1999) by defining the target component values as the corresponding X-11 estimators that would be obtained if the series were free of sampling errors and long enough to permit the use of the symmetric filters embedded in the program. In other words, the target components are real entities defined as linear combinations of finite population means or totals over time, in close correspondence to the target values in classical finite population sampling. In particular, under definition GE2 of the signal, the target component values are just linear combinations of the unadjusted finite population values. Interestingly, while the programme X-11 for seasonal adjustment and its previous and subsequent versions have been in wide use for many decades, the target estimated values were never defined in a precise form. This is rather unusual in statistics, where an estimator is defined but not what is estimated. This problem does not exist when using model-dependent methods where the targets are defined by the model, such as in the BSM, the Tramo and Seats program (Gómez and Maravall 1996) and in one of the modules of

X-13ARIMA-SEATS, but purely model dependent estimators are not in common use, at least not in national statistical offices.

A second notable advantage of our procedure is that for definition *GE*2 of the signal, the procedure is basically automatic and does not require new programs or external intervention beyond what is required for the production of the component estimators themselves. Thus, the X-13ARIMA-SEATS programme produces the models for the trend and seasonal components and hence for the signal. These models are then used to estimate the signal within the observation period and to predict the signal outside the observation period. The weights required to define the X-11-ARIMA estimators and the bias estimators (Eq. 9) can be obtained by repeated runs of X-11-ARIMA, as described in Section 3 and in Burck and Sverchkov (2001). In the case of definition *GE*1 of the signal, the application of our procedure additionally requires the estimation of the variances and covariances of the combined errors or at least the variance and covariances of the irregular terms (Subsection 2.3), for which an additional program has to be used.

A third important advantage of the procedure is its flexibility in terms of the target values and the estimators used. It is applicable to the case where the signal consists of only the trend and the seasonal effect and the time series irregular component is part of the error (definition GE1 of the signal and error), and to the case where the irregular component is part of the signal, as under the B-K approach. It is up to the user to decide which definition of the signal is more appropriate. In addition, the procedure is applicable to any linear estimator with known coefficients.

Finally, and most importantly, we have illustrated the good performance of the procedure in estimating the true unknown MSEs, as defined in this article.

Taking into account the clear interpretation of the target values and the estimated MSE and the other advantages listed above, we hope that our proposed procedure will be experimented with by other users and we shall be happy to receive questions arising from these experiments.

7. References

Bell, W.R. and M. Kramer. 1999. "Toward Variances for X-11 Seasonal Adjustments." *Survey Methodology* 25: 13–29.

Burck, L. and M. Sverchkov. 2001. "A General Method for Estimating the Variances of X-11-ARIMA Estimators." *Federal Committee on Statistical Methodology Research Conference* 3: 1–11. November 14–16, 2001, Washington DC, U.S.A. Available at: http://fcsm.sites.usa.gov/files/2014/05/2001FCSM_Burck.pdf (accessed October 2014).

Chen, Z.G., P. Wong, M. Morry, and H. Fung. 2003. *Variance Estimation for X-11 Seasonal Adjustment Procedure: Spectrum Approach and Comparison*. Statistics Canada (Report BSMD-2003-001E).

Findley, D.F. and D.E.K. Martin. 2006. "Frequency Domain Diagnostics of SEATS and X-11/12-ARIMA Seasonal Adjustment Filters for Short and Moderate-Length Time Series." *Journal of Official Statistics* 22: 1–34.

Gómez, V. and A Maravall. 1996. *Programs TRAMO and SEATS, Introduction for User (Beta Version)*. Banco de España: Banco de España Working Papers 9628.

Harvey, A.C. 1989. *Forecasting Structural Time Series With the Kalman Filter*. Cambridge: Cambridge University Press.

Hilmer, S.C. and G.S. Tiao. 1982. "An ARIMA-Model-Based Approach to Seasonal Adjustment." *Journal of the American Statistical Association* 77: 63–70. DOI: http://dx.doi.org/10.1080/01621459.1982.10477767

Pfeffermann, D. 1994. "A General Method for Estimating the Variances of X-11 Seasonally Adjusted Estimators." *Journal of Time Series Analysis* 15: 85–116. DOI: http://dx.doi.org/10.1111/j.1467-9892.1994.tb00179.x

Pfeffermann, D., M. Morry, and P. Wong. 1995. "Estimation of the Variances of X-11 ARIMA Seasonally Adjusted Estimators for a Multiplicative Decomposition and Heteroscedastic Variances." *International Journal of Forecasting* 11: 271–283. DOI: http://dx.doi.org/10.1016/0169-2070(94)00573-U

Pfeffermann, D. and S. Scott. 1997. "Variance Measures for X-11 Seasonally Adjusted Estimators: Some Developments with Application to Labor Force Series." In *Proceedings of the Section on Business & Economic Statistics: American Statistical Association*, 211–216. August 10–14, 1997, Anaheim, California, U.S.A.

Pfeffermann, D., M. Feder, and D. Signorelli. 1998. "Estimation of Auto-correlations of Survey Errors with Application to Trend Estimation in Small Areas." *Journal of Business and Economic Statistics* 16: 339–348.

Pfeffermann, D., S. Scott, and R. Tiller. 2000. "Comparison of Variance Measures for Seasonally Adjusted and Trend Series." In *Proceedings of the 2nd International Conference on Establishment Surveys*, 755–764. June 17–21, 2000, Buffalo, New York, U.S.A.

Pfeffermann, D. and R. Tiller. 2005. "Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters." *Journal of Time Series Analysis* 26: 893–916. DOI: http://dx.doi.org/10.1111/j.1467-9892.2005.00448.x

Scott, S., M. Sverchkov, and D. Pfeffermann. 2012. "Estimating Variance in X-11 Seasonal Adjustment." In *Economic Time Series: Modeling and Seasonality*, edited by William R. Bell, Scott H. Holan, and Tucker S. McElroy, 185–210. London: Chapman and Hall.

Tiller, R.B. 1992. "Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey." *Journal of Official Statistics* 8: 149–166.

Tiller, R.B. 2012. "Frequency Domain Analysis of Seasonal Adjustment Filters Applied to Periodic Labor Force Survey Series." In *Economic Time Series: Modeling and Seasonality*, edited by William R. Bell, H. Holan Scott, and Tucker S. McElroy, 135–158. London: Chapman and Hall.

Wecker, W.E. 1979. "A New Approach to Seasonal Adjustment." In *Proceedings of the Section on Business and Economic Statistics: American Statistical Association*, 322–323. August 13–16, 1979, Washington DC, U.S.A.

X-13A-S Reference Manual, Version 0.1 (Beta). Time Series Staff, Statistical Research Division, Room 3000-4, U.S. Census Bureau, Washington, DC 20233-9100. Available at: https://www.google.com/?gws_rd=ssl#q=X-13-A-S+Reference+Manual%2c+Version+0.1+(Beta).

Received December 2012
Revised July 2014
Accepted July 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 839–857, http://dx.doi.org/10.2478/JOS-2014-0050

Data Smearing: An Approach to Disclosure Limitation for Tabular Data

Daniell Toth[1]

Statistical agencies often collect sensitive data for release to the public at aggregated levels in the form of tables. To protect confidential data, some cells are suppressed in the publicly released data. One problem with this method is that many cells of interest must be suppressed in order to protect a much smaller number of sensitive cells. Another problem is that the covariates used to aggregate and level of aggregation must be fixed before the data is released. Both of these restrictions can severely limit the utility of the data. We propose a new disclosure limitation method that replaces the full set of microdata with synthetic data for use in producing released data in tabular form. This synthetic data set is obtained by replacing each unit's values with a weighted average of sampled values from the surrounding area. The synthetic data is produced in a way to give asymptotically unbiased estimates for aggregate cells as the number of units in the cell increases. The method is applied to the U.S. Bureau of Labor Statistics Quarterly Census of Employment and Wages data, which is released to the public quarterly in tabular form and aggregated across varying scales of time, area, and economic sector.

Key words: Cell suppression; contingency tables; synthetic data; confidentiality; multiple imputation; nearest neighbor.

1. Introduction

Statistical agencies often collect data under a confidentiality agreement and are bound to protect the identity and/or the provided information of individual respondents. To accomplish this, a disclosure limitation method (DLM) is chosen to protect the sensitive data while allowing the provided data set to retain as much of the utility of the original data as possible. Because quantifying the level of protection and the utility a given DLM provides is difficult (Lambert 1993), comparing DLMs (and thus choosing a method) is not straightforward. Indeed, the level of protection offered by a DLM usually depends on characteristics of the data being published and is usually only quantified with certain restrictions on how the data can be accessed (see, for example Wasserman and Zhou 2010). Measures of the utility, on the other hand, often depend on the intended purpose of the data.

[1] Bureau of Labor Statistics, Office of Survey Methods Research, Suite 1950, Washington, DC 20212, U.S.A. Email: toth.daniell@bls.gov

Acknowledgments: The author would like to thank the editors and referees for their careful review of this article. Their helpful comments and suggestions have materially improved this article. I also thank Michail Sverchkov for many discussions and insights while developing the method, Michael Buso for his expert help with the QCEW data, Polly Phipps and John Eltinge for their helpful comments on the article and especially Randall Powers for his hard work during the evaluation of the method on the QCEW data set.

Sometimes, the sensitive information is collected with the intent to provide data only at certain aggregated levels in a way that still protects the sensitive data. For instance, income may be collected at the household level, while only the mean wages by geographic location such as state or county are reported; or an individual's opinion on a given topic is collected, but only percentages by gender and age category are reported.

When disseminating the data through published tables, cell suppression (CS) is one DLM that is often used by statistical agencies to protect the data of individual respondents. This method requires that cell entries deemed risky be withheld (usually because they represent only a few units or have estimates dominated by one or two large units). Protecting the privacy of responders using CS comes at the cost of withholding values of aggregated cells for which the data was intended to provide information. Often, this results in statistics for the gross aggregates being published, while more refined aggregates are suppressed. Depending on the sample size and level of refinement desired, this usually leads to tables with many holes, reducing the utility of the published data. In addition to the holes in the table resulting from the cell suppressions, CS requires an even larger number of secondary cell suppressions when the data is published as hierarchical contingency tables with more than one dimension.

Take, for example, complex data releases such as the Quarterly Census of Earnings and Wages (QCEW) published by the Bureau of Labor Statistics (BLS). The QCEW aims to provide time-series data with multiscale aggregations (by area and industry classifications). In order to protect against disclosure risks that arise from additive relationships within a table, additional (secondary) cell suppressions are required. Cox (1995) provides background on secondary cell suppressions and a solution to the problem of selecting these cells. Though these secondary suppressions are necessary, they further reduce the utility of the provided data. Over sixty percent of the possible QCEW table cells are suppressed.

Additionally, assuming that all of the risks of disclosure are accounted for through primary and secondary cell suppressions is problematic. For example, the BLS consistently applies both primary and secondary cell suppressions, yet additional risks still arise from the additive relationships in the table along with serial correlation. Holan et al. (2010) showed that it was possible to impute many of the suppressed values within one percent accuracy. Their approach takes advantage of the additive relationships of the QCEW tables (multiscale aggregations) and the serial correlation of the longitudinal data.

Another limitation of the CS method is that the cells defined by the published contingency tables must be fixed by the statistical agency in advance. This limits the potential utility of the data by preventing the release of different cells when other variables are available for further conditioning. For example, an agency may release tables of wages aggregated across only industry and occupation while area data is also available.

For these reasons, the BLS has considered replacing CS with another method for protecting the data (Yang et al. 2012). Any chosen DLM would have to protect the sensitive values (employment count and wages) while allowing for the publishing of total estimates for cells defined by industry, area, and ownership. The published estimates for the main cells with high-level of aggregation should be close to the true collected totals, while sufficiently protecting cells representing few establishments. Ideally, the new method would not require any cell suppressions and would allow

estimates for user-defined cells. In addition, the employment and wage trends, which are very important for users of QCEW data, would be preserved by the estimates obtained from the new method.

One way to accomplish this might be to use a synthetic data approach on the microdata where the synthetic data is generated from random draws from a specified distribution. Using synthetic data to deal with disclosure limitation was proposed by Rubin (1993). Using a synthetic approach, the agencies can provide (or allow users to produce) any requested slice of the data, allowing them to produce any contingency table, without fear of disclosing confidential information.

Fully synthetic data approaches usually focus on trying to produce a data set with a distribution that matches the distribution of the observed microdata as closely as possible in order to allow valid inference while protecting sensitive information. A model is estimated using the sampled data and then values for the entire population (including sample values) are produced using draws from the estimated model distribution. The data obtained for either the entire population or for a sample from these random draws is released to the public (Reiter 2002; Reiter 2004; Reiter and Raghunathan 2007; and Graham et al. 2009).

Since the identity of units contained in the sample is generally unknown, the synthetic values could be as close to the true values as possible without risk of disclosure. However, the QCEW is a census of establishments, the location and identity of most establishments is already public knowledge, so the chosen DLM will have to protect the data without the benefit of anonymity. Because all population values are known, a model could be obtained that produces synthetic values very close to the true values, providing good utility, but not much protection. In addition, these synthetic approaches have the potential to impose associations between the data that do not exist, while reducing or eliminating legitimate associations (Graham et al. 2009). Eliminating this possibility or even the perception of this possibility is a major concern for the production of official statistics.

A related approach is to instead publish the true values with values masked by adding a random noise factor (Fuller 1993; Evans et al. 1998). A complication to applying this approach to establishment data is that the distribution of establishment wages and employment are extremely skewed, making it impossible to use the same noise factor for all establishments. Yang et al. (2012) determine that it is not possible to directly apply a standard noise model of Evans et al. (1998) to the QCEW data because of the inherent skewness of establishment data. In an attempt to modify the method, they propose three different noise factors (multiplicative as well as additive). Unlike the original method of additive noise, this new procedure results in biased marginal totals. To correct for this, a raking procedure is used to guarantee unbiased marginal totals. The cumulative effect of these adjustments on each value becomes unclear and could potentially result in removing the noise from some sets of establishments.

We propose a simple, more specialized DLM (data smearing), which is guaranteed to protect an individual's sensitive data by replacing it with an average value of surrounding units. This allows users to obtain aggregated estimates for any cell, which under a set of given conditions is shown to be asymptotically consistent. To accomplish this, the proposed method relies on a sampling scheme and a weighted estimator to divide the data for a sampled unit among its nearest neighbors. Essentially the method acts to "smear" the

data of each unit around an "area" defined by a unit's characteristics so that each individual unit's data is replaced by data that represents an area's average.

Advantages of this method include those of the synthetic approaches since all microdata values will be replaced, without the risk of disclosure or inducing nonexistent relationships among variables. However, the data released under this method no longer represents the microdata, but instead an average of the data of surrounding units, so the data no longer has the distribution of the original data but can be used only for providing statistics that are functions of totals.

The remainder of the article is organized as follows. Section 2 presents the proposed method and contains a discussion of some properties of the method. Section 3 contains results from an application of the method to QCEW microdata, while Section 4 includes a discussion of the results and mentions future areas of research.

2. The Proposed Disclosure Limitation Method

Suppose a data set consists of elements $\mathbf{u}_i = (\mathbf{Y}_i, \mathbf{X}_i)$ and is indexed by the set $U = \{i = 1. \; . \; .N\}$, where $\mathbf{Y}_i$ are the protected variables and $\mathbf{X}_i$ is a vector of unprotected auxiliary data. We assume these vectors include the data used to form cells of a table for release to the public. For example, the sensitive variables in the QCEW are the total employment and total wages paid, while the auxiliary information includes the establishment's industry and geographic location.

In this article, we assume that $\mathbf{Y}_i$ contains the sensitive information of the i-th unit. Since QCEW represents a census of establishments, the establishment's identity and inclusion in the sample must be considered known. Therefore a disclosure limitation procedure for the QCEW must account for this. This is in contrast to the situation of protecting sample data, where a disclosure limitation procedure can often exploit the fact that the identity of units that have been included in the sample is unknown. Therefore, sensitive sample data can be released for an individual unit as long as there are enough units in the population with similar characteristics to mask their identity or if characteristics are changed slightly.

2.1. Description of the Method

The first step in the procedure, is to define a metric $\|\cdot\|$ on the data elements which will determine the distance between each unit $d(\mathbf{u}_i, \mathbf{u}_j) = \|\mathbf{u}_i - \mathbf{u}_j\|$. We use this distance function to find the k-nearest neighbors for each element in the population. In case of ties, we include a small real-valued noise variable to be used in the distance function. These neighbors define the units the method will use to select a sample in order to produce an average value. A neighborhood is found for each unit in the population. Neighborhoods are defined so that the units contained within them are likely to be included in the aggregate cells to be produced from the estimates. For example, the metric may include geographic location and industry when applying the procedure to business surveys.

Dummy variables are used to handle categorical variables like industry and political borders by assigning "penalties" to units not in the same category. For instance, one could add a ν-mile "penalty" to the geographic distance between units that are not in the same state. That is, if $state_i$ is the state in which unit i is located and $geo(\mathbf{u}_i, \mathbf{u}_j)$ is the geographic

distance between units i and j in miles, then the distance between units i and j defined by the metric is

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu \mathbb{1}_{\{state_i \neq state_j\}}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\nu \in [0, \infty)$. A value of $\nu < \infty$ would allow the "smearing" over categories while $\nu = \infty$ would require that neighbors be in the same category.

The next step of the procedure is to find the k-nearest neighbors for each unit. That is, for each $i \in U$, define r_i as the smallest real number such that the set

$$\{j \neq i \in U \mid \|\mathbf{u}_j - \mathbf{u}_i\| \leq r_i\}$$

has k elements. Note that r_i exists for every $i \in U$, as long as $k \leq N$, and we assume that in most practical situations $k \ll N$. We define $K(i)$ as the k-nearest neighborhood of unit i,

$$K(i) = \{j \neq i \in U \mid \|\mathbf{u}_j - \mathbf{u}_i\| \leq r_i\}.$$

To make sure that the data for each element is spread out among enough other units, we extend the k-nearest neighborhood $K(i)$ to be the k-network, $\overline{K(i)}$, defined by also including every unit j for which unit i is a k-nearest neighbor. Formally,

$$\overline{K(i)} = K(i) \cup \{j \mid i \in K(j)\}.$$

Figure 1 illustrates why extending the network could be necessary for some establishments. In this example, units j, k, and l are in $K(i)$, but i is not in $K(j)$, $K(k)$, or $K(l)$. In fact, there are no units in the population shown that contain unit i in their k-nearest neighborhood. The completed network ensures that the information of unit i gets represented in the other synthetic units produced by the method.

For each $i \in U$, draw a random sample of size $n \leq k$ from unit $i's$ network, $\overline{K(i)}$. For example, for the applications of the method in this article, we used a simple random sample without replacement (SRSWOR). Let $\delta_j(i) = 1$ if unit j is selected in the sample from $\overline{K(i)}$ which will be used to protect element i, and 0 otherwise. To produce a fully synthetic data set, we replace $\mathbf{Y}_i$ for each unit $i \in U$ with the weighted average

$$\tilde{\mathbf{Y}}_i = w_i \mathbf{Y}_i + \sum_{j \in \overline{K(i)}} w_j \delta_j(i) \mathbf{Y}_j, \tag{1}$$

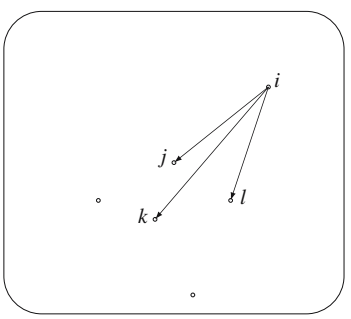


Fig. 1. Illustration shows K(i), the k-nearest neighbor of unit i in the population, where k = 3.

for a given fixed set of weights $\{w_i \mid i \in U\}$. The properties of the method depend on the choice of weights. We will present a choice of weights which are shown to produce asymptotically consistent estimates for cells satisfying certain conditions.

Note that the sampling is done to give an extra level of protection by not allowing users to guess the members included in the unit's network; however if the network is defined large enough, this would not be necessary. Instead, defining the synthetic value as a weighted average of every unit in the network would remove this uncertainty from the synthetic value. Another way to remove some of this uncertainty is to produce a number ($m > 1$) of synthetic values for each unit i ($\tilde{\mathbf{Y}}_{i1}, \tilde{\mathbf{Y}}_{i2}, \ldots, \tilde{\mathbf{Y}}_{im}$) independently and use the average of these as the synthetic value

$$\tilde{\mathbf{Y}}_i = \frac{(\tilde{\mathbf{Y}}_{i1} + \tilde{\mathbf{Y}}_{i2} + \ldots + \tilde{\mathbf{Y}}_{im})}{m}. \tag{2}$$

Alternatively, the multiple sets of imputed values can be released directly to allow the user to estimate the variance of any estimates produced using the synthetic value. By releasing multiply-imputed synthetic values for each establishment, the agency would be giving a data user some indication of the reliability of each estimate being produced.

In addition, a referee pointed out that the variance of the synthetic values could be controlled by defining $\tilde{\mathbf{Y}}_i = \alpha Y_i + (1 - \alpha)\bar{Y}_i$. This would give the data providers another option to provide more accurate estimates, and all the consistency properties described below hold for synthetic values defined this way or by Equations (1) or (2). However, α would have to be chosen carefully to balance the added utility with a loss of protection.

2.2. *Properties of the Method*

The required aggregated data for the released tables are produced using these new synthetic values. The differences in the values and the properties of the synthetic data that is produced depend on the distance function defined by the agency and the moment structure of $\mathbf{Y}_i$ in the k-nearest networks. Therefore, the protection that is afforded the individual units depends on the distance function and the set of networks it produces. The protection also depends on the other parameters of the method, including the value of k, the size of the sample n selected from the k-networks, and the set of weights.

Now we define the notation used to investigate some of the theoretical properties of the synthetic data produced by the proposed method. First, define any subset of units $C \subseteq U$ to be a closed area if it is equal to

$$\bar{C} = \bigcup_{i \in C} \overline{K(i)}. \tag{3}$$

The circle in Figure 2 displays a hypothetical user-defined area that is an example of a closed area. Note that given any subset C, there exists a closed area that contains C. We will use $\bar{C}$ to denote the smallest of these.

Let $|\overline{K(i)}|$ denote the total number of units in $\overline{K(i)}$. The following property of the method states that if we define the weights in Equation (1) correctly, then the cells in the table will be unbiased for large enough levels of aggregation. That is, the expected value of the aggregated value of a cell produced from the synthetic values will be equal

to the aggregated value of the cell using the original microdata, if the cell defines a closed network using the given distance function. The original microdata are considered fixed values and the expectation is with respect to the random samples from the k-networks.

The number of times a given unit's value is used to produce different synthetic values depends on the size of the unit's network and the probability of selection used in the sampling process. With this in mind, we show in the next result, that if we select a weight for each unit that is the inverse of the expected number of times the unit's value will appear in other synthetic values, then we will get unbiased cell totals for those cells defined by closed networks.

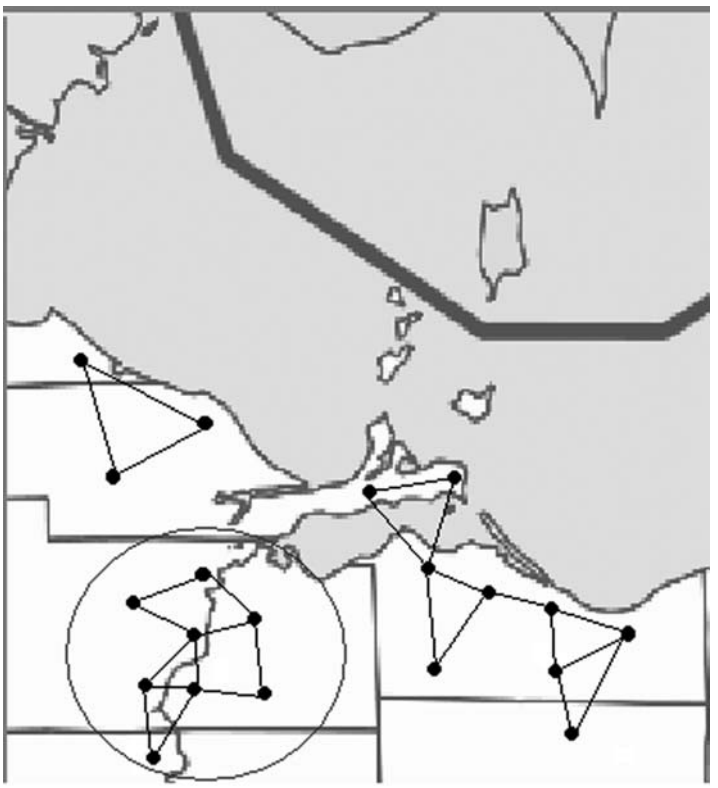


Fig. 2. Illustration with hypothetical establishments in a given location and the k-nearest network that would result from these establishments if k = 2. The circle representing a selected area is an example of a closed area. Every establishment in the closed k-network is included in the selected area.

Lemma 2.1 *If a cell C is a closed area and the weights, w_i used in Equation (1) are defined as*

$$w_i = \left(1 + n\sum_{j \in \overline{K(i)}} \frac{1}{|\overline{K(j)}|}\right)^{-1}, \tag{4}$$

then the synthetic data produced from the method satisfies

$$E\left[\sum_{i\in C}\tilde{\boldsymbol{Y}}_i\right]=\sum_{i\in C}\boldsymbol{Y}_i.$$

Proof

From Equation (1)

$$\sum_{i\in C}\tilde{\mathbf{Y}}_i=\sum_{i\in C}w_i\mathbf{Y}_i+\sum_{i\in C}\sum_{j\in\overline{K(i)}}\delta_j(i)w_j\mathbf{Y}_j. \tag{5}$$

Note that for any i and j if $j\in\overline{K(i)}$, then $i\in\overline{K(j)}$, by the definition of a nearest network. Also, since C is a closed area, if $i\in C$, then for all $j\in\overline{K(i)}$, $j\in C$. Therefore, we can re-write the sum in Equation (5) as

$$\sum_{i\in C}\tilde{\mathbf{Y}}_i=\sum_{i\in C}w_i\mathbf{Y}_i+\sum_{i\in C}\sum_{j\in\overline{K(i)}}\delta_i(j)w_i\mathbf{Y}_i=\sum_{i\in C}\left(1+\sum_{j\in\overline{K(i)}}\delta_i(j)\right)w_i\mathbf{Y}_i.$$

Since each sample from the k-networks is drawn using a SRSWOR,

$$E[\delta_i(j)]=n|\overline{K(j)}|^{-1},$$

so the expectation of Equation (5) is

$$\sum_{i\in C}\left(1+\sum_{j\in\overline{K(i)}}E[\delta_j(i)]\right)w_i\mathbf{Y}_i=\sum_{i\in C}\left(1+n\sum_{j\in\overline{K(i)}}\frac{1}{|\overline{K(j)}|}\right)w_i\mathbf{Y}_i.$$

The proof follows by substituting Equation (4) for w_i.

□

Note, that if the neighborhood of unit i does not contain any units that have an extended neighborhood, then $\forall j\in\overline{K(i)},\ \ |\overline{K(j)}|=k$. This means that the weight for unit i given by Equation (4) is simply $1/(n+1)$.

The result of Lemma 2.1 applies only to table cells that are closed areas. In general we can expect that many cells of interest will not necessary be closed areas. The next result states that we can still expect to obtain reasonable estimates for any area as long as most of the data of the area being estimated are contained in a closed area.

Define the boundary of area C as the set $\partial C=\overline{C}-C$. This is the set of elements that contribute information to the estimate of area C, but are not located in the area. Property 2.1 states that the ratio of the area total estimated using the synthetic data over the total estimated using the real data asymptotically goes to one as long as the data in the interior of the area of interest increases sufficiently fast compared to the data in the boundary.

Property 2.1 Assume $|Y_i-E[\tilde{Y}_i]|<M<\infty$ *for all* i. *If* $|\partial(C)|=o(\sum_{i\in C}\boldsymbol{Y}_i)$ *and that the weights are defined by Equation (4), then the synthetic data produced from the*

method satisfies

$$\lim_{|C|\to\infty}\left(\sum_{i\in C}\mathbf{Y}_i\right)^{-1} E\left[\sum_{i\in C}\tilde{\mathbf{Y}}_i\right] = 1.$$

Proof By the definition of the boundary and Lemma 2.1

$$\begin{aligned} E\left[\sum_{i\in C}\tilde{\mathbf{Y}}_i\right] &= E\left[\sum_{i\in \overline{C}}\tilde{\mathbf{Y}}_i\right] - E\left[\sum_{i\in \partial C}\tilde{\mathbf{Y}}_i\right] \\ &= \sum_{i\in \overline{C}}\mathbf{Y}_i - E\left[\sum_{i\in \partial C}\tilde{\mathbf{Y}}_i\right] = \sum_{i\in C}\mathbf{Y}_i + \sum_{i\in \partial C}\mathbf{Y}_i - E\left[\sum_{i\in \partial C}\tilde{\mathbf{Y}}_i\right] \\ &= \sum_{i\in C}\mathbf{Y}_i + \sum_{i\in \partial C}\left(\mathbf{Y}_i - E[\tilde{\mathbf{Y}}_i]\right). \end{aligned}$$

Since

$$\sum_{i\in \partial C}\left(\mathbf{Y}_i - E[\tilde{\mathbf{Y}}_i]\right) \le M|\partial C|$$

we can divide by $\sum_{i\in C}\mathbf{Y}_i$ to get the result.

□

Figures 3 and 4 give examples of two different user-defined areas. Figure 3 is an illustration of an area that is likely to satisfy the condition $|\partial(C)| = o\left(\sum_{i\in C}\mathbf{Y}_i\right)$. On the other hand, the area shown in Figure 4 has a boundary that would likely grow faster than the contained area as the area expands. The difference is that the first area is a sphere in the coordinates used to define the metric whereas the second area is a very elongated shape with respect to those coordinates.

3. Application to QCEW Data

The Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW) program aims to publish a near census of wage and employment data for every industry at the national, state, county and metropolitan statistical area (MSA) levels. Industry is defined by the establishment's assigned six-digit code from the North American Industrial Classification Systems (NAICS). The codes are organized hierarchically, where higher digit codes aggregate to fewer digit codes. For instance, the three-digit industry codes 423 (merchant wholesales, durable goods), 424 (merchant wholesalers, nondurable goods), and 425 (electronic wholesale markets) aggregate to the two-digit industry code 42 (wholesale trade).

The QCEW collects the number of employees on the payroll of an establishment each month and the total payroll of an establishment every quarter. Every quarter, QCEW publishes employment and wage data in tabular form aggregated across varying cells defined by these location and industry categories. Less aggregated-level data can only be published if disclosure restrictions are met. Currently, over 60% of the possible cells are

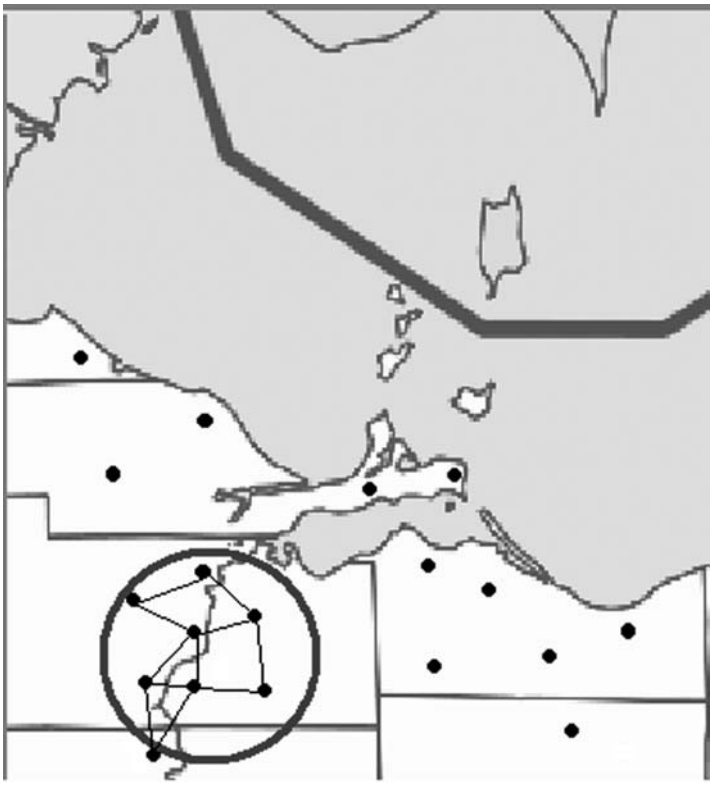


Fig. 3. The circle is an example of a selected area that is not closed but likely to satisfy the conditions of Property 2.1. Though there is one establishment in the closed k-network that has been excluded from the selected area, the number of establishments from the closed k-network that are included in the selected area are likely to dominate the total estimate for the selected area.

suppressed as a result of the current use of CS as the DLM for QCEW. In addition, requested aggregate estimates for areas not published cannot easily be accommodated under CS without risk of disclosure. Using the proposed data smearing DLM, all currently produced cells as well as any requested cells could be published with varying degrees of accuracy without risk of disclosure.

Table 1 shows an example table of one month employment totals (second month of the quarter) for four quarters of QCEW employment data. The table was produced for one industry comprised of three sub-industries over a given MSA. The original QCEW table (top) was produced using the original data for the given MSA for the three industries and their aggregates. The table represents data for roughly 80, 2, and 58 establishments, respectively, for the three industries each quarter.

The same table (bottom) was produced using synthetic values obtained from the data-smearing method with parameter values of $k = 3$, $n = 3$, $m = 5$. The method provided synthetic data that produced a table with values close to the original (all within 1% of the true values) for cells represented by more than two establishments and for the aggregate series and the annual totals. Unsurprisingly, the cells that differ the most are for the middle sub-series, which of course are composed of the smallest number of establishments. This row would be suppressed under the CS method currently used

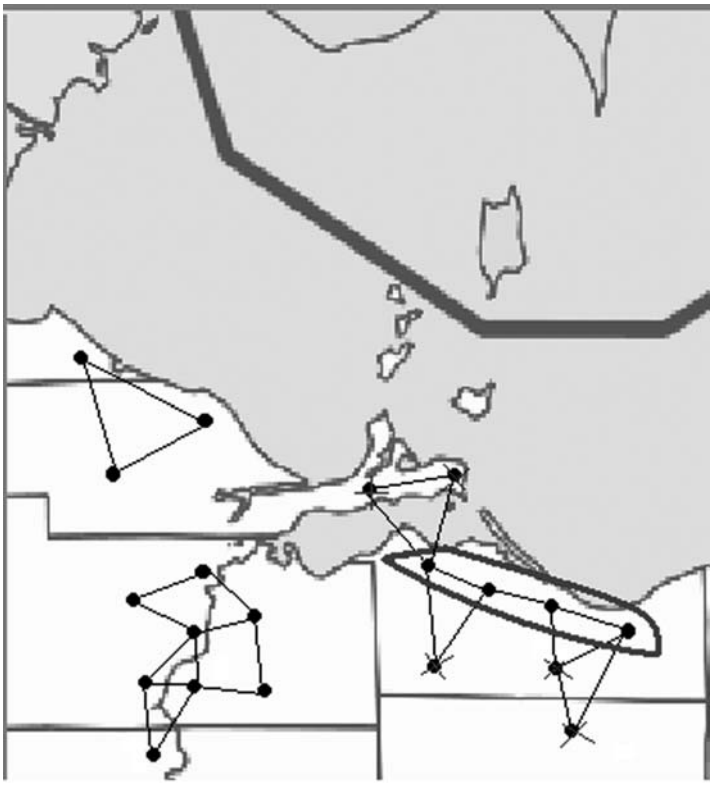


Fig. 4. The selected area is an example of an area that is not closed and unlikely to satisfy the conditions of Property 2.1 because the number of units in the network located outside the selected area is larger than the number of units contained in the area. The values from the establishments in the closed k-network that have been excluded from the selected area are likely to be at least of equal magnitude to the values from the establishments that are included in the selected area. Therefore, the estimated total for the selected area could be biased using the synthetic data.

by the BLS. In addition, another row (probably row three) would be suppressed as the secondary suppression.

The metric used to produce Table 1 used longitude and latitude of each establishment to find the geographical distance $geo(\cdot)$ between establishments and the six-digit industry classification code,

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu \mathbb{1}_{\{ind6_i \neq ind6_j\}}, \tag{6}$$

where $\nu = \infty$, and $ind6_i$ is the six-digit industry code for establishment i. Defining the metric in this way, we are forcing the algorithm to pair establishments with the same industry classification in close geographic proximity to one another. This could also be achieved by applying the algorithm to industries with the same six-digit industry code separately and using only the geographical distance between establishments.

Next we illustrate the method by applying it to one month of QCEW employment data for all (non-government-owned) establishments, over all industries, across the entire country. Again we use parameter values of $k = 3$, $n = 3$, $m = 5$, and the metric given by Equation (6). The weights are defined by (4). As we mentioned earlier, the statistical agency could produce multiple ($m > 1$) synthetic data sets for publication, but the results

Table 1. Example of a 2010 QCEW employment table for one MSA for establishments in a given industry code composed of three sub-series. The first table (Top) was produced using the true values while the second table (Bottom) used the synthetic values. Totals for each of quarter-1 through quarter-4 are displayed for the series and each sub-series along with the annual totals. For this MSA, the table is based on data from roughly 180, 2, and 58 establishments in the three industrial sub-series, sub1, sub2, and sub3, respectively

Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2,600	2,899	3,022	2,599	11,120
Sub1	1,981	2,256	2,382	1,957	8,576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2,409
Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2,622	2,929	3,062	2,589	11,202
Sub1	1,989	2,271	2,420	1,947	8,627
Sub2	42	38	40	34	154
Sub3	591	620	602	608	2,421

below are focused on one synthetic data set using Equation (2), the average of the five independent draws. A comparison of the true and the synthetic values presented in Figure 5 shows that synthetic values produced are highly correlated to the true values. The very small values are inflated while larger values tend to be decreased by the method.

The data-smearing approach acts like a synthetic approach to disclosure limitation in the sense that it replaces each value at the microlevel with a synthetic value. Unlike many synthetic data approaches to disclosure limitation, this current method does not attempt to match the distribution of the synthetic data to that of the original data. Because we are replacing individual values with the mean value of a surrounding area, extreme values are replaced by values closer to the middle of the distribution. Though the two distributions are similar, this figure illustrates the tendency of the method to shift the true values toward the mean. For instance, the new synthetic distribution has a smaller proportion of units with the smallest value. As an example, Figure 6 displays the distribution of total employment

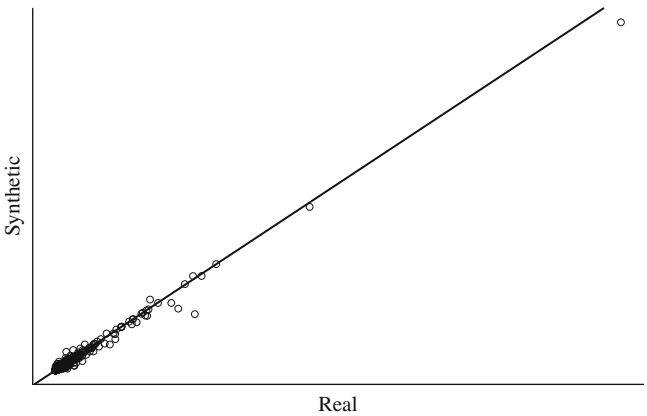


Fig. 5. Relationship between true values (x-axis) and synthetic values (y-axis) with the line $y = x$.

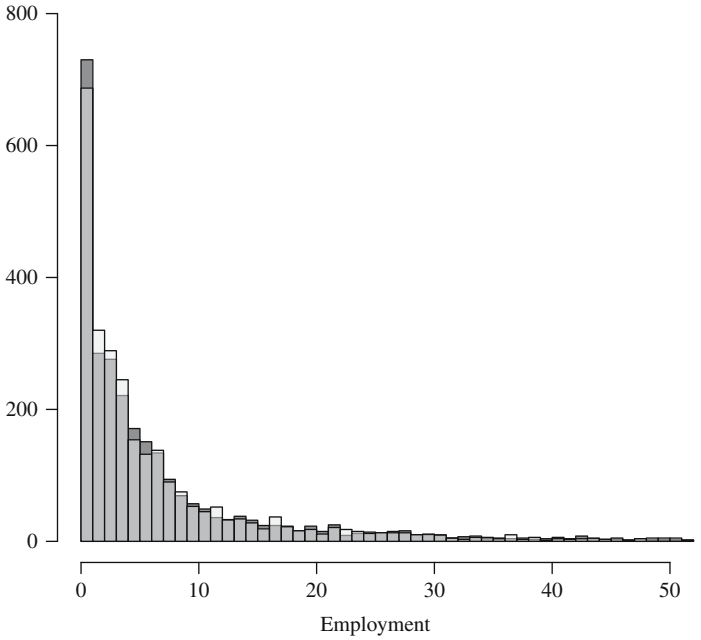


Fig. 6. Histogram of true (dark grey) and synthetic (light grey) employment values for establishments in a given industry and state.

for all establishments in the population in a specific industrial classification. As the figure shows, the synthetic values do not have the same distribution as the real values. Therefore, relying on individual microlevel data for statistical analysis would be very problematic. This is as it should be, since we are protecting the individual values.

The choice of parameters, k, n, and m affects the level of protection as well as how closely the synthetic values represent the real values. Larger k values will smear the value of a given establishment over a wider area. The value of n, and more particularly n/k will affect the variance between each of the m imputed values, with larger values giving smaller variances. By using a large value for m (which we recommend) and the estimator defined by Equation (2), this variance (and the protection derived from the sampling) could be virtually eliminated. The data of individual units would still be protected as long as $n \geq 2$. A value of $n \geq 2$ ensures that the data published for an individual establishment will be the average of at least two other units. In our evaluations of the method, we found that when using even moderate values of m, varying parameter values had a relatively small impact on the overall estimates compared to changing the metric. We proceed by investigating the impact of the metric on the method.

Using the synthetic values obtained from the smearing method, we computed aggregated employment counts e_j for every two, four, and six-digit industry level. The metric given by Equation (6) was designed to give accurate answers for all industrial classifications, so we would expect estimates of total employment aggregated by industry classifications to be close to the true estimates. For each cell estimate produced, we calculate the percent relative difference (PRD) $100 * (\tilde{e}_i - e_i)/e_i$ between the synthetic value $\tilde{e}_i$ and the true value e_i.

Figure 7 displays boxplots of the PRD for each cell estimate over different quantiles of cell sizes, where size is the number of establishments. The top graph gives the results for the two-digit industry level aggregates, the middle graph shows the four-digit level and the bottom graph the result for the six-digit level. As expected, the estimates produced using the synthetic values are all close to the true cell totals. The cells aggregated to the two-digit industry level are within 0.5% of the true value for all 24 cells. This is not surprising

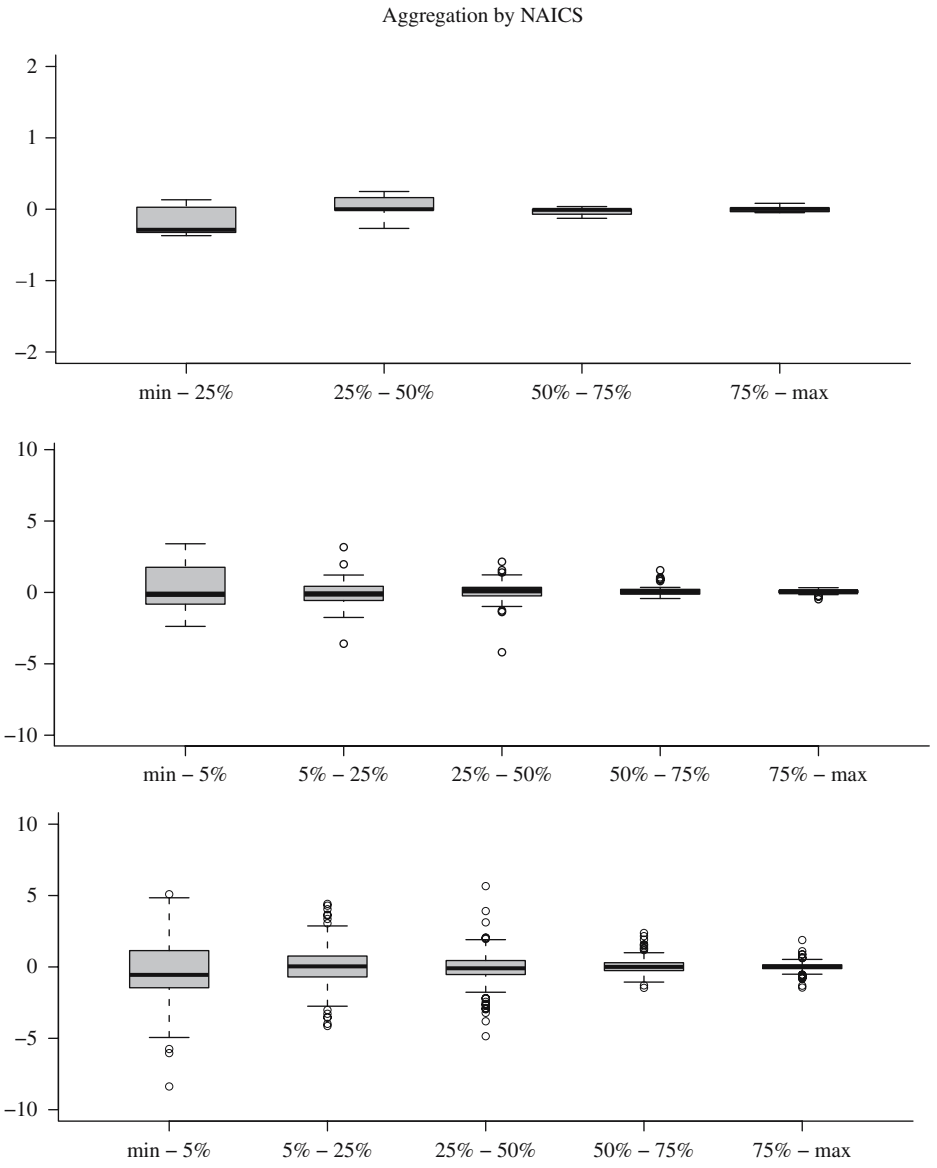


Fig. 7. Boxplots of the percent relative differences for the synthetic values of industry totals compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the industry total). The top graph is of quantiles of errors for all the two-digit industry code totals, the middle graph is of quantiles of errors for all four-digit industry code totals, and the bottom graph is for all six-digit totals. All graphs represent percent relative errors for the synthetic values using the metric given by Equation (6).

since the smallest cell size is 14,652 at this level of aggregation. However, the method produced estimates close to the true values, even at the six-digit industry level of aggregation. Indeed, 99% of the cells have less than a 4.5% difference. This is despite the fact that the smallest 1% of the cells have fewer than 22 establishments.

The situation is very different however, when we consider cells aggregated by state and industry. The top graph in Figure 8 shows the same boxplots of the PRD by quantile of cell size for cells aggregated by state to the two-digit industry level. The cell estimates are not nearly as good for state estimates, even at this high level of aggregation. Though a handful of these cells have fewer than five establishments, 99% have more than 18 so the method should be expected to produce reasonable estimates for most of these cells.

Because state was not included in the metric, there is no penalty for choosing neighbors that are across a state border. Therefore, it should be expected that many synthetic data values represent averages of establishments over more than one state, which biases state-level estimates. However, after adding state to the metric (6), the state-level estimates were still not very good, even though the penalty was rather high (100 miles) for being in a different state.

This is because of the hard restriction that the establishments in a neighborhood must all be in the same industry to the six-digit level. At the six-digit level of aggregation more than 18.1% of state industry cells have fewer than four establishments. This means that the method is forced to use at least one establishment from another state to produce the synthetic values for each of these cells (no matter how large the penalty). This biases the estimates of both states.

Instead we replace the metric given by (6) with

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu_1 \mathbb{1}_{\{ind4_i \neq ind4_j\}} + \nu_2 \mathbb{1}_{\{ind5_i \neq ind5_j\}} + \nu_3 \mathbb{1}_{\{ind6_i \neq ind6_j\}} + \nu_4 \mathbb{1}_{\{state_i \neq state_j\}}, \tag{7}$$

where $(\nu_1, \nu_2, \nu_3, \nu_4) = (\infty, 50, 10, 100)$, and *indt* is the *t*th-digit industry code. This new metric replaces the hard restriction that all industries match to the six-digit level with one at the four-digit level. In addition, there are 50 and ten-mile penalties for not matching at the five and six-digit industry levels respectively. There is a 100-mile penalty for being in a different state.

The percentage of state industry cells with less than 4 establishments drops from 18.1% at the six-digit level to 7.4% at the four-digit level. Therefore, we would expect the values given by the method using the metric (7) to continue to give accurate cell estimates aggregated to the four-digit industry level, while giving improved state-level industry estimates. The bottom graph in Figure 8 shows the same boxplots for cells aggregated by state to the two-digit industry level as the top graph, but instead using this new metric. This shows that the cell estimates for the state two-digit industry level are indeed improved; 95% of all the estimates are within 4% of the true value. There are still a number of estimates that are significantly off, but this is to be expected given that there are a number of small cells for some states even at the two-digit industry level.

Figure 9 gives the results for the same two, four and six-digit industry level aggregates as Figure 7 for the new metric given by Equation (7). The results show that the estimates for the two and four-digit industry level aggregates remain close to the

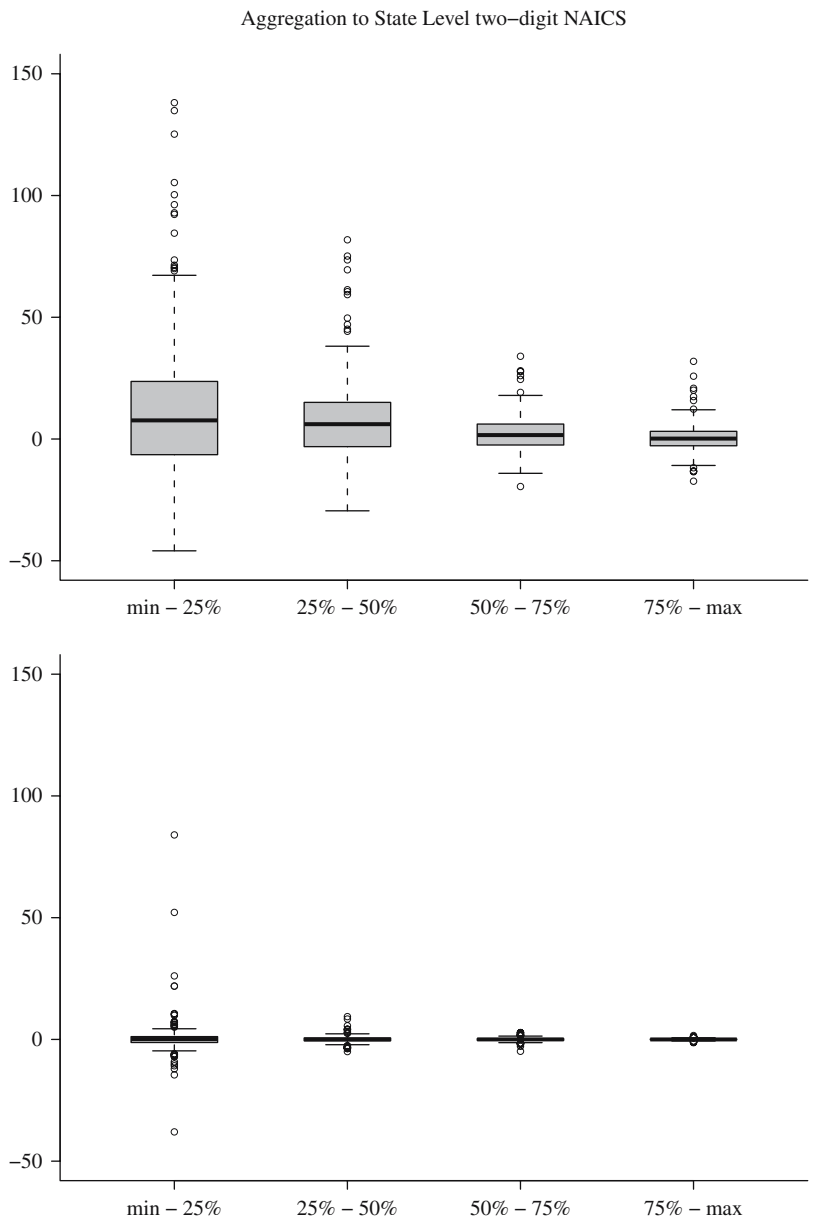


Fig. 8. Boxplots of the percent relative differences for the synthetic values of two-digit industry totals by state compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the two-digit industry state total). The top graph is of quantiles of percent relative errors for the synthetic values using the metric given by Equation (6) while the bottom graph uses metric given by Equation (7).

true values under this new measure. However, as we would expect, since the penalty for not matching industry code at the six-digit level is small, many of the estimates at the six-digit industry level are no longer accurate. This demonstrates that the data provider would only be able to give assurances for marginals being controlled for by the metric. However, as long as the interior of the cells being estimated were large

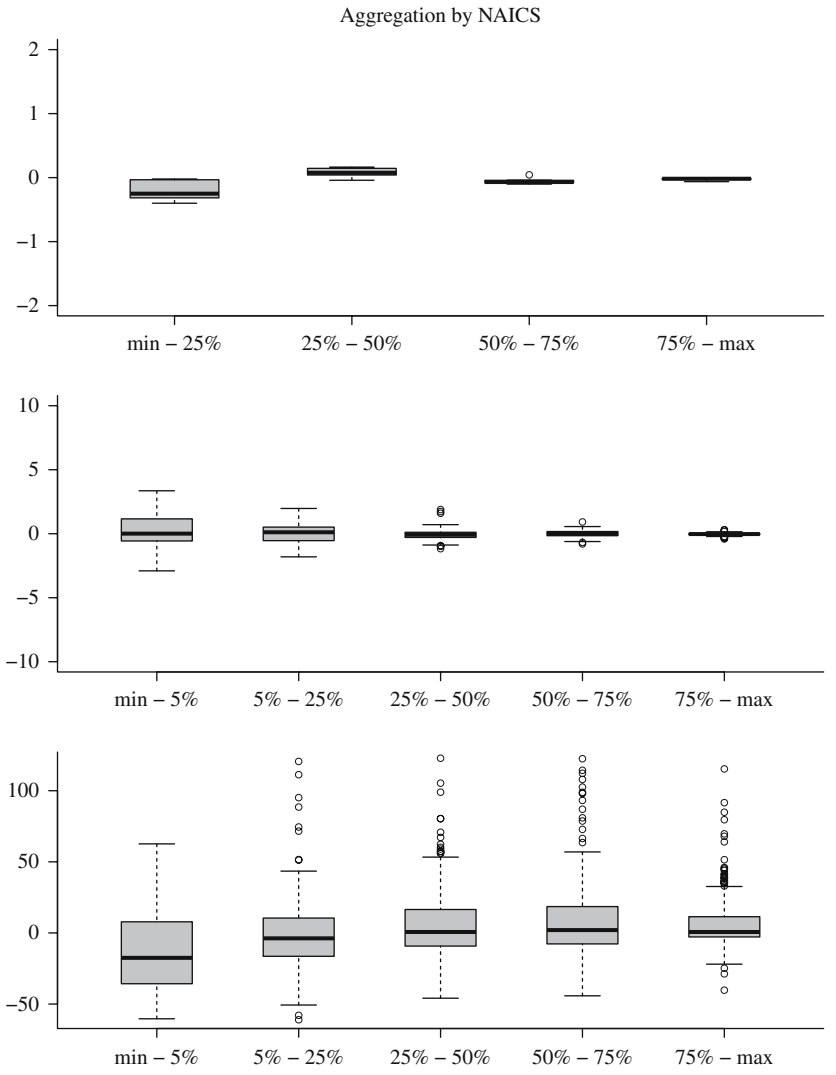


Fig. 9. Boxplots of the percent relative differences for the synthetic values of industry totals compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the industry total). The top graph is of quantiles of errors for all the two-digit industry code totals, the middle graph is of quantiles of errors for all four-digit industry code totals, and the bottom graph is for all six-digit totals. All graphs represent percent relative errors for the synthetic values using the metric given by Equation (7).

compared to its boundary, the estimates produced should be increasingly accurate the larger the cell, as stated in Property 2.1.

4. Discussion

We have introduced a new disclosure limitation method, "data smearing." The method is intended to allow the release of a synthetic data set that can be used to produce accurate contingency tables while protecting the data of individual units. Though the method

was demonstrated on census data in this article, the method will work equally well for sample data.

Unlike other synthetic data approaches, the method focuses on producing accurate contingency tables rather than trying to match the distribution of the original data. The released data for each unit has the intuitive interpretation of representing the average value for the units in the surrounding neighborhood. Neighborhoods are defined by the metric chosen by the agency releasing the data and can be shared with the data users. Importantly, the tables can be user defined after the data set has been released. Additionally, the data from each unit is guaranteed to be protected in that the value assigned to every unit is the average value of at least $n + 1$ units.

We demonstrate the method using QCEW employment data using two different metrics. One metric is defined to connect units within the same six-digit industrial classification that are in close geographical proximity. The second metric tries to connect units within the same state that are in close geographical proximity and have matching industrial classification codes to at least the first four digits. The relative performance of the two metrics shows that the accuracy of a contingency table produced using the synthetic data from this method is highly dependent on the variables included in the metric.

The proposed DLM has performed well during the initial testing on the QCEW data set. It has been shown to produce accurate aggregated cell estimates on cells for which the metric was designed. However, this article attempts only to introduce the method. There is much further testing to be done and properties of the method yet to be explored as well as a number of possible extensions of the method. These and other questions are sure to be the subject of future research.

5. References

Cox, L. 1995. "Network Models for Complementary Cell Suppression." *Journal of the American Statistical Association* 90: 1453–1462.

Evans, T., L. Zayatz, and J. Slanta. 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics* 14: 537–551.

Fuller, W. 1993. "Masking Procedures for Microdata Disclosure Limitation." *Journal of Official Statistics* 9: 383–406.

Graham, P., J. Young, and R. Penny. 2009. "Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models." *Journal of Official Statistics* 25: 245–268.

Holan, S., D. Toth, M. Ferreira, and A. Karr. 2010. "Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality." *Journal of the American Statistical Association* 105: 564–577.

Lambert, D. 1993. "Measures of Disclosure Risk and Harm." *Journal of Official Statistics* 9: 313–331.

Reiter, J. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18: 531–544.

Reiter, J. 2004. "New approaches to data dissemination: A glimpse into the future (?)." *Chance* 17: 12–16.

Reiter, J. and T. Raghunathan. 2007. "The Multiple Adaptations of Multiple Imputation." *Journal of the American Statistical Association* 102: 1462–1471.

Rubin, D. 1993. "Discussion: Statistical disclosure limitation." *Journal of Official Statistics* 9: 462–468.

Wasserman, L. and S. Zhou. 2010. "A statistical framework for differential privacy." *Journal of the American Statistical Association* 105: 375–389.

Yang, M., S. Pramanik, A. Mushtaq, F. Scheuren, M. Buso, S. Butani, and D. Hiles. 2012. "Evaluation of Three Disclosure Limitation Models for the QCEW Program." In Proceedings, Joint Statistical Meeting, American Statistical Association. San Diego, July 28–August 2. 4217–4229.

Received December 2012
Revised September 2014
Accepted September 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 859–864, http://dx.doi.org/10.2478/JOS-2014-0051

Editorial Collaborators

The editors wish to thank the following referees who have generously given their time and skills to the Journal of Official Statistics during the period October 1, 2013 – September 30, 2014. An asterisk indicates that the referee served more than once during the period.

Abbott, Owen, Office for National Statistics, Fareham, Hampshire, UK
Adua, Lazarus, Ohio State University, Columbus, OH, U.S.A.
Aizcorbe, Ana, Virginia Bioinformatics Institute, Arlington, VA, U.S.A.
Al Baghal,Tarek, University of Essex, Colchester, UK
Alsuhail*, Faiz, Statistics Finland, Helsinki, Finland
Altintas, Evrim, University of Oxford, Oxford, UK
Anagnostopoulos*, Christoforos, Imperial College London, London, UK
Aurizio, Leandro D., Bank of Italy, Rome, Italy
Axelson, Martin, Statistics Sweden, Orebro, Sweden
Bakker*, Bart F.M., Statistics Netherlands, The Hague, Netherlands
Banjak, Frans, FHNW, Olten, Switzerland
Banks, Randy, University of Essex, Colchester, UK
Barcena-Martin, Elena, University of Málaga, Málaga, Spain
Basel*, Wesley, Census Bureau, Washington, DC, U.S.A.
Baskin, Robert, Agency for Health Care, Rockville, MD, U.S.A.
Bavdaž, Mojca, University of Ljubljana, Ljubljana, Slovenia
Beckers, Tilo, University of Düsseldorf, Düsseldorf, Germany
Bergdahl, Heather, Statistics Sweden, Örebro, Sweden
Berger, Yves G., University of Southampton, Southampton, UK
Bianconcini, Silvia, University of Bologna, Bologna, Italy
Bilgen, Ipek, University of Chicago, Chicago, IL, U.S.A.
Bishop, Glenys R., Australian National University, Acton, Australian Capital Territory, Australia
Biffignandi, Silvia, University of Bergamo, Bergamo, Italy
Blom, Annelies G., University of Mannheim, Mannheim, Germany
Bolger, Niall, Columbia University, New York, NY, U.S.A.
Boostra*, Harm Jan, Statistics Netherlands, Heerlen, Netherlands
Bouchard*, Martin, Simon Fraser University, Burnaby, British Columbia, Canada
Brick, J. Michael, Westat, Rockville, MD, U.S.A.
Brown, James, University of Technology, Sydney, Australia
Bruil, Arjan, Statistics Netherlands, Den Haag, Netherlands
Bryant,John Robert, Statistics New Zealand, Christchurch, New Zealand
Buelens*, Bart, Statistics Netherlands, Heerlen, Netherlands
Buono, Dario, Eurostat, Luxemburg, Luxemburg
Cantor*, David, Westat, Inc., Rockville, MD, U.S.A.
Carley-Baxter, Lisa, Research Triangle Institute, Research Triangle Park, NC, U.S.A.
Carton, Ann, Vlaamse Overheid, Brussels, Belgium
Chandra, Hukum, Indian Agricultural Statistics Research Instt, New Delhi, India
Chang, LinChiat, San Francisco, CA, U.S.A.

Gubman, Yury, Central Bureau of Statistics, Jerusalem, Israel
Gulyá, Ágnes, Canterbury Christ Church University, Canterbury, UK
Haraldsen, Gustav, Statistics Norway, Kongsvinger, Norway
Hawala, Sam, U S Census Bureau, Washington, DC, U.S.A.
Haziza, David, University of Montréal, Montreal, Canada
Hedlin*, Dan, Stockholm University, Stockholm, Sweden
Hitchcock, David B., University of South Carolina, Columbia, SC, U.S.A.
Hochguertel,Tim, Federal Statistical Office, Wiesbaden, Germany
Hoffmeyer-Zlotnik, Jürgen, Justus-Liebig-University Giessen, Ludwigshafen, Germany
Hogan*, Howard R., U.S. Census Bureau, Washington, DC, U.S.A.
Holbrook*, Allyson L., University of Chicago, Chicago, IL, U.S.A.
Hua*, Jianjun, Dartmouth College, Hanover, NH, U.S.A.
Hundepool, Anco, Statistics Netherlands, Voorburg, Netherlands
Imai*, Kosuke, Princeton University, Princeton, NJ, U.S.A.
Israel, Glenn D., University of Florida, Gainesville, FL, U.S.A.
Jäckle, Annette, University of Essex, Colchester, UK
Janssen*, Eric, OFDT, La Plaine Saint Denis, France
Johnson, Timothy P., University of Illinois, Chicago, IL, U.S.A.
Jones*, Jacqui, Office for National Statistics, Newport, UK
Junker, Christoph, Federal Statistical Office, Neuchâtel, Switzerland
Kao, Fei-Fei, Ming Chuan University, Gwei-Shan, Taoyuan County,Taiwan
Keefe, Christine O, CSIRO, Clayton South, Australia
Kennedy, Courtney, Abt SRBI, Massachusetts, U.S.A.
Kenett*, Ron S., KPA, Raanana, Israel
Kennickell, Arthur B., Federal Reserve Board, Washington, DC, U.S.A.
Khan*, M.G.M., University of the South Pacific, Suva, Fiji
Kiesl, Hans, University of Regensburg, Regensburg, Germany
Kim*, Jae-Kwang, Iowa State University, Ames, IA, U.S.A.
King, Thomas, Newcastle University, Newcastle, UK
Kinney, Satkartar, NISS, Research Triangle Park, NC, U.S.A.
Kirchner, Antje, Institute for Employment Research, Nuremberg, Germany
Kirkendall, Nancy, The committee on National Statistics, Washington, DC, U.S.A.
Klein, Martin, U.S. Census Bureau, Washington, DC, U.S.A.
Knies, Gundi, University of Essex, Colchester, UK
Knoef, Marike, Leiden University, Leiden, Netherlands
Knottnerus, Paul, Statistics Netherlands, The Hague, Netherlands
Koerner, Thomas, Statistisches Bundesamt, Wiesbaden, Germany
Koyuncu, Nursel, Hacettepe University, Ankara, Turkey
Kozak, Marcin, Warsaw Agricultural University, Warsaw, Poland
Krumpal, Ivar, University of Leipzig, Leipzig, Germany
Larraz, Beatriz, University of Castilla-La Mancha, Castilla-La Mancha, Spain
Lee, Geoff, Canberra, Australia
Lee, Hyunshik, Westat, Inc., Rockville, MD, U.S.A.
Lee, Sunghee, University of Michigan, Ann Arbor, MI, U.S.A.
Leon, Carlos, Statistics Canada, Ottawa, Ontario, Canada
Lewis*, Daniel, Office for National Statistics, Newport, UK
Li, Feng, Stockholm University, Stockholm, Sweden
Liao*, Dan, RTI International, Washington, DC, U.S.A.

Pavillon, Gerard, Inserm, Le Kremlin Bicëtre, France
Perron, Francois, University of Montreal, Montreal, Canada
Pforr, Klaus, Leibniz Institute for the Social Sciences, Mannheim, Germany
Piersimoni, Federica, ISTAT, Rome, Italy
Pinter, Robert, Corvinus University of Budapest, Hungary
Pratesi, Monica, University of Pisa, Pisa, Italy
Preston, John, Australian Bureau of Statistics, Fortitude Valley, Brisbane, Australia
Proietti, Tommaso, University of Rome, Rome, Italy
Qualité, Lionel, University of Neuchâtel, Neuchâtel, Switzerland
Ralphs*, Martin, Office for National Statistics, London, UK
Ranalli, M. Giovanna, University of Perugia, Perugia, Italy
Read, Janet, University of Central Lancashire, Preston, UK
Reist, Benjamin, U.S. Census Bureau, Washington, DC, U.S.A.
Reiter, Jerome P., Duke University, Durham, NC, U.S.A.
Rendtel, Ulrich, Freie Universität Berlin, Berlin, Germany
Rey del Castillo, Pilar, Eurostat, Luxemburg
Rivest, Louis-Paul, University of Laval, Quebec City, Quebec, Canada
Robison*, Edwin L., Bureau of Labor Statistics, Washington, DC, U.S.A.
Rocchetti, Irene, Istat, Rome, Italy
Rodgers, Willard L., University of Michigan, Ann Arbor, MI, U.S.A.
Safir, Adam, U.S. Bureau of Labor Statistics, Washington, DC, U.S.A.
Sakshaug, Joseph W., University of Michigan, Ann Arbor, MI, U.S.A.
Sala, Emanuela, University of Milano-Bicocca, Milano, Italy
Salgado*, David, National Statistical Institute, Madrid, Spain
Salvati*, Nicola, University of Pisa, Pisa, Italy
Sampath, S., University of Madras, Chennai, India
Sánchez-Fernández, Juan, University of Granada, Granada, Spain
Scanu, Mauro, ISTAT, Rome, Italy
Scheuren, Fritz J., NORC, Alexandria, VA, U.S.A.
Schmid, Timo, Freie Universität Berlin, Berlin, Germany
Schober, Michael F., New School for Social Research, New York, NY, U.S.A.
Schofield, Lynne, Swarthmore College, Swarthmore, PA, U.S.A.
Scholtus, Sander, Statistics Netherlands, The Hague, Netherlands
Schonlau, Matthias, University of Waterloo, Waterloo, Ontario, Canada
Shabbir, Javid, Quaid-i-Azam University, Islamabad, Pakistan
Sikkel, Dirk, Sixtat, Leidschendam, Netherlands
Silver, Mick, IMF, Washington, DC, U.S.A.
Singer, Eleanor, University of Michigan, Ann Arbor, MI, U.S.A.
Singh, Avi C., University of Chicago, Chicago, IL, U.S.A.
Sinha, Bimal, University of Maryland, Baltimore, MD, U.S.A.
Sinibaldi, Jennifer, Institute for Employment Research, Nuremberg, Germany
Skinner*, Christopher J., London School of Economics and Political Science, London, UK
Slud, Eric V., University of Maryland, College Park, MD, U.S.A.
Smith, David D., Tennessee Technological University, Cookeville, TN, U.S.A.
Smith, Paul A., University of Southampton, Southamton, UK
Smith, Peter W.F., University of Southampton, Southampton, UK
Spreen, Marinus, Applied University Stenden, Leeuwarden, Netherlands
Steorts, Rebecca, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 865 – 868, http://dx.doi.org/10.2478/JOS-2014-0052

Index to Volume 30, 2014

Contents of Volume 30, Numbers 1–4

Author Index

Book Reviews

Printed in December 2014