Journal of Official Statistics
vol. 34, 3 (septiembre 2018)

Responsive and Adaptive Design for Survey Optimization ... p.581
Asaph Young Chun, Steven G. Heeringa and Barry Schouten

A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census. ... p. 599
Andrew Keller, Vincent T. Mule, Darcy Steeg Morris and Scott Konicki

Transitioning a Survey to Self-Administration using Adaptive, Responsive, and Tailored (ART) Design Principles and Data Visualization ... p. 625
Joe Murphy, Paul Biemer and Chip Berry

A Study of Interviewer Compliance in 2013 and 2014 Census Test Adaptive Designs ... p. 649
Gina Walejko and James Wagner

Estimation of True Quantiles from Quantitative Data Obfuscated with Additive Noise ... p. 671
Debolina Ghatak and Bimal Roy

Accounting for Spatial Variation of Land Prices in Hedonic Imputation House Price Indices: a Semi Parametric Approach ... p. 695
Yunlong Gong and Jan de Haan

Accounting for Complex Sampling in Survey Estimation: A Review of Current Software Tools ... p. 721
Brady T. West, Joseph W. Sakshaug and Guy Alain S. Aurelien

Generalized Method of Moments Estimators for Multiple Treatment Effects Using Observational Data from Complex Surveys ... p. 753
Bin Liu, Cindy Long Yu, Michael Joseph Price and Yan Jiang

Book Review ... p. 785-796

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 581–597, http://dx.doi.org/10.2478/JOS-2018-0028

Responsive and Adaptive Design for Survey Optimization

Asaph Young Chun[1], *Steven G. Heeringa*[2], *and Barry Schouten*[3]

We discuss an evidence-based approach to guiding real-time design decisions during the course of survey data collection. We call it responsive and adaptive design (RAD), a scientific framework driven by cost-quality tradeoff analysis and optimization that enables the most efficient production of high-quality data. The notion of RAD is not new; nor is it a silver bullet to resolve all the difficulties of complex survey design and challenges. RAD embraces precedents and variants of responsive design and adaptive design that survey designers and researchers have practiced over decades. In this paper, we present the four pillars of RAD: survey process data and auxiliary information, design features and interventions, explicit quality and cost metrics, and a quality-cost optimization tailored to survey strata. We discuss how these building blocks of RAD are addressed by articles published in the 2017 JOS special issue and this special section. It is a tale of the three perspectives filling in each other. We carry over each of these three perspectives to articulate the remaining challenges and opportunities for the advancement of RAD. We recommend several RAD ideas for future research, including survey-assisted population modeling, rigorous optimization strategies, and total survey cost modeling.

Key words: Responsive design; adaptive design; survey errors; survey costs; optimization; paradata; auxiliary data; total survey error.

1. Introduction

For over a century, survey researchers have faced continual change in the essential conditions under which surveys are designed and data collection is conducted. Time and again, survey practitioners have “responded” or “adapted” to arising challenges and opportunities they faced using innovations in the statistical design and analysis of their studies or in the methods used to collect study data. Many of the changes in the essential conditions of surveys have been clearly recognized as transformational in the field: the idea that samples could represent a population; a theory of inference based on probability samples; the formalization of household sampling and survey methods; the advent of RDD telephone surveys; computerization of survey design, data collection and analysis;

[1] U.S. Census Bureau. 4600 Silver Hill Road, Washington, D.C., U.S.A. Email: Asaph.Young.Chun@census.gov.
[2] University of Michigan, 500 S. State Street, Ann Arbor, MI 48109, U.S.A. Email: sheering@umich.edu.
[3] Statistics Netherlands, Henri Faasdreef 312, CBS-weg 11, 2492 JP The Hague, The Netherlands. Email: jg.schouten@cbs.nl.

Acknowledgments: We would like to extend our thanks to the contributors to this Special Section on Responsive and Adaptive Design for their dedicated work and to all peer reviewers for their critical review. We are also so grateful to the JOS editorial staff who provided essential support.

Disclaimer: The views expressed are the authors’ and not necessarily those of the institutions with which the authors are affiliated.

and internet and web-based data collection. In addition to benefitting from these transformational opportunities, survey researchers have been increasingly challenged by several inter-related and chronic trends – the increasing number and complexity of surveys, rising costs of traditional survey designs and methods, and steady decline in respondent participation rates.

In light of these changes and opportunities to seize, the remainder of Section 1 describes the genesis and evolution of responsive and adaptive design (RAD), including the four pillars of RAD that are essential to survey design. Section 1 closes by presenting the overarching research questions addressed by articles published in the 2017 JOS special issue and this special section. Section 2 presents three perspectives, discussing how core elements of RAD are addressed by these articles. Section 3 outlines what open research questions and challenges remain, following each of three perspectives as presented in Section 2. Sections 4 and 5 include a discussion of future directions for RAD and conclusions. This discussion article is a companion to the article by Chun et al. (2017) providing an overview of the 2017 JOS special issue devoted to RAD.

1.1. The Notion and Evolution of Responsive and Adaptive Design

The idea of responsive and adaptive design (RAD) arose in response to the survey challenges that were enumerated at the outset of our discussion. By way of definition, RAD is a data-driven scientific approach to controlling survey design features in real-time data collection by monitoring explicit costs and errors of survey estimates that are informed by auxiliary information, paradata, and multiple sources of data; RAD works toward a goal of survey optimization based on cost-error tradeoff analysis and evidence-driven design decisions, including the most efficient allocation of resources to survey strata. The concept of using RAD for conducting surveys is not new. Some clear antecedents include survey practices like replication (i.e., phases) in sample release, the embedding of experiments in survey data collections, and double sampling (two-phase sampling). These practices are inclusive of subsampling for nonresponse (Hansen and Hurwitz 1946), sequential analysis or adaptive trials (Wald 1947), and total survey error and total quality management (Morganstein and Marker 1997).

Groves (2011) described the three decades spanning 1960–1990 as the “Era of Expansion” in the application of survey methods. He cites Eleanor Singer, who labeled this same period the Golden Era of survey research. Many of us who lived and worked through the day-to-day challenges of that time might view it more as “gilded and shiny” than truly golden; it was an era of stability – scientifically designed surveys, using highly standardized and uniform methods, resulted in high-quality surveys with relatively high response rates and acceptable cost structures. Late in this period of relative stability, the challenge of rising costs for large scale in-person data collections was buffered by the advent of new telephone survey methodology – a buffer that remained effective through at least the late 1990s.

By the new millennium, the continual change in key survey conditions presented survey researchers with new challenges. Scientific and government surveys became more complex and often posed great uncertainty in design parameters and operational features. Survey populations’ resistance to survey participation continued to increase. Survey cost structures were becoming even more dependent on decisions being made in the field or data collection centers, often with no evidentiary basis to measure or respond to cost fluctuations. Cost metrics, which are inherently multi-dimensional, remained as elusive as data quality. As these challenges grew, new opportunities also arose. Due to advances in computing and technology, there was improved access to sample frames, administrative data sources, global positioning (GPS), and geographic information systems (GIS).

Groves and Heeringa (2006) coined the term "Responsive Design," following the oft-quoted advice of their mentor, Leslie Kish, who said, "If you want recognition for an idea, put a proper name on it" (e.g. "Design Effect"). Leslie's advice in this regard was sage. The term "responsive design" stuck as did the companion term, "adaptive survey design." As a catalyst for the initial growth of RAD, foremost among the new opportunities emerging in the first years of the new millennium were sophisticated new "real time" systems for sample management, data acquisition, and paradata capture in interviewing systems.

New developments in survey designs and methods can often take a decade or more to develop and mature with respect to research, publication, and applications in the field. This was certainly true for RAD. Initial RAD applications often emphasized simple nonresponse subsampling features to address the cost and effort of "end game" data collection at the conclusion of the survey period (Tourangeau et al. 2016). With time, the breadth and sophistication of RAD research and applications have expanded, resulting in the diverse body of knowledge and experience that is in evidence in publications such as the 2017 JOS special issue, this 2018 JOS special section, and a book recently published by Schouten et al. (2017).

RAD, in youth slang, means wonderful, fantastic, or extraordinary. We observe the rise of variants of RAD ideas, turning these ideas into survey practice in various contexts. RAD seems to be coming of age since the explicit implementation of responsive design or adaptive design during the mid-2000s (e.g., Groves and Heeringa 2006; Wagner 2008).

1.2. Four Pillars of Responsive and Adaptive Design

Though precedents and variants of RAD have been embedded in survey practice over decades (Groves and Heeringa 2006; Wagner 2008; Calinescu and Schouten 2016), we argue that RAD has four pillars for constructing survey design: 1) use of auxiliary information to stratify the heterogeneous population under study, 2) design features and interventions to adapt treatment, 3) explicit quality and cost metrics and functions to evaluate the efficacy of adaptation to strata, and 4) a quality-cost optimization strategy to find optimal allocations of treatments to strata. RAD is essentially a form of adjustment by *design* in the data collection as opposed to adjustment by estimation, that is adjustment introduced in the design and data collection stage in contrast to adjustment in the estimation stage. As a consequence, similar to estimation, auxiliary data should relate to nonresponse and other sources of survey errors under investigation, as well as to the key survey variables. Design features should be effective in reducing survey errors for the relevant strata. Quality and cost functions quantifying effort and errors should be properly defined and measurable, but, above all, should be accepted by the stakeholders involved.

The quality-cost optimization strategy should be transparent, reproducible, and easy to implement.

The first two pillars – auxiliary data and design features – emphasize data collection and a behavioural social sciences component, whereas the other two pillars – cost/quality metrics and optimization – are tied to estimation and statistics component. Between 2000 and 2015, there was renewed interest in paradata, or auxiliary data coming from the data collection process (e.g., Kreuter 2013). For example, call record data, audit trails, and interviewer observations were increasingly used in dashboards to monitor data collection. This might have resulted from increasing digitization of communication. The real-time paradata were instrumental to developing evidence-driven models to understand the process of response and nonresponse and to creating statistical interventions to control for potential nonresponse bias.

Survey design features obviously go as far back as surveys themselves; however, there has been renewed interest in mixed-mode surveys with the emergence of online devices (e.g., Dillman et al. 2014; Klausch 2014). The survey mode appears to be the strongest quality-cost differential of all design features. Between 2005 and present, various articles have been published about indicators for nonresponse (e.g., Chapter 9 in Schouten et al. 2017). It has been declining response rates, we observe, that drove the development of alternative indicators; not necessarily to replace response rates but to supplement them and to provide a more comprehensive picture of data quality. Notable in data quality metrics is the development of response propensity measures (e.g., Chun 2009; Chun and Kwanisai 2010; Toureangeau et al. 2016). It is unfortunate that efforts to develop and implement cost metrics remain quite limited – probably due to practical constraints of quantifying or modelling cost parameters.

Optimization strategies remain an underexplored area. This may be, in part, because they are the final step of RAD. In other words, they require that choices in the other elements have been made and implemented. For instance, a consensus is necessary on quality and cost indicators. We observe, however, that it is also because optimization requires accurate estimates of survey design parameters, such as response propensities and survey costs. Survey cost metrics are multi-dimensional like data quality; optimization strategies, therefore, remain incomplete as long as cost estimates as input variables are neither reliable nor valid indicators of survey costs.

1.3. Overarching Research Questions

We present that the overarching research questions addressed by the 2017 JOS special issue and the 2018 JOS special section are as follows: 1) what approaches can be used to guide the development of cost and quality metrics in RAD and their use over the survey life cycle? 2) which methods of RAD are able to identify phase boundaries or stopping rules that optimize responsive designs? and 3) what would be best practices for applying RAD to produce high quality data in a cost-effective manner?

In response to these core questions of RAD, the JOS special issue and special section sought to address the following topics of adaptive design: theoretical contributions and applications, innovations, and comparisons of different methods of adaptive design that leverage the strengths of administrative records, big data, census data, and paradata as well as survey response data. For instance, what cost-quality tradeoff paradigm can be operationalized to guide the development of cost and quality metrics and their use around the survey life cycle? Under what conditions can administrative records or big data be *adaptively* used to supplement survey data collection? How are paradata in multi-mode data collection conceptualized, pretested and collected to inform survey design decisions?

The articles included in the JOS special issue and special section address *interdisciplinary* dimensions of adaptive design, which encompass the following survey drivers: cost, response burden, data quality such as representativeness and response propensity, multiple sources of data, multiple modes of data collection, paradata, and new technologies. For instance, what indicators of data quality can be combined to monitor the course of the data collection process? Under what scenarios can the rules of switching from one mode to another be cost-effective? What stopping rules of data collection can be used across major phases of the survey life cycle?

We reiterate that the JOS articles involve experimental designs or simulations of adaptive design in household surveys, business surveys, and censuses. For instance, how could adaptive design be effectively designed and executed, especially in surveys involving multiple data sources and mixed modes of data collection? How could adaptive design guide web surveys while controlling for multiple sources of survey errors, such as nonresponse, measurement errors, and sampling errors?

2. Critical Pillars of RAD Addressed in the JOS Special Issue and Special Section

In this section we present three perspectives, discussing how the four pillars of RAD that are essential to survey design are addressed by articles published in the 2017 JOS special issue and in this 2018 JOS special section. Perspective A presents discussion points by leveraging the four pillars of RAD above. Perspective B articulates five key elements of RAD, or variants of the four pillars of RAD, to make a coherent discussion. Perspective C focuses on elaborating on cost measures and cost modeling, the missing half of cost-quality tradeoff analysis and optimization strategy, as tied to the third and fourth pillars of RAD.

2.1. Perspective A

Looking at the special issue and special section, overall, the articles lean towards the more statistical pillars: indicators and optimization. This is understandable as it is less reliant on costly experiments or pilots and more on feasible simulations. These articles are well written, useful, and creative in articulating contributions that introduce new perspectives and approaches to the existing literature. As mentioned, considering the optimization strategy, more specifically the translation of quality-cost to intervention and adaptation, there has been a gap in the literature. The advances in this direction are very welcome in the special issue and section. It is, to be sure, a pleasant surprise to see scholars from a variety of survey settings work on these methods.

Nonetheless, the greater statistical focus is also somewhat of a missed chance. This is for three reasons. First, being more statistical, the articles often employ simulations to demonstrate utility rather than real applications, which will bring problems of their own. In many cases the simulations do have a link to real surveys, but not always. Such case studies may not convince survey designers and data collection departments, as they ignore practical and logistical constraints. Furthermore, the outcomes of case studies appeal less to experience about achievable quality and costs. Second, in the end, we need effective auxiliary variables – design feature combinations to find differentiation and leverage to adapt. RAD works only if we manage to find strategies that work better than what we have done before or proof that for some strata we need to spend more. And third, the more statistically oriented articles tend to be further removed from data collection personnel and, as such, are more difficult to implement. They need translation to daily practice. This holds true for the article by Burger et al. (2017) about the robustness of RAD to inaccurate design parameters. A few articles are exceptions, especially the article authored by the US Census Bureau researchers about the Annual Survey of Manufactures (Thompson and Kaputa 2017) and the article by NASS about the Crops APS survey (McCarthy et al. 2017); these articles are much closer to implementation. They do use realistic auxiliary variable – design feature combinations and seem to be driven by a practical need to improve uniform survey designs.

Thus, the special issue and special section represent strong articles that advance the statistical foundation, but are somewhat removed from survey practice. In Chapter 12 of Schouten et al. (2017), nine areas are enumerated and discussed that require progress and further research. The first area is that of empirical evidence that RAD works, or in other words, success stories. These are still relatively thin; see also the discussion of Tourangeau et al (2016). In the special issue, two of the original authors, Brick and Tourangeau (2017), explore and discuss how success stories may be achieved. It would be of great benefit if the authors of the other articles return with follow-up articles describing the spin-off and results of their future work. It is strongly recommended that they do.

2.2. Perspective B

At the risk of misclassification or over-simplification, the articles published in the 2017 JOS special issue and included in this 2018 supplement address five key elements of RAD, or variants of the four pillars as articulated above.

A first and foundational element of any RAD approach is the recognition by the research team that the survey population is *heterogeneous* with respect to the orientation to the survey topic, incentive for survey participation and preference concerning the timing and mode of data collection. Equally important to acknowledging this heterogeneity is the ability to predict where it occurs in the population so that RAD features can be tailored accordingly either prior to or during the survey data collection. In the 2017 special issue, the articles by Kaminska and Lynn (2017); Durrant et al. (2017): Thompson and Kaputa (2017) and McCarthy et al. (2017) address topics related to this element.

The feasibility and effectiveness with which features can be operationalized and actively managed during data collection represent a second key feature of any RAD survey. An “elegant” design may have tremendous appeal but is of little use if it does not work when put into operation. In the 2017 special issue, the articles by Vandenplas et al. (2017); Early et al. (2017); Plewis and Shlomo (2017), and Burger et al. (2017) address challenges in operationalizing and managing desired features of RAD designs. In this 2018 special section, the article by Walejko and Wagner and the article by Murphy and his coauthors both address operationalization and survey management and monitoring issues encountered in tests of RADs for tailored designs: the former for the 2020 U.S. Census tests and the latter for the U.S. Energy Information Agency's Residential Energy Consumption Survey.

Even in today's world where RAD concepts are widely accepted and practiced, we as survey practitioners find it hard not to focus all possible efforts on maximizing the response to the survey. It is in our genes. Consequently, we struggle with the RAD concept that a rigorously applied survey protocol will ultimately reach a phase capacity at which additional effort and expenditures will not add any significant information content to the data that have been collected. Even more foreign to our traditional view is the idea that the design itself has reached the point at which further effort should be stopped. But if we can get over that hurdle, how do we as practitioners decide when a phase has reached capacity or the RAD study data collection should stop? Lewis (2017) and Paiva and Reiter (2017) both in the 2017 special issue describe quantitative tools that can guide phase transition or stoppage in RAD data collections.

As noted above, a principal aim of RAD is to achieve an optimal balance of cost and errors in survey populations where individuals' orientation to the survey request, incentive to participate, or data collection preferences vary. To fully achieve this aim, there must be reliable metrics for assessing both costs and errors. Nonresponse and associated selectivity in the composition of the observed population sample are a potentially important source of bias in the survey data. But nonresponse bias can be difficult to quantify, especially for surveys where the sample frame provides little information on the characteristics of respondents and nonrespondents. In the 2017 special issue, Brick and Tourangeau (2017) present models for survey nonresponse and investigate just how effective responsive designs might be at attenuating the bias associated with those nonresponse mechanisms. Särndal and Lundquist (2017) investigate whether actively controlling the "balance" of the observed sample during the RAD data collection should be preferred to standard methods in which post-survey calibration weighting adjustments are used. Closely related to the topic of weighting calibration using sample frame and administrative data is the option to use large scale administrative data sources as a substitute for direct survey or census enumeration. In this special section, Keller et al. describe a U.S. census investigation into the metrics for evaluating when external administrative data may be a suitable substitute for assigning vacancy status to addresses in the forthcoming 2020 Census enumeration.

2.3. Perspective C

Going forward, the litmus test of RAD success depends heavily on the extent to which the third and fourth pillars of RAD are properly assembled and tested against the pressure of total survey errors and total survey costs – both anticipated and unanticipated. The critical gap remaining in these two pillars of RAD is more due to under-development of the framework of cost metrics and lack of its implementation in real-world survey applications. Cost-quality optimization, by definition, would suffer inasmuch as cost metrics are not properly implemented. The underlying questions to ask include but are not limited to the following fundamentals: What are measurable survey costs? What would be desirable properties of cost modeling? What methods are available and feasible to measure survey costs, inform cost-quality tradeoff analysis, and develop cost-quality optimization strategies?

Costs and errors are reflections of each other; increasing one tends to reduce the other (Groves 1989). Thus cost-quality optimization strategies would be neither feasible nor complete unless there is rigorous development and examination of the cost functions of various survey designs that offer error properties (Groves 1989; Chun 2012; Mulry and Spencer 2012). It bears reminding that a viable cost model is a function of fixed costs and variable costs as follows (Groves 1989):

Total Cost = Fixed Costs + Variable Stratum Costs

$$C = C_0 + \sum_{1}^{H} C_h n_h$$

Where C = total cost;

C_0 = fixed cost, incurred regardless of selected sample size;
C_h = variable stratum cost of the n_h sample cases in the hth stratum, namely cost of selecting, measuring, and processing each of the n_h sample cases in the hth stratum.

Fixed costs are costs that remain fairly constant in a survey, such as costs for survey system design, IT, and survey management. Variable costs are costs that vary as a function of the sample cases in various strata. Variable costs may include costs of frame construction, interviewing, nonresponse followup, data entry, and editing, which incur over the survey life cycle.

In practice, the pragmatic cost models need to be inclusive of *nonlinear, discontinuous and stochastic* properties of survey costs (Felligi and Sunter 1974; Groves 1989). They deserve discussion. Groves observes that existing cost models tend to be linear functions of survey parameters like the number of interviews, although nonlinear cost models often apply to practical survey administration. Most cost models are continuous in those parameters; however, he points out that discontinuities in costs often arise when administrative changes accompany certain design changes. While cost models tend to be deterministic, costs can vary extensively because of chance occurrences in probability sample selection, or choice of interviewers. Groves argues that because closed-form solutions to complex design problems do not exist, simulation approaches are useful to measure the sensitivity of results to changes in various design, cost, and error parameters. He offers several simulation examples of cost and error models that demonstrate that gathering better cost data must be given priority in order to develop reasonable cost models accounting for cost-error tradeoff.

The cost models proposed by Groves remain useful and viable today. Cases in point are the articles by Paiva and Reiter (2017) and Kaminska and Lynn (2017) in the 2017 JOS special issue and by Murphy and his colleagues in this special section. Using data from the 2007 U.S. Census of Manufactures, Paiva and Reiter show how to compute and compares measures of cost for various sample sizes by applying the traditional cost model. Kaminska and Lynn provide and test explicit cost metrics to determine pros and cons of alternative methods for allocating sample elements to data collection protocols, particularly in a longitudinal survey setting. Extending the cost model by Groves, Kaminska and Lynn demonstrate how variants of adaptive and non-adaptive designs can be appraised in terms of relative costs as well as multiple measures of data quality for each proposed scenario of RAD. In a discussion of adaptive, responsive, and tailored (ART) design principles, Murphy and his colleagues make a smart move of presenting relative cost per case by interview protocol. They also provide data visualization of percentage of cases requiring editing, one that is tailored to the needs of cost metrics in an energy consumption survey sponsored by the U.S. Energy Information Administration. None of these articles, however, has taken a major step yet towards nonlinear, discontinuous, and stochastic properties of cost modeling.

3. What Open Research Questions and Challenges Exist for Implementing RAD?

Following each of the three perspectives as presented above, we turn to discussing what major questions and challenges remain to be addressed for advancing RAD.

3.1. Perspective A

In line with Perspective A as articulated above, the other eight areas of challenges addressed by Schouten et al. (2017) are as follows: 2) best practices for implementation, 3) clear and flexible quality-cost objectives, 4) versatile data collection systems, 5) skillsets for data collection staff, 6) relevant designed paradata, 7) application to longitudinal settings, 8) a total survey error approach, and 9) optimization strategies.

Areas 2 to 6 all relate to prerequisites for actual implementation, some of which are methodological and some of which are logistical and IT-related. They do not translate directly to academic research questions but do pose very interesting challenges for which there is an outlet in more practical journals and in conference proceedings. The real challenge here is to bridge the gap between theory and practice. RAD, perhaps more than ever before, has strong implications for how surveys are actually done. In its most strict form, RAD prescribes what sample units get what treatments to optimize what specified objectives under what decision rules. This has traditionally been the mandate of data collection departments and staff, and not of methodology. In order to get closer to implementation, however, data collection staff need to become co-researchers and co-authors. The case studies need to be driven by real-life issues with decreasing response and increasing costs. Areas 7 and 8 present new settings and applications that are mostly unexplored. Here, research questions could easily be formed and the attendant challenges are very exciting but also complicated.

How to implement RAD in longitudinal settings is a very interesting avenue to explore. The Kaminska and Lynn (2017) special issue article is one of the first to dive into this area. In panels, there are rich data about respondents, obviously, but also new challenges such as attrition, panel refreshment and conditioning. How informative is the previous-wave survey data about participation and measurement data quality? Is their explanatory power strong enough to overcome time lags and attrition? Are panel respondents consistent in terms of participation, costs and measurement, that is do they show the same behaviour in subsequent waves? How can RAD be combined with panel refreshment and could RAD be part of panel refreshment strategies? The measurement and answering behaviour is very interesting in terms of RAD optimization. Since longitudinal studies are often about change, how can RAD be embedded over multiple waves?

The other broad research area is total survey error. RAD has mostly had a nonresponse perspective, probably because it has been driven by declining response rates and increasing costs. However, the most powerful design feature, the survey mode, has impacts on all survey errors. Nowadays, many survey designs are sequential mixed-mode and go from cheaper to more expensive modes – the rationale being that part of the sample is empathic to surveys and will respond under all strategies. RAD optimization, then, concerns decisions about the allocation of more expensive modes. The obvious questions are whether measurement is equivalent and whether a possible gain in participation is offset by a loss in comparability; and in RAD terms, whether these questions are answered differently for different sample subgroups. This is a discussion that goes beyond that of the mandate of data collection departments, as questionnaire content and survey estimates are typically produced by substantive departments. Similar total survey error impact may come from other design features such as interviewer allocation, split-questionnaire designs, central question follow-up procedures, and mobile devices. This will become even more prominent when survey data are combined with big data or mobile device sensor data into hybrid forms of data collection. RAD must, therefore, have a total survey error view. In such a setting, the number of quality indicators and constraints may increase or may require experimental components such as repeated measurements or randomization in question ordering.

Area 9, optimization strategies, is a key element of RAD. When posing RAD as a mathematical optimization problem, one finds that the number of possible designs quickly grows to a level beyond the reach of naïve/brute force optimization. The large number of options is not necessarily a problem as long as the optimization problem is (nearly) linear, but the most interesting problems are not linear and, even worse, not convex. These are, generally, complex problems to solve, such that clever and intuitive strategies are needed. Another approach may be to accept that suboptimal designs are good enough as long as they are better than uniform designs. Optimization strategies go hand in hand with strategies to learn and update. Mobile device data collection, for example, may be promising but it is a relatively unknown area. How can we optimize design when promising yet new design features emerge? Most survey designers and survey users are averse to constant change in design and with good reason. So how do we include optimization in terms of time series continuity?

3.2. *Perspective B*

RAD is a design tool that researchers can apply to potentially reduce both costs and errors of a one-off survey or a longer term program of surveys. RAD is not a panacea, capable of solving all problems of nonresponse, budget, or other survey errors. Considerable research and empirical work have demonstrated that not all RAD applications will succeed in optimizing the costs and errors of survey data collections.

There are several researchable questions that the study team should answer before a RAD is considered. First, given what is known from prior or similar surveys, is the survey population truly heterogeneous with respect to the orientation to the survey topic, their incentive for survey participation, and timing or mode of data collection? Is it possible to design and implement adaptive or responsive features in the survey design that are matched to these different motivations, incentives, and preferences? Second, if such heterogeneity is present in the survey population and alternative design features can be identified that are responsive to these differences, can a RAD that incorporates these features be successfully operationalized during the field period? For example, attempts at a multi-phase responsive design for a telephone survey that spans a three-week data collection period will be limited by the time available to implement and evaluate the initial phase before transitioning to a second phase with alternative design features. The time constraint imposed by this same survey might be addressed using an adaptive design with pre-allocated features (e.g., contact materials, mode, and incentives). However, to be effective, such an adaptive design will require information from past experiences or experimental testing to guide the presurvey assignment of alternative design features to individual sample members. Finally, before deciding on a RAD for a future survey, the research team should carefully consider the added implementation and management costs for a RAD design that may entail multiple phases, multiple modes and other variations in design features. Will the fixed costs of the RAD implementation be offset by a reduction in the variable costs of collecting the survey responses? The criterion by which a RAD (or any statistical design) should be judged is that it should minimize mean squared error for key statistical aims subject to the budget and time period allocated for the project. It is relatively easy, given a fixed budget, to construct a RAD-like design that will result in increased nominal sample sizes and possibly even higher weighted overall response rates. However, that same design may be subject to large losses in effective sample size or differential selection biases for total sample and subpopulation estimates.

3.3. Perspective C

When it comes to the development and implementation of cost metrics that realize cost-quality tradeoff analysis and optimization, RAD seems to have a long way to go. Cost estimates, to us as survey practitioners, are day-to-day concerns to take into account in building a reasonable cost-quality tradeoff analysis and in developing optimization strategies that are rigorous enough to design optimal allocations of treatments to the population strata under study. Yet good and best practices of survey cost modelling are quite limited. Articles in the JOS special issue and special section move in the right direction of nailing down cost metrics and integrating cost metrics together with data quality metrics. Cost functions, however, remain to be rudimentary, not reaching the pragmatic cost models that need to be inclusive of nonlinear, discontinuous, and stochastic properties of survey cost estimates.

“Survey researchers have given much less attention to survey cost models than to survey error models,” wrote Groves (1989, 79) three decades ago. Unfortunately, we are facing the same issue today even as survey costs are increasingly becoming the major driver of survey design. The U.S. National Children’s Study stopped in large part due to problems with its design and management, as well as a huge survey cost that had already cost USD 1.2 billion by the time of its termination (Altman et al. 2014). With a life-cycle cost of about USD 13 billion, the 2010 U.S. Census was the most expensive U.S. census in history; it was 56 percent more costly than the 2000 Census in constant 2010 dollars (Government Accountability Office 2015). Costs are the main driver of design changes for the 2020 U.S. Census, spurring innovations, such as the use of administrative records and third-party data, to materialize notable cost savings and sustain data quality (U.S. Census Bureau 2017). Unless the missing half of the cost-quality optimization is rigorously examined and fixed, RAD looks to be facing an uphill battle. What questions and challenges remain to be addressed by the community of RADers?

First, we should probably draw lessons from the Total Survey Error framework, based on which the error function has been well specified and developed over the last few decades. Can we conceive of the total survey cost model? Can we pair total survey cost model to Total Survey Error model? We are pointing to the model that may account for the traditional notion of fixed and variable costs and that is adaptive enough to be tailored to the pragmatic needs of cost modeling – nonlinear, discontinuous, and stochastic in terms of survey cost properties. We should work closely with accountants and field managers who may have the first-hand experience in cost estimating, computing, and cost modeling – top down or bottom up followed by converging cost estimates. Second, we as survey practitioners and survey designers need to maintain and archive cost data together with data quality information and make proprietary cost information available, whenever possible, in the collective interests of cost-quality optimization research. Probably the first step is to steadily maintain and share survey cost data based on a reasonable framework of total survey cost in the vein of total survey error. We, as a collective community of survey cost-quality modelers, should make concerted efforts to implement what is measurable and what is collectable when it comes to fixed and variable costs. These efforts should be followed by collective standardizing, if possible, as the American Association for Public Opinion Research has established standards of various response rates. Finally, cost-quality tradeoff analysts and optimization strategists should answer the question of examining the cost implications of designs that offer different survey error properties as echoed by Groves (1989). Several examples of cost analysis were provided by Groves when it comes to sampling error, nonresponse error and a few of the measurement errors associated with mode of data collection. Studies on total survey cost, cost per interview, and relative cost are on the rise as reported by Wagner et al. (2016). However, research on cost modeling remains thin relative to a coherent framework of cost-error properties.

4. Future Directions for Research and Application in RAD

The articles published in the 2017 JOS special issue and those included in this section are early indicators of the directions that future research and applications of RAD for surveys will be taking. The original concept of responsive design presented by Groves and Heeringa (2006) was in many ways overly structured and formulaic for the wide range of RAD applications that we now see described in the pages of JOS and elsewhere. The diversity of design and methodological developments that now fall under the RAD banner is so much larger than the original ideas. Some of these developments have been such major departures that they earned new labels: adaptive survey design (ASD) and, most recently, adaptive, responsive, and tailored (ART) design (e.g., see Murphy, Biemer, and Berry in this issue). Furthermore, a richer application of RAD has been found, for instance, in a framework of dynamic question ordering (DQO) where question order is adaptively altered to improve response rate and imputation quality (Early et al. 2017).

Without question, the future of survey research will continue to bring new challenges and opportunities. The past decade's development of RAD approaches to population surveys has prepared us to adapt to expected changes in survey and other population data environments. Although the future demand for surveys and primary data collections will remain high, there will certainly be a growing emphasis on data collections to augment existing systems of administrative systems and other structured and unstructured sources of "big data". "Survey-assisted" population modeling, one that integrates large big data systems and streams with carefully designed survey observations, has long been used in the fields of agriculture, forestry and environmental sciences, and small area estimation. It is rapidly being extended to medicine and epidemiology as well as to economics, demography, and other social sciences. In this integrated role, survey data collection will assist in several ways:

> "Model training" – providing timely estimates of models parameters relating the outcomes of interest to the covariate information available in the big data systems;
>
> "Model refinement" – by supplying more complete information on multivariate associations, mediating and moderating effects and chronological or spatial variation in big data models;
>
> "Compensation" – for population noncoverage, nonobservation or missing data in the large data systems;
>
> "Insight" – into the error structure of large scale data systems that can only be obtained through direct survey measurement.

RAD will increasingly be called upon to support systems of model-based estimation, inference and prediction. Research to develop optimal, cost-efficient designs for such applications will need to be developed in the context of the statistical information that is present in existing sources of data and the specific statistical aims of the survey-assisted modeling system. Survey statisticians and methodologists leading this research will need to transfer their general knowledge of RAD principles and total survey error to these problems of survey-assisted modeling. Accountants and field managers proficient in cost estimation and cost modeling should be paired with survey designers and survey methodologists to develop, test and scale up pragmatic cost modeling aligned with a reasonable framework of total survey costs.

5. Conclusions

RAD is not an entirely new concept; tailoring and targeting have been part of survey design for quite some time. But it has never drawn such attention nor possessed such a formal structure. Given the strong cost-quality differential of self-administered versus interviewer-assisted surveys, the rise of all kinds of new devices and forms of communication, the option to form hybrid data collections using sample surveys and big data and the general individualization of societies, we believe it is inevitable that RAD will be a natural component of many data collections. Perhaps this will be in relatively simple and rudimentary forms, but it is not logical to be inflexible and apply a uniform strategy. Furthermore, innovation in communication technology seems to be accelerating, with the likely consequence that it will be a moving target and it will always be different subpopulations that have adopted older and newer forms of communication. RAD principles may also expand to the big data arena by differentiating what sources are used for whom. We hope that the gap between data collection and methodology will become smaller so that data collection experts will not be misunderstood and they will polish the very promising ideas and anticipated yield of RAD to feasible designs.

The development of cost functions that may explicitly account for complex survey design features remains to be pursued by leveraging simulations and experimental studies that control for different cost properties. Cost-quality optimization strategies of RAD may be realistically developed only if cost functions are rigorously designed, tested, and implemented, while multiple metrics of data quality are being further matured.

6. References

Altman, R., P. Pizzo, R. Gibbons, K. Hudson, R. Jenkins, B. Lee, M. Lichtveld, M.L. Miranda, C. Perry, H. Zoghbi, and L. Jorgenson. 2014. *National Children's Study*. Working Group NIH Advisory Committee to the Director Final Report. Bethesda, MD: National Institutes of Health. https://acd.od.nih.gov/documents/reports/NCS_WG_FINAL_REPORT.pdf.

Brick, M. and R. Tourangeau. 2017. "Responsive Survey Designs for Reducing Nonresponse Bias." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 735–752. Doi: https://doi.org/10.1515/jos-2017-0034.

Burger, J., K. Perryck, and B. Schouten. 2017. "Robustness of Adaptive Survey Designs to Inaccuracy of Design Parameters." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 687–708. Doi: https://doi.org/10.1515/jos-2017-0032.

Calinescu, M. and B. Schouten. 2016. "Adaptive Survey Designs for Nonresponse and Measurement Error in Multi-Purpose Surveys." *Survey Research Methods* 10(1): 35–47. Doi: http://dx.doi.org/10.18148/srm/2016.v10i1.6157.

Chun, A.Y. 2009. Nonparticipation of the 12th graders in the National Assessment of Educational Progress: Understanding Determinants of Nonresponse and Assessing the Impact on NAEP Estimates of Nonresponse Bias According to Propensity Models. University of Maryland, College Park, USA. http://hdl.handle.net/1903/9916.

Chun, A.Y. 2012. "What Counts as Group Quarters? – A Glimpse of Census Cost-Data Quality Models." Paper presented at the Joint Statistical Meetings, San Diego, U.S.A. (accessed July 30, 2012).

Chun, A.Y. and M. Kwanisai 2010. "A Response Propensity Modeling Navigator for Paradata." Proceedings of the Survey Research Methods Section of the American Statistical Association, Joint Statistical Meetings, Vancouver, Canada, 356–369. http://ww2.amstat.org/sections/SRMS/Proceedings/y2010/Files/306125_55196.pdf.

Chun, A.Y., B. Schouten, and J. Wagner. 2017. JOS Special Issue on Responsive and Adaptive Design: "Looking Back to See Forward – Editorial." In A.Y. Chun, B. Schouten, J. Wagner (Eds), JOS Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics*, 33(3): 571–577. Available at: https://www.degruyter.com/downloadpdf/j/jos.2017.33.issue-3/jos-2017-0027/jos-2017-0027.pdf.

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

Durrant, G., O. Maslovskaya, and P. Smith. 2017. "Using Prior Wave Information and Paradata: "Can They Help to Predict Response Outcomes and Call Sequence Length in a Longitudinal Study?" In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 801–833. Doi: https://doi.org/10.1515/jos-2017-0037.

Early, K., J. Mankoff, and S. Fienberg. 2017. "Dynamic Question Ordering in Online Surveys." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 625–657. Doi: https://doi.org/10.1515/jos-2017-0030.

Felligi, I.P. and A.B. Sunter. 1974. "Balance Between Different Sources of Survey Errors – Some Canadian Experiences." *Sankhya* Series C Vol. 36(3): 119–142.

Government Accountability Office 2015. 2020 Census: *Progress Report on Using Administrative Records to Control Enumeration Costs. Testimony before the Subcommittees on Government Operations and information Technology*, Committee on Oversight and Government Reform, House of Representatives, Washington DC. https://oversight.house.gov/wp-content/uploads/2015/11/Goldenkoff-GAO-Statement-11-3-Census-2020.pdf.

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

Groves, R.M. 2011. "Three Eras of Survey Research." *The Public Opinion Quarterly* 75(5): 861–871. https://doi.org/10.1093/poq/nfr057.

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: Statistics in Society 169(3): 439–457. Doi: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Hansen, M.H. and W.N. Hurwitz 1946. "The Problem of Nonresponse in Surveys." *Journal of the American Statistical Association* 41(236): 517–529. Doi: 10.1080/01621459.1946.10501894.

Kaminska, O. and P. Lynn 2017. "The Implications of Alternative Allocation Criteria in Adaptive Design for Panel Surveys." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 781–799. Doi: https://doi.org/10.1515/jos-2017-0036.

Klausch, L.T. 2014. "Informed Design of Mixed-Mode Surveys: Evaluating Mode Effects on Measurement and Selection Error." PhD Thesis, University Utrecht, The Netherlands. https://dspace.library.uu.nl/handle/1874/300673.

Kreuter, F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. New Work, U.S.A.: John Wiley & Sons.

Lewis, T. 2017. "Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol." In A.Y. Chun, B. Schouten, and J. Wagner

(Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 601–624. Doi: https://doi.org/10.1515/jos-2017-0029.

McCarthy, J., J. Wagner, and H. Sanders 2017. "The Impact of Targeted Data Collection on Nonresponse Bias in an Establishment Survey: A Simulation Study of Adaptive Survey Design." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 857–871. Doi: https://doi.org/10.1515/jos-2017-0039.

Morganstein, D. and D.A. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality* edited by L.E. Lyberg, P.P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 475–500. New York, U.S.A.: John Wiley & Sons.

Mulry, M. and B. Spencer. 2012. "A Framework for Empirical Cost Modeling Relating Cost and Data Quality." Paper presented at the 2012 International Total Survey Error Workshop (accessed September 3, 2012).

Paiva, T. and J. Reiter. 2017. "Stop or Continue Data Collection: A Nonignorable Missing Data Approach for Continuous Variables." 579–599. In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 579–599. Doi: https://doi.org/10.1515/jos-2017-0028.

Plewis, I. and N. Shlomo 2017. "Using Response Propensity Models to Improve the Quality of Response Data in Longitudinal Studies." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 753–779. Doi: https://doi.org/10.1515/jos-2017-0035.

Särndal, C. and P. Lundquist 2017. "Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 709–734. Doi: https://doi.org/10.1515/jos-2017-0033.

Schouten, B., A. Peytchev, and J. Wagner 2017. *Adaptive Survey Design*. Chapman and Hall.

Thompson, K.J. and S. Kaputa 2017. "Investigating Adaptive Nonresponse Follow-up Strategies for Small Businesses through Embedded Experiments." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 835–856. Doi: https://doi.org/10.1515/jos-2017-0038.

Tourangeau, R., M. Brick, S. Lohr, and J. Li. 2016. "Adaptive and Responsive Survey Designs: a Review and Assessment." *Journal of the Royal Statistical Society:* Series A 180: 203–223. Doi: http://dx.doi.org/10.1111/rssa.12186.

U.S. Census Bureau. 2017. 2020 Census Operational Plan v3.0: *A New Design for the 21st Century*. Washington D.C. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan3.pdf.

Vandenplas, C., G. Loosveldt, and K. Beullens. 2017. "Fieldwork Monitoring for the European Social Survey: an illustration with Belgium and the Czech Republic in Round 7." In A.Y. Chun, B. Schouten, and J. Wagner (Eds), Special Issue on Responsive and Adaptive Design, *Journal of Official Statistics* 33(3): 659–686. Doi: https://doi.org/10.1515/jos-2017-0031.

Wagner, J. 2008. "Adaptive Survey Design to Reduce Nonresponse Bias." PhD thesis, University of Michigan, Ann Arbor, USA.

Wagner, J., K. Olson, and R. Anderson. 2016. "Survey Costs: The Missing Half of the "Cost-Error." Tradeoff, Paper presented at the Joint Statistical Meetings, Chicago, IL, USA.

Wald, A. 1947. *Sequential Analysis*. Toronto, Canada: General Publishing Co.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 599–624, http://dx.doi.org/10.2478/JOS-2018-0029

A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census

Andrew Keller[1], *Vincent T. Mule*[1], *Darcy Steeg Morris*[1], *and Scott Konicki*[1]

The U.S. Census Bureau is conducting research on using administrative records to reduce the cost while maintaining the quality of the 2020 Census Nonresponse Followup (NRFU). Previous census tests have implemented approaches that use predictive models and optimization procedures to identify vacant and occupied housing units using administrative records. This article details a modification to previous approaches, introducing a simple distance metric to define a quality ranking of housing units to enumerate using administrative records. The distance approach is illustrated, assessed, and compared to a previous approach via a retrospective study of the 2010 U.S. Census.

Key words: 2020 Census; administrative records; nonresponse followup.

1. Introduction

Sample surveys and censuses are historically the primary source for producing official statistics. In order to deal with increasing operational costs and decreasing response rates, national statistical organizations are researching how and when to use administrative records in the census and survey life cycle (Bakker et al. 2015; Fienberg 2015; Wallgren and Wallgren 2007; Brackstone 1987; Federal Committee of Statistical Methodology 1980). Administrative records are data "generated for a different purpose" that "arise organically through administrative processes" (Japec et al. 2015), whether collected through administering a program of a federal government agency or a service of a commercial business. The U.S. Office of Management and Budget has defined administrative records as data held by agencies and offices of the government that has been collected for other than statistical purposes to carry out basic administration of a program (U.S. Office of Management and Budget 2014). This article also considers nonpublic, commercial data similar to administrative records, which is consistent with the wider definition proposed by the United Nations Economic Commission for Europe (UNECE 2011). With respect to surveys, Groves and Harris-Kojetin (2017) outline potential beneficial ways to use administrative records in various stages of the survey life cycle. These include being used as a survey frame, as a replacement for survey data collection, for editing and imputation of missing responses, or for survey evaluation. With

[1] U.S. Census Bureau, Washington, DC 20233, U.S.A. Emails: andrew.d.keller@census.gov, vincent.t.mule.jr@census.gov, darcy.steeg.morris@census.gov, and scott.m.konicki@census.gov

Disclaimer: 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.

respect to censuses, Steffey and Bradburn (1994) note possible uses of administrative records including for coverage improvement, census evaluation, operational efficiency improvement, or to replace traditional census-taking wholly or partially with an administrative records (i.e., register-based) census. Many countries have indeed adopted full register-based (Thygesen 2015; van Zeijl 2014) or partial register-based censuses (Maris et al. 2012). Using administrative records in such a way offers a cost-saving opportunity in a changing census environment of escalating costs; however, it is equally important to consider the quality implications to guide when the use of administrative records is appropriate.

The goal of the 2020 U.S. Census is to count each person once in their correct location at a lower cost per household (adjusted for inflation) than the 2010 Census while maintaining data quality. To meet this goal, the Census Bureau is researching fundamental changes to the design, implementation, and management of the 2020 Census. One major innovation research area noted in the 2020 Operational Plan (U.S. Census Bureau 2017a) is the development of methodologies to incorporate administrative records (AR) into the census design. The U.S. Census Bureau proposes using administrative records in various parts of the operation including to update the address frame, for effective advertising, and to validate respondent addresses for Internet responses to prevent fraud. The 2020 Operational Plan (U.S. Census Bureau 2017a) also specifically recognizes using administrative records to reduce contacts in the Nonresponse Followup (NRFU) operation.

In the 2010 Census, the NRFU operation sent enumerators to about 50 million addresses in all areas of the country to verify the status for every non-self-responding address. Each NRFU address was allowed up to six enumerator contacts. After over 90 million personal visit attempts across the country with field costs of about USD 1.6 billion (Walker et al. 2012), each address was determined to be occupied, vacant, or nonexistent. The occupied units were assigned a person count and person roster including basic demographic characteristics such as name, age, date of birth, race, Hispanic origin, and relationship to householder.

Modernizing the U.S. decennial census using administrative records to supplement or replace traditional census-taking has been a topic of interest since the 1980s (Alvey and Scheuren 1982; Scheuren 1999). However, unlike other countries that implement full or partial register-based censuses, the U.S. has not had a single administrative records system with a high coverage of the entire population (Mulry 2014). For example, the Census Bureau is provided conditional access to data from organizations such as the Internal Revenue Service (IRS), Social Security Administration (SSA), Center for Medicare and Medicaid Services (CMS), and commercial data vendors. Even though each of these data sources covers just a segment of the entire U.S. population, they provide information relevant to census enumeration such as a person's tax-filing address from IRS and birth date from SSA. Previous research has developed methods to combine and use several administrative sources to identify occupied and vacant units prior to or after minimal NRFU fieldwork, thus reducing the number of enumerator visits (Mule and Keller 2014). The administrative sources are used as an input to decision rules about mode switching in NRFU. In this article, we describe an approach to classify units as vacant or occupied at the beginning of NRFU to enable census field operations to reduce costs, thereby allowing resources to focus on units where administrative data are unreliable or unavailable. Our approach is developed as a way to classify administrative record data as high quality or not in order to selectively substitute for field responses early in NRFU activities. However, the approach can also be used adaptively throughout the data collection phase or in other census operations, such as for imputation in the data-processing phase (see Subsection 5.2). Or, more generally, the approach can be applied to sample surveys with little alteration when appropriate data are available (see Subsection 5.3).

Through a series of census field tests, various approaches for determining vacant and occupied housing units via administrative records have been tested and refined with increasing levels of complexity and integration with other census operations. In the 2013 and 2014 Census Tests, rules-based approaches were implemented (Walejko et al. 2014; Keller et al. 2016), followed by a predictive modeling approach used in the 2015 Census Test based on linear optimization of logistic regression model predictions (Morris et al. 2016). Most recently, the 2016 Census Test used an adaptation of the modeling approach that is based on a Euclidean distance function (Chapin and Keller 2017). In this article, we present the distance function approach to determine high-quality administrative records in a way that simplifies implementation, while maintaining similar quality to the procedure used in the 2015 Census Test. We retain the same underlying predictive modeling structure, as it naturally incorporates information from multiple administrative records sources and other auxiliary data, but offers a new way to synthesize the model information. We illustrate the utility of this advancement of the predictive modeling methodology in a retrospective study of the 2010 Census. The distance function approach is a direct alternative to the optimization approach in Morris et al. (2016). Comparing the two methods, we find that our method has a high overlap with the linear optimization approach in identifying cases of sufficient quality that can be enumerated using administrative records. At the same time, the distance function yields similar quality metrics (measured through a retrospective study of the 2010 Census) while being easier to implement. Furthermore, the classification mechanism of the distance function approach selects housing units based on their own merit rather than relative to a predetermined set of housing units.

2. Administrative Records Data for the Study of the Decennial Census

The Census Bureau receives separate administrative record data files from various government agencies and private companies for statistical use. To enable the linking of these diverse data sets, an anonymized identifier is assigned to each person record in each administrative record file. The Census Bureau's Person Identification Validation System (PVS) determines the protected identification key (PIK) via a probabilistic matching algorithm between the administrative record source data and a series of reference files. See Wagner and Layne (2014) for details on the PVS algorithm. For simplicity we assume that the PIK assignment is correct and match the files accordingly, however, we acknowledge the importance of linkage error associated with the PVS methodology. See, for example, Layne et al. (2014) for discussion of error associated with PIK assignment given the use of the various reference files.

In our study of the 2010 Census, we use these linkable and anonymized administrative record files to compile a household roster composed of administrative record persons for all housing units in the 2010 NRFU universe in the United States. The 2010 vintage administrative record sources used to create 2010-level administrative record household rosters are:

- IRS Individual Tax Returns (Form 1040)
- IRS Informational Returns (Form 1099)
- Indian Health Service (IHS) Patient Database
- CMS Medicare Enrollment Database

The resulting administrative record household roster – the collection of PIKs found in any of the selected administrative record files at a given address – is unique by person and address. That is, no persons are duplicated within a housing unit. We use person-level administrative record data, as well as an aggregated housing unit-level administrative record data set that includes characteristics such as administrative record household count and general characteristics of the people in the household. The Social Security Numerical Identification (Numident) File is used to obtain age and sex information for each person in the administrative record household roster.

Rastogi and O'Hara (2012) compared several administrative record and third-party sources to the 2010 Census. For federal files, IRS 1040 individual tax returns had the highest match rate to the 2010 Census. This is due to the magnitude of persons and the fact that tax filings start in February with a deadline of April 15, close to the April 1 Census Day. The analysis showed that CMS' Medicare Enrollment Database had a high match rate for the elderly population. The IHS Patient Database is chosen to address potential undercoverage of the American Indian population. The Social Security Numident file has been shown to have very high coverage and reliable data for age and sex.

It should be noted that not all housing units have information in the selected administrative record files. Conversely, there are people in the administrative records files that are not enumerated in the census. Hence, undercoverage and overcoverage exists when comparing between a census roster and an administrative record roster for the same unit. Because we are not assuming that the administrative records files have sufficient coverage of the entire population, our approach is to eliminate NRFU visits to addresses for which we are confident in the administrative record data. That is, we are trying to limit the use of administrative records to cases where coverage differences between administrative records and fieldwork are minimized, provided that fieldwork would generate the correct Census Day roster.

In addition to the administrative sources, information from commercial files, is used to inform the models. Variables derived from these data are used as independent variables in the models. We also incorporate data from the United States Postal Service (USPS) Delivery Sequence File (DSF), the American Community Survey (ACS), the Master Address File (MAF), census operational information, and USPS Undeliverable as Addressed (UAA) reason codes obtained from census mailings delivered around Census Day.

3. Models and Methodology

The administrative records data described in Section 2 contains a wealth of timely information about the characteristics of addresses. We employ a modeling approach to extract predictive information from the administrative records to identify housing units with sufficiently reliable vacancy and roster information. The predictive models described in Subsections 3.1 and 3.2 to follow are the same as those used in Morris et al. (2016). A cursory description of the models is provided here; see Morris et al. (2016) for further details. These models estimate various measures of administrative record quality that are subsequently used to rank housing units based on their likelihood of vacancy or their likelihood of correct enumeration for occupied housing units. In Subsection 3.3, we present the distance function approach as a way to use the predicted probabilities from the models to define a quality ranking and identify high-quality housing units that can be removed from the NRFU workload and enumerated using existing administrative records. We refer to units identified as having sufficiently good information from administrative records to accurately predict a vacant housing unit as *AR Vacant*; we define *AR Occupied* units analogously.

3.1. Model for Determining Vacant Housing Units

To identify vacant units via administrative record information, we rely on a statistical model to estimate predicted probabilities of Census Day housing unit status. We fit a multinomial logit model on the housing unit-level administrative record data to predict the three possible values of housing unit status: occupied $\left(y_h^{unocc} = 1\right)$, vacant $\left(y_h^{unocc} = 2\right)$, or nonexistent $\left(y_h^{unocc} = 3\right)$, where the *unocc* superscript denotes the model used for administrative records removal of unoccupied or vacant housing units, and the *h* subscript indexes the housing unit. From this model, we estimate the probability of each unit status type in the 2010 Census data (i.e., the training data):

$$\hat{p}_{h,occ}^{unocc} = P\left(y_h^{unocc} = 1\right), \quad \hat{p}_{h,vac}^{unocc} = P\left(y_h^{unocc} = 2\right), \quad \hat{p}_{h,del}^{unocc} = P\left(y_h^{unocc} = 3\right).$$

The predicted probabilities, $\hat{p}_{h,occ}^{unocc}$ and $\hat{p}_{h,vac}^{unocc}$, are passed to the distance function to determine which cases are identified as AR Vacant.

The use of a statistical model naturally allows the incorporation of information from multiple sources. For example, vacancy information from a USPS mailing around Census Day is strongly associated with Census Day vacancy (Keller et al. 2016), however it is not a perfect proxy and is not the only strong predictor. This model combines information from USPS mailing data and persons associated with a housing unit present in, for example, tax returns or the Medicare enrollment database. Specifically, housing unit status – as determined by the training data (2010 Census data in our application) – is modeled as a function of independent variables from administrative records, field collection paradata, and survey information. Such covariate information includes the UAA data from the USPS for each of the census mailings, persons from the administrative record sources listed in Section 2, characteristics associated with the block group as determined by the ACS, and other address-level information. The appendix contains a complete list of independent variables for the vacant model.

3.2. Models for Enumerating Occupied Housing Units

To identify and enumerate occupied units via administrative record information, we rely on two statistical models to measure the quality of the administrative records information for enumerating households accurately.

3.2.1. Person-Place Model

The person-place model estimates the probability of enumerating a person on the administrative records at the same address as the 2010 Census data (i.e., the training data). We fit a logistic regression model on the person-level administrative record data to predict the outcome:

$$y_{ih}^{occ1} = \begin{cases} 1 & \text{if person } i \text{ is found in AR and 2010 Census at the same address } h \\ 0 & \text{otherwise} \end{cases}$$

where the *occ1* superscript denotes the person-place model for determining occupied units, the h subscript indexes the housing unit, and the i subscript indexes the administrative record person. Morris (2014) and Morris (2017) study a version of the person-place model comparing alternative estimation approaches (logistic regression, classification trees, and random forests). The choice of estimation procedure has little impact on the findings, thus logistic regression is used here for consistency with the other models used in this research. This model assigns to all person-place pairs in administrative record files a predicted probability, $\hat{p}_{ih}^{occ1} = P\left(y_{ih}^{occ1} = 1\right)$, that the 2010 Census and the administrative record roster data place the person at the same address. The person-place model includes all administrative record person records associated with the address from the sources in Section 2. The 2010 Census person records are assigned PIKs with the methodology discussed in Section 2. Note that a person in administrative record and not the Census is coded as $y_{ih}^{occ1} = 0$. This category could include possible census omissions. Conversely, a person not in administrative records and in the census is excluded from the modeling universe.

Person-place match is modeled as a function of independent variables from person-level administrative record information (e.g., indicators of the presence of the administrative records person in each source at the address, indicators of presence of the administrative records person at a different address within the same administrative records source), address-level administrative record information (e.g., number of administrative records people associated with an address), field operations information (e.g., USPS mailing information, number of NRFU neighbors), and information from other survey sources (e.g., characteristics of the local geography – such as poverty rate, renter rate, Hispanic rate, vacancy rate – from the ACS). The person- and address-level administrative record information is of particular importance. For example, Morris (2014) finds that the presence of an IRS 1040 record at given address, and conversely, the presence of an IRS 1040 at a different address, are strong predictors in the person-place model. The former is associated with an increased probability of the administrative records placing the person at the census address, whereas the latter is associated with a decreased probability. The appendix contains a complete list of independent variables for person-place model.

The person-place model is fit at the person-level, but decisions are made at the housing unit-level. Therefore, the person-level predicted probabilities, $\hat{p}_{ih}^{occ1}$, are summarized for each address such that the housing unit-level predicted probability for address h was defined as:

$$\hat{p}_{h}^{occ1} = \min\left(\hat{p}_{1h}^{occ1}, \ldots, \hat{p}_{n_h h}^{occ1}\right)$$

where n_h is the number of people at address h. This minimum criterion assigned to the housing unit the predicted probability for the person in the housing unit for which we had the lowest confidence – a relatively conservative approach. The administrative record household count is defined as the sum of all individuals associated with the administrative record address, and each address has the associated predicted probability of having an administrative record/census address match. These predicted probabilities, $\hat{p}_h^{occ1}$, are passed to the distance function to determine which cases are identified as AR Occupied.

3.2.2. Household Composition Model

The household composition model is used to estimate the probability that the sample address has the same household composition (number of adults and children) determined by NRFU fieldwork as its pre-identified administrative record household composition. We fit a multinomial logistic model on the housing unit-level administrative record data to predict the outcome from the 2010 Census (i.e., the training data):

$$y_h^{occ2} = \begin{cases} 0 & \text{if unit } h \text{ is vacant in 2010 Census} \\ 1 & \text{if unit } h \text{ has 1 adult and 0 children in 2010 Census} \\ 2 & \text{if unit } h \text{ has 1 adult and } \geq 1 \text{ children in 2010 Census} \\ 3 & \text{if unit } h \text{ has 2 adults and 0 children in 2010 Census} \\ 4 & \text{if unit } h \text{ has 2 adults and } \geq 1 \text{ children in 2010 Census} \\ 5 & \text{if unit } h \text{ has 3 adults and 0 children in 2010 Census} \\ 6 & \text{if unit } h \text{ has 3 adults and } \geq 1 \text{ children in 2010 Census} \\ 7 & \text{if unit } h \text{ has } \geq 4 \text{ adults in 2010 Census} \end{cases}$$

where the *occ2* superscript denotes the household composition model for determining occupied units, and the h subscript indexes the housing unit. For every address, this model assigns a predicted probability of each household composition type, $\hat{p}_{h,k}^{occ2} = P\left(y_h^{occ2} = k\right)$ for $k = 0,1,2,3,4,5,6,7$. Note that the construction of the dependent variable assumes that age is nonmissing for all housing units. This assumption is satisfied in our application because we use an edited file that includes imputed age for any nonresponse.

The household composition dependent variable y_h^{occ2} is modeled as a function of independent variables from housing unit-level administrative record information (e.g., count of all administrative records person records associated with the address from each of the administrative records sources), person-level administrative record information (e.g., indicators of whether any administrative records person was found at a different address within the same administrative records source), and housing unit-level information from other survey sources (e.g., flags indicating that young children, elderly, Black or White persons from administrative records were associated with the household). The appendix contains a complete list of independent variables for household composition model.

We are solely interested in the predicted probability associated with the household composition observed in the administrative records. That is, for each housing unit we extract the household composition predicted probability associated with the administrative record household composition, defining $\hat{p}_h^{occ2} = \hat{p}_{h,k^*}^{occ2}$ where k^* is the administrative record household composition. For example, $\hat{p}_h^{occ2} = \hat{p}_{h,3}^{occ2}$ for a housing unit with an administrative record household composition type of two adults and zero children. These predicted probabilities, $\hat{p}_h^{occ2}$, are passed to the distance function to determine which cases are identified as AR Occupied.

3.3. *Identifying Administrative Record Vacant and Occupied Housing Units Using a Distance Function*

We study a direct alternative for the approach described in Morris et al. (2016) that was implemented in the 2015 Census Test. Morris et al. (2016) use linear programming techniques to combine information from the previously described models to determine AR Vacant and AR Occupied housing units. The optimization approach requires setting multiple threshold parameters that are not straightforward to select and interpret. Furthermore, the constraints in the optimization routine involve averages of probabilities over select workloads, where a workload is a set of housing units that requires enumeration.

Specifically, for identifying AR Vacant units, Morris et al. (2016) set constraints that (1) the average vacant predicted probability must exceed a prespecified threshold and (2) the sum of the occupied predicted probability did not exceed a certain percentage of the estimate of occupied housing units from the American Community Survey. With respect to identifying AR Occupied units, the authors set constraints that (1) the average person-place predicted probability must exceed a prespecified threshold and (2) the average household composition predicted probability must also exceed a different prespecified threshold. This is potentially problematic for two reasons: (1) it allows housing units other than the housing unit of interest to contribute to the identification of that unit as AR Vacant or AR Occupied, and (2) the workload over which to take the average must be predefined and has an effect on each housing unit's identification.

Consider a simple example of determining AR Vacant units in two NRFU workloads, each of four addresses with the following vacant probabilities:

$$\text{Workload 1: } \hat{p}_{1,vac}^{unocc} = 0.81, \quad \hat{p}_{2,vac}^{unocc} = 0.81, \quad \hat{p}_{3,vac}^{unocc} = 0.75, \quad \hat{p}_{4,vac}^{unocc} = 0.50$$

$$\text{Workload 2: } \hat{p}_{1,vac}^{unocc} = 0.90, \quad \hat{p}_{2,vac}^{unocc} = 0.90, \quad \hat{p}_{3,vac}^{unocc} = 0.72, \quad \hat{p}_{4,vac}^{unocc} = 0.50$$

Focusing solely on the average predicted probability constraint for illustrative purposes, the optimization approach of Morris et al. (2016) identifies AR Vacant addresses as those contained in the subset of housing unit-level predicted probabilities that maintains an average that exceeds a specified cutoff. Using the cutoff of 0.8 used in Morris et al. (2016), in this example averaging would identify housing units $h = 1$ and $h = 2$ as AR Vacant in Workload 1, and housing units $h = 1$, $h = 2$, and $h = 3$ as AR Vacant in Workload 2. Due to the nature of averaging, the third household ($h = 3$) is identified as AR Vacant in Workload 2 despite that it has a lower predicted probability of vacancy in Workload 2 as compared to Workload 1. This simplistic example illustrates how the averaging of predicted probabilities allows other cases to contribute to identification of AR Vacant units. In the same vein, the AR Vacant determination depends crucially on the set of predicted probabilities included in the average. Average predicted probabilities are computed over a predefined area; therefore a decision has to be made about over what areas the averaging is done. One possibility would be to run the linear optimization over the entire nation. This could cause a disproportionate amount of cases to be removed in one area, resulting in unbalanced workloads. Another alternative could be to run the linear optimization for each state or county. Doing this would require running the optimization 50 or 3,000 times, which could increase the computational time and complexity. In an environment where field operations are waiting on results from the administrative records models, the days it would take to run the optimization routine would make timing more challenging.

We study a simpler approach using a distance function that avoids the concerns of the optimization approach – in particular, the distance method evaluates each housing unit on its own merit – and relies on a more transparent and interpretable threshold parameter. Furthermore, the distance method is easier to implement in that real-time workload adjustments can be determined by simply changing the threshold parameter rather than rerunning the optimization procedure. This alternative is partially motivated by the use of a decision criterion for identifying cases to enumerate using administrative records based on distances measured via Receiver Operator Characteristic (ROC) graphs (Morris 2014, 2017). We define distance functions that take multiple measures of the quality of the administrative records, with respect to determining vacancy and for enumeration of occupied housing units, as inputs to output a single measure. This scalar distance measure combines multiple predicted probabilities – which are themselves based on the combination of multiple sources of information via the statistical models – to allow (1) a ranking of the housing units by quality and (2) a definition of a subset of the highest quality housing units by choosing a threshold.

With regard to vacancy determination, we define the housing unit-level *vacant distance* based on the vacant probability, $\hat{p}_{h,vac}^{unocc}$, and occupied probability, $\hat{p}_{h,occ}^{unocc}$, estimated via the housing unit status model discussed in Subsection 3.1. These predicted probabilities can be thought of as a two-dimensional plane with each probability on one dimension with values between 0 and 1. Based on the two probabilities, each address would have a point in this two-dimensional space. The most likely vacant cases would be those that have shortest distance to the point where the occupied probability equals 0 and the vacant probability equals 1 (i.e., the (0,1) point). As a result, we define the Euclidean vacant distance, d_h^{vac}, for each unit h, as

$$d_h^{vac} = \sqrt{\left(1 - \hat{p}_{h,vac}^{unocc}\right)^2 + \left(\hat{p}_{h,occ}^{unocc}\right)^2}.$$

With regard to identifying occupied housing units for administrative record enumeration, we define the housing unit-level *occupied distance* based on predicted probabilities from the two occupied models: the minimum person-place probability for the address, $\hat{p}_h^{occ1}$, and the household composition probability associated with the observed administrative record household composition, $\hat{p}_h^{occ2}$. Both of these probabilities are measures of quality (count match and household composition match, respectively) such that the housing units with higher quality administrative records are associated with higher estimated probabilities. Even though the predictions from these two models are correlated, Morris et al. (2016) show higher agreement in population count and household composition when both models are used together as compared to using one or the other. Accordingly, we use results from both the person-place and household composition model as inputs for the distance function. Similar to the construction of the vacant distance, the most likely occupied and correct enumeration cases would be those that have shortest distance to the point where the predicted probability from both models equals 1 (i.e., the (1,1) point). Based on this idea, we use the Euclidean distance to define the occupied distance, d_h^{occ}, for each unit h as

$$d_h^{occ} = \sqrt{\left(1 - \hat{p}_h^{occ1}\right)^2 + \left(1 - \hat{p}_h^{occ2}\right)^2}.$$

The distances d_h^{vac} and d_h^{occ} are used to determine AR Vacant and AR Occupied housing units, respectively. That is, we define a given distance cutoff targeting a certain rate of removal of cases from the face-to-face follow-up. We then treat those administrative records as a reasonably correct representation of the true status for those addresses.

4. Application: 2010 Decennial Census Data

We apply the distance function methodology for determining AR Vacant and AR Occupied housing units in a retrospective study of the NRFU operation of the 2010 Census. In this analysis, the vacant model and two occupied models are fit to a sample of the NRFU housing units in the 2010 Census. The fitted coefficients are then applied to all NRFU housing units to obtain the predicted probabilities ($\hat{p}_{h,vac}^{unocc}$ and $\hat{p}_{h,occ}^{unocc}$ for the vacant model, $\hat{p}_h^{occ1}$ and $\hat{p}_h^{occ2}$ for the occupied models) and the associated distances (d_h^{vac} and d_h^{occ}) for each housing unit h.

4.1. Identifying Administrative Record Vacant Housing Units

Figure 1 plots the estimated vacant probability, $\hat{p}_{h,vac}^{unocc}$, and occupied probability, $\hat{p}_{h,occ}^{unocc}$, for the 50 million NRFU housing units in the 2010 Census. The vacant distance measure, d_h^{vac}, is used to create percentile bands generated by assuming varying cutoffs. The upper

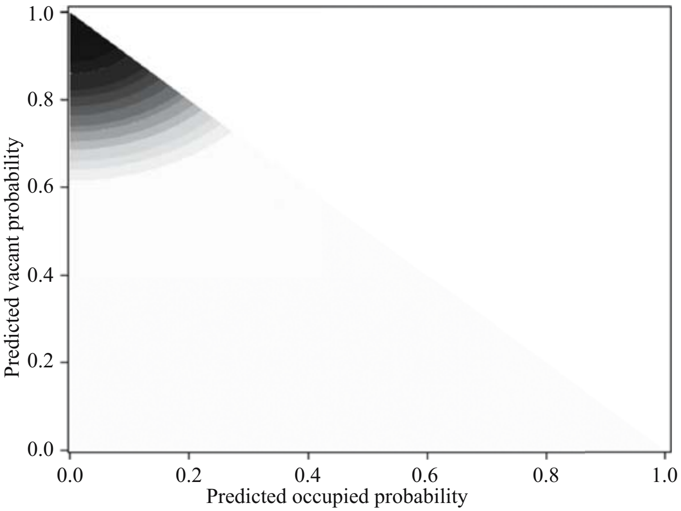


Fig. 1. AR vacant predicted probabilities by vacant distance percentile (source: 2010 Census simulation).

leftmost area, denoted by the black shading, represents the top one percent of NRFU cases with the smallest vacant distance. If we were to restrict our AR Vacant identification total to 500,000 cases, removing these cases from the NRFU workload would reduce the number of visits during NRFU at the smallest predicted expense of quality. The band just below the upper leftmost area, denoted by the darkest gray, are those housing units between the top one percent and two percent of NRFU housing units with the smallest vacant distance. Dividing the data by percentile bands yields the partial concentric circles in Figure 1 depicting various scenarios of target NRFU workload reduction.

To assess accuracy for varying vacant distance cutoffs, we treat the 2010 NRFU housing unit status as the gold standard and compare field vacancy determination to administrative record vacancy determination. Figure 2 shows the true positive rate – the percent of AR Vacant cases that were resolved as vacant during the 2010 NRFU – for each mutually exclusive percentile band up to the 15th percentile, with the lowest vacant distance starting at the first percentile. We see in Figure 2 that, for the top one percent of cases (500,000 NRFU cases) with the shortest vacant distance between the (0,1) point and $\left(\hat{p}^{unocc}_{h,occ}, \hat{p}^{unocc}_{h,vac}\right)$, the true positive rate is 90.8 percent – indicating that among the 500,000 NRFU cases identified as AR Vacant using the distance function approach, 90.8 percent were resolved as vacant through NRFU fieldwork. For the second best one percent of cases (i.e., cases of rank 500,001 to 1,000,000), the true positive rate is 84.9 percent. There is a gradual decrease in the true positive rate as the percentiles increase, depicting the decrease in the quality of administrative records information for cases with a vacant distance that is further from the optimal (0,1) point.

Based on the analysis and the tradeoff between cost reduction and quality, a decision can be made about how many bands to designate as being AR Vacant. The tradeoff exists because by identifying more AR Vacant cases, thereby reducing costs incurred by NRFU followup, we see a larger percentage of cases return as occupied.

Morris et al. (2016) use linear optimization processing of the same predicted probabilities – $\hat{p}^{unocc}_{h,occ}$ and $\hat{p}^{unocc}_{h,vac}$ – to determine about ten percent of the NRFU universe (5,132,613 addresses) as AR Vacant. We are interested in comparing the performance of the linear optimization approach with the simpler distance function approach presented in

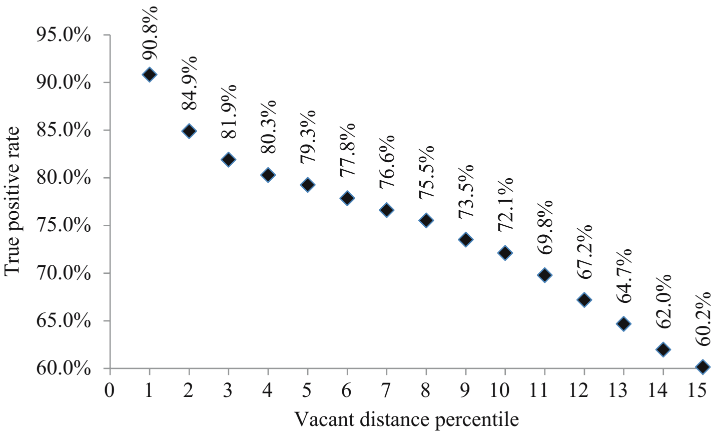


Fig. 2. AR vacant true positive rate by vacant distance percentile (source: 2010 Census simulation).

this article. To do this, we sort the housing units from smallest to largest vacant distance and identify the 5,132,613 addresses with the smallest vacant distance to be AR Vacant. Among these AR Vacant cases, the smallest vacant distance value is $d_h^{vac} = 0.0078$ and the largest vacant distance value is $d_h^{vac} = 0.3559$. We find that 91 percent of the addresses determined to be AR Vacant using the distance function are also identified as AR Vacant by the linear optimization approach. The two methods largely identify the same AR Vacant cases. However the distance function is easier to operationalize.

We further evaluate the distance function and linear optimization AR Vacant cases compared to their 2010 NRFU results. Table 1 shows the results from contrasting the optimization approach versus the distance approach for the same workload. We find similar observed 2010 distributions between the two identification approaches. The distance approach does slightly better in terms of agreement with the NRFU result – the percentage of AR Vacant cases with a vacant NRFU status is higher for the distance approach versus the optimization approach (79.0% vs. 78.1%). Regardless of the approach, not all cases identified as AR Vacant were vacant in the 2010 NRFU. Some of the misclassification between administrative records and census may be due to errors in the 2010 Census. Keller and Konicki (2016) show that approximately ten percent of persons enumerated in these AR Vacant and field occupied units are erroneous enumerations and 20 percent are imputed.

To further assess quality implications, we can look to other 2010 coverage results. Cresce (2012) showed that the 2010 Census continued the trend from the 1990 Census and 2000 Census of underestimating the vacancy rate as compared to other estimates like the American Housing Survey and the Current Population Survey. The Census Coverage Measurement program found that vacant housing units were undercounted by 4.8 percent in 2010 (Mule and Konicki 2012). These evaluation results suggest that by conducting interviews between March and August to assess the population on April 1, the decennial census may have enumerated people in units that were vacant on Census Day.

4.2. *Identifying Administrative Record Occupied Housing Units*

Our assessment of the identification of AR Occupied units is analogous to that of identifying AR Vacant units in the previous section; however, the distance function for identifying AR Occupied units depends on predicted probabilities from two separate models rather than one model. Figure 3 plots the predicted probability from the person-place model, $\hat{p}_h^{occ1}$, and for the household composition model, $\hat{p}_h^{occ2}$, for the eligible NRFU housing units. Only those NRFU addresses with an associated administrative record person are eligible to be AR Occupied. The occupied distance measure, d_h^{occ}, is used to create percentile bands generated by assuming varying cutoffs. The upper rightmost

Table 1. AR vacant versus NRFU status assigned – optimization approach versus distance approach (source: 2010 Census simulation).

AR vacant approach	Workload removal	Occupied (%)	Vacant (%)	Nonexistent (%)	Unresolved (%)
Optimization	5,132,613	9.1	78.1	11.9	0.9
Distance	5,132,613	8.8	79.0	11.3	0.9

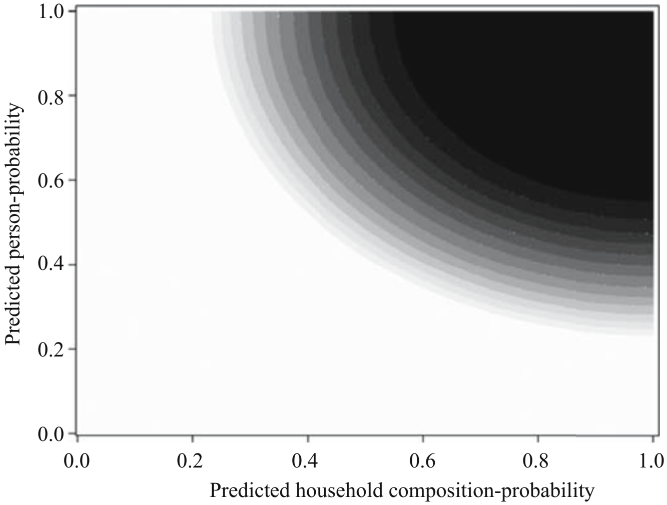


Fig. 3. AR occupied predicted probabilities by occupied distance percentile (source: 2010 Census simulation).

area, denoted by the black shading, represents the top one percent of NRFU cases with the smallest occupied distance. Dividing the data by percentile bands yields the concentric circles in Figure 3 depicting various scenarios of target NRFU workload reduction.

To assess accuracy for varying occupied distance cutoffs, we again treat the 2010 NRFU housing unit status as the gold standard and compare field occupancy determination to administrative record occupancy determination. Figure 4 shows the true positive rate – the percent of AR Occupied cases that were resolved as occupied during the 2010 NRFU – for each mutually exclusive percentile band up to the 15th percentile, with the lowest occupied distance starting at the first percentile. We see in Figure 4 that, for the top one percent (500,000 NRFU cases) with the shortest occupied distance between the (1,1) point and $\left(\hat{p}_h^{occ1}, \hat{p}_h^{occ2}\right)$, the true positive rate is 94.6 percent – indicating that among the 500,000 NRFU cases identified as AR Occupied using the distance function approach, 94.6 percent were resolved as occupied through NRFU fieldwork. There is a gradual decrease in

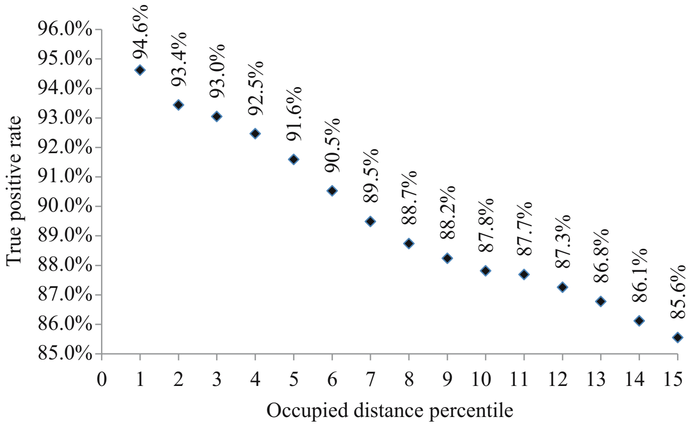


Fig. 4. AR occupied true positive rate by occupied distance percentile (source: 2010 Census simulation).

the true positive rate as percentile increases, similar to administrative records vacant identification.

In addition to determining occupancy and the household count, the decennial census collects information on the characteristics of the people in occupied housing units. It is important to recognize the ramifications on characteristics for cases that are enumerated via administrative records rather than fieldwork. In the previous example of implementing administrative record enumeration for the top one percent based on the occupied models, 500,000 housing units are assigned persons from administrative records. However, because no interviews are completed, characteristics for people in these housing units must be obtained from the administrative records or imputed.

Some characteristics are readily available from the administrative records sources: age is a necessary requirement to be AR Occupied as the household composition model depends on age by definition. Obtained from the Numident file, sex is also usually a nonmissing characteristic. Other characteristics are less straightforward, namely race and Hispanic origin. We use administrative record data from various sources to identify race and Hispanic origin for persons enumerated in AR Occupied units. See Ennis et al. (2015) for a full explanation of how race and Hispanic origin are assigned to persons in the administrative record data. Figure 5 shows the housing unit-level missing data rate for race and Hispanic origin for housing units identified as AR Occupied by each percentile of the occupied distance (starting at the first percentile). For example, of the 500,000 NRFU units identified as AR Occupied in the second percentile, about 0.50 percent of housing units are missing Hispanic for all persons. This would necessitate assigning Hispanic origin for all persons in these housing units via an imputation procedure.

Similar to the vacant cases, we are interested in comparing the performance of the linear optimization approach with the simpler distance function approach presented in this article. We sort the housing units from smallest to largest occupied distance and identify the 7,292,195 addresses with the smallest occupied distance to be AR Occupied. In this case, about 15 percent of the NRFU universe is identified as AR Occupied corresponding to a occupied distance threshold of $d_h^{occ} = 0.7140$, where the smallest observed occupied distance value is $d_h^{occ} = 0.1907$. We find that 93 percent of the addresses determined to be

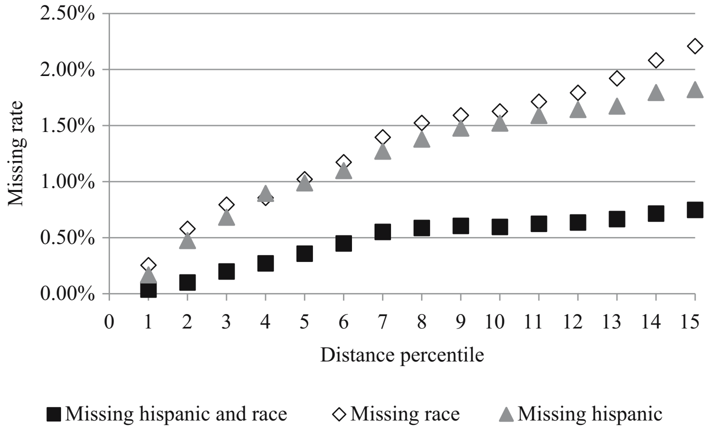


Fig. 5. Housing unit missing data rates by occupied distance percentile (source: 2010 Census simulation).

Table 2. AR occupied versus NRFU status assigned – optimization approach versus distance approach (source: 2010 Census simulation).

AR occupied approach	Workload removal	Occupied (%)	Vacant (%)	Nonexistent (%)	Unresolved (%)
Optimization	7,292,195	90.2	7.9	1.6	0.3
Distance	7,292,195	89.7	8.4	1.7	0.3

AR Occupied using the distance approach are also identified as AR Occupied by the linear optimization approach. Hence, the two methods largely identify the same cases to be removed from the operation and enumerated as occupied via administrative records.

Table 2 shows similar observed distributions of 2010 housing status when contrasting the optimization approach and the distance approach for the same workload amount. The optimization approach does slightly better in terms of agreement with the NRFU result – the percentage of AR Occupied cases with an occupied NRFU status is higher for the optimization approach versus the distance approach (90.2% vs. 89.7%). Note that not all cases identified as AR Occupied were occupied in the 2010 NRFU. Some of the misclassification between administrative records and census may be due to omissions in the census.

5. Conclusion and Discussion

To prepare for the 2020 Census, the Census Bureau is researching cost-saving changes to NRFU. The use of administrative records to reduce field contacts in NRFU is one cost-saving measure specifically noted in the 2020 Operational Plan (U.S. Census Bureau 2017a). We propose a modeling approach for assessing the quality of administrative records for enumerating housing units in conjunction with a distance function to identify AR Vacant and AR Occupied units. The results from the retrospective study of the 2010 Census provide evidence of the internal validation of the model and methodology as the distance function approach accurately recognizes vacancy and occupancy in the vast majority of AR Vacant and AR Occupied cases, respectively. Similarly, the 2016 Census Test provided external validation of the distance approach (Chapin and Keller 2017). We contrast the distance function approach with the optimization approach discussed in Morris et al. (2016) and implemented in the 2015 Census Test. Even though we find that the two methods perform similarly on the 2010 Census data, we favor the distance function approach for its simplicity and operational ease to document in a production environment. This new approach provides a more objective way to define thresholds that dictate the cost and quality tradeoff. The choice of the distance measure cutoff implies a cost reduction in that the addresses identified would receive fewer visits during NRFU, but quality metrics such as true positive rates must be factored in as well.

5.1. Contact Strategy

The proposed distance approach for identifying AR Vacant and AR Occupied cases can be used operationally in the context of a broader contact strategy. Here we provide an overview of a NRFU field visit strategy related to units identified as AR Vacant or AR

Occupied. This contact strategy – which was implemented in the 2016 Census Test (Chapin and Keller 2017) – illustrates how and when administrative records may substitute for face-to-face interviews, thus reducing costs of field operations. Research and testing programs continue to adapt and refine this contact strategy leading up to the 2020 Census, but generally suggest using administrative records in a reasonably similar manner (U.S. Census Bureau 2017b).

Prior to the start of the 2016 Census Test NRFU operation, each address was eligible to receive up to four mailings before and after Census Day. If the address did not respond to these mailings, the Census Bureau decided how many times to visit the address during the NRFU operation in accord with the quality of administrative record data. Figure 6 shows the flowchart of the visit strategy for NRFU housing units in the 2016 Census Test. The distance function methodology was used to identify the AR Vacant and AR Occupied housing units shaded in the flowchart in Figure 6 (Chapin and Keller 2017).

Housing units identified as AR Vacant did not receive any visits during NRFU. In general, the AR Vacant units were those units with Undeliverable as Addressed reason codes returned from the initial census mailings and an absence of administrative record presence (i.e., no sign of life in the administrative records). As part of the NRFU contact strategy, a postcard was mailed to the AR Vacant units to allow an additional opportunity for self-response.

The cases not identified as AR Vacant received one field visit. This visit allowed cases to be resolved in several ways: completion of an interview with the household member, field determination of vacancy, or field determination that the address was not a housing unit. If the enumerator did not make contact with anybody at the housing unit, the enumerator left a notice of visit regardless of whether the unit was AR Occupied or not. This notice of visit included self-response information to encourage the household to respond by going online, dialing the questionnaire assistance number, or returning the paper questionnaire sent earlier. Units determined to be AR Occupied received only this one visit in the 2016 Census Test. After one visit, if the housing unit remained unresolved then AR Occupied housing units received an additional postcard mailing with self-response information. All other unresolved housing units (those not identified as AR Occupied) were contacted via the usual protocol (i.e., additional contacts). As shown, there were several ways before and during NRFU that the Census Bureau attempted to obtain and use self-responses before enumerating cases via administrative record information.

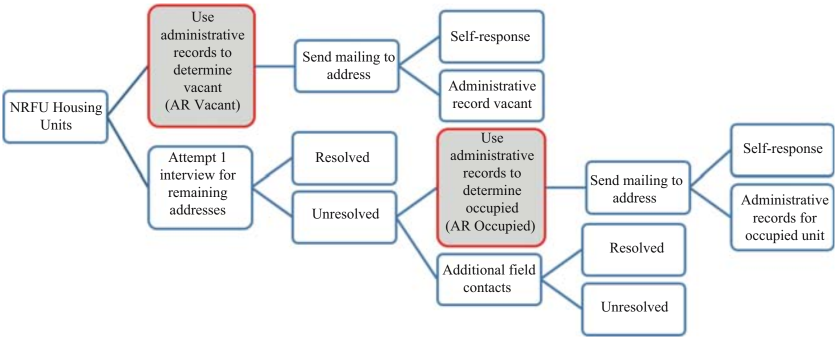

NRFU Housing Units
Use administrative records to determine vacant (AR Vacant)
Send mailing to address
Self-response
Administrative record vacant
Attempt 1 interview for remaining addresses
Resolved
Unresolved
Use administrative records to determine occupied (AR Occupied)
Send mailing to address
Self-response
Administrative records for occupied unit
Additional field contacts
Resolved
Unresolved

Fig. 6. Nonresponse followup visit strategy (2016 Census test).

5.2. Adaptive Uses of the Distance Function Method

The contact strategy described in Subsection 5.1 shifts AR Vacant and AR Occupied units from an approach solely reliant on enumerator visits. In practice, the AR Occupied units are only allowed a maximum of one enumerator visit before having to respond via another mode. This tailored contact strategy results from identifying cases for removal based on administrative record information available at the start of NRFU. The distance function approach assumes a fixed set of data on which the underlying models are fit. However, the approach can be implemented adaptively as new administrative record information is obtained during the NRFU operation. In the context of the 2015 Census Test, Keller (2016) describes multi-phased integration of administrative record modeling as an adaptive component throughout NRFU. For this test, after initial AR Occupied and AR Vacant cases were identified, the Census Bureau received additional IRS 1040 and IRS 1099 information. After processing these data, the administrative record models were refit. Additional units were identified as AR Occupied and subsequently enumerated via the new administrative record data. Although this was not preplanned, this adaptation enabled the resolution of cases using administrative record data that had not been available at the start of NRFU. Doing so in real-time allowed the Census Bureau to shift resources to units that had proven to be more difficult to enumerate.

The distance function methodology can also be used after data collection is complete, as an alternative to unit imputation of status and population size for unresolved housing units. In the context of the 2015 Census Test, Keller (2016) documents a modification of the optimization approach: refitting and determining AR Vacant and AR Occupied cases by lowering the average constraint values in the optimization approach – thus identifying more AR Vacant and AR Occupied cases. The new cases that remained unresolved addresses after the full visit strategy are assigned occupancy status and enumerated using administrative records rather that via an imputation. In the same fashion, rather than relying on the optimization approach, the new distance function approach can be extended to allow additional unresolved addresses to be assigned an AR Vacant or AR Occupied status by lowering the threshold.

To elaborate on this scenario, Figure 7 shows a hypothetical example for the AR Occupied determination. A distance threshold can be specified to identify the dark gray area in the upper right corner of the figure. Addresses with predicted probabilities in this area will receive no more than one visit. A second distance threshold can be specified to identify the medium gray area. These cases would receive the full visit strategy during NRFU; however, if they are unresolved after fieldwork is completed, then administrative records information would be used to determine occupancy status and a roster, if occupied, instead of using count imputation for these cases. The administrative records for the remaining housing units in the light gray area would not be utilized, as they are not of sufficient quality. This figure and hypothetical scenario exemplify the clarity and ease of communicating the distance function approach for reducing visits or avoiding imputation.

5.3. Implementing on Surveys

We have focused exclusively on using administrative records to replace household responses specifically for the decennial census. Using administrative records to curtail

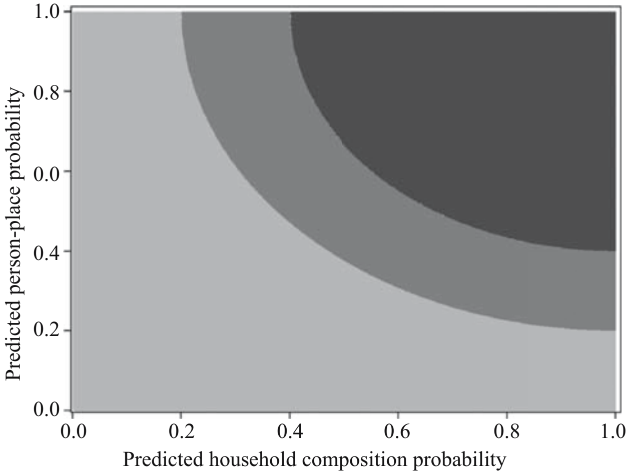


Fig. 7. Hypothetical example of using different occupied distance thresholds (source: 2010 Census simulation).

contacts or reduce respondent burden can be generally useful in surveys. However, admittedly, the use of administrative records in the decennial census is a less arduous problem due to the limited number of interview questions. The decennial census is only charged with forming a Census Day household roster of persons, to include minimal demographic data such as age, sex, Hispanic origin, race, and relationship for persons in occupied units. However, surveys such as the ACS have more data items with more complex topics. Nevertheless, provided that the administrative data relevant to the subject of measurement is available to the survey administrator, the methodology presented in this article can potentially be adapted for survey use.

For example, the Census Bureau has been researching the use of administrative records to reduce the difficulty and length of the American Community Survey to address concerns about respondent burden (Stempowski 2015). Ruggles (2015) identified potential administrative record sources for replacing or supplementing field response data. The American Community Survey Office (ACSO) at the Census Bureau has an active research program to further study topics and variables suggested in Ruggles (2015), for example, income (O’Hara et al. 2016), year built (Moore 2015), and housing value (Kingkade 2013). The preliminary work from ACSO has addressed the potential for an all-or-nothing use of administrative records to eliminate ACS questions. However the distance method could serve as an intermediate solution for reducing respondent burden by tailoring the survey questions based on the administrative record data availability and quality for each respondent. Specifically, historic survey data and the relevant administrative record data could be used to model the quality of administrative record data for a given question. The quality measures resulting from applying the model fit to a current round of data collection could then be used as in the distance method to repurpose the administrative record information via item substitution only for those units with quality exceeding some threshold. Such an adaptive strategy would reduce respondent burden, particularly those with consistently high-quality administrative record data. To our knowledge, such an implementation has not been studied or operationalized in any surveys. However, relatedly, Chesnut (2013) has studied modeling approaches for adaptive mode switching – model-based tailoring of contact strategies – to reduce respondent burden in the ACS.

5.4. Future Work

The definition of the distance measure for determining vacant and occupied units assumes equal weighting on the two corresponding predicted probabilities. We conjecture that this may be an issue for our AR Occupied identification because it uses two predictions with different census quality ramifications. The person-place model concerns counting people in the right place, whereas the household composition model concerns the agreement between administrative record household composition and census household composition. Differential weighting may be desired if it makes practical and empirical sense to emphasize one model over the other. Alternatively, transformations of the predicted probabilities may have an impact on the conclusions. The distance function uses the raw predicted values; however, the two dimensions each have a different dependent variable such that the distribution of predicted probabilities for each are not likely the same. Further work will examine if transformations including standardizations of the two probabilities can be useful in the determination.

An underlying assumption of the models in this research is that the relationships between the administrative records and the 2010 Census will remain consistent for the 2020 Census. The approach assumes the estimated relationships from the training data (e.g., 2010 Census data) can be reasonably applied for predicting the test data (e.g., 2020 Census data). Additionally, the approach assumes the 2010 Census data as "truth," even though there exists inherent error in Census results. Although the 2010 Census data is a reasonable basis for model development, the Census Bureau is actively researching the feasibility of using alternate training data to fit the administrative record models. For example, the use of more current ACS data as training data in conjunction with 2015 Census Test data could be treated as the gold standard (Chow et al. 2017).

Appendix

Table A1. List of independent variables for vacant and occupied models.

			Occupied models	
	Variable	Vacant model (Section 3.1)	Person-place (Section 3.2.1)	HH composition (Section 3.2.2)
	American community survey block group level variables			
% of	persons in block group (BG) between 25 and 44 years old	X	X	X
	persons in BG greater than 64 years old	X	X	X
	persons in BG identifying as Black	X	X	X
	persons in BG identifying as Hispanic	X	X	X
	occupied housing units in BG with at least 2 related HH members	X	X	X
	persons over 4 in BG speaking language other than English at home	X	X	X
	housing units in BG considered as mobile homes	X	X	X
	housing units in BG where householder/spouse are members of HH	X	X	X
	occupied housing units in BG that are not owner occupied	X	X	X
	housing units in BG vacant at time of interview		X	X
	housing units in BG occupied at time of interview	X		
	persons in BG living below poverty level	X	X	X
	Housing unit characteristics			
	# of neighbors in Non Response Followup (NRFU)	X		
	USPS Undeliverable As Addressed (UAA) reason (two mailings)	X	X	X
	USPS UAA reason agreement – Kappa Coefficient	X		
	housing unit type (e.g., multi-family)	X	X	X
	within structure description	X		

Table A1. Continued.

			Occupied models	
	Variable	Vacant model (Section 3.1)	Person-place (Section 3.2.1)	HH composition (Section 3.2.2)
	has Delivery Sequence File "X" status and both neighbors are in NRFU	X		
	on fall Delivery Sequence File of 2009	X	X	X
	apartment with Unable to Forward UAA reason code on 1st mailing	X		
	Housing unit characteristics from administrative records			
> = 1 person in HU is...	White			X
	Black	X		X
	Hispanic	X		
	missing ethnicity	X		
	age <2	X		X
	age <10	X		X
	age 10–17	X		X
	age 18–24			X
	age 25–44			X
	age 65+	X		X
	Housing unit level administrative record source information			
> = 1 person in HU is placed at this HU according to...	Internal Revenue Service (IRS) 1040 Tax Year (TY) 2009	X		X
	IRS 1099 TY 2009	X		X
	Indian Health Service Patient Database (IHS)			X
	Medicare			X
	Commercial data	X		X
	IRS 1040 TY 2008	X		
	Administrative Records (AR) HH count		X	
	AR HH composition	X	X	X
	HH with IRS 1040 TY 2008 persons, no AR persons in current year	X		
	IRS 1040 TY 2009 persons also in IRS 1040 TY 2008 at same unit			X

Table A1. Continued.

			Occupied models	
	Variable	Vacant model (Section 3.1)	Person-place (Section 3.2.1)	HH composition (Section 3.2.2)
> = 1 person in HU is placed at another HU according to...	IRS 1040 TY 2009	X		X
	IRS 1099 TY 2009	X		X
	Medicare	X		
	Commercial data	X		
Person level administrative record source information				
Person is placed at this HU according to...	IRS 1040 TY 2009		X	
	IRS 1099 TY 2009		X	
	IHS		X	
	Medicare		X	
	Commercial data		X	
	IRS 1040 TY 2008		X	
Person is placed at another HU according to...	IRS 1040 TY 2009		X	
	IRS 1099 TY 2009		X	
	IHS		X	
	Medicare		X	
	Commercial data		X	

6. References

Alvey, W. and F. Scheuren. 1982. “Background for an Administrative Record Census.” in JSM Proceedings, Social Statistics Section, American Statistical Association, Cincinnati, OH, August 16–19, 1982. Washington, DC: American Statistical Association. 137–152.

Bakker, B.F.M., P.G.M. van Heijden, and S. Scholtus. 2015. “ “Preface” to a Special Issue on Coverage Problems in Administrative Sources.” *Journal of Official Statistics* 31(3): 349–355. Doi: http://dx.doi.org/10.1515/jos-2015-0021.

Brackstone, G.J. 1987. “Issues in the Use of Administrative Records for Statistical Purposes.” *Survey Methodology* 13(1): 29–43. Available at: http://www.statcan.gc.ca/pub/12-001-x/1987001/article/14467-eng.pdf (accessed November 2017).

Chapin, M. and A. Keller. 2017. “Administrative Records Research and Planning, 2018 End-to-End Census Test: Nonresponse Followup.” from 2020 Census Program Management Review–April 21, 2017. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/pmr-materials/04-21-2017/pmr-update-testing-2017-04-21.pdf (accessed November 2017).

Chesnut, J. 2013. “Model-Based Mode Switching from Internet to Mail in the American Community Survey.” DSSD 2013 American Community Survey Memorandum Series

#ACS13-MP-01. Available at: https://census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Chesnut_01.pdf (accessed November 2017).

Chow, M.C., H.P. Janicki, M.J. Kutzbach, L.F. Warren, and M. Yi. 2017. "A Comparison of Training Modules for Administrative Records Use in Nonresponse Followup Operations: The 2010 Census and the American Community Survey." Center for Economic Studies Working Paper #CES 17-47. Washington, DC: U.S. Census Bureau. Available at: https://www2.census.gov/ces/wp/2017/CES-WP-17-47.pdf (accessed November 2017).

Cresce, A. 2012. "Evaluation of Gross Vacancy Rates From the 2010 Census Versus Current Surveys: Early Findings from Comparisons with the 2010 Census and the 2010 ACS 1-Year Estimates." Federal Committee on Statistical Methodology 2012 Research Conference, Washington, DC, January 10–12, 2012. Available at: https://www.census.gov/housing/files/FCSM%20paper.pdf (accessed November 2017).

Ennis, S.R., S.R. Porter, J.M. Noon, and E. Zapata. 2015. "When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records and Third Party Sources: Exploring Methods to Assign Responses." Center for Administrative Records Research and Applications Working Paper #2015-08. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2015/adrm/carra-wp-2015-08.pdf (accessed November 2017).

Federal Committee of Statistical Methodology. 1980. "Report on Statistical Uses of Administrative Records." *Working Paper 6*, Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/spwp6.pdf (accessed November 2017).

Fienberg, S.E. 2015. "Discussion" of a Special Issue on Coverage Problems in Administrative Sources. *Journal of Official Statistics* 31(3): 527–535. Doi: http://dx.doi.org/10.1515/jos-2015-0032.

Groves, R.M. and B.A. Harris-Kojetin (editors), National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington DC: The National Academies Press.

Japec, L., F. Kreuter, M. Berg, P. Biemer, P., Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. 2015. "AAPOR Report on Big Data: Report of the AAPOR Big Data Task Force." Available at: http://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf (accessed November 2017).

Keller, A. 2016. "Imputation Research for the 2020 Census." *Statistical Journal of the International Association of Official Statistics* 32: 189–198. Doi: http://dx.doi.org/10.3233/SJI-161009.

Keller, A., T. Fox, and V.T. Mule. 2016. "2014 Census Test – Analysis of Administrative Record Usage." U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-analysis-2014-census-test-ad-rec.pdf (accessed November 2017).

Keller, A. and S. Konicki. 2016. "Using 2010 Census Coverage Measurement Results to Better Understand Possible Administrative Records Incorporation in the Decennial Census." in JSM Proceedings, Survey Research Methods Section, American Statistical Association, Chicago, IL, July 30–August 4, 2016. Alexandria, VA: American

Statistical Association. 701–710. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2016/files/389544.pdf (accessed November 2017).

Kingkade, W. 2013. "Self-assessed Housing Values in the American Community Survey: An Exploratory Evaluation Using Linked Real Estate Records." in JSM Proceedings, Government Statistics Section, American Statistical Association, Montreal, Quebec, Canada, August 3-8, 2013. Alexandria, VA: American Statistical Association. 990–1004. Available at: https://ww2.amstat.org/MembersOnly/proceedings/2013/data/assets/handouts/308063_80215.pdf (accessed November 2017).

Layne, M., D. Wagner, and C. Rothhaas. 2014. "Estimating Record Linkage False Match Rate for the Person Identification Validation System." Center for Administrative Records Research and Applications Working Paper #2014-02. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-02.pdf (accessed November 2017).

Maris, M., E.S. Nordholt, and J. van Zeijl. 2012. "Comparing Approaches of Different (Partly) Register-based Countries." from the United Nations Economic Commission for Europe Conference of European Statisticians: UNECE-Eurostat Expert Group Meeting on Censuses Using Registers. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2012/use_of_register/WP_3_Netherlands.pdf (accessed November 2017).

Moore, B. 2015. "Preliminary Research for Replacing or Supplementing the Year Built Question on the American Community Survey with Administrative Records." U.S. Census Bureau Working Paper. Available at: www.census.gov/library/working-papers/2015/acs/2015_Moore_02.html (accessed November 2017).

Morris, D.S. 2014. "A Comparison of Methodologies for Classification of Administrative Records Quality for Census numeration." in JSM Proceedings, Survey Research Methods Section, American Statistical Association, Boston, MA, August 2–7, 2014. Alexandria, VA: American Statistical Association. 1729–1743. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2014/files/311864_88281.pdf (accessed November 2017).

Morris, D.S. 2017. "A Modeling Approach for Administrative Records Enumeration in the Decennial Census." *Public Opinion Quarterly: Special Issue on Survey Research, Today and Tomorrow* 81(S1): 357–384. Doi: http://dx.doi.org/10.1093/poq/nfw059.

Morris, D.S., A. Keller, and B. Clark. 2016. "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the International Association of Official Statistics* 32: 177–188. Doi: http://dx.doi.org/10.3233/SJI-161002.

Mule, V.T. and A. Keller. 2014. "Using Administrative Records to Reduce Nonresponse Followup Operations." in JSM Proceedings, Survey Research Methods Section, American Statistical Association, Boston, MA, August 2–7, 2014. Alexandria, VA: American Statistical Association. 3601–3608. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2014/files/313148_90659.pdf (accessed November 2017).

Mule, T. and S. Konicki. 2012. "2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Housing Units." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-02. https://www.census.gov/coverage_measurement/pdfs/g02.pdf (accessed November 2017).

Mulry, M.H. 2014. "Measuring Undercounts for Hard-to-Survey Groups." In *Hard-to-Survey Populations* (Chapter 3), edited by R. Tourangeau, N. Bates, B. Edwards, T. Johnson, and K. Wolter. Cambridge, England: Cambridge University Press. 37–57.

O'Hara, A., A. Bee, and J. Mitchell. 2016. "Preliminary Research for Replacing or Supplementing the Income Question on the American Community Survey with Administrative Records." U.S. Census Bureau Working Paper. Available at: www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Ohara_01.pdf (accessed November 2017).

Rastogi, S. and A. O'Hara. 2012. "2010 Census Match Study Report." 2010 Census Planning Memorandum Series. Available at: https://www.census.gov/2010census/pdf/2010_Census_Match_Study_Report.pdf (accessed November 2017).

Ruggles, P. 2015. "Review of Administrative Record Sources Relevant to the American Community Survey." U.S. Census Bureau Working Paper. Available at: www.census.gov/library/working-papers/2015/acs/2015_Ruggles_01.html (accessed November 2017).

Scheuren, F. 1999. "Administrative Records and Census Taking." *Survey Methodology* 25(2): 151–160. Available at: http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4878-eng.pdf?contentType=application%2Fpdf (accessed November 2017).

Steffey, D.L. and N.M. Bradburn (editors), National Research Council. 1994. *Counting People in the Information Age*. Washington DC: The National Academies Press.

Stempowski, D. 2015. "Agility in Action: A Snapshot of Enhancements to the American Community Survey." ACS Information Series Memorandum Number 2015-05. Available at: https://www.census.gov/content/dam/Census/programs-surveys/acs/operations-and-administration/2015-16-survey-enhancements/Agility%20in%20Action%20v2.0.pdf (accessed November 2017).

Thygesen, L. 2015. "The Use of Administrative Sources for Censuses: Merits and Challenges." *Statistical Journal of the IAOS* 31(3): 381–389. Doi: http://dx.doi.org/10.3233/SJI-150909.

United Nations Economic Commission for Europe (UNECE). 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. United Nations Publication. www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed November 2017).

United States Census Bureau. 2017a. *2020 Census Operational Plan: Version 3.0*. Washington DC: Census Bureau. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan3.pdf (accessed November 2017).

United States Census Bureau. 2017b. "Administrative Records Modeling Update for the Census Scientific Advisory Committee." Presented at the Census Scientific Advisory Committee Meeting–March, 30, 2017. Available at: https://www2.census.gov/cac/sac/meetings/2017-03/admin-records-modeling.pdf (accessed November 2017).

United States Office of Management and Budget. 2014. *M-14-06: Guidance for Providing and Using Administrative Data for Statistical Purposes*. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf (accessed November 2017).

van Zeijl, J. 2014. “From Traditional to Register-Based Censuses in the Netherlands.” from the National Academies of Science: International Conference on Census Methods. Available at: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_088800.pdf (accessed November 2017).

Wagner, D. and M. Layne. 2014. “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software.” Center for Administrative Records Research and Applications Working Paper #2014-01. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf (accessed November 2017).

Walejko, G., A. Keller, G. Dusch, and P.V. Miller. 2014. “2020 Research and Testing: 2013 Census Test Assessment.” U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/programs-surveys/decennial/2020-census/2013_Census_Test_Assessment_Final.pdf (accessed November 2017).

Walker, S., S. Winder, G. Jackson, and S. Heimel. 2012. “2010 Census Nonresponse Followup Operations Assessment.” 2010 Census Planning Memoranda Series, No. 190, April 30, 2012. Available at: https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf (accessed November 2017).

Wallgren, A. and B. Wallgren. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley and Sons.

Received March 2016
Revised November 2017
Accepted March 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 625–648, http://dx.doi.org/10.2478/JOS-2018-0030

Transitioning a Survey to Self-Administration using Adaptive, Responsive, and Tailored (ART) Design Principles and Data Visualization

Joe Murphy[1], *Paul Biemer*[2], *and Chip Berry*[3]

This article discusses the critical and complex design decisions associated with transitioning an interviewer-administered survey to a self-administered, postal, web/paper survey. Our approach embeds adaptive, responsive, and tailored (ART) design principles and data visualization during a multi-phased data collection operation to project the outcomes of each phase in preparation for subsequent phases. This requires rapid decision making based upon experimental results using a data visualization system to monitor critical-to-quality (CTQ) metrics and facilitate projections of outcomes from the current phase of data collection to inform the design of the subsequent phase. We describe the objectives of the overall design, the features designed to address these objectives, components of the visual adaptive total design (ATD) system for monitoring quality components and relative costs in real time, and examples of the visualization elements and functionalities that were used in one case study. We also discuss subsequent initiatives to develop an interactive version of the monitoring tool and applications for other studies, including those employing adaptive, responsive, and tailored (ART) designs. Our case study is a series of pilot studies conducted for the Residential Energy Consumption Survey (RECS), sponsored by the U.S. Energy Information Administration (EIA).

Key words: Responsive design, adaptive design, monitoring, data collection, visualization.

1. Introduction

Interviewer-administered survey modes, such as face-to-face and telephone, have traditionally been viewed as the standard for collecting high quality data on nonsensitive topics. The presence of an interviewer can help foster cooperation with the respondent and the interaction between interviewer and respondent can help assure that key questions are interpreted and answered correctly. While they may set the standard for quality, interviewer-administered modes are typically more costly than modes that do not involve interviewers, such as web and paper surveys. Costs for recruiting and training interviewers, their salaries, and their transportation costs (for face-to-face surveys) are major investments for a survey project. From a total survey error perspective, allocating such a large share of the survey budget to face-to-face interviewing may be suboptimal for some

[1] RTI International, 230 W Monroe St. Suite 2100, Chicago, IL, U.S.A, 60606. Email: jmurphy@rti.org
[2] RTI International, 3040 East Cornwallis Road, Post Office Box 12194, Research Triangle Park, NC 27709-2194 U.S.A. Email: ppb@rti.org
[3] U.S. Energy Information Administration; 1000 Independence Ave., SW, Washington, DC 20585, U.S.A. Email: james.berry@eia.gov

surveys (Groves 1989; Biemer 2010). For example, using a less expensive data collection mode could allow a larger sample size, more extensive nonresponse follow-up, more questionnaire pretesting, and the elimination of interviewer error. In addition, using a mail delivery mode obviates the need for cluster sampling that is often required in face-to-face surveys to reduce interviewer travel costs.

In recent years, it has only become more difficult to efficiently collect data, regardless of mode. Response rates have continued to decline and, to achieve acceptable response rates, costs have increased. Some well-established surveys in the U.S. that have employed interviewers in the past such as the Longitudinal Survey of Adolescent Health (Biemer et al. 2017a), the Behavior Risk Factor Surveillance Survey (Link and Mokdad 2005), the National Health Care Interview Survey (Howden et al. 2015) and the Residential Energy Consumption Survey (RECS) (Eddy and Marton 2012) have investigated or are now considering changes to incorporate self-administered designs for cost reduction. However, the change to self-administration is not simple or straightforward when comparable, high quality data are desired using these less expensive modes.

For surveys that are conducted periodically, such as repeated cross-sectional or longitudinal studies, the decision to change modes is only the first in a series of design decisions that must be made before implementing a specific self-administered data collection protocol. To inform this rather complex and challenging transition, a series of pilot studies can be designed to experimentally test a range of alternative designs and identify the best data collection approaches for self-administration, including web and paper modes. Such a design may require that a series of experiments be conducted within a highly compressed schedule with little time between studies for data analysis. Yet, it is essential that the results and lessons learned from each experiment be thoroughly understood and transferred across experiments. As such, design decisions need to be made for the next set of experiments before data collection and analysis for the previous set of experiments are fully completed. Key to decision making is a system for monitoring and visualizing the results of data collection while the survey is in progress. Groves and Heeringa (2006) discuss this problem in the context of a three-phase responsive design for the National Survey of Family Growth where the first phase constituted an experiment that was followed immediately by the main data collection phase. That phase was then followed immediately by a nonresponse follow-up phase. Decisions based upon incomplete data were made during each phase that affected the design of the subsequent phase. To facilitate this process, the study team reviewed daily updates on the results for each treatment or design feature being monitored. This approach involved defining multiple quality components and their metrics as well as a system to compile a vast amount of information for quick, clear, presentation and a minimum of burden on the survey managers.

In this article, we discuss the key design decisions required for the transition of an interviewer-administered survey to self-administration via paper and web using a series of data collection phases. Our approach embeds adaptive, responsive, and tailored (ART) design principles and data visualization during a multi-phased data collection operation to project the outcomes of each phase in preparation for subsequent phases. Key to this process is identifying critical-to-quality (CTQ) metrics to monitor and a data visualization system to meet the requirements of the research strategy. We describe the objectives of the design and system, the features required to address these objectives, and the implementation of a visualization approach for monitoring costs and quality components in real time. We also describe and provide examples of the visualization elements we created and applied as well as their functionalities. Looking forward, we discuss current initiatives to develop an interactive version of the monitoring tool with applications in subsequent studies especially those employing ART designs.

Our case study is a series of pilot studies conducted for the U.S. Energy Information Administration (EIA) for the RECS. The goal of these studies was to assess the operational feasibility, data quality and costs of converting the RECS to a web and paper mixed mode design. The RECS has traditionally been conducted by face-to-face interviewing; however, self-administration via web and paper questionnaires represents an opportunity to lower costs, gather more timely and frequent data, and expand sample sizes to meet ever-expanding user precision requirements.

2. Embedding ART Principles in the Study Design

When considering a change from interviewer- to self-administration, several questions emerge. Key among these are the following:

- Will questions in the interviewer-administered setting translate to provide comparable data in the self-administered setting?
- Will sample members respond to the survey at an acceptably high rate? Will those who respond represent the population of interest?
- Can the survey collect high quality data (e.g., low measurement error) while leveraging the efficiencies of self-administered modes?
- What data collection protocol will yield the best overall quality given the survey goals?
- What data collection protocol will be most cost efficient?

The first question can be evaluated using a variety of pretesting methods, including cognitive interviews and online pretesting (Murphy et al. 2016; Edgar et al. 2016). To answer the remaining questions, we can design experiments in which one or more features of the data collection protocol is altered. For example, we may conduct an experiment using different incentive levels to determine which is most appropriate given the goals of the survey. Or we may randomly assign some sample members to a version of the survey with a shorter completion time and others to a longer one to determine the tradeoffs between information gained overall and from individual cases. Often, there are more issues than can be feasibly investigated in a single round of experiments. Sometimes the design options to be tested are interdependent. For example, whether incentives should be guided by response propensity models depends upon how those propensity models perform. In this case, it may be advantageous to conduct experiments iteratively, where the "best" protocol identified in one phase of the survey is carried into the subsequent phase, while the protocols that did not yield good results are excluded.

Our ability to draw conclusions from such experiments and answer questions to inform the "best" design for a survey depends on 1) the data available, 2) the specific components of quality to be monitored and 3) the interventions at our disposal to affect design features to improve quality. These three requirements can vary greatly depending on the survey mode or modes employed in the survey. For example, for a face-to-face household survey, we can measure the effectiveness of different contact strategies and interviewers, the timing and level of effort devoted to contacting respondents, the physical characteristics of the household, interviewer performance metrics and other paradata. These and other data can be monitored in real-time, analyzed to determine if an intervention is warranted and, if so, to deploy whatever intervention is indicated as quickly as possible.

Several design strategies offer the potential to help determine the best fit in terms of approaches for a survey or individual sampled units. The *tailored design* method (Luiten and Schouten 2013) advocates varying the survey protocol across population subgroups rather than using a "one size fits all" approach. This approach attempts to customize the survey design to individual preferences in order to minimize the total error for the entire sample. Another form of tailoring uses a combination of survey design features demonstrated to be effective in the literature to construct a single, optimal survey design that, when applied to the entire sample, will provide excellent results across a wide range of survey topics and populations (Dillman et al. 2014). *Responsive design* was introduced by Groves and Heeringa (2006) as an approach to adjust the data collection protocol for a single survey based on the outcomes of an initial set of cases. By continuously assessing the results of the data collection process and remaining resources, strategies can be modified for the remaining cases to be pursued (Laflamme and Wagner 2016). *Adaptive survey design* similarly proposes different approaches within the same survey, but focuses on the heterogeneity of sample cases and identifying the optimal survey protocol for each individual. For example, adaptive survey design recognizes that some sample members may be swayed to participate in a survey by incentives where others will not. Design-specific response propensities can be calculated based on paradata for each individual sample member (Schouten et al. 2017; Chun et al. 2017).

The successful ART design should adhere to these simple but key principles:

1. identify a few, critical factors that drive costs and quality (i.e., CTQs) and focus attention on these throughout the process,
2. create and monitor metrics that are strongly associated with CTQ outcomes and intervene when these metrics deviate beyond their acceptable limits and,
3. verify that the interventions were successful and that the aberrant CTQ metrics return to their acceptable limits.

A fourth overarching principle is to simplify the quality management strategy to the extent possible using informative graphical displays, parsimony in the selection of CTQs and their corresponding metrics, and a focused strategy for continual improvement of the CTQs.

These general principles are common to all three approaches – that is, A, R, and T – where the specific approach can be viewed as a variant in the way these principles are applied. For example, responsive designs may incorporate experimental phase and may use the concept of phase capacity to signal the end of a phase. These features address the principles of monitoring metrics and intervening when they meet certain prespecified criteria. Likewise, the tailored designs may attempt to adapt the data collection protocol to specific subgroups of the population. This feature can be viewed as application of

Principle 1, where metrics are defined at the subgroup level and interventions can vary by subgroup. Finally, adaptive design can contain elements of both responsive and tailored designs, but may focus more broadly on total survey error and costs, clearly in the spirit of Principles 1, 2, and 3.

ART considerations for in-person surveys are, overall, rather complex and may represent a wide array of potential CTQs for reducing error risks and costs. By comparison, the considerations for self-administered, postal surveys are relatively straightforward. For example, the typical web/paper survey involves a series of participation requests and reminders sent by mail or, if available, email according to a prespecified contact schedule with little room for deviation. The absence of interviewers and control over the timing of contact results in fewer variables to consider. However, problems can occur that may require rapid intervention when the observed results deviate substantially from expectations. In that situation, the interventions may be limited to actions such as: using additional contacts, increasing the sample size, altering the wording of the invitations, or something similar, none of which represent a substantial departure from the planned protocol. Mail invitations are typically sent in large batches and while the U.S. Postal Service (USPS) returns letters that were undeliverable, the outcome of the contact attempt is unknown unless the respondent actively participated or contacts the survey organization to refuse participation. As such, identifying, monitoring, and using data for ART designs in a web and/or mail survey environment presents a unique challenge.

Our recommended approach for surveys transitioning to self-administration is to incorporate elements of ART designs where appropriate. For instance, a design may be responsive by including multiple phases of data collection, each drawing from the successful strategies of the previous phase. The design can be adaptive in the sense that it uses paradata metrics to monitor quality during each phase of data collection to consider the appropriate treatment for each individual case. And it can be tailored in sense that it attempts to vary the survey protocol according to the (often predicted) preference of the sample member; as an example, using a paper-questionnaire-only protocol for sample members who do not have internet access.

When designing a protocol for a sequence of experiments to be conducted in rapid succession, it is vitally important to identify the goals and metrics for success from the outset. While all surveys strive for high quality in the data collected and estimates produced, the exact definition of "quality" may differ from project to project. For this reason, it is crucial for the survey stakeholders (sponsors, data collectors, data users, etc.) to discuss the definitions of quality and success from the very beginning of the survey planning stages. Once quality is defined, it is a matter of operationalizing this definition by selecting metrics that can be tracked during data collection that can serve as CTQs. These indicators can reflect quality dimensions such as successful study recruitment (response rates), the extent to which respondents represent a benchmark measurement of the population of interest (e.g., demographic characteristics that match Census estimates), success in obtaining responses at the item level from respondents, the ability to push respondents to respond via web rather than paper, and so on. Assuming a rapid development schedule, once the CTQs are identified, a system needs to be put in place to track these metrics during data collection so design decisions for the subsequent phase can be made before the current phase is complete.

The task of identifying metrics or indicators for a CTQ is seldom straightforward. As an example, the response rate is commonly employed for the CTQ to minimize nonresponse bias; yet, other metrics maybe better indicators of nonresponse bias. Often the best solution is to employ several metrics that may reflect different dimensions of the CTQ but that can add complexity to the monitoring task. Thus, it is important to strike a balance between parsimony and completeness. In the RECS pilot studies, we monitored response rates as well as a measure of representativity known as Cramér's V. In addition, it may be futile to define a CTQ for which there is no opportunity or plan to intervene on behalf of the CTQ. As an example, in the RECS pilot studies, obtaining accurate reports of household appliances was certainly an objective of questionnaire design; we refined these questions during pretesting and checked the data for anomalies at several points during data collection, but no steps were taken to monitor this indicator on a daily basis.

In the following section, we present a case study to illustrate a rapid sequence of experiments for the RECS that incorporate various ART design principles. We describe our process for identifying CTQs and monitoring them during data collection using a visualization system designed for such scenarios.

3. The Need for Rapid Decision-Making and Role of Visualization for the RECS Pilots

The RECS originated in 1978 and has been conducted periodically by the EIA since then. The RECS program is responsible for collecting and disseminating timely, detailed information about how energy is being used within the residential sector of the economy. This includes data on the fuels used in homes, equipment and appliance stocks, household behaviors, and disaggregated consumption and expenditures. In this article, we discuss the RECS that was conducted in 2015. Prior to this, RECS was conducted by face-to-face interviewing in 2009. As the opportunities and challenges associated with survey research have changed over the years, the planning for each RECS has required reflection on how best to meet the goals of the survey and needs of the data users while maintaining comparability with past rounds, and adhering to schedule and budget constraints.

As previously noted, the RECS has been conducted by field interviewers using computer-assisted personal interviewing (CAPI). The costs of this data collection mode are relatively high, averaging nearly USD 400 per completed interview for the 2009 study, which limits other important quality initiatives, such as more frequent data collection, larger sample sizes, and precise estimates for more geographic areas. EIA commissioned an expert panel study of the National Research Council of the National Academy of Sciences (NAS) to examine its energy demand surveys, identify gaps in substantive coverage, and make recommendations for EIA's priorities for data collection (Eddy and Marton 2012). One suggestion from the panel was to explore alternative data collection approaches for the RECS, specifically incorporating a self-administered web mode. EIA followed the NAS recommendation with the goal of assessing whether self-administered modes could result in the collection of high-quality data in an environment where field data collection was continuing to face increasing challenges.

RECS had always achieved response rates at or above 80 percent. It was apparent that such high response rates would not be feasible in today's environment, especially using a mail-delivered questionnaire protocol. Thus, EIA faced many questions about the trade-off between cost and quality with this radical shift to a web/paper design. The schedule for conducting the next RECS meant that EIA would need to test several different promising protocols in a very short period of time before selecting the one that would serve as the most appropriate production system for the future. In light of these challenges, EIA determined that, using the ideas of responsive design, a multi-phased pilot study design would best meet their needs considering costs, timing and goals. With a phased approach, a well-selected set of design features could be tested at each phase that took advantage of the lessons learned and the results gleaned from the prior phase's experiments. Then the final phase of testing could incorporate the best features of the prior phases.

The following sections describe each RECS pilot phase's timing, design, and results. We also describe how early results from each phase were used to inform subsequent phases, as well as how the official 2015 RECS CAPI study was ultimately impacted by the pilot tests.

3.1. Phase 1: The Cities Pilot

The first RECS phase, referred to as the Cities Pilot Test, collected responses from an address-based sample of households in five U.S. cities. Planning and design for the Cities Pilot Test began in December 2014 and data collection began in March 2015, continuing into July 2015. Planning for this survey involved extensive cognitive interviewing and pretesting to determine how best to shorten the 40-minute traditional face-to-face questionnaire to a 20–30 minute self-administered instrument (Murphy et al. 2016). Further, extensive analysis was conducted on the energy characteristics of U.S. cities in order to identify five cities that together could sufficiently represent the diverse and challenging issues to be confronted in the redesign of the national RECS.

In addition to assessing the viability of a self-administered RECS, the Cities Pilot included two experiments to evaluate options for key design components. These components were (1) questionnaire length and (2) initial mode assignment. We found evidence that the 30-minute self-administered RECS achieved a similar response rate to the 20-minute version and deemed it feasible and efficient for both web and paper modes. The Cities Pilot Test achieved a higher-than-expected response rate overall (38%) within budget and demonstrated that data collection can be accomplished within a 14-week field period. We also found that tailoring the initial mode assignment (either by web only or by paper) based upon a model predicting the propensity to respond by each mode was not effective primarily because our working hypothesis did not hold. We hypothesized that households that do not have broadband Internet access would prefer the paper mode. Thus, we developed a model for the probability a household has Internet access as described in Zimmer et al. (2016). But while the propensity model was reasonably accurate for predicting Internet access, Internet access was apparently not a good indicator of mode preference and thus response rates for the model-guided protocol were not significantly higher than the control group which used a static web-first mode assignment (Zimmer et al. 2016). Given these results, mode tailoring based on internet access propensity was not used in the subsequent phases.

As shown in Figure 1, the Cities Pilot Test data collection phase overlapped significantly with planning for the next phase, referred to as the National Pilot Test. Daily tracking of key Cities Pilot quality metrics was instrumental in determining design and experiment options for the subsequent phase. Monthly, or even weekly, status reports would have been insufficient if the project were to stay on schedule.

3.2. *Phase 2: The National Pilot Test*

The RECS National Pilot Test was planned to run more or less concurrently with the official 2015 RECS CAPI study and, thus, planning and decision making for the National Pilot needed to take place while the Cities Pilot data were still being collected. Planning and design for the National Pilot Test began in April 2015; data collection ran from September 2015 to January 2016. The RECS National Pilot Test collected responses from a national address-based sample of households and expanded on lessons learned from the RECS Cities Pilot, carrying over the 30-minute questionnaire length and materials developed for that previous pilot.

Like the Cities Pilot Test, the National Pilot Test phase included experiments to continue to explore the most successful protocol for web and paper administration of the RECS according to the criteria outlined in Section 4 of this article. Because the Cities Pilot Test showed evidence of the superiority of the web mode for data collection in terms of cost and data quality, we aimed for a design that would effectively "push" respondents to the web (Dillman 2016). Since respondents in the Cities Pilot generally preferred to respond via paper rather than web, a key question for the National Pilot Test was whether participants could be incentivized to respond by web rather than paper.

The pilot evaluated eight treatment combinations of equal sample size formed by crossing two factors: respondent incentives (Factor A) with two levels and mode protocols (Factor B) with four levels forming a two-by-four factorial design. The two factors and their levels are defined as follows:

- **A1.** A USD 5 prepaid incentive included in the first questionnaire mailing; USD 10 was promised for response under the response protocol specified by Factor B.
- **A2.** A USD 5 prepaid incentive was included in the first questionnaire mailing; USD 20 was promised for response under the response protocol specified by Factor B.
- **B1.** Web Only Protocol–only the web response option was offered for all survey response invitations.

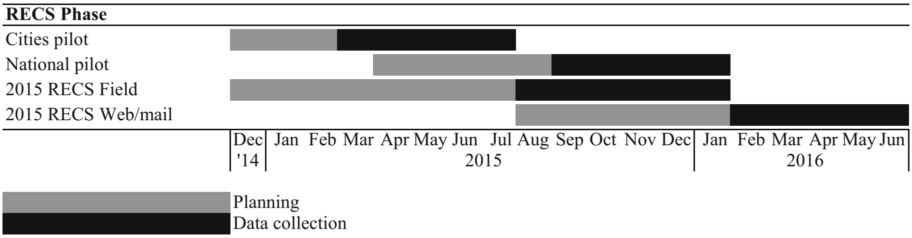


Fig. 1. RECS pilots timeline.

- **B2.** Web/Paper Protocol – the web response option was offered in the first invitation and first nonresponse invitation; both web and paper were offered in all subsequent invitations.
- **B3.** Choice Protocol – response by either paper or web questionnaire was requested by each survey response invitation.
- **B4.** Choice+ Protocol – response by either paper or web questionnaire was requested by each survey response invitation. However, a USD 10 promised bonus incentive was provided in addition to the incentives specified by Factor A if the respondent chose to respond by the web option rather than by paper.

The National Pilot Test found that the most successful protocol included a USD 5 unconditional cash pre-incentive plus a USD 10 cash promised incentive for participation. Those electing to respond via web rather than paper were offered an additional USD 10 for completing. This protocol, termed "Choice Plus" (Choice+) is discussed in detail in Biemer et al. (2017b).

Following the main data collection period of the National Pilot, all non-respondents except refusals received an extended nonresponse followup (xNRFU). A single UPS high-priority mailing was sent to these addresses containing the offer letter, an abbreviated, one-page questionnaire and a postage-paid return envelope. A random half of the xNRFU sample was offered an additional (that is, in addition to the incentives they would have received under Factors A and B) USD 10 if they completed the abbreviated questionnaire and returned it in the stamped envelope.

We calculated response rates using American Association for Public Opinion Research formula RR3 (AAPOR 2015). The final overall response rate for the main phase of the National Pilot was 40.4 percent. The overall rate rose to 54.9 percent after the completion of the nonresponse follow-up phase.

3.3. Phase 3: 2015 RECS CAPI Remediation

While the National Pilot was being conducted, the 2015 RECS CAPI was also in the field with interviewers making face-to-face visits to selected households. The National Pilot Test ran concurrently with the 2015 RECS CAPI so we could compare the results of data collection and estimates obtained by survey mode. In January 2016, it became apparent that the 2015 RECS would not achieve its goals in terms of cost and response using interviewer administration. In contrast, the National Pilot had concluded and demonstrated that self-administered modes could be used to achieve good data quality, acceptable response rates, and costs several times lower per case than CAPI. Given the success of the National Pilot Test to that point, and in particular the Choice+ protocol, EIA made the decision to transition most unresolved or unreleased sample cases in the 2015 RECS CAPI to self-administration by web and paper using the Choice+ protocol beginning in February 2015. The 2015 RECS CAPI remediation phase continued through June 2016.

4. Identifying and Visualizing CTQ Metrics

Historically, RECS had a fairly stable design with a predictable range of outcomes. The response rate for the CAPI studies, for example, was consistently 80 percent or higher.

To monitor field progress during data collection, RECS project staff relied on summary field and cost tables which were compiled at regular intervals. These static tables, which included weekly and cumulative labor and travel costs for the entire sample and by geography, were sufficient. The use of more detailed metrics was limited to field supervisors as a means to appropriately assign interviewer work. Given its quadrennial cycle, there was also ample time to analyze field results following data collection and plan for any protocol changes for the next round.

The objectives of the RECS Pilot Test required rapid decision making in order to plan the subsequent round, and it was necessary to track data collection metrics from the outset and at more frequent intervals. As the process evolved, it was essential to assess the current state of progress for CTQs in a way that did not require a significant time investment from staff. Making design decisions on the schedule presented in Figure 1 required real-time (daily) monitoring of the performance across multiple quality indicators.

To develop a quality monitoring system, the first step was to agree upon the definition of quality for the pilot tests. A number of design features were predetermined such as the data collection modes (web/paper), overall sample sizes, the use of an address-based sampling frame, number of mailings and incentivized response. Given these fixed assumptions, the definition of quality and what constituted a successful outcome from the survey sponsor's perspective drove the remaining design decisions. Thus, quality was essentially defined as a balance across the following four CTQs:

- participation rates (higher is better, all else being equal),
- web response rate (higher response by web compared to response by paper is better),
- respondent sample representativeness relative to external benchmarks (higher concordance with benchmark is better), and
- relative costs per completed case (lower is better).

Another important quality goal for any survey transitioning from face to face to web/mail mixed mode data collection should be the evaluation and control of mode effects which are essentially the methodological differences in the estimates due to the change in mode. The RECS is certainly not immune to mode effects; in fact, significant mode effects were expected for some characteristics, most notably the ascertainment of housing unit square footage. In the face to face mode, interviewers can explain the "official" concept of housing unit square footage and even assist the respondent in estimating it. In self-administered modes, respondents are only aided by the instructions embedded in the instrument which can be quite technical. Unless they happen to know the square footage of their home, respondents often err in its estimation. Unfortunately, measurement errors are quite difficult to monitor and control in real-time during web/mail data collection. For the RECS, only post-survey evaluations of measurement error were conducted. In particular, a post-survey analysis of RECS square footage data can be found in Amaya et al. (2017).

While we did not set a quantitative value for each of these CTQs to identify what worked "best," we monitored the results as the pilot tests were conducted and frequently discussed the trends in the process of selecting methods appropriate for the subsequent pilot. Biemer et al. (2017b) provides a discussion of the specific rationale for the chosen survey experiments for the RECS National Pilot design to identify the "best" design given the definition of quality above.

The CTQs we identified for real-time monitoring can be classified into the following categories:

1. *Participation.* We sought high rates of participation across key domains defined by housing unit/household characteristics. We also sought high rates of participation in the early stages of data collection to minimize the cost of multiple follow-up mailings. We also tracked submission rate metrics for each of the experimental conditions. Here, submission rate refers to the count of cases submitted via web or paper form divided by the total number of sampled cases. The rate served as a simple proxy for response rate, which was not calculated until the end of data collection due to the timing of defining criteria for the estimation of eligibility among cases of unknown eligibility (e.g., cases with no evidence of receipt of contact with a respondent, USPS and UPS undeliverables). Cumulative daily submission rates were monitored overall and by: experimental treatment, mode protocol, promised incentive amount, geographic region, and urbanicity.
2. *Rate of response via web (rather than paper).* We sought to minimize cost by encouraging response via web survey rather than paper, since paper included extra costs for printing, return postage, receipt, data entry, and data review. We also sought to minimize measurement error from item nonresponse, out-of-range responses, and errors in following skip patterns by encouraging web vs. paper response. We monitored the rate of web submission overall and by the same factors noted for participation.
3. *Respondent representativeness compared with the sample and an external benchmark.* We sought a balanced unweighted distribution of respondents relative to benchmark data sources. To assess representativeness, we compared RECS Pilot responding housing unit/household distributions to the corresponding distributions for the 2014 U.S. American Community Survey (ACS) 1-year estimates (U.S. Census Bureau 2015). The variables compared included: type of housing unit, main heating fuel used, household income, and age of respondent/householder. Additionally, we compared respondents to sampling frame distributions using a variable available for both: housing unit building type (single family vs. multi-unit). We also considered comparing RECS Pilot respondents with 2015 RECS CAPI respondents or 2009 weighted estimates on energy-specific metrics, such as water heating fuel and number of refrigerators, as a means to track the bridge between old and new survey methods. These metrics were not part of the CTQ tracking system, but rather were evaluated at the conclusion of the data collection.
4. *Relative cost per case.* We calculated the costs associated with printing materials, mailings, receipt of completed questionnaires, and incentive payments for each protocol. Depending on whether and when each sample member responded, the cost per case varied. By tracking costs at the case level, we could determine the overall costs at the protocol summary level over the course of data collection. We measured costs relative to the simplest and expected least costly protocol, Web Only. We also considered the costs associated with data editing needs for each protocol. Data editing involved necessary recodes to reported values, such as the assignment of values to open-ended responses or edits to ensure consistency of responses. The need

for data editing did not necessarily suggest lower quality data, but did require staff resources, and so we discuss it as a cost metric.

With our CTQs for the pilot tests identified, the next challenge was selecting the best system for monitoring progress data collection. In selecting an approach for CTQ visualization, we identified several system requirements specific to our purposes:

- The system should be simple enough to be quickly accessed and understood by a wide range of project staff.
- Charts should limit the number of data series plotted (no more than five lines on a single chart at once; as needed, "small multiples" of charts by specific dimensions should be employed (Cleveland 1993; Tufte 2001).
- The chart axes, legends, and labels should be large enough to read easily and include descriptive labels to minimize the effort for a user to gain information.
- Charts should use a consistent format and consistently use patterns, colors and symbols so users do not have to re-orient when examining multiple charts.
- The charts should be fully interpretable when printed in black and white.
- All charts should be accompanied by full data tables so users can reference exact values when needed.
- The charts should not require purchase or licensing of special software not already available to users.
- The system had to be cost efficient and not require extensive use of IT resources.

We determined that the goals stated above could best be met using software already at our disposal – SAS and Microsoft Excel. The creation process began with a nightly export of data from the survey control system into a single SAS database containing sample, frame, case history, web response data, paper response data, and auxiliary data. SAS version 9.4 was used to compute CTQs in counts per day, cumulative counts per day, and cumulative rates per day for each metric. The same automated SAS program then prepared data tables in the format necessary for export to an Excel 2013 workbook with preformatted chart shells.

The process was automated to run on a daily basis for sharing with the project team on a shared secure web portal. By containing all the output in a single Excel workbook, we had available a self-contained, convenient, and widely familiar format for sharing pilot progress data across the project team and with EIA managers, as needed. And although the charts were not necessarily "interactive," it should be noted that Excel does include the default option to hover the cursor over a data point to view its exact x and y values.

For chart designs we looked to the literature and our own experience and intuition about the simplest and most effective displays. For example, Tufte (2001) advocates for maximizing the "data-ink ratio" or "proportion of a graphic's ink devoted to the non-redundant display of data-information" in data visualization. As a result, we sought to eliminate any elements that were not helpful for quick interpretation of the data such as excessive gridlines, non-meaningful uses of color, redundant labels, and so on. We consciously adhered to other evidence and advice from documented best practices of data visualization regarding the choice of chart types. For instance, for our cumulative charts, we used line charts connecting individual numeric data points, which have been advocated as a “simple, straightforward way to visualize a sequence of values. Their primary use is to display trends over a period of time” (Hardin et al. 2012).

Figure 2 presents a “snapshot” of the RECS National Pilot monitoring display referred to as an Adaptive Total Design (ATD) monitoring chart (see, for example, Biemer 2010), formatted using a dashboard-type design. The legend at the top of the figure describes the markers used on the x-axis in this and subsequent figures to identify key dates in the data collection protocol. Having this information consistently displayed daily with various metrics in close proximity allowed our team to quickly ascertain and keep apprised of the various CTQs and their performance throughout data collection. The charts included a mixture of line (for time-dependent, cumulative rates), bar charts, and maps. Though it is not a standard option for Excel, maps are not difficult to add to a workbook using Visual Basic for Applications (VBA) (Camoes 2008).

To illustrate an individual chart from the dashboard, in Figure 3 we present cumulative submission rates by protocol during the National Pilot Test data collection period. At a glance, it is obvious that submission rates rose most rapidly at the beginning of data

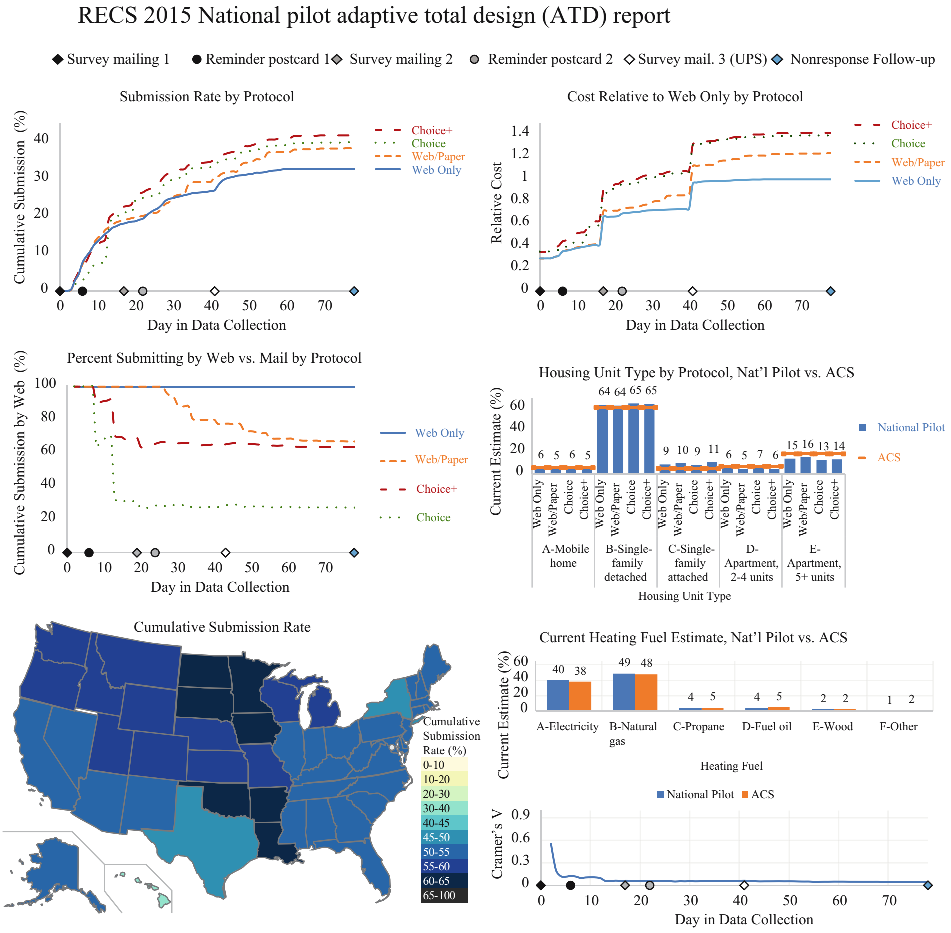


Fig. 2. National pilot ATD monitoring charts (partial view).

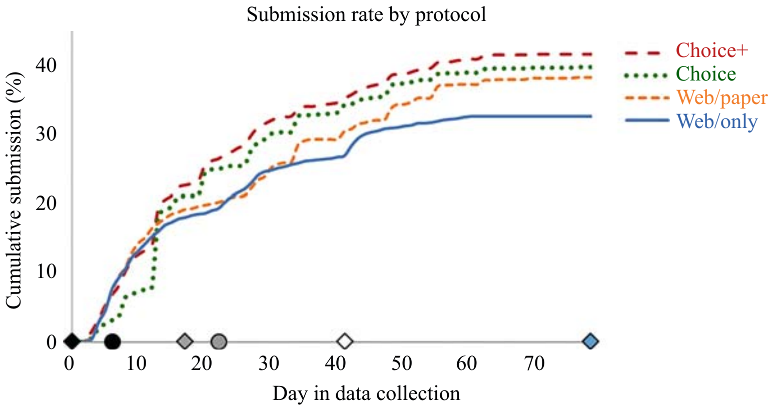


Fig. 3. National pilot submission rates by protocol.

collection in response to several mailings to sampled households. A second spike occurred just after the nonresponse followup mailing (diamond at day 78) among all experimental protocol groups, but the increase was most dramatic for the Web Only group that had not previously been offered a paper option for response. Figure 3 exemplifies the format of our charts. Reviewing this chart regularly during data collection was critical for identifying the best protocol for use in the second phase of the 2015 RECS.

As evident in Figure 3, all protocols performed similarly in terms of submission rate prior to the nonresponse follow-up period, with the possible exception of Web Only. While this version of the daily monitoring chart does not, reflect the statistical uncertainty (e.g., standard errors) around estimates, we recognize the importance of including this information when comparing protocols. We designed the experiment such that both groups had a robust and equal sample size (2,412 in each protocol). This means that a practical difference of a few percentage points would be statistically significantly different as well. We calculated and check for statistically significant differences at certain “check points” during data collection. In Figure 4, we present a version of the chart monitoring Choice+

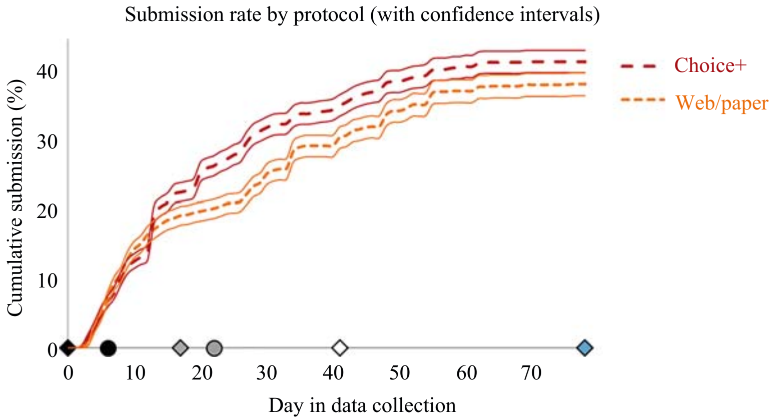


Fig. 4. National pilot choice+ and web/paper submission rates with 90 percent confidence intervals.

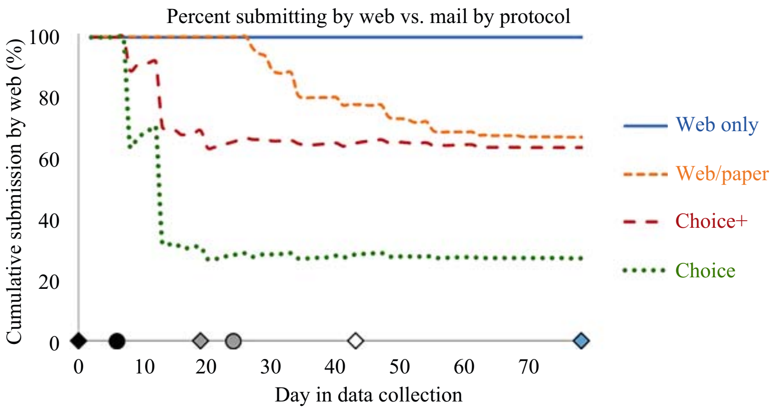


Fig. 5. RECS National pilot cumulative proportion of interviews completed via web.

and Web/Paper submission rate with the inclusion of the 90 percent confidence intervals for each (to test for whether the Choice+ submission rate was significantly higher than Web/Paper). To render these lines, we computed for each protocol and day in the data collection period using the formula for confidence intervals for a one sample dichotomous outcome.

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{1}$$

The upper and lower confidence interval values for each protocol were added to the chart. This made it possible to see where those intervals overlapped for the two protocols. Figure 4 shows the difference between the protocols' submission rates becoming significant around day 20 in the data collection period. They remained significantly different through the end of the main data collection phase, though the rates were only different by a few percentage points.

Submission rates represent one dimension of quality, but given that a key concept of ART design is monitoring multiple indicators of quality and cost, we cannot rely on overall submission rates alone. For instance, it was noted that a significant cost and quality driver was the proportion of interviews that would be completed via web surveys as opposed to paper. As shown in our next chart example (Figure 5), both Web/Paper and Choice+ had a majority of interviews completed via web during Phase 1 of the National Pilot. The Choice group resulted in only about a third of cases being completed by web. The Web Only group, by definition, had 100 percent of cases completed by web during Phase 1. Taken together with submission rate, these two metrics begin to paint a more complete picture of the quality and cost tradeoffs of the different treatments. It was only because we were tracking these trends closely that we could make the rapid decision to implement a self-administered protocol (Choice+) for the 2015 RECS.

Another CTQ metric that was closely monitored reflected the ability of each protocol to elicit responses from key respondent domains. One such characteristic of interest was the age of the householder – a characteristic we expected to have some correlation with mode preference as web access and use tends to be higher among younger individuals. As a benchmark, we used estimates from the 2014 ACS for householder age in our sampled areas. In Figure 6 we present our chart for monitoring the distribution of respondent

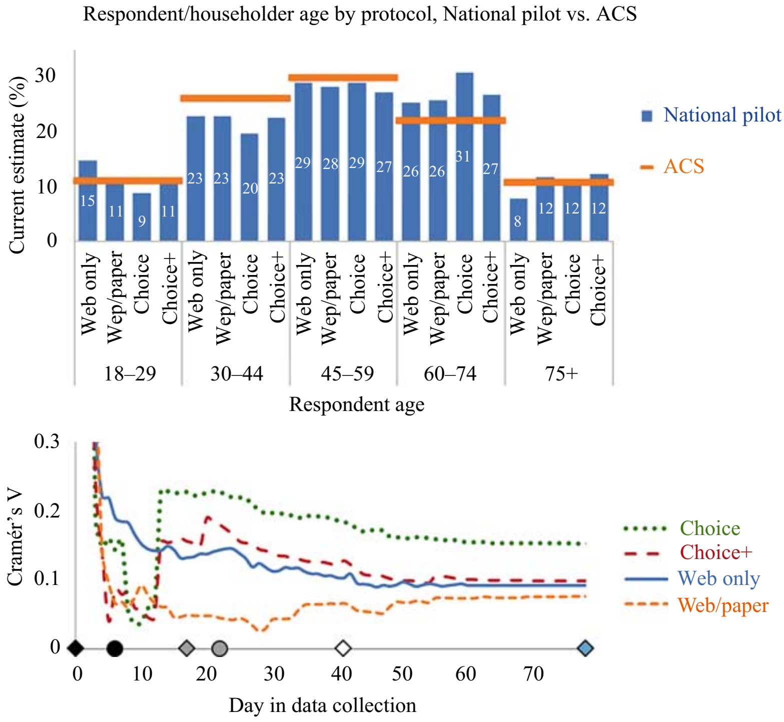


Fig. 6. RECS National pilot age of respondent vs. ACS Age of householder.

(assumed to be the householder) age during data collection. The bar chart at the top shows that no protocol resulted in a distribution matching the ACS exactly, but the differences were not extreme.

To monitor householder age and representativity trends over the course of data collection, we opted to compute a single index to communicate the ability of the protocol to realize a sample matching the ACS. We used the Cramér's V measure, which can be used to determine the degree of association between nominal variables (Cramér 1946). The value ranges from 0 (no relationship) to 1 (perfect relationship) with values under 0.2 indicating a very weak relationship. In this case, the relationship is between the age distribution and the surveys (National Pilot Test and ACS) so a weak relationship suggests little difference between the surveys (i.e., lower is better).

Several other measures of representativeness were considered including the dissimilarity index (see, for example, https://en.wikipedia.org/wiki/Index_of_dissimilarity) and various sample balance indicators such as the R-indicator (Schouten et al. 2009). The former provides essentially the same information as Cramér's V and its advantages over V is only a matter of personal preference and a similar discussion of the post-hoc use of the dissimiliary index can be found in Biemer et al. (2017b). However, the latter measures are inappropriate for comparing a sample variable distribution to an external benchmark distribution which was the objective of our representativeness criterion.

It should be noted that high values of V do not necessarily mean the protocol will result in biased estimates. The respondent sample was ultimately adjusted for nonresponse and coverage error, correcting some of the non-representativity. In addition, perfectly representative samples do not always generate V = 0. The 2014 ACS was conducted nearly two years before the National Pilot and any changes in the target population during that time could cause an increase in V. Also, both the National Pilot and ACS are subject to sampling error and this has not been taken into account in this analysis. Finally, minor differences in the wording of questions, eligibility criteria for housing units, and target populations for the two surveys could impact V. Regardless, we found Cramér's V useful for highlighting practical differences in representativity among the protocols.

The bottom half of Figure 6 shows the cumulative Cramér's V values for age distribution by protocol. Once the survey was several days into data collection, there was little difference between the National Pilot Test and ACS age distributions and the difference became smaller over time as more interviews were completed. The largest difference between the National Pilot Test and ACS by protocol was seen in the Choice group. A review of the top half of Figure 6 suggests that National Pilot Test Choice respondents skewed towards the older age groups, suggesting that the Choice protocol, relative to the other protocols, was on the whole a more attractive option for older respondents. A review of mode choice suggests that older respondents were much more likely than younger respondents to select the paper mode and the Choice protocol did little to dissuade respondents from choosing paper over web. As shown earlier in Figure 5, however, Choice+ offered an additional incentive for web response and was much more effective at attaining a high proportion of completed via web compared to paper during the main data collection phase.

Regarding costs, we compared the average cost per case across the data collection period by protocol. We began with the understanding that any self-administered protocol would be several times less expensive than CAPI, but were interested to compare costs between self-administered protocols to inform future designs. While cost was not the primary concern in comparing self-administered protocols, it was an important dimension. In Figure 7, we present the cost per case of each protocol, relative to Web Only (with a value of 1 at the end of the data collection period). For each protocol, there was a large increase in costs with each subsequent mailout (sent only to nonrespondents). Between mailouts, costs

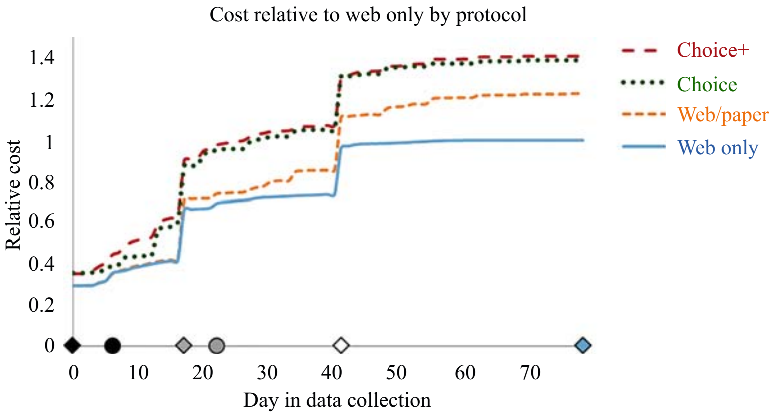


Fig. 7. RECS National pilot relative cost per case by protocol.

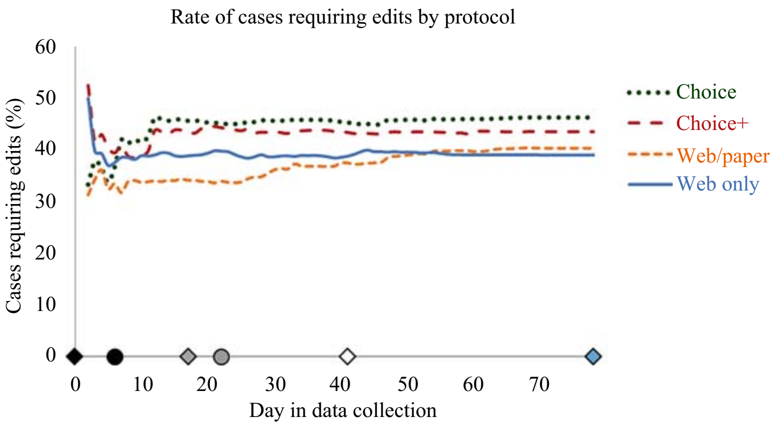


Fig. 8. RECS National pilot rate of cases requiring edits by protocol.

were incurred with the receipt and entry of paper questionnaires and payment of incentives. In the end, we found Web/Paper to be about 20 percent more expensive than Web Only and Choice and Choice+ to be about 40 percent more expensive.

Much of the content of the RECS questionnaire is technical, therefore considerable staff time can be devoted to reviewing inconsistent or improbable responses, such as lack of heating equipment in cold climates or extremely large housing unit measurements. Editing, therefore, is considered as much of a cost metric as it is a quality one. Figure 8 presents the rate of completed interviews requiring data edits by protocol. All responses, regardless of mode, were subjected to the same edit specifications. However, the rate for Web Only was lowest since the web allows for greater restriction of response options and programmatic skips through the questionnaire. The rate of data edits required for the other protocols, which included paper responses, were higher. Web/Paper initially had the lowest edit rate, but as the later period allowed for paper questionnaires to be submitted, this rate increased.

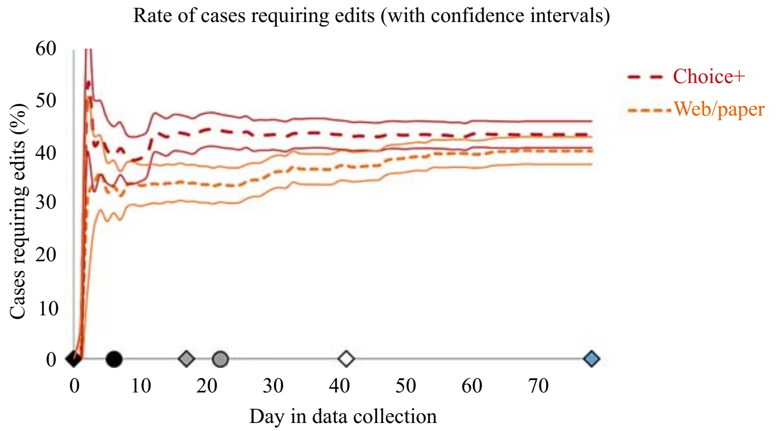


Fig. 9. National pilot choice+ and web/paper editing rates with 90 percent confidence intervals.

In Figure 9, we add one-sided 90 percent confidence intervals to compare the editing rate for cases in the Choice+ and Web/Paper protocols, testing to determine whether the Web/Paper editing effort was statistically significantly less expensive than Choice+ over the course of data collection. The figure suggests that very early in the data collection period, the difference in rates was not statistically significant, but that after the second week, enough cases had been completed to determine a difference. However, later in the data collection period, after the Web/Paper protocol introduced the paper option, the rate of cases requiring editing rose for that protocol so that the rate was no longer statistically significantly lower than that for Choice+.

Taken together these metrics illustrated in Figures 3 through 9 begin to paint a more complete picture of the quality and cost tradeoffs of the different treatments. Reviewing these charts regularly during data collection was an important step in identifying the best protocol for use in the 2015 RECS CAPI remediation. It was only because we were tracking these trends closely that we could make the rapid decision to implement a self-administered protocol (Choice+) for the 2015 RECS CAPI. Choice+ demonstrated the ability to achieve the highest level of response, a majority of cases responding by web vs. paper, good comparability with external benchmarks, and costs that were reasonable for the needs of RECS.

As a final example, we include in Figure 10 a monitoring chart helpful for decision making during the RECS 2015 CAPI Remediation Phase. By monitoring data collection progress regularly, we could follow the submission rate trend in all phases of the Pilots to determine that the self-administered protocol was achieving a similar or higher rate in a shorter amount of time. This chart helped identify and communicate the impetus for switching to web/paper for the remediation.

Once the charts were produced, it was important to get them in the users' hands to facilitate discussion and planning. Our nightly process published the charts in a single file on the secure project web site where all users could access and download the file. We referred to the charts in day-to-day planning and included a copy with the materials for each of our weekly planning meetings. We found this approach minimized the burden on individual users while maximizing the reference to and use of the charts for decision making.

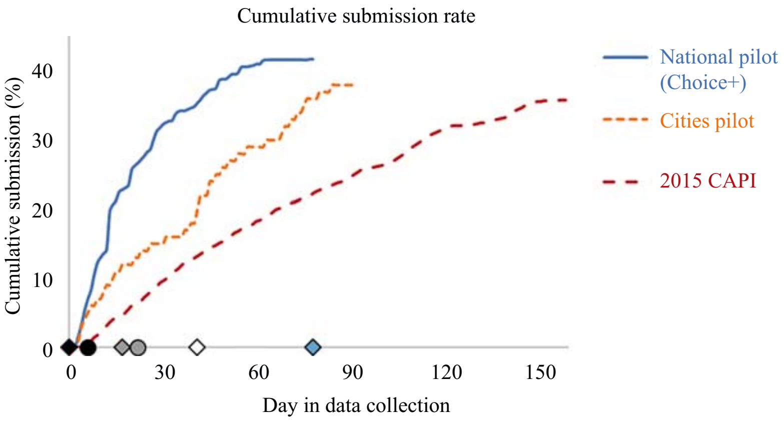


Fig. 10. RECS CAPI Remediation submission rates by phase.

5. Discussion

The ART approaches described in this paper proved to be quite powerful for the RECS and were essential for guiding the experimentation and field work during the piloting of self-administered modes, then the transformation from face-to-face to web/paper administration. There are several key lessons learned, however, and it may take multiple survey cycles to develop the right mix of CTQ metrics for RECS. With the RECS Pilot, we focused primarily on data quality because the approximate cost savings of moving from CAPI to web/paper were easy to predict. In future rounds, cost will be a much more critical metric to track, as EIA continues to explore the optimal mix of web/paper or web/paper/CAPI modes. In particular, project staff might trade lower costs for greater data quality in the design of experiments to determine the proper mix and sequencing of modes, whether to use non-English survey instruments, when to implement stopping rules, and the scope of nonresponse follow-up. In addition, suggestions for CTQs should be solicited from staff involved in downstream processes such as data editing, weighting, imputation, and energy modeling Finally, as suggested in the previous section, the scope of the CTQ metrics should include more energy-specific comparisons using the prior RECS estimates as benchmarks. These comparisons would balance the metrics tracking demographic representativity with the energy characteristics representativity of RECS respondents.

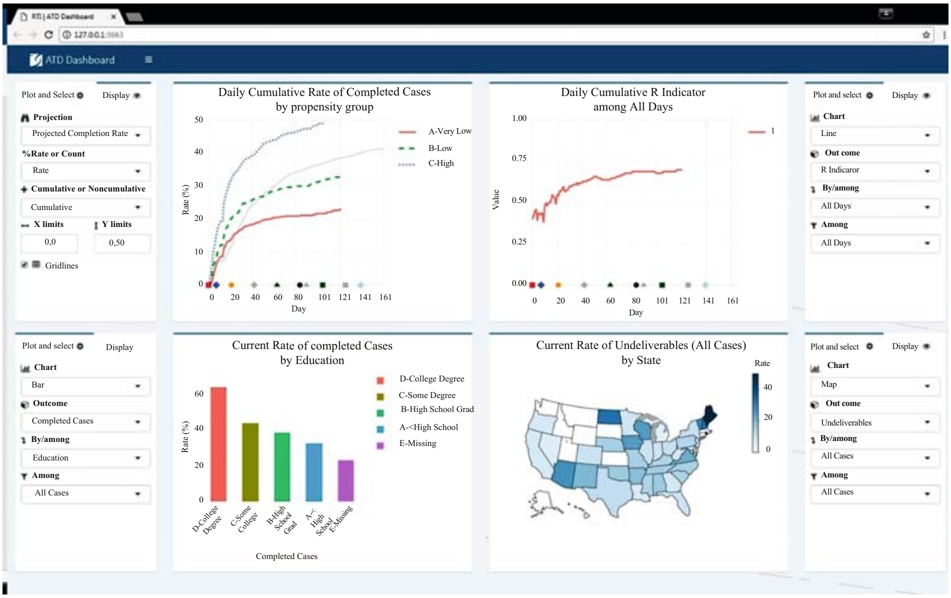
As previously noted, the successful implementation of ART designs requires monitoring critical metrics in real-time and extrapolating current trends to accurately predict future outcomes. These predictions become the basis for designing effective and timely interventions that minimize survey costs, mitigate the highest error risks and avoid major schedule delays. Visual displays of trends in the performance data supplemented by statistical tests of significance allows survey managers to detect the indications of anomalies that require action early on and in real-time when such actions are the most effective.

Our basic approach is generalizable to virtually any survey facing similar transformative decisions based upon a sequence of experiments that must be conducted in rapid succession with little or no time for analytic pauses between data collection phases. Notwithstanding the success of our current approach, there are several important ways visual ATD can be improved by the addition of features, options and tools that would enhance its utility while improving its functionality.

1. **Interactivity.** We are currently embedding interactive functionality in the ATD system (Murphy et al. 2017; Duprey et al. 2017). In particular, interactive visualizations are very useful to detect data anomalies and/or interactions among error sources and to search for their probable causes. Users are presented with an array of display options and mechanisms for categorizing, subsetting, and aggregating data, as well as overlaying projections, survey outcomes from prior rounds, or model-derived predictions. Given that data inputs may be derived from disparate systems and may exist at multiple units of analysis (e.g., sample-member level, interviewer-level, day level), a data taxonomy embedded in the display and selection menus that restrict combinatorial structures to only logical instantiations is also being implemented. Thus, CTQ indicators can be prominently displayed while extraneous information is minimized, using best practices of visual design (see, for

example, Cleveland 1993). Figure 11 provides a snapshot of the interactive system under development.

2. **More Predictive Metrics.** Ideally, a metric for a CTQ is one that can accurately indicate when the CTQ falls below a quality level where some remedial intervention is required to achieve a desired or specified output quality level. Although good metrics exist for some CTQs such as response rates, standard errors and sample balance, this is not true for other sources such as mode effects and other measurement errors data validity/reliability and questionnaire design flaws. For field studies, we are adding visualization metrics based upon computer assisted recorded interviewing (CARI) to detect interviewer errors due to poor interviewing performance, fabrication, violations of protocols and the like. Similarly, CARI metrics can be devised to detect respondent comprehension issues or questionnaire flaws that cause confusion during the interview. Going beyond CARI, it may be possible to embed a limited number of replicate measurements in the instrument to detect response reliability and validity issues. Consistency checks can also be used to detect some types of measurement errors. For example, a model derived estimate of square footage based upon number of rooms, floors, inclusion of attics, basements, etc. could be used to identify gross errors in the estimation of housing unit square footage. These metrics would supplement and enhance the CARI metrics and other traditional metrics based upon response patterns (such as straight-lining) and response latency.
3. **Interpreting Variation**. An important issue in the interpretation of visual information is separating variation that is inherent in the data collection process (referred to as "common cause") from variation that is due to anomalous stimuli (referred to as "special cause"). It is important to distinguish between common and



Plot and Select
Display
Projection: Projected Completion Rate
%Rate or Count: Rate
Cumulative or Noncumulative: Cumulative
X limits: 0,0
Y limits: 0,50
Gridlines

Plot and select
Display
Chart: Line
Out come: R Indicaror
By/among: All Days
Among: All Days

Plot and select
Display
Chart: Bar
Outcome: Completed Cases
By/among: Education
Among: All Cases

Plot and select
Display
Chart: Map
Out come: Undeliverables
By/among: All Cases
Among: All Cases

Fig. 11. *Interactive monitoring dashboard example.*

special cause variation because their mitigation strategies are distinctly different. Special cause variation can be addressed by targeted interventions while common cause variation is mitigated by redesigning the process. Methods for interpreting variation are well-known in the quality control literature (see, for example, Breyfogle 2003). Morganstein and Marker (1997) and Biemer (2010) describe how these methods can be applied to survey processes. Adding these features to the ATD system is a priority because of the risks to survey costs and data quality of misinterpreting and inappropriately mitigating temporal and spatial variation.

4. **Automatic Detection of Anomalies**. The age of "big data" has brought about an explosion in the volume, velocity and variety of data available for anomaly detection. We have already seen an explosion of paradata and their associated metrics for detecting a wide variety of cost, quality and data timeliness anomalies. These will increase exponentially as the search for anomalies extends to interviewers and respondents at varying levels of geography, for a variety of questionnaire items, cross-classified by interviewer, respondent and geographic characteristics. The search for anomalies in the data is made even more complex by the need to identify special versus common cause variation. Fortunately, artificial intelligence provides a solution for competently managing these data at lightning speeds to detect data problem early in real time. We believe the automatic detection of data anomalies is a high priority because managing these data complexities, detecting actionable patterns in the data and prioritizing apparent anomalies according to their error costs and error risks all in real-time and with high accuracy will not be possible without it.
5. **Usability Research**. We have observed that the visual ATD system worked well for the goals of the RECS project. However, we have yet to carefully evaluate the process by which users interpret the charts and whether those interpretations are accurate. It is important to avoid the situation where users rely on fast, instinctive and emotional thinking to draw conclusions from the graphics (Kahneman's (2011) "System 1") and support the slower, more deliberative, and more logical thought process of users ("System 2"). By evaluating users' interactions with the visualizations and assessing their usability relative to alternative visualizations (Hornbaek and Frokjaer 2003), we can improve the design, resulting in even more effective interpretation and decision making.

6. References

AAPOR (The American Association for Public Opinion Research). 2015. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 8th edition. Oakbrook Terrace, IL: AAPOR.

Amaya, A., P. Biemer, and D. Kinyon. 2017. "Total Error in a Big Data World with Applications to the Residential Energy Consumption Survey." Presented at the American Association for Public Opinion Research Annual Conference, New Orleans, LA.

Biemer, P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–848. Doi: https://doi.org/10.1093/poq/nfq058.

Biemer, P., K.H. Harris, B. Burke, K. Considine, C. Halpern, and C. Suchindran. 2017a. "Transitioning an In-Person Longitudinal Survey to a Mixed-Mode, Two-Phase Survey Design: Preliminary Results." Presented at the Annual Conference of the American Association for Public Opinion Research. New Orleans, LA.

Biemer, P., J. Murphy, S. Zimmer, C. Berry, G. Deng, and K. Lewis. 2017b. "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys." *Journal of Survey Statistics and Methodology*. Doi: https://doi.org/10.1093/jssam/smx015.

Breyfogle, F. 2003. *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. Hoboken, NJ: John Wiley & Sons.

Camoes, J. 2008. *How to Create a Thematic Map in Excel*. Available at: http://www.excelcharts.com/blog/how-to-create-thematic-map-excel/ (accessed November 26, 2017).

Chun, A.Y., B. Schouten, and J. Wagner. 2017. "JOS Special Issue on Responsive and Adaptive Survey Design: Looking Back to See Forward – Editorial." *Journal of Official Statistics* 33(3): 571–577. Doi: http://dx.doi.org/10.1515/JOS-2017-0027.

Cleveland, W. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

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

Dillman, D.A. and M.L. Edwards. 2016. "Designing a Mixed-Mode Survey." In Wolfe, Christof, Joye, Dominique, Smith, Tom W. and Fu, Yang-chih, Sage Handbook of Survey Methodology. Sage Publications Wolf, Joye, Smith and Fu. Thousand Oaks. CA, 255–268.

Duprey, M., J. Murphy, P. Biemer, and R. Chew. 2017. "Veni, Vidi, Vici: Interactive Data Visualizations for Adaptive Total Design." Presented at the 5th Workshop on Adaptive and Responsive Survey Design. Ann Arbor, MI.

Eddy, W.F. and Marton, K., Editors. 2012. *Effective Tracking of Building Energy Use: Improving the Commercial Buildings and Residential Energy Consumption Surveys*. Washington D.C.: The National Academies Press.

Edgar, J., J. Murphy, and M. Keating. 2016. "Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions." *SAGE Open* 6(4): 1–14. Doi: https://doi.org/10.1177/2158244016671770.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. 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* 169(3): 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Hardin, M., D. Horn, R. Perez, and L. Williams. 2012. *"Which Chart or Graph is Right for You? Telling Impactful Stories with Data."* Tableau Software. Available at: http://theathenaforum.org/sites/default/files/WHich%20chart%20is%20right%20for%20you.pdf (accessed November 26, 2017).

Hornbaek, K. and E. Frokjaer. 2003. "Reading Patterns and Usability in Visualizations of Electronic Documents." *ACM Transactions on Computer-Human Interaction* 10(2): 119–149. Doi: https://doi.org/10.1145/772047.772050.

Howden, L., S. Joestl, and R. Cohen. 2015. Improving Response Rates using a Mixed-Mode Approach: Results from the National Health Care Interview Survey. Presented at the 2015 FedCASIC Conference. Available at: https://www.census.gov/fedcasic/fc2015/ppt/27_howden.pdf (accessed November 21, 2017).

Kahneman, D. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

Laflamme, F. and J. Wagner. 2016. "Responsive and Adaptive Designs." In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T. Smith, and Y. Fu. Los Angeles: Sage.

Link, M. and A. Mokdad. 2005. "Alternative Modes for Health Surveillance Surveys: an Experiment with Web, Mail, and Telephone." *Epidemiology* 16: 701–704. Doi: 10.1097/01.ede.0000172138.67080.7f.

Luiten, A. and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: an Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society A* 176: 169–189. Doi: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Morganstein, D.R. and D.A. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality*, edited by L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (pp. 475–500). New York: John Wiley & Sons.

Murphy, J., D. Mayclin, A. Richards, and D. Roe. 2016. "A Multi-method Approach to Survey Pretesting." In *2015 FCSM Research Conference Proceedings*. Available at: https://fcsm.sites.usa.gov/files/2016/03/D3_Murphy_2015FCSM.pdf. (accessed November 26, 2017).

Murphy, J., P. Biemer, M. Duprey, and R. Chew. 2017. "Interactive Adaptive Total Design Reports for Near Real-Time Survey Monitoring." Presented at the 2017 Conference of the European Survey Research Association. Lisbon, Portugal.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indictators of Representativeness of Survey Nonresponse." *Survey Methodology* 35: 101–113.

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. Boca Raton, FL: Chapman and Hall/CRC.

Tufte, E. 2001. *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press. ISBN 0-9613921-4-2.

U.S. Census Bureau. 2015. American Community Survey (ACS) 2014 Data Release New and Noteable. Available at: https://www.census.gov/programs-surveys/acs/news/data-releases/2014/release.html#par_textimage_12. (accessed November 21, 2017).

Zimmer, S., P. Biemer, P. Kott, and C. Berry. 2016. "Testing a Model-Directed, Mixed Mode Protocol in the RECS Pilot Study." In *2015 FCSM Research Conference Proceedings*. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2016/03/G2_Zimmer_2015FCSM.pdf. (accessed November 26, 2017).

Received March 2016
Revised March 2018
Accepted June 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 649–670, http://dx.doi.org/10.2478/JOS-2018-0031

A Study of Interviewer Compliance in 2013 and 2014 Census Test Adaptive Designs

Gina Walejko[1] *and James Wagner*[2]

Researchers are interested in the effectiveness of adaptive and responsive survey designs that monitor and respond to data using tailored or targeted interventions. These designs often require adherence to protocols, which can be difficult when surveys allow in-person interviewers flexibility in managing cases. This article describes examples of interviewer noncompliance and compliance in adaptive design experiments that occurred in two United States decennial census tests. The two studies tested adaptive procedures including having interviewers work prioritized cases and substitute face-to-face attempts with telephone calls. When to perform such procedures was communicated to interviewers via case management systems that necessitated twice-daily transmissions of data. We discuss reasons when noncompliance may occur and ways to improve compliance.

Key words: Computer-assisted personal interviewing; decennial census.

1. Introduction

Researchers are interested in measuring the effectiveness of adaptive and responsive survey designs that monitor frame data, paradata, and survey response data and react to this information using tailored or targeted interventions (Groves and Heeringa 2006; Kirgis and Lepkowski 2013). While several studies have successfully evaluated adaptive design experiments that call cases at specific times or stop effort on unproductive cases in computer-assisted telephone interviewing (CATI) systems (e.g., Coffey 2013; Luiten and Schouten 2013; Wagner 2013a), those that measure the effectiveness of adaptive designs in computer-assisted personal interviewing (CAPI) environments are scarce.

This article suggests that few in-person adaptive design studies have been executed and reported because interviewer noncompliance can limit the effectiveness of these interventions, making them difficult or impossible to evaluate. In contrast, when interviewers follow intervention protocols, researchers can evaluate their effectiveness. In-person adaptive design experiments often rely on computerized case management systems that allow interviewers much flexibility in managing their workload, including the number of calls made to each case and the timing of those calls (Morton-Williams 1993). Overhauling these case management systems completely to test an adaptive design

[1] U.S. Census Bureau, 4600 Silver Hill Road, Suitland, Maryland 20746, U.S.A. Email: gina.k.walejko@census.gov
[2] University of Michigan, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48105, U.S.A. Email: jameswag@isr.umich.edu

Acknowledgments: The authors would like to thank Susanne Johnson with her help in programming preliminary data analysis.

experiment may be unfeasible. Thus – in CAPI adaptive design experiments – randomization often relies on in-person interviewers working cases precisely as communicated by their case management systems, which interviewers may choose not to follow. This decision may or may not be for legitimate reasons, such as a change in their schedule making it inefficient to drive to an address across town or having personal knowledge about a housing unit that increases its likelihood to be interviewed. Whether for valid reasons or not, noncompliance limits the ability of survey organizations to implement centralized decision rules typical of adaptive designs.

This article describes examples of interviewer noncompliance and compliance in adaptive design experiments that occurred in two decennial census tests – tests that required in-person interviewers to follow field procedures specified in their case management systems to implement adaptive designs. Conducted by the United States Census Bureau, the two tests occurred at different sites and used different interviewers, supervisors, trainings, monitoring infrastructure, and case management systems. We briefly present the test results from the first experiment and then describe how interviewers did not implement the treatment. We next discuss results from the second experiment and then show that interviewers largely were compliant and applied the treatment. Where possible in both experiments, we attempt to explain why interviewers did or did not comply. Our analysis illustrates the challenges associated with controlling field procedures when implementing adaptive designs in CAPI environments and may be of value to survey staff interested in experimenting with or implementing field procedures that rely on interviewers following instructions.

2. Background

Responsive or adaptive designs attempt to alter survey protocols either by targeting particular cases or subgroups to receive differential treatment or by changing protocols over time. Often the goal of these interventions is to optimize the allocation of resources such that total survey error is minimized for a fixed cost.

Several web and telephone surveys have used computerized case management systems to implement responsive or adaptive interventions that achieve improvement in cost or data quality by prioritizing cases, calling cases at specific times, or assigning groups of cases to specific interviewers. Statistics Canada prioritized cases in several CATI surveys – including the Households and the Environment Survey and the Survey of Labour and Income Dynamics – and found this adaptive approach led to lower total system time (i.e., cost savings) and similar response rates in both surveys when compared to control methods (Laflamme and St-Jean 2011). The Survey of Consumer Attitudes (SCA) experimentally altered CATI calling algorithms to call during time windows when cases had the highest estimated probability of contact and found calls made during such time windows had a higher contact rate in the experimental group (Wagner 2013a). Statistics Netherlands assessed a pilot test designed to increase representativeness and reduce cost in the Survey of Consumer Sentiment. During the first survey wave, they grouped cases into high, low, or medium cooperation based on predicted scores. Depending on group, cases were sent one of three invitations to participate: web, mail, or choice (i.e., web or mail). In CATI follow-up to nonrespondents, the same survey assigned different call schedules to groups with high, medium, or low contact propensities, and the highest-performing CATI interviewers were assigned to the lowest cooperation group and vice versa. They found tailored strategies increased representativeness at comparable – although slightly higher – cost (2.6 percent) and obtained similar response rates (Luiten and Schouten 2013).

CAPI surveys attempting adaptive survey interventions have found mixed results due, in part, to a lack of compliance with requests from central office staff. For example, the National Survey of Family Growth (NSFG) experimented with interviewers working prioritized cases. In only two of 16 separate experimental interventions were response rates significantly higher in the experimental group, which may have been due to lack of compliance. Interviewers made more calls on prioritized cases in all 16 experiments but in only seven were call attempts significantly higher (Wagner et al. 2012).

Other adaptive survey designs that relied on in-person interviewers to implement experimental manipulations could not be evaluated since control and treatment interviewers behaved the same. Similar to the SCA, each day the NSFG estimated time windows during which cases had the highest probability of contact. CAPI case management systems stored and showed the recommended call time to in-person interviewers in the treatment condition. Interviewers in the control who were not shown such call times happened to coincide attempts with recommended windows 23.0 percent of calls while treatment interviewers who were shown recommended call times made attempts during suggested windows only 23.6 percent of the time. In debriefings, interviewers said they did not follow recommendations because geographically clustered cases did not always have the same suggested time windows. The authors note that – rather than attempt to balance the efficiencies of clustered cases and predicted time windows – interviewers stuck with typical behavior, calling cases at time windows of their convenience (Wagner 2013a).

No evidence suggests that interviewer compliance is worse in adaptive survey designs than other types of field surveys. In fact, other survey experiments have experienced issues analyzing results because interviewers did not follow procedures. NSFG interviewers were asked to leave a "Sorry I Missed You" card at households where such a notice was estimated to increase the probability of contact. Interviewers ignored these instructions (Wagner 2013b) leaving researchers unable to evaluate the effects of the card. Biemer et al. (2013) report interviewers admitted they did not record every call attempt as required to avoid having cases hit a specified cap on the number of allowable call attempts or because interviewers were unclear about what constituted a call attempt (e.g., a "drive-by" sighting that no one is home).

To advance our understanding of interviewer compliance and its effects on evaluating adaptive interventions, this article examines the results of interviewer behavior associated with the 2013 and 2014 Census Tests. These decennial census field studies tested adaptive procedures including having interviewers (1) work prioritized cases and (2) supplement face-to-face attempts with telephone calls to specified sample units. The first intervention also depended upon interviewer compliance with a requested twice-daily transmission of data made from laptop computers to databases maintained in the central office.

The focus of the article is on interviewer compliance, an important issue for adaptive designs in CAPI settings. The experimental results themselves are less interesting since they are difficult to interpret in the presence of noncompliance, and the methods may not be useful for surveys other than the decennial census. First, we describe a case prioritization intervention. Other surveys have successfully implemented this type of intervention (Wagner et al. 2012; Peytchev et al. 2010). In this article, interviewers did not implement the intervention as designed. Further, it did not increase contact and completion rates. We then compare this experiment to an intervention that – largely – interviewers correctly implemented. This intervention led to a reduction in personal visit attempts per case. Our discussion concludes with reasons for noncompliance and how requests to CAPI interviewers in experiments might be improved.

3. Data and Methods

3.1. 2013 Census Test

The 2013 Census Test piloted subsequent decennial census test procedures between October and December 2013. Census Bureau staff selected 2,077 sample addresses from six block group pairs in Philadelphia. One block group in each matched pair was assigned randomly to "No Priority Condition" interviewers and the other to "Adaptive Condition" interviewers. No Priority interviewers served as the control group for the test.

3.1.1. Interviewers

Eighteen interviewers who had recently finished working on another survey were selected to work on this pilot because supervisors recommended them, and they had better histories of recording contact attempt information in a previous survey. Eight interviewers were assigned randomly to work Adaptive Condition cases and ten to work No Priority Condition cases. Two supervisors from the field office managed each condition separately. More detail on the 2013 Census Test can be found in Walejko et al. (2014).

3.1.2. Intervention Goal

One goal of the 2013 Census Test was to measure the effect of case prioritization on efficiency. Up to seven cases with the highest predicted propensity to respond on the next contact attempt were prioritized on each Adaptive interviewer's case list. (Adaptive interviewers may have received more or fewer than seven "high priority" cases due to reassignments between interviewers, interviewers not transmitting, or other anomalies.) Cases were rescored, and priority cases were updated daily. Geography was not used in creating this prioritization. Prioritized cases could fall anywhere within the six block groups assigned to the Adaptive condition. For this intervention to be implemented, interviewers needed to attempt all seven high priority cases every day they worked. Success metrics for this intervention included higher contact and completion rates on prioritized cases.

3.1.3. Training, Supervising, and Monitoring

Supervisors instructed Adaptive and No Priority interviewers separately over the course of a two-day training. Supervisors instructed all interviewers that the test was about following instructions provided to them through their case management systems. Trainings, training manuals, and job aids highlighted the importance of Adaptive interviewers attempting all priority cases every day they worked. Adaptive interviewers were only to attempt a “regular” case (i.e., not high priority) if it was nearby or had an appointment. No Priority interviewers were instructed to work cases using Census Bureau survey guidelines that allow flexibility in which cases they visited and when they made contact attempts. After monitoring in-person interviewers’ performances and observing poor compliance, supervisors performed a half-day refresher training on the 17th day of data collection for Adaptive interviewers to increase understanding of 2013 Census Test procedures including working prioritized cases.

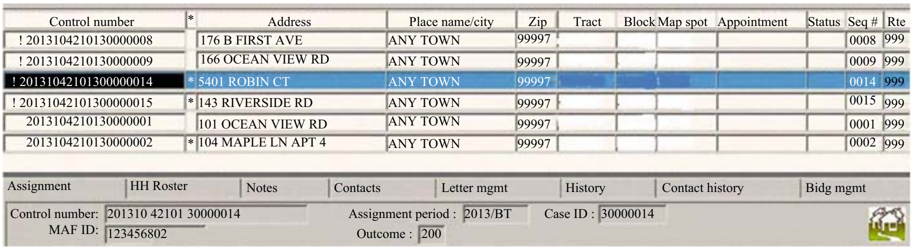
An interviewer performance report monitored all interviewers daily on specific field procedures. The report tracked data transmission compliance as well as daily counts of attempted high priority cases for each of the Adaptive interviewers. Headquarters and field staff conducted a daily meeting during which they discussed this report and other interviewer performance topics. Supervisors were instructed to address noncompliance observed in the report by talking to interviewers.

3.1.4. Case Management System

The 2013 Census Test used many existing Census Bureau information technology resources including a computerized case management system located on interviewer laptops. (See Figure 1 for a screenshot of the 2013 Census Test Adaptive Condition case management system.) Each high priority case, designated with a unique control number, was preceded by an exclamation point, and high priority cases were sorted to the top of the case list. Cases did not have a priority indicator on No Priority interviewers’ case lists.

3.1.5. Data Transmissions

Interviewers in the 2013 Census Test needed to transmit data from their case management systems to the operation control system twice daily, once before they started work and once after they completed work for the day. Before-work transmissions pulled any updated interviewer instructions from the central control system to interviewers’ case lists. Daily instructions updated which cases were prioritized. After-work transmissions pushed contact history information and outcome codes from interviewer laptops to the control system so that instructions for the next day could be calculated by business rules. Transmissions needed to occur after and before set times – not too late at night or early in the morning. Due to the six-time zone span of the U.S. (three in the continental U.S.), a decennial census would need transmissions to occur so work transmitted late at night in the



Control number	*	Address	Place name/city	Zip	Tract	Block	Map spot	Appointment	Status	Seq #	Rte
! 2013104210130000008		176 B FIRST AVE	ANY TOWN	99997						0008	999
! 2013104210130000009		166 OCEAN VIEW RD	ANY TOWN	99997						0009	999
! 20131042101300000014	*	5401 ROBIN CT	ANY TOWN	99997						0014	999
! 20131042101300000015	*	143 RIVERSIDE RD	ANY TOWN	99997						0015	999
2013104210130000001		101 OCEAN VIEW RD	ANY TOWN	99997						0001	999
2013104210130000002	*	104 MAPLE LN APT 4	ANY TOWN	99997						0002	999

Control number: 201310 42101 30000014
MAF ID: 123456802
Assignment period : 2013/BT
Outcome : 200
Case ID : 30000014

Fig. 1. Screenshot of 2013 Census test adaptive condition case management system.

west could be read in by business rules that assigned instructions and would be available early in the morning for interviewers who transmitted in the east.

In the 2013 Census Test, all transmissions required interviewers to log into their laptop case management system, connect to the internet, and click a "transmission" button (i.e., transmissions were manual). Interviewers were instructed to transmit once before working, no earlier than 8:00 a.m., and once in the evening after they were done working, no later than 10:00 p.m.

3.2. 2014 Census Test

In 2014, the Census Bureau carried out a larger field test of adaptive procedures. The CAPI portion of this test included 46,247 sample addresses located in Washington, DC and Montgomery Country, Maryland and ran during August and September 2014. More detail on the 2014 Census Test can be found in Poehler and colleagues (2016). These sample addresses were assigned in geographic clusters to one of three interviewing conditions: a "Control Condition," an "Experimental Contact Strategy Condition," and an "Adaptive Condition." (We do not analyze the Adaptive Condition in this article because it employed CATI interviewers to call sample addresses rather than CAPI interviewers.) Control Condition interviewers had much flexibility as to how they contacted cases, similar to 2010 Census enumerators. They were instructed to perform a personal visit first and then up to two more personal visits and three telephone attempts at their discretion. Experimental Contact Strategy interviewers were instructed to make no more than three total contact attempts (although the case management system allowed more than three) with the first being a personal visit, the next being a telephone contact, and the final a personal visit. Although not an example of "dynamic" adaptive design, the Experimental Contact Strategy can, in our opinion, be considered a "static" adaptive design (see Schouten et al. 2013) because the protocol instructing interviewers to make the second attempt by telephone was applied to only addresses that had been successfully matched to telephone numbers, 81.2 percent of cases.

3.2.1. Interviewers

The Census Bureau recruited interviewers from the area specifically to work on the 2014 Census Test. Interviewers were new hires who were not required to have past interviewing experience, although many had worked on previous decennial censuses. Crew leaders, also new hires, supervised these interviewers, and were, in turn, supervised by higher-level managers. The Control and Experimental Contact Strategy had 304 and 389 employees who recorded making at least one call attempt or transmitting, respectively. (Employees were not spread evenly across conditions because conditions did not have the same number of sample cases. Two additional interviewers transmitted but did not make any contact attempts.)

3.2.2. Intervention Goal

One goal of the 2014 Census Test was to measure the cost effect of substituting costlier in-person visits in place of telephone call attempts made by CAPI interviewers on second contact attempts for addresses with matched telephone numbers. The goal of this approach was to decrease nonresponse follow up costs by reducing the total number of attempts and, specifically, the number of personal visit contact attempts. For this intervention to be implemented correctly, Experimental Contact Strategy interviewers needed to attempt a phone call to all open cases with matched telephone numbers on the second contact attempt. (Using the Census Bureau's Master Address File ID, 81.2 percent of cases were matched to phone numbers available from several commercial data files.) In contrast with the 2013 intervention, this intervention did not require transmission since it could be implemented algorithmically via the case management system (i.e., using a programmed rule such as "if the case has a telephone number and one attempt, the next attempt should be made via the telephone"). Success metrics for this intervention included lower costs, measured by fewer contact attempts per case, fewer personal visits per case, and fewer average attempts – notably personal visit attempts – per completion.

3.2.3. Training, Supervising, and Monitoring

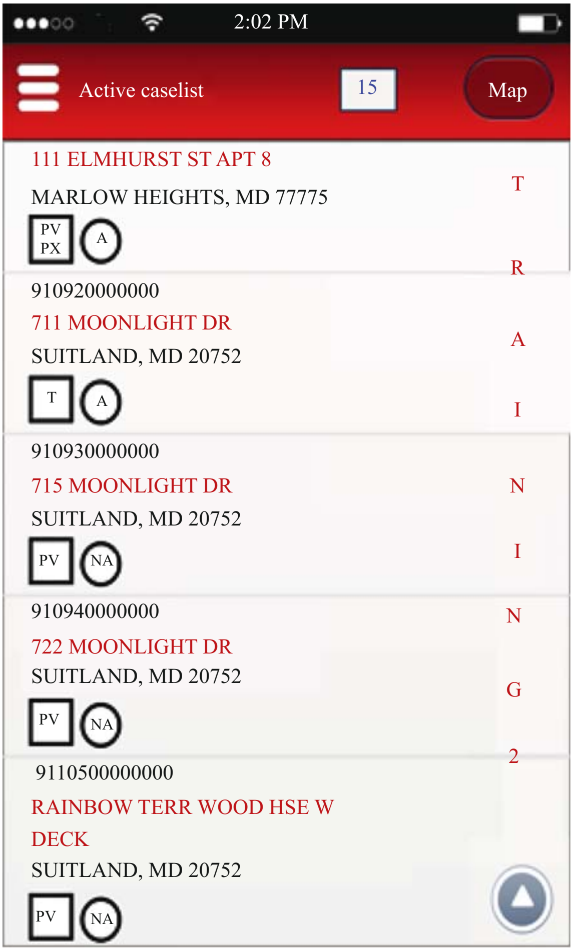
Over the course of three days, crew leaders instructed interviewers how to plan their day, follow field procedures, record contact history information, transmit their data, and perform interviews. Trainings, training manuals, and job aids instructed Experimental Contact Strategy interviewers to call all cases with matched telephone numbers after first attempting a personal visit. One half-day of training was devoted to interviewers performing production interviews and supervisors reviewing this work. Supervisors were instructed to use reports that monitored their interviewers' activities including interviewers' transmissions. (Reports did not monitor whether second contact attempts were done by phone or in-person.)

3.2.4. Case Management System

For the 2014 Census Test, the Census Bureau developed a new computerized case management system available to interviewers as an iPhone application. This system functioned similarly to that used in 2013, providing interviewers with a list of their cases and instructions on how to work each case as well as allowing interviewers to collect interview data and record contact attempt information. The Experimental Contact Strategy interviewer's case management system indicated when to do a telephone attempt and provided these interviewers with the matched telephone numbers. (See Figure 2 for a screenshot of this case management system. The box with a "T" inside it indicates that the interviewer should make a telephone attempt on the indicated case.)

3.2.5. Data Transmissions

In contrast with the technical systems used for the 2013 Census Test, the 2014 Census Test case management system was designed to manage data transmissions automatically, and transmissions were not necessary for interviewers to be displayed the correct mode. The system attempted automatic transmissions when two hours had passed since the last successful transmission, when an interviewer logged into the app or completed a contact attempt, and when an interviewer completed a case or logged out of the app. The case



111 ELMHURST ST APT 8
MARLOW HEIGHTS, MD 77775
PV PX A

910920000000
711 MOONLIGHT DR
SUITLAND, MD 20752
T A

910930000000
715 MOONLIGHT DR
SUITLAND, MD 20752
PV NA

910940000000
722 MOONLIGHT DR
SUITLAND, MD 20752
PV NA

9110500000000
RAINBOW TERR WOOD HSE W
DECK
SUITLAND, MD 20752
PV NA

Fig. 2. Screenshot of 2014 Census test experimental contact strategy condition case management system.

management system itself kept track of whether a case had a matched phone number as well as the number, mode, and outcome of each contact, allowing mode to be displayed correctly without a transmission.

Some protocols not analyzed in this article necessitated twice-daily transmissions, and automated transmissions would not work if interviewers became disconnected from the Census Bureau's network, for example, by driving or walking through an area without cell coverage. For this reason, 2014 Census Test interviewers were instructed to transmit manually twice each day that they worked, once before working no earlier than 7:00 a.m. and again after working but no later than midnight. (After 2013 Census Test results and debriefings uncovered interviewers had difficulty transmitting between 8:00 a.m. and 10:00 p.m., the time period within which interviewers were instructed to transmit on days they worked was expanded to between 7:00 a.m. and midnight for the 2014 Census Test.) Interviewers were able to view when their last successful transmission occurred using the application.

4. 2013 Census Test Results

Table 1 summarizes the design of both 2013 and 2014 Census Tests. Results of the 2013 Census Test did not support the hypothesis that the Adaptive Condition would have higher contact and completion rates than the No Priority Condition in the 2013 Census Test. Instead, contact rates on personal visits were significantly higher in the No Priority Condition than the Adaptive Condition. Furthermore, completion rates on personal visits were the same (Table 2). To help understand these results, we examine interviewer compliance with implementing two necessary actions, transmitting data twice daily and attempting prioritized cases daily. These results showed interviewers did not comply in either transmitting data or working prioritized cases. The fact that the Adaptive Condition interviewers did worse in terms of contact and completion rates seems to indicate that this intervention would not achieve the stated aims. However, given the selective nature with which it is applied, higher or lower rates may have been achieved if the intervention had been applied to the full sample. In any event, poor compliance with the requested actions discussed in the next two sections limits the ability of the central office to implement case prioritization schemes aimed at controlling which cases respond.

4.1. *2013 Test: Interviewers Transmit Data Twice Daily*

In the 2013 Census Test, interviewers transmitted as instructed (i.e., once before working and once after working between 8:00 a.m. and 10 p.m.) over 71 percent

Table 1. 2013 and 2014 Census test designs.

	2013		2014	
	No priority	Adaptive	Control	Experimental contact strategy
Interviewers	10	8	304	389
Location	Philadelphia, PA		Washington, DC and Montgomery county, MD	
Case management system	Modified existing survey system using laptops		New system using cell phone application	
Training	Separate for each condition; 2-day training on procedures; half-day refresher training; training manual, job aid		Separate for each condition; 3-day training on procedures with 1/2 of day for supervisor review of work; training manual, job aid	
Monitoring and supervision	Performance monitoring report; daily meetings; feedback to interviewers via supervisors		Performance monitoring reports; feedback to interviewers via supervisors	

Table 2. 2013 Census test contact and completion rates on personal visits between adaptive and no priority condition interviewers.

Condition	Number	Contact percent	Standard error of percent	p-value	Completion percent	Standard error of percent	p-value
Adaptive	1,283	24.50	3.20	0.03	18.97	2.77	0.86
No Priority	1,354	31.73	2.35		19.69	3.13	

Note: Standard errors and significance take into account clustering by interviewer.
Note: Includes both compliant and non-compliant transmissions.
Note: Excludes personal visit attempts where an appointment was set.

(standard error, 7.6 percent) of days worked (i.e., all days each interviewer worked summed over all interviewers). Figure 3 shows the number of interviewers grouped by five categories of percent compliant daily transmissions. Compliant transmission days ranged between 14 and 100 percent by interviewer, with nine of 18 having over 80 percent compliant transmission days. Five interviewers had 40 percent or fewer compliant transmission days.

Transmission compliance varied across time ranging between 0 and 100 percent over the 2013 Census Test field period. Figure 4 shows the percent of interviewers who worked and transmitted correctly each day, where compliance is measured as transmitting as instructed – once before working no earlier than 8:00 a.m. and once in the evening after they were done working, no later than 10:00 p.m. Small numbers of working interviewers explain peaks in low compliance and high compliance. On December 2, a day with no compliant transmissions, one interviewer worked, and on November 28, only five interviewers worked. On November 24 and December 3, days with 100 percent compliance, fewer than three interviewers worked. On other days with poor compliance, interviewers often transmitted earlier or later than instructed. For example, on November 7, fifteen of sixteen working interviewers transmitted before 8:00 a.m., but only three

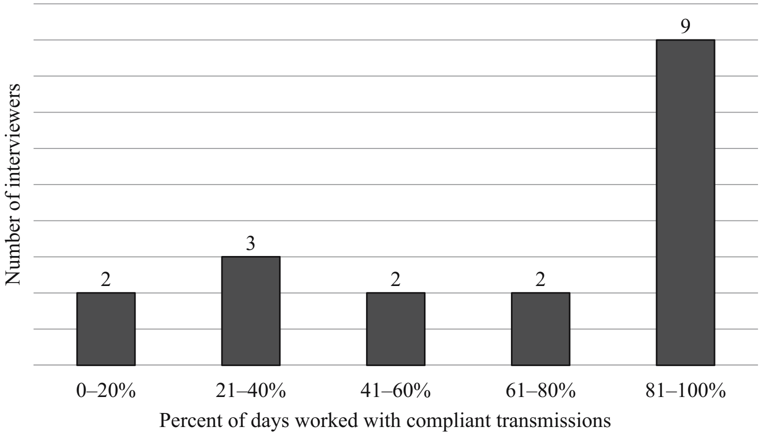


Fig. 3. 2013 Census test compliant daily transmissions.

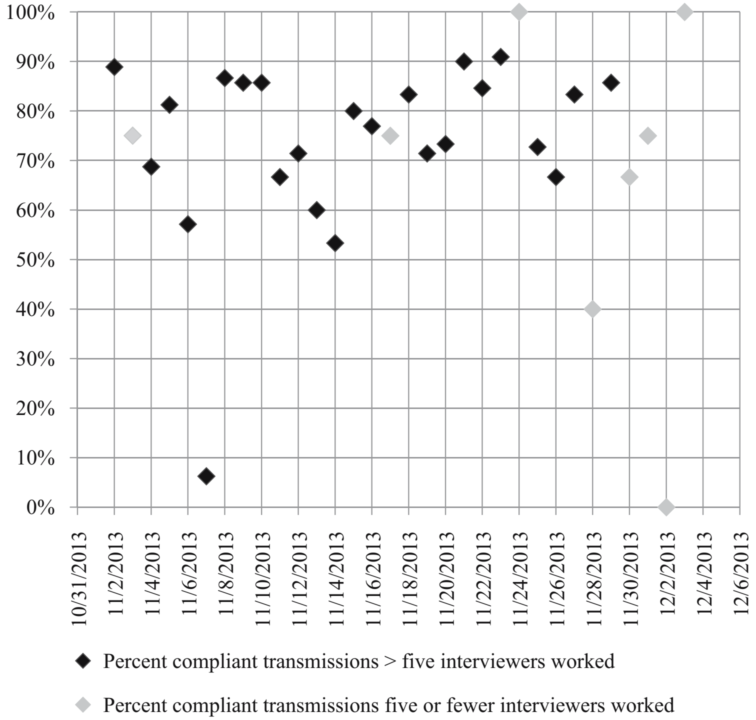


Fig. 4. 2013 Census test percent compliant transmissions by day.

transmitted after 8:00 a.m. as instructed. On the five days with the smallest percent compliant transmissions, 73 percent of interviewers transmitted between 6:00 a.m. and midnight, either earlier or later than instructed (i.e., between 8:00 a.m. and 10:00 p.m.). These results and debriefing led the 2014 Census Test to expand the time period within which interviewers were instructed to transmit on days worked.

4.2. 2013 Test: Interviewers Attempt Prioritized Cases Daily

In order to provide a basis for assessing compliance, 2013 Census Test interviewers needed to receive high priority cases each day they worked. Because 2013 Census Test interviewers did not transmit correctly on 29 percent of days they worked, we broaden the definition of successful transmissions to include those that occurred between 6:00 a.m. and midnight, which includes more days for analysis. To avoid confusion, we will call these "reliable transmissions."

In 2013, the eight adaptive interviewers worked all high priority cases on 45 percent of days with reliable transmissions. These interviewer days are compliant. On seven percent of days with reliable transmissions, interviewers did not attempt all high priority cases but also did not attempt other, "regular" cases, which may indicate they ran out of time before attempting all high priority cases. These interviewer days are potentially compliant. Interviewers did not attempt all high priority cases and worked regular cases on nearly 48 percent of days they transmitted reliably. These days are not compliant.

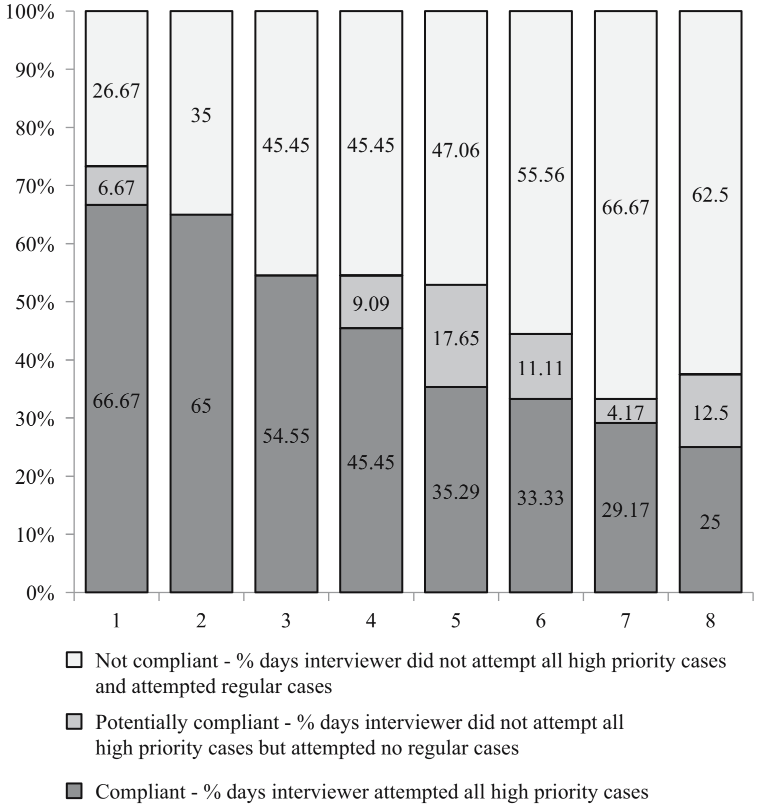


Fig. 5. Percent compliant or potentially compliant days for CAPI interviewers with reliable transmissions in 2013 Census test (Adaptive condition).

Across all eight interviewers, the days they worked all high priority cases ranged between 25 and 67 percent (Figure 5). Two interviewers worked all their high priority cases over 65 percent of the days on which they worked and made reliable transmissions, while two attempted regular cases even though they did not attempt all their high priority cases over 60 percent of the days they worked. This lack of compliance is an interesting result in its own right, as it hampers the ability of data collection operations to implement centrally directed interventions. In this case, it appears that the intervention would not have met its goals, but the lack of compliance is an important finding for other field surveys attempting to prioritize cases.

5. 2014 Census Test Results

In contrast to the 2013 Census Test, results from the 2014 Census Test support the hypothesis that Experimental Contact Strategy interviewers performed actions leading to a reduction in cost; they made fewer average contacts and personal visits per case than Control interviewers. They also had a lower average attempts per complete than Control interviewers and a notably lower average number of in-person attempts per complete than Control interviewers (Table 3). However, the mean number of attempts for the

Table 3. 2014 Census test mean attempts and average attempts per complete between control and experimental contact strategy conditions.

Condition	Number of cases	Mean attempts	Mean in-person attempts	Percent complete	Average attempts per complete	Average in-person attempts per complete
Control	7,394	4.29	3.14	0.62	6.95	5.07
Exp. Contact Strategy	8,873	3.25	2.41	0.57	5.75	4.26

Note: Includes only cases with matched telephone numbers.

Experimental Contact Strategy (3.25) is still greater than 3, indicating some noncompliance. This reduction in effort also appears to have reduced the completion rate relative to the Control Condition (0.57 for the Experimental Contact Strategy Condition and 0.62 for the Control Condition). The Experimental Contact Strategy Condition had about 76 percent of the effort measured as attempts relative to the Control Condition and produced 92 percent of the completion rate relative to the Control Condition. However, the poor compliance observed in the 2013 Census Test led researchers to investigate in more detail the extent to which interviewer compliance may have affected the 2014 data.

5.1. 2014 Test: Interviewers Perform Telephone Calls when Instructed

As shown in Table 4, Experimental Contact Strategy interviewers for the 2014 Census Test followed mode instructions on over 88 percent of contact attempts to cases with matched numbers. (See Table 5 for a summary of compliant procedures.) They performed personal visits as instructed over 99 percent of the time on the first contact attempt. Compliance in attempting contact by telephone on the second attempt was 82 percent, which differs starkly from the control interviewers who performed personal visits 72 percent of the time on the second attempt. In this experiment, case management systems directed interviewers to perform a contact attempt in a particular mode, and it appears this directive changed interviewer behavior when comparing the second contact attempt of the Experimental Contact Strategy and Control interviewers. (The interpretation of results does not change when analysis includes cases without matched numbers.)

5.2. 2014 Test: Interviewers Transmit Data Twice Daily

Unlike the 2013 Census Test, 2014 Experimental Contact Strategy interviewers did not need to transmit twice daily to receive updated mode instructions. However, other protocols not analyzed in this article did rely on data transmissions, so researchers analyzed whether automated transmissions helped interviewers to transmit their data twice daily, once before and once after work.

In the 2014 Census Test, transmissions – either automated or manual – occurred once before working no earlier than 7:00 a.m. and once after working before midnight on 43

Table 4. Number and percent of compliant and non-compliant mode contact attempts by contact attempt number for the 2014 Census test (Experimental contact strategy and control conditions).

	Experimental contact strategy					Control				
	Compliant mode		Non compliant mode		Standard error of percent	Personal visit mode		Telephone mode		Standard error of percent
Contact attempt	*n*	Percent	*n*	Percent		*n*	Percent	*n*	Percent	
1st	8,787	99.03	86	0.97	0.28	7,335	99.20	59	0.80	0.29
2nd	5,722	81.67	1,284	18.33	1.40	4,284	72.15	1,654	27.85	2.23
3rd	4,942	84.94	876	15.06	0.90	3,121	65.73	1,627	34.27	2.15
4th+	6,010	83.83	1,159	16.17	1.05	8,440	61.75	5,228	38.25	1.86
Total	**25,461**	**88.20**	**3,405**	**11.80**	**0.60**	**23,180**	**73.01**	**8,568**	**26.99**	**1.24**

Note: Includes only cases with matched telephone numbers.

Note: The Experimental Contact Strategy called for the first and third attempts to be in-person and the second by telephone. Although interviewers were instructed to perform three attempts total, the case management system allowed for more.

Note: Excludes records of inbound telephone call attempts.

Note: Standard errors take into account clustering by interviewer.

Table 5. 2013 and 2014 Census test procedures.

	2013		2014	
	No priority	Adaptive	Control	Experimental contact strategy
Data transmissions	Manual; 2 times on work days before and after attempts; no earlier than 8:00 a.m.; no later than 10 p.m.		Automatic and manual; 2 times on work days before and after attempts; no earlier than 7:00 a.m.; no later than midnight	
High priority cases	Cases attempted at interviewer discretion; no high priority	7 highest propensity cases prioritized daily; interviewers attempt all daily	*Not analyzed here*	
Telephone calls	*Not analyzed here*	Ring cases with numbers at interviewer discretion		Ring cases with numbers on 2nd attempt
Total contact attempts allowed	3 in-person before proxy attempts; interviewers ring cases with numbers up to 2 times	3 in-person before proxy attempts; CATI before field	6 total before proxy attempts with 3 in-person and 3 phone if number	3 total before proxy with 2 in-person and 1 phone if number

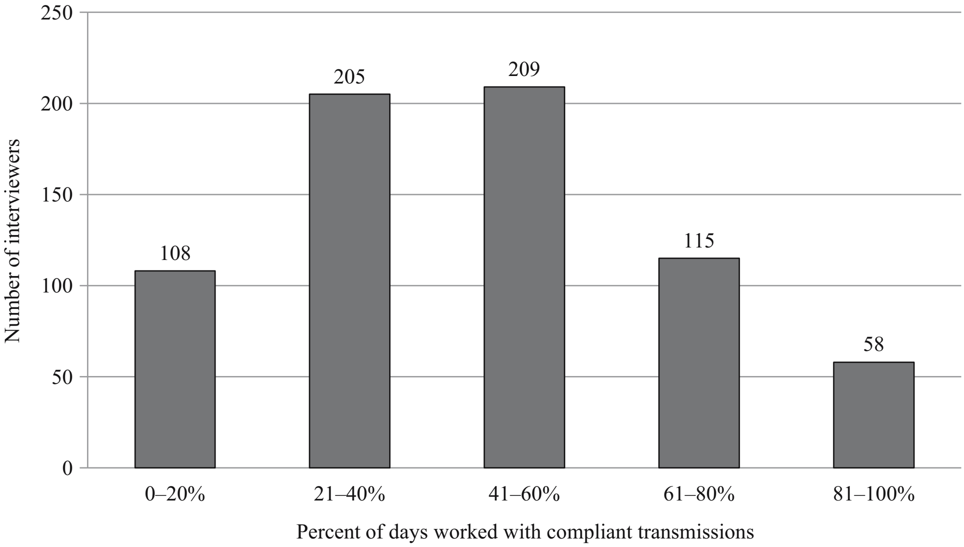


Fig. 6. 2014 Census test compliant daily transmissions.

percent (standard error, 0.9 percent) of days worked (i.e., all days each interviewer worked summed over all interviewers). Compliant transmission days ranged between 0 and 100 percent by interviewer, with 58 of 695 interviewers (8.3 percent) having over 80 percent compliant transmission days and 108 interviewers (15.5 percent) having less than 21 percent compliant transmission days (Figure 6).

6. Discussion

6.1. Limitations

Results should be considered in conjunction with several study limitations. First, the 2013 and 2014 Census Tests recruited interviewers from just two geographic sites. As a result, the tests are not generalizable to the broader United States. Second, the 2013 test contained a sample size of only 18 interviewers. A larger interviewer workforce could have led to different results.

6.2. Summary

Interviewers were somewhat compliant in transmitting data and receiving updates twice daily during the 2013 Census Test. Overall, interviewers transmitted correctly on 71 percent (standard error, 7.6 percent) of days worked. For this test, transmissions were manual. We found lower compliance during the 2014 Census Test where case management systems transmitted automatically, but – as a backup – trainings and training materials instructed interviewers to transmit manually twice daily. Interviewers did not perform this backup transmission, and compliant manual or automated transmissions occurred only 43 percent (standard error, 0.9 percent) of the days interviewers worked.

Lack of compliance and improper functioning of automated transmissions meant instructions (i.e., prioritized cases) were not updated every day interviewers worked during the 2013 Census Test.

In the 2013 Census Test, prioritized cases did not have higher contact and completion rates than nonprioritized cases running counter to our hypothesized result. However, the request to attempt contact on prioritized cases met with low compliance. During the 2013 Census Test, Adaptive interviewers worked all high priority cases on fewer than half (45 percent) of the days they worked, and this percentage is limited to days with reliable transmissions only. While acknowledging 100 percent compliance is unrealistic, noncompliance observed in the 2013 Census Test affected our ability to analyze an adaptive intervention by limiting the number of days we could evaluate interviewers working *all* prioritized cases on their case lists to only a nonrandom 32 percent of days. Regardless of the potential benefit, this noncompliance limits the ability of the central office to intervene by prioritizing cases.

The 2014 Census Test results showed Experimental Contact Strategy interviewers performed actions that led to cost reductions including having lower average attempts per complete than Control interviewers. In contrast with the 2013 Test, we observed quite high compliance with the request that in-person interviewers attempt telephone calls rather than personal visits at certain points in 2014 data collection.

6.3. Reasons for Noncompliance

There are several reasons interviewers may have been noncompliant in transmitting and working prioritized cases. First, as with any kind of job, interviewers may have life circumstances such as a sudden change in their planned schedule due to a sick family member or a safety concern with approaching a sampled housing unit that prevents them from carrying out their assigned tasks. For example, one Adaptive interviewer's high priority case was a house where illegal drug trade occurred, so they did not visit it. Under such circumstances during the 2013 Census Test, it was unrealistic for the interviewer to carry out contact attempts following algorithmic rules.

Second, instructions relayed via the case management system allowed interviewers the flexibility to be noncompliant. Years ago, case management for CAPI interviewers constituted a pen-and-paper system that communicated which addresses to work – usually located nearest to where an interviewer lived – with space to fill out contact information. A historical artifact of paper, most digital CAPI case management systems today supply interviewers with a list of cases to work and leave much to their discretion, including when to work, which cases to attempt, and how frequently to make attempts. The 2013 Census Test interviewers could and did choose to work regular, nonprioritized cases. For example, in debriefings some interviewers did not like having to return to the same block the next day. Although CAPI sample management systems designed to constrain interviewers to follow instructions would be preferred for testing experimental manipulations, reprogramming such systems for a test is cost prohibitive for most organizations that perform CAPI surveys. This finding led to the development of new systems that constrained interviewers to attempt contact on sets of cases selected daily by the central office, as these were the only cases displayed to interviewers.

Third, technical issues could explain at least some of the discussed noncompliance in transmitting data to and from laptop or smartphone case management systems. For example, during the 2013 Census Test, a few interviewers whose data showed continuous transmission issues claimed they were transmitting twice daily as instructed or that they could not transmit as instructed. In debriefings, several interviewers said they needed to transmit more than once before the server would connect to their laptop. Because the 2014 Census Test used smartphones rather than laptops, it is possible that interviewers attempted but were unable to transmit because they lost reception causing them to no longer be connected to the network. During 2014 debriefings, a majority of interviewers and supervisors brought up issues with cellphone reception.

It is also possible that interviewer noncompliance happened when following new procedures competed with other interviewer activities. For example, interviewers in the 2013 Census Test mentioned it was difficult to transmit before 10 p.m. on days when interviewing lasted late into the evening. Supervisors in the 2013 test also reported it was difficult to balance managing interviewer noncompliance with other supervising responsibilities and that a rolled-up report of potential problems to discuss with interviewers could help to alleviate this time pressure. The 2013 and 2014 test interviewers remarked that, when they saw a respondent near their address, they attempted an in-person interview with that respondent, even if their address was not a high priority case or was supposed to be attempted via telephone.

Fifth, the nature of the intervention itself may have led to interviewers to be more accepting of 2014 Census Test procedures than 2013 procedures. In debriefings, all 2014 Census Test Experimental Contact Strategy interviewers reported understanding the test procedures for conducting telephone calls – attempt a contact in the mode that the case management system instructed. However, at least one Adaptive interviewer in the 2013 Census Test admitted it was unclear why case management systems deemed cases as high priority, indicating confusion regarding the nature of the intervention itself.

Finally, it may be that Adaptive interviewers' experience with previous surveys, where they had wider discretion, may have made it more difficult to train them to work under a new centrally directed approach. A common theme in debriefings with experienced interviewers were differences between previous data collection procedures including the 2010 Census. For example, a few Adaptive interviewers did not like planning their route in the morning after an early data transmission provided them new instructions, as they were accustomed to doing it the night before. The interviewers used in the 2013 Census Test were experienced. While we do not have data on 2014 Census Test interviewers' past interviewing experience, being less seasoned and – thus – less inclined to recall past protocols, may have played a role in interviewers following the suggested mode.

Current practice allows interviewers wide latitude for deciding how to conduct their work, but the experimental adaptive design interventions described in this article restrict this range. A tension exists between the centralized, data-driven control of interviewers and decentralized decision making by interviewers who rely upon their expertise and local knowledge. While experienced, expert interviewers with local knowledge may perform at a higher level than if centrally directed, interviewers vary in their ability to plan efficient trips and recruit respondents in practice (Wagner and Olson 2011; O'Muircheartaigh and

Campanelli 1997; Purdon et al. 1999; Pickery and Loosveldt 2002; Durrant and Steele 2009). Further, local interviewers are unable to make decisions that balance response across more cases than their own sample. While centralized, data-driven interventions like those described in this article may go against current interviewer practices, they improve the ability of data collection organizations to control important aspects of the response process, including balanced respondent pools and overall data collection costs. Finding a balance between centralized and decentralized procedures remains a complex function involving the available interviewing staff, the capabilities of the data collection organization, the particulars of the survey design, and the overall goals of the survey. Finding the correct equilibrium may also require considering training, explaining the intervention purpose to interviewers, and other actions considered in the next section.

6.4. Ways to Improve Compliance

Although not possible in the experiments described here, researchers could construct some protocols to constrain CAPI interviewers into compliance. Kreuter and colleagues (2014) found setting prespecified appointments based on the prior wave interview date in the Medical Expenditure Panel Survey Household Component significantly decreased the number of attempts (e.g., phone, in-person, letter) to get an interview. To increase the likelihood that interviewers kept appointments, they mailed sample addresses a postcard with the appointment date and time so "interviewers could not simply ignore the treatment without the risk of upsetting respondents who expected the appointment to be kept" (page 212).

Survey organizations may also develop computerized case management systems to constrain interviewers to follow protocols. For the 2015 and 2016 Census Tests, partially in response to the findings from 2013 and 2014, the Census Bureau redesigned decennial test case management systems to give interviewers cases to be worked on a daily basis. This new system also asked interviewers what their schedule would be and took into account how many and at what times interviewers would be working. Such a design allowed interviewers no flexibility in whom they visited and attempted to constrain when they worked (Blumerman et al. 2015).

It may be that better case management system designs could further improve compliance. The field of decision support systems examines how to construct systems that enable informed decision making, including following requested actions. Much of this work aims at enabling medical professionals to implement treatments following evidence-based best practices. Kawamoto and colleagues (2005) summarize the lessons learned from this literature regarding approaches that ensure compliance.

Improved interviewer training may also increase compliance. Fowler and Mangione (1988) demonstrated that extended interviewer training could improve compliance with standardized interviewing practices. In the realm of nonresponse, Groves and McGonagle (2001) showed that training interviewers with methods for tailoring survey introductions could improve response rates. Indeed, in 2013 Census Test debriefings supervisors recommended self-assessments that would test interviewers' understanding of procedures while allowing supervisors to gauge interviewer knowledge. Interviewers also called for more training with additional role-playing situations.

Other approaches – such as incentives – may improve compliance. Tourangeau and colleagues (2012) found offering incentives to interviewers for every identified eligible person led to higher eligibility rates. However, Rosen et al. (2011) offered incentives to interviewers who completed cases with a low estimated propensity of response. They found incentives did not change interviewer behavior, and low propensity cases in an experimental group did not receive more effort than low propensity cases in a control group. Peytchev and colleagues (2010) offered interviewers incentives for converting cases with low response propensities but found completion rates between low propensity control and treatment cases to be the same, possibly due to high response rates for low propensity cases (i.e., 90.3 percent). Evidence that incentives prompt interviewers to follow field procedures is mixed, and we need more research to determine if and when such approaches can increase interviewer compliance, thereby improving our ability to test adaptive designs in CAPI environments.

Lastly, aligning performance standards with adaptive protocols could increase the likelihood that interviewers follow procedures. In widely cited research on what motivates individuals at their jobs, Hackman and Oldham (1976) argue workers need "knowledge of results" in the form of feedback that clearly aligns with the effectiveness of their job performance. Interviewers, too, may benefit from not only feedback on how well they are doing at following protocols, such as working prioritized cases, but also explanation as to how following such instructions ties to their overall job performance.

7. Conclusion

In sum, we view a lack of in-person interviewer compliance as an obstacle to the implementation of adaptive designs, which hinders our ability to evaluate their successes in CAPI settings. When interviewers do not comply with data transmissions or working all prioritized cases each day, analysis is limited to nonrandom subsets of days or cases. Thus, we cannot say whether an observed difference between treatment and control is due to the adaptive design or the interviewer choosing when to comply – or not comply – with protocols.

Fortunately, lack of CAPI interviewer compliance is a problem with solutions and – as illustrated here – not a barrier to all in-person adaptive designs. Future adaptive design research needs to strengthen both the actions requested from interviewers and the ways in which these requests are delivered. Furthermore, the field would benefit from a study designed specifically to understand reasons for in-person interviewer compliance and noncompliance with a variety of protocols.

Adaptive designs that rely on in-person interviewers to implement protocols must consider the balance between flexibility and prescription. Survey methodologists and systems programmers should deliberately acknowledge when interviewer's local, accumulated knowledge outweighs the prescriptiveness that can be built into case management systems. These decisions likely depend on the survey. For data collections like the decennial census that, in 2010, hired over 500,000 employees across the United States, many of whom had limited interviewing experience, the balance might best tip towards prescription.

8. References

Biemer, P., P. Chen, and K. Wang. 2013. "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society Series A* 176(1): 147–168. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01058.x.

Blumerman, L., E. Moffett, M. Bentley, T. Boone, and M. Chapin. 2015. "2020 Census Operational Plan Overview and Operational Areas." Presentation to the Census Bureau's National Advisory Committee. October 8, 2015.

Campanelli, P., P. Sturgis, and S. Purdon. 1997. "*Can You Hear Me Knocking? An Investigation into the Impact of Interviewers on Survey Response Rates*. London, GB, National Centre for Social Research.

Coffey, S. 2013. "Implementing Adaptive Design for the National Survey of College Graduates." Article presented to FedCASIC, Suitland, Maryland, March 20, 2013.

Durrant, G. and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(2): 361–381. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00565.x.

Fowler, F. and T. Mangione. 1988. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Sage Publications: Newbury Park, CA.

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.

Groves, R. and K. McGonagle. 2001. "A Theory-Guided Interviewer Training Protocol Regarding Survey Participation." *Journal of Official Statistics* 17(2): 249–265.

Hackman, J. and G. Oldham. 1976. "Motivation through the Design of Work: Test of a Theory." *Organizational Behavior and Human Performance* 16: 250–279. Doi: http://dx.doi.org/10.1016/0030-5073(76)90016-7.

Kawamoto, K., C. Houlihan, E. Balas, and D. Lobach. 2005. "Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success." *British Medical Journal* 330(7494): 765. Doi: http://dx.doi.org/10.1136/bmj.38398.500764.8F.

Kirgis, N. and J. Lepkowski. 2013. "Design and Management Strategies for Paradata-Driven Responsive design: Illustrations from the 2006–2010 National Survey of Family Growth." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter, 121–144. Hoboken, New Jersey: John Wiley and Sons.

Kreuter, F., A. Mercer, and W. Hicks. 2014. "Increasing Fieldwork Efficiency through Prespecified Appointments." *Journal of Survey Statistics and Methodology* 2(2): 210–223. Doi: http://dx.doi.org/10.1093/jssam/smu005.

Laflamme, F. and H. St-Jean. 2011. "Highlights from the First Two Pilots of Responsive Collection Design for CATI Surveys." In Proceedings of the Joint Statistical Meetings, American Statistical Association. Available at: https://www.amstat.org/sections/srms/proceedings/y2011/Files/301087_66138.pdf (accessed March 2016).

Luiten, A. and B. Schouten. 2013. “Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction.” *Journal of the Royal Statistical Society, Series A* 176(1): 169–189. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01080.x.

Morton-Williams, J. 1993. *Interviewer Approaches*. England: Dartmouth Publishing Company Limited.

O’Muircheartaigh, C. and P. Campanelli. 1999. “A Multilevel Exploration of the Role of Interviewers in Survey Non-Response.” *Journal of the Royal Statistical Society, Series A* 162(3): 437–446. Doi: http://dx.doi.org/10.1111/1467-985X.00147.

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. “Reduction of Nonrespose Bias in Surveys through Case Prioritization.” *Survey Research Methods* 4(1): 21–29. Doi: http://dx.doi.org/10.18148/srm/2010.v4i1.3037.

Pickery, J. and G. Loosveldt. 2002. “A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse.” *Quality and Quantity* 36(4): 427–437. Doi: http://dx.doi.org/10.1023/A:1020905911108.

Poehler, E., D. Cronkite, P. Sanchez, A. Wakim, G. Dusch, H. Walrath, R. King, and J. Jones. 2016. “2020 Research and Testing: 2014 Census Test Nonresponse Followup Panel Comparisons and Instrument Analysis.” Washington, DC: U.S. Census Bureau.

Purdon, S., P. Campanelli, and P. Sturgis. 1999. “Interviewers Calling Strategies on Face-to-Face Interview Surveys.” *Journal of Official Statistics* 15(2): 199–216.

Rosen, J., J. Murphy, A. Peytchev, S. Riley, and M. Lindblad. 2011. “The Effects of Differential Interviewer Incentives on a Field Data Collection Effort.” *Field Methods* 23(1): 24–36. Doi: http://dx.doi.org/10.1177/1525822X10383390.

Schouten, B., M. Calinescu, and A. Luiten. 2013. “Optimizing Quality of Response through Adaptive Survey Designs.” The Hague: Statistics Netherlands.

Tourangeau, R., F. Kreuter, and S. Eckman. 2012. “Motivated Underreporting in Screening Interviews.” *Public Opinion Quarterly* 76(3): 453–469. Doi: http://dx.doi.org/10.2307/41684579.

Wagner, J. and K. Olson. 2011. “Where Do Interviewers Go When They Do What They Do? An Analysis of Interviewer Travel in Two Field Surveys.” In Proceedings of the Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, Miami, July 30-August 4, 2011.

Wagner, J. 2013a. “Adaptive Contact Strategies in Telephone and Face-to-Face Surveys.” *Survey Research Methods* 7(1): 45–55. Doi: http://dx.doi.org/10.18148/srm/2013.v7i1.5037.

Wagner, J. 2013b. “Using Paradata-Driven Models to Improve Contact Rates.” In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter, 145–170. Hoboken, New Jersey: John Wiley and Sons.

Wagner, J., B. West, N. Kirgis, J. Lepkowski, W. Axinn, and S. Ndiaye. 2012. “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection.” *Journal of Official Statistics* 28(4): 477–499.

Walejko, G., A. Keller, G. Dusch, and P. Miller. 2014. “2020 Research and Testing: 2013 Census Test Assessment.” Washington, DC: U.S. Census Bureau.

Received April 2016
Revised October 2017
Accepted February 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 671–694, http://dx.doi.org/10.2478/JOS-2018-0032

Estimation of True Quantiles from Quantitative Data Obfuscated with Additive Noise

Debolina Ghatak[1] *and Bimal Roy*[1]

Privacy protection and data security have recently received a substantial amount of attention due to the increasing need to protect various sensitive information like credit card data and medical data. There are various ways to protect data. Here, we address ways that may as well retain its statistical uses to some extent. One such way is to mask a data with additive or multiplicative noise and revert to certain desired parameters of the original distribution from the knowledge of the noise distribution and masked data. In this article, we discuss the estimation of any desired quantile of a quantitative data set masked with additive noise. We also propose a method to choose appropriate parameters for the noise distribution and discuss advantages of this method over some existing methods.

Key words: Data obfuscation; quantile estimation; additive noise.

1. Introduction

In official statistics, the main goal of most studies is to analyze a data set to extract different statistics like mean, median, variance and so on, which may help in various statistical analyses. However, in case the data is sensitive (e.g., income data, medical data, marksheet data, etc.), it may be completely impossible to publish it in its raw form. In such cases, statistical agencies often release a masked version of the original data, sacrificing some information. Data obfuscation refers to the type of data masking where some useful information about the complete data set remains even after hiding the individual piece of sensitive information. Therefore, the main objectives of data obfuscation are (i) to minimize the risk of disclosure resulting from providing access to the data, and (ii) to maximize the analytic usefulness of the data.

There are various ways of obfuscating data, such as “Top-coding”, “Grouping”, “Adding Noise”, “Rank Swapping”, and so on. A detailed discussion on various ways of obfuscating sensitive data may be found in the papers by Fuller (1993) and Kim and Karr (2013). Here, we deal with the obfuscation of data using multiplicative or additive noise. A typical problem involves a true quantitative data set $X_1, X_2, \ldots, X_n$; $Y_1, Y_2, \ldots, Y_n$ is a random sample from some known continuous distribution $F(\cdot)$, drawn independent of $\{X_i,\ 1 \leq i \leq n\}$. Then the noised data looks as follows:

$$Z_i = X_i + Y_i,\quad i = 1, 2, \ldots, n \quad \text{(Additive Noise Model), or} \tag{1}$$

$$Z_i = X_i Y_i,\quad i = 1, 2, \ldots, n \quad \text{(Multiplicative Noise Model)} \tag{2}$$

[1] Indian Statistical Institute, Applied Statistics Unit, p. 8. Basudebpur Sarsuna Main Road, Kolkata 700108, India. Emails: deboghatak@gmail.com and bimal@isical.ac.in

In case $\{X_i, 1 \leq i \leq n\}$ is known or assumed to follow a certain distribution, it is enough to estimate the parameters of the distribution as discussed in the papers by Fuller (1993), Mukherjee and Duncan (1997), and Kim and Karr (2013). If there is no distributional assumption on $\{X_i, 1 \leq i \leq n\}$, except that it is continuous, estimating statistics like mean, variance or raw moments from a multiplicative noise model were studied by Zayatz et al. (2011). However, the estimation of nonpolynomial statistics like quantiles may be a problem of concern. Some Bayesian methods to do the same were discussed in the article by Sinha et al. (2011). In the article by Poole (1974), he discussed the estimation procedure of the Distribution Curve of the true population from the data collected through randomized response, randomized with multiplicative noise of a particular form.

However, in all the above cases, authors have mainly concentrated on estimating the quantiles from data, obfuscated with multiplicative noise. In our problem, we work on estimating the quantiles in case the noise is additive instead of multiplicative. The goal of our study is to suggest a procedure with "reasonable" masking of the data set that may return a "good" guess of the quantiles, (one would prefer if estimation procedures of other statistics like mean, variance and so on, are also not harmed by the suggested method). We find an estimate of the distribution function for Normal, Laplace and Uniform errors that may be equated to $0 < \alpha < 1$ to find the required quantiles. A similar problem was discussed by Fan (1991) on a more general basis, popularly known as the deconvolution problem. However, we present an alternative way to look at the problem. We also propose (see subsec. 2.5) a technique for choosing the parameter for the noise distribution (statement may be found in Proposition 2.4). This is a modest attempt at solving the problem stated in the first paragraph of the introduction.

In Section 2, we describe our procedure with required proofs in the Appendix section, and in Section 3, we give some simulation results in support of our procedure. In Section 4, we give a real life example for further illustration. Finally, in Section 5, we conclude with some discussions on the whole procedure.

2. Additive Noise Model: Obfuscation and Estimation

We have a data set $\{X_i, 1 \leq i \leq n\}$ that is sensitive and hence cannot be released. We add an error $\{Y_i, 1 \leq i \leq n\}$ to each value in the data set that comes from some known distribution with a cumulative distribution function $F(\cdot)$. $Z_i = X_i + Y_i$ is the released data known as obfuscated or masked data. $F(\cdot)$ is the obfuscating distribution.

Let $G(\cdot)$, $H(\cdot)$ be the cumulative distribution functions of X and Z, respectively. We assume that (i) X and Y are independent, and (ii) X and Y (and hence Z) are continuous random variables.

Our aim is to find the quantiles of X from the knowledge of Z and $F(\cdot)$. Since we are interested in all the quantiles, we may try estimating the whole distribution curve $G(\cdot)$ of X, which can be used to find the required quantiles.

2.1. Basic Problem

Since the problem is to estimate the distribution function of X, one may first think of writing the cumulative distribution function of X, $G(\cdot)$ in terms of $H(\cdot)$ and $F(\cdot)$. But that will not be convenient, since Z and Y are not independent. Instead, we try writing $H(\cdot)$ in terms of the others. For any real number z,

$$\begin{aligned} H(z) &= P(Z \leq z) \\ &= P(X+Y \leq z) \\ &= \int_{-\infty}^{\infty} P(X+Y \leq z | Y = y) f(y) dy \end{aligned}$$

where $f(\cdot)$ denotes the probability density function of Y. Since X and Y are independent, we may write

$$\begin{aligned} H(z) &= \int_{-\infty}^{\infty} P(X \leq z - y) f(y) dy \\ &= \int_{-\infty}^{\infty} G(z-y) f(y) dy \end{aligned}$$

Thus our main equation is,

$$H(z) = \int_{-\infty}^{+\infty} G(z-y) f(y) dy. \tag{3}$$

This is an integral equation with an infinite range, where $G(\cdot)$ is the unknown function to be solved, for f is known and $H(\cdot)$ is to be estimated from the data. Note that our equation says that H is a convolution of f and G. It can alternatively be written as

$$H(z) = \int_{-\infty}^{+\infty} f(z-y)\, G(y) dy \tag{4}$$

Various methods are known to solve integral equations of different kinds. In the following subsections, we will deal with some special cases that arise in practical life. Forms of estimated $G(x)$ are given for Uniform, Normal and Laplace Error (all assumed to have zero mean). Gaussian Kernel and Silverman's Rule of Thumb bandwidth were used to estimate the densities. Then these forms of $\hat{G}(x)$ are equated to $0 < \alpha < 1$, to find the αth quantile of X. Moreover, we discuss (see subsec. 2.5) the choice of appropriate parameters of the Error Distributions, which minimize the risk of disclosure and error in estimation. As far as we know, this is a novel approach to the stated purpose.

2.2. *Uniform Error*

The following result holds if Y is *Uniform(0,a)*; that is, if the density function of Y is of the following form,

$$f(y) = \begin{cases} 1/a, 0 < y < a \\ 0, \; otherwise. \end{cases}$$

Lemma 2.1. *If $h(\cdot)$ is the density function of the obfuscated variable Z, then $\forall x \in R$*

$$G(x) = ah(x) + ah(x - a) + ah(x - 2a) + \cdots$$

In our problem, $h(\cdot)$ is unknown; instead, we can use the kernel density estimate of $h(\cdot)$ to get an estimate $\hat{G}(x)$ of $G(x)$ for all $x \in R$. Then, equating $\hat{G}(x) = \alpha$ for $0 < \alpha < 1$ we get the αth quantile of X.

Note: If Y has 0 mean, i.e., $Y \sim Uniform\left(-\frac{a}{2}, \frac{a}{2}\right)$, the form of $G(x)$ becomes

$$G(x) = ah\left(x - \frac{a}{2}\right) + ah\left(x - \frac{3a}{2}\right) + ah\left(x - \frac{5a}{2}\right) + \cdots$$

in a similar way.

2.3. Normal Error

Here $f(x) = \phi_\sigma(x) = \phi\ (x, 0, \sigma^2)$ for $x \in R,$ where $\phi(x, \mu, \sigma^2)$ is the normal density at point x with mean μ and variance σ^2.

Note that if the mean is $\mu \neq 0$ then

$$Z = X + Y \Rightarrow Z - \mu = X + (Y - \mu),\ Y - \mu \text{ has mean } 0,\ Z - \mu \text{ is known.}$$

So without loss of generality, the mean can be assumed to be zero. The following Lemma 2.2 gives an estimated form of the distribution function of X.

Before stating the next Lemma, we introduce the following assumption

(A1) The probability densities of X and Y are bounded.

We also let $\Phi(x,\mu,\sigma^2)$ denote the cumulative distribution function of the normal distribution with mean μ and variance σ^2, evaluated at the point x.

Lemma 2.2. *Assume that assumption (A1) holds. Then if $Y \sim N(0, \sigma^2)$, an estimate of $G(x)$ is,*

$$\hat{G}(x) = \frac{1}{n}\sum_{j=1}^{n} \Phi\left(x - Z_j, 0, \sqrt{b^2 - \sigma^2}\right), \quad \forall x \in R,\ b > \sigma$$

where $b = 1.06n^{-1/5}A$,

$$A = Min\left(\sqrt{\widehat{Var}(Z), \frac{IQR(Z)}{1.34}}\right)$$

$$\widehat{Var}(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2, \quad \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

and,

$$IQR(Z) = Interquantile\ range\ of\ Z = Third\ quartile\ of\ Z - First\ quartile\ of\ Z.$$

Note: The restriction on σ makes the result very weak since in most cases $b > \sigma$ is not likely to happen. However if one uses a different Kernel to estimate the density, the restriction may not hold in such cases. In the next subsection, we would like to suggest an alternative way to deal with this problem such that there is no bound on the choice of σ.

2.4. Laplace Error

The main reason behind the choice of such Error distribution is because the Laplace distribution has an "ordinary smooth density" (as defined by Fan (1991)), unlike the Normal or Cauchy distributions that possess the supersmooth density, which results in an easy solution to the problem of estimating $G(x)$ with Gaussian Kernel without any restriction on the choice of parameter.

Lemma 2.3. *An estimate of* $G(x)$, *under assumption* (*A1*) *if* $Y \sim$ *Laplace* $(0, \sigma^2)$, *i.e.,*

$$f(x) = \frac{1}{2\sigma} e^{-\left|\frac{x}{\sigma}\right|} \ \forall x \in R,$$

is given by,

$$\hat{G}(x) = \frac{1}{n}\sum_{j=1}^{n}\left\{\left(1+\frac{\sigma^2}{b^2}\right)\Phi(x, Z_j, b) - \frac{\sigma^2}{b^2}\int_{-\infty}^{(x-Z_j)/b} u^2\Phi(u)du\right\} \tag{5}$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left\{\left(1+\frac{\sigma^2}{b^2}\right)\Phi(x, Z_j, b) - \frac{\sigma^2}{b^2}0.5\left(1 + sign(x-Z_j)\mathcal{G}_{\left(\frac{3}{2},1\right)}\left(\frac{(x-Z_j)^2}{2b^2}\right)\right)\right\} \tag{6}$$

where $\mathcal{G}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{x})$ *is the cumulative distribution function of Gamma distribution with parameters* (α, β) *at x.*

Note: The density function of a Gamma distribution with parameters (α, β) *is given below:*

$$g_{(\alpha,\beta)}(y) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\lambda^{\alpha}} y^{\alpha-1} e^{-\lambda y}, y > 0 \\ 0, \textit{otherwise}. \end{cases}$$

where $\Gamma(\cdot)$ *denotes the Gamma function.*

2.5. Choice of Parameters of Error Distribution

It is to be noted that if the variance of the Error distribution is very small compared to the range of X, then the error behaves like a known constant which can be easily subtracted from Z_j to get a value very close to corresponding X_i. Hence a very small variance means no obfuscation at all. On the other hand, a very large variance may increase the error in estimation to a large extent. Hence, we need a perfect choice of the parameters of the Error Distribution to efficiently deal with the whole problem. Towards that, we make the following observation.

After obfuscating a particular value X_i we cannot get it back from $Z_i = X_i + Y_i$, but since we know the distribution of Y_i, we will get a confidence interval for each X_i. Assuming the mean of Y_i is zero, that is, Z_i and X_i has same mean, suppose for each X_i we want a minimum spread of ε with confidence $100(1 - \delta)\%$.

Proposition 2.4. *For fixed* $\delta > 0$ *and* $\varepsilon > 0$ *suppose we want a* $100(1-\delta)\%$ *Confidence Interval to be* ($Z_i - \varepsilon$, $Z_i + \varepsilon$) (ε moderately large), *then the parameter* σ *of the Error distribution can be taken as the solution of the equation*

$$F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}$$

under the condition that $F_\sigma(\cdot)$ *is the cumulative distribution function of a random variable symmetric about* 0.

Proof. Since ($Z_i - \varepsilon$, $Z_i + \varepsilon$) is $100(1-\delta)$ % *Confidence Interval* for X_i,

$$P[X_i \in (Z_i - \varepsilon, Z_i + \varepsilon)] = 1 - \delta$$

$$\Rightarrow P(|Z_i - X_i| < \varepsilon) = 1 - \delta$$

$$\Rightarrow P(|Y_i| < \varepsilon) = 1 - \delta$$

Since $F(\cdot)$ *is symmetric around 0, we can write*

$$2F_\sigma(\varepsilon) - 1 = 1 - \delta$$

$$\Rightarrow 2F_\sigma(\varepsilon) = 2 - \delta$$

$$\Rightarrow F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}.$$

Hence given ε and δ, we can find a value of σ from the equation

$$F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}.$$

Special Cases

Laplace(0, $\boldsymbol{\sigma^2}$) The c.d.f. is given by,

$$F_\sigma(x) = 0.5 + 0.5 sign(x)\left(1 - e^{-\frac{|x|}{\sigma}}\right)$$

Hence from Proposition 2.4 the solution of σ is

$$\sigma = -\frac{\varepsilon}{\log \delta}.$$

Uniform($-\frac{\sigma}{2}, \frac{\sigma}{2}$) The c.d.f. is given by $F_\sigma(x) = \frac{x+\frac{\sigma}{2}}{\sigma}$. Hence from Proposition 2.4 the solution of σ is

$$\sigma = \frac{2\varepsilon}{1 - \delta}$$

Note. For Normal Error the process only works if the solution is less than the bandwith of Z. With 95% confidence, a choice of σ is approximately $\varepsilon/1.65$.

3. Some Simulation Results

In order to apply the above problem, we simulate a non-normal sample of size $n = 2{,}000$, with $IQR/1.34 \approx 1{,}000$, and then add an error Y_i to each sample unit X_i, such that ($Z_i - \varepsilon$, $Z_i + \varepsilon$) is a 95% C.I. for X_i. The parameter for the error distribution is chosen by the formula in Proposition 2.4. For small ε, we apply Uniform, Normal and Laplace Errors to the sample, while for larger ε, Normal is not applicable. We therefore check results for Uniform and Laplace only. First, we check if the obfuscation is good enough. It is obvious that obfuscation improves as ε increases. In addition, for increasing ε, we also check how the estimation procedure works.

A sample of ten data points is taken from the data set and the corresponding obfuscated values are given for different errors. In the following Table 1, ε is assumed to be 200 (which is very small, since it is much smaller compared to the measure of dispersion of X).

Figure 1 shows the graph of the true distribution curve $\{G(x),\ x \in R\}$ along with the ones estimated from obfuscated data. Table 2 will show estimates of the true quantile values which is computed from the knowledge of $G(x)$ (Here, $G(x)$ is $Laplace(\mu = 10;\ \sigma = 1{,}000)$ using the function *qlaplace* under package {*rmutil*} of *R 3.3.2*. The quantile values are calculated from data $X_1, X_2; \ \ldots, X_n$ using function *quantile*. Also, estimated values of the quantiles are shown which we get by equating $\hat{G}(x)$ with (α: $0 < \alpha < 1$) by an iterative search method using the function *uniroot*; found in the package {*stats*} of *R 3.3.2*.

Note that the true and obfuscated values in Table 1 are quite close, which makes it easier for an intruder to guess the original value based on the obfuscated one. However, the estimation works quite well.

Now, we try increasing the value of ε. However, as the value increases, the Normal distribution is no longer an option; larger ε will make σ larger than the bandwidth of the corresponding Z.

The following Table 3 shows the true and obfuscated values of the same data points from Table 1 for increasing ε. Figure 2 will show how the estimated curve of $G(x)$ deteriorates with increasing ε. Table 4 gives the estimated and true quantiles for increasing ε.

Note that as ε increases the obfuscation improves but the estimation deteriorates. This is quite intuitive, since small ε implies no masking at all. As ε increases, both Uniform and Laplace gives result unlike Normal, but from the graph (Fig. 2), we can clearly see that for

*Table 1. Showing true and obfuscated values for ten data points selected from the 2,000 data points, **ε = 200**.*

No.	Data point	Uniform	Laplace	Normal
1	606.768	671.915	651.491	678.75
2	3139.892	3078.08	3166.548	3230.559
3	987.809	891.076	990.928	1023.493
4	2912.623	3120.068	2864.294	2714.819
5	- 1425.763	- 1369.556	- 1470.395	- 1518.552
6	- 185.086	- 305.841	- 68.098	- 205.403
7	- 940.958	- 1097.012	- 897.075	- 804.884
8	- 955.503	- 964.716	- 979.366	- 1005.702
9	- 224.565	- 46.007	- 228.214	- 304.326
10	- 511.614	- 470.031	- 469.044	- 597.995

Table 2. Estimated quantiles from obfuscated data, ε = 200.

α	TRUE	Original	Uniform	Laplace	Normal
“0.1”	- 1599.438	- 1476.929	- 1525.415	- 1534.134	- 1512.133
“0.2”	- 906.291	- 847.771	- 895.061	- 900.945	- 893.431
“0.3”	- 500.826	- 491.793	- 521.976	- 522.429	- 525.321
“0.4”	- 213.144	- 224.8	- 240.816	- 243.329	- 245.115
“0.5”	10	- 9.7	3.925	2.659	6.166
“0.6”	233.144	242.808	257.094	260.244	267.592
“0.7”	520.826	533.289	552.537	559.502	564.615
“0.8”	926.291	922.478	954.164	966.336	963.852
“0.9”	1619.438	1655.947	1687.02	1697.753	1698.098

larger quantiles the Uniform distribution gives very bad estimates, since the estimate of $G(x)$ at times even becomes decreasing, which is not at all desirable. However, Laplace seems to give comparatively better results compared to the Uniform. A theoretical explanation of the drawback of using Uniform Error is discussed in Section 5. Hence, we here prefer the use of Laplace Error over Uniform and Normal for reasonably large ε.

Hence, to investigate deeper into the statistical properties of such estimates, we note that the estimate is consistent, as is the estimate by Fan (1991). To evaluate other properties, such as the bias and mean square error in estimation, we find the Monte Carlo estimates of the bias and root-mean-squared-error (RMSE) over a simulation of S error samples (We take $S = 500$, 800 and 1,000). The Tables 5-8 present estimates of bias and RMSE for growing ε.

Compared to the dispersion of the data set (IQR = 1:34 ≈ 1,000), the RMSE does not seem to be very large for $\varepsilon = 200$, 500 or 1,000. $\varepsilon = 2{,}000$ gives very large bias and RMSE but that large ε is rarely needed.

It can be easily observed that the bias and RMSE were consistent in the sense that 500, 800, and 1,000 simulations resulted in approximately similar values for all the cells in the above tables 5-8.

Observing the tables 5-8, we note that the main error in estimation comes from the bias of the estimate. Hence, an estimation of bias for the above problem can be a very interesting problem and a useful result for future research work.

But from this scenario, it is not clear whether the estimator is consistent, that is, with increasing n whether the bias decreases, although from Fan (1991) we can easily see that theoretically the estimate of $G(x)$ is consistent for all $x \in R$. So, to investigate, we simulate some other samples $X_1, X_2, \ldots, X_n$ using the same distribution as before, but larger n (we take $n = 5{,}000$, 10,000), and obfuscate using Laplace error similarly to find the Monte Carlo estimates of bias and RMSE, using $S = 1{,}000$.

One may easily observe from the tables (Table 7 and Table 8) that there is a decrease in the value of absolute bias and RMSE with larger n. Hence, with increasing n, ideally, the error tends to vanish.

4. A Real Life Example

For further illustration, we consider a real life application of the problem. We collect a data set of grades achieved by 445 students in the second year of the Masters of Statistics

Table 3. True and obfuscated values for ten data points selected from the 2,000 data points with increasing ε.

		$\varepsilon = 500$		$\varepsilon = 1{,}000$		$\varepsilon = 2{,}000$	
No.	Data point	Uniform	Laplace	Uniform	Laplace	Uniform	Laplace
1	606.768	777.005	697.307	1549.425	-54.751	-866.566	326.243
2	3139.892	3414.243	3134.548	4152.921	3718.174	3635.376	2679.141
3	987.809	1210.838	936.253	52.861	1216.174	3055.399	1140.865
4	2912.623	2760.988	2985.984	2376.626	2442.173	1178.522	3182.984
5	-1425.763	-1521.242	-1451.637	-1908.008	-1720.502	-530.379	-1017.015
6	-185.086	330.237	-245.401	676.281	796.201	-1985.132	-163.254
7	-940.958	-420.662	-960.868	-1199.835	-1051.968	-1665.925	-936.007
8	-955.503	-948.84	-1040.34	-1429.07	-1083.252	-1027.686	-860.724
9	-224.565	146.065	-299.93	-901.975	-786.876	-1592.798	-1222.795
10	-511.614	-216.055	-532.046	-568.219	381.672	404.65	-1145.256

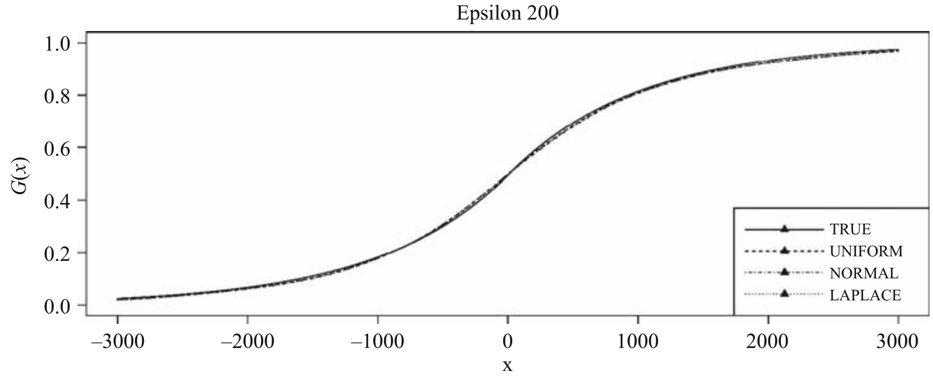


Fig. 1. True and estimated distribution curve with ε = 200.

program at the Indian Statistical Institute Kolkata over ten years, from 2006–2015. As grades are sensitive data, they cannot be released in raw form. We therefore apply the above problem to this data and try to find the results. Standard variation of the data was checked to be approximately 100, so we assumed an $\varepsilon = 200$. The bandwidth values from Uniform and Laplace data was found to be 48.68 and 41.15. The following Table 9 represents true and obfuscated values of ten data points to show how the values are masked with Uniform and Laplace Errors. From the obfuscated values, the true distribution and quantiles are then estimated as shown in Figure 3 and Table 10 respectively.

In this problem, σ was chosen according to Proposition 2.4 with $\varepsilon = 200$. Without access of the obfuscated data, all one knew about the marks of an individual was that it ranged between 0 to 1,000. Consider the first individual in Table 9. Its masked value after masking with $Laplace(0,\sigma^2)$ is 733.93. Now, we can say $X_i \in (533.93; 933.93)$ with 95% confidence. Hence, a disclosure takes place here. Note that, as per our knowledge, Z_i is the best estimator of X_i, based on the available information. However, if the intruder has an algorithm that can be used to find a better estimator of X_i using the knowledge of the obfuscating distribution and obfuscated data, this disclosure risk may not be valid (it can easily be shown that if true variance of Y is greater than $\frac{n}{n-1}$ times the true variance of X, then $\hat{Z}$ is a better estimator of X_i than Z_i; that is, the mean squared error of $\hat{Z}$ about X_i is less than that of Z_i about X_i but such a case is rare as σ usually does not need to be so large).

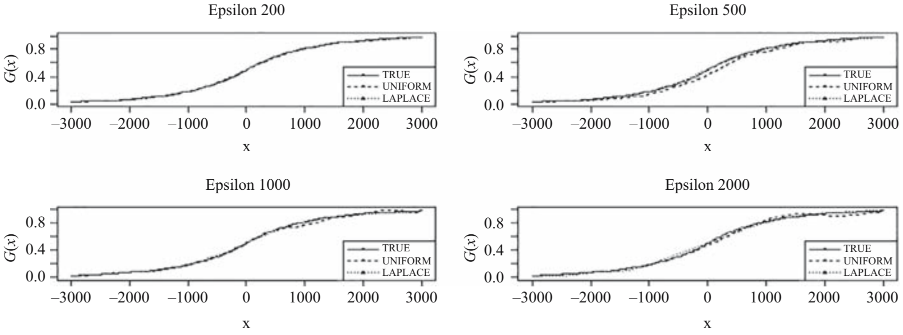


Fig. 2. True and estimated distribution curves with increasing ε.

Table 4. Estimated quantiles from obfuscated data with increasing ε.

			$\varepsilon = 500$		$\varepsilon = 1{,}000$		$\varepsilon = 2{,}000$	
α	TRUE	No error	Uniform	Laplace	Uniform	Laplace	Uniform	Laplace
“0.1”	- 1599.438	- 1476.929	- 1270.597	- 1540.032	- 1464.955	- 1695.023	- 1633.896	- 1892.266
“0.2”	- 906.291	- 847.771	- 742.253	- 925.49	- 880.811	- 1014.747	- 917.164	- 1090.942
“0.3”	- 500.826	- 491.793	- 360.159	- 554.813	- 487.193	- 615.94	- 496.031	- 730.971
“0.4”	- 213.144	- 224.8	- 55.626	- 264.023	- 221.973	- 266.593	- 179.127	- 337.419
“0.5”	10	- 9.7	178.56	- 12.464	- 5.545	- 1.624	91.225	6.49
“0.6”	233.144	242.808	389.331	257.945	213.649	296.902	339.212	349.937
“0.7”	520.826	533.289	644.638	580.443	540.851	610.137	590.126	717.073
“0.8”	926.291	922.478	1168.204	989.891	1179.53	1065.751	876.444	1155.097
“0.9”	1619.438	1655.947	1679.645	1745.236	1730.618	1827.97	1284.765	1902.892

Table 5. *Showing true values of quantiles of a data set and the corresponding* ***bias*** *of the estimate for* ***Laplace error*** *with increasing* ε.

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Org. data	$S = 500$	7.172	3.874	1.122	0.457	-0.235	-0.145	0.912	0.853	0.637
	$S = 800$	5.025	1.564	-0.239	0.137	0.111	0.187	0.217	0.258	-0.881
	$S = 1{,}000$	5.224	1.278	0.083	0.507	0.293	0.73	0.642	0.971	-0.164
$\varepsilon = 200$	$S = 500$	-23.802	-26.068	-27.564	-23.289	0.19	24.065	29.521	30.645	31.316
	$S = 800$	-26.253	-28.116	-28.747	-23.852	0.102	24.06	29.041	29.885	30.019
	$S = 1{,}000$	-26.096	-28.036	-28.432	-23.574	0.46	24.571	29.69	30.612	30.632
$\varepsilon = 500$	$S = 500$	-25.19	-27.786	-30.065	-24.923	-0.009	24.816	30.874	32.841	32.347
	$S = 800$	-27.962	-30.432	-30.921	-25.012	0.249	24.973	30.592	32.277	31.562
	$S = 1{,}000$	-28.165	-30.218	-30.566	-24.798	0.601	25.499	31.235	32.769	32.494
$\varepsilon = 1{,}000$	$S = 500$	-27.934	-35.081	-35.918	-28.76	0.431	30.062	38.386	40.433	40.181
	$S = 800$	-31.278	-36.546	-37.2	-29.646	-0.157	29.498	37.29	39.21	39.93
	$S = 1{,}000$	-32.331	-36.816	-37.039	-29.157	0.419	29.86	37.521	39.493	40.519
$\varepsilon = 2{,}000$	$S = 500$	-52.89	-54.773	-52.252	-39.383	-2.804	35.897	52.904	58.526	59.112
	$S = 800$	-56.447	-54.972	-53.402	-38.78	-1.55	36.939	52.953	58.96	57.323
	$S = 1{,}000$	-53.661	-56.074	-53.609	-38.725	-0.888	37.88	54.206	59.954	59.896

*Table 6. Showing true values of quantiles of a dataset and the corresponding **root mean square error** of the estimate for **Laplace error** with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Org. data	$S = 500$	68.292	44.792	33.593	27.865	23.739	28.55	34.165	44.095	64.891
	$S = 800$	67.933	44.249	33.581	27.893	23.425	28.543	34.946	45.653	65.1
	$S = 1,000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
$\varepsilon = 200$	$S = 500$	67.882	48.976	41.846	35.056	24.495	35.696	43.637	52.063	69.307
	$S = 800$	68.329	49.882	42.652	35.395	24.516	35.976	43.876	52.454	68.891
	$S = 1,000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
$\varepsilon = 500$	$S = 500$	69.828	50.778	44.136	36.893	25.412	36.825	45.225	53.997	71.167
	$S = 800$	71.076	52.533	44.756	36.962	25.6	37.374	45.691	54.721	71.555
	$S = 1,000$	71.256	52.176	44.357	36.542	25.266	37.605	46.008	54.802	72.471
$\varepsilon = 1,000$	$S = 500$	81.559	61.189	52.276	42.909	30.259	44.654	55.829	65.916	84.804
	$S = 800$	81.106	61.823	53.471	43.672	30.466	44.809	55.617	65.361	83.706
	$S = 1,000$	80.892	61.798	53.174	43.091	29.979	44.58	55.214	64.933	84.427
$\varepsilon = 2,000$	$S = 500$	124.104	93.579	77.468	62.005	45.251	60.263	78.433	94.74	128.227
	$S = 800$	126.806	93.281	78.115	62.302	46.944	61.471	78.53	95.605	127.758
	$S = 1,000$	125.164	93.145	77.559	62.169	46.741	61.918	79.275	96.219	128.16

Table 7. Showing true values of quantiles of three data sets with sample size 2,000, 5,000, 10,000 and the corresponding estimated ***bias*** *of the quantile estimate with $S = 1{,}000$ simulations for Laplace error with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Original	$n = 2{,}000$	5.224	1.278	0.083	0.507	0.293	0.73	0.642	0.971	-0.164
	$n = 5{,}000$	-0.277	-1.503	-1.164	-0.523	-0.183	-0.437	-1.164	-2.326	-0.914
	$n = 10{,}000$	-0.584	-0.231	-0.688	-0.671	-0.313	-0.635	-0.802	-0.698	-2.466
$\varepsilon = 200$	$n = 2{,}000$	-26.096	-28.036	-28.432	-23.574	0.46	24.571	29.69	30.612	30.632
	$n = 5{,}000$	-21.416	-21.935	-21.277	-18.278	-0.361	17.343	19.126	18.568	19.985
	$n = 10{,}000$	-16.184	-15.692	-15.939	-14.558	-0.467	13.341	14.505	14.66	12.957
$\varepsilon = 500$	$n = 2{,}000$	-28.165	-30.218	-30.566	-24.798	0.601	25.499	31.235	32.769	32.494
	$n = 5{,}000$	-23.077	-22.767	-22.223	-19.07	-0.291	18.139	20.025	19.7	21.558
	$n = 10{,}000$	-16.73	-16.634	-17.02	-15.612	-0.783	14.054	15.741	15.824	14.79
$\varepsilon = 1{,}000$	$n = 2{,}000$	-32.331	-36.816	-37.039	-29.157	0.419	29.86	37.521	39.493	40.519
	$n = 5{,}000$	-26.49	-26.753	-26.414	-22.493	-0.72	21.409	24.189	23.618	26.626
	$n = 10{,}000$	-21.879	-20.424	-19.784	-17.714	-0.925	16.116	18.12	18.263	18.515
$\varepsilon = 2{,}000$	$n = 2{,}000$	-53.661	-56.074	-53.609	-38.725	-0.888	37.88	54.206	59.954	59.896
	$n = 5{,}000$	-45.808	-40.622	-39.889	-29.43	0.823	30.261	39.033	37.867	38.36
	$n = 10{,}000$	-28.213	-30.479	-29.453	-24.909	-1.367	22.811	28.866	29.327	26.131

Table 8. *Showing true values of quantiles of three data sets with sample size 2,000, 5,000, 10,000 and the corresponding estimated* ***root mean square error*** *of the quantile estimate with* $S = 1{,}000$ *simulations for* ***Laplace error*** *with increasing* ε.

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Original	$n = 2{,}000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
	$n = 5{,}000$	41.639	27.444	21.334	16.825	14.143	17.589	21.532	28.819	42.719
	$n = 10{,}000$	29.708	20.233	15.28	12.255	9.894	11.977	15.526	19.61	29.455
$\varepsilon = 200$	$n = 2{,}000$	68.099	49.741	42.235	34.963	24.266	36.193	44.147	52.711	69.264
	$n = 5{,}000$	44.931	34.362	29.217	24.229	14.738	23.881	28.026	32.84	44.692
	$n = 10{,}000$	32.599	24.863	21.596	18.638	10.382	17.568	20.496	23.802	30.9
$\varepsilon = 500$	$n = 2{,}000$	71.256	52.176	44.357	36.542	25.266	37.605	46.008	54.802	72.471
	$n = 5{,}000$	47.129	36.027	30.832	25.65	15.838	25.245	29.45	34.327	46.771
	$n = 10{,}000$	34.199	26.524	23.124	20.063	11.156	18.565	21.985	25.269	33.25
$\varepsilon = 1{,}000$	$n = 2{,}000$	80.892	61.798	53.174	43.091	29.979	44.58	55.214	64.933	84.427
	$n = 5{,}000$	57.132	43.227	36.534	30.642	18.875	30.116	36.236	41.937	56.043
	$n = 10{,}000$	43.107	32.83	27.969	24.14	14.94	22.751	26.719	31.432	41.413
$\varepsilon = 2{,}000$	$n = 2{,}000$	125.164	93.145	77.559	62.169	46.741	61.918	79.275	96.219	128.16
	$n = 5{,}000$	98.15	66.757	57.042	44.217	31.295	45.695	56.853	67.185	93.56
	$n = 10{,}000$	71.966	52.769	44.693	36.178	25.85	35.557	43.169	50.703	71.999

Table 9. True and obfuscated values for ten data points selected from the 445 data points, ε = 200.

No.	TRUE	Uniform	Laplace
"1"	814	960.562	733.931
"2"	750	695.214	829.526
"3"	764	656.395	591.158
"4"	574	704.041	599.055
"5"	614	670.67	586.944
"6"	669	595.926	670.136
"7"	616	553.873	533.097
"8"	674	748.607	677.74
"9"	714	595.295	658.648
"10"	740	883.885	764.591

In this case, Y_i is the error in estimation, and there is no risk of disclosure. However, there is a probability that the error is very small. Hence, the risk of disclosure with error less than d, is given by,

$$P[|Z_i - X_i| < d] = P[|Y_i| < d]$$

For $S = 1{,}000$ simulations, an estimate of this risk is

$$\frac{\sum_{s=1}^{S} I_{[Z_{si} \in (X_i - d, X_i + d)]}}{S}$$

where Z_{si} is the masked value of X_i for sth simulation and $I_{[A\ ]} = 1$, if event A occurs and zero otherwise. The following Table 11 shows estimates of disclosure risk for growing error values at ten selected points (the points in Table 9), and also a column giving the true risk value. We see the estimated risks are quite close to the theoretically determined risk at all the selected points.

5. Conclusion

Given the simulation results and also the real life example one can easily see that an increase in the value of ε, that is, an increase in obfuscation, results in weakly reliable

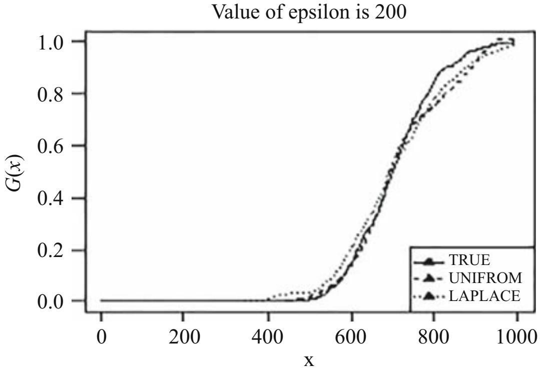


Fig. 3. Showing estimated distribution curve from TRUE and obfuscated data sets.

Table 10. Showing estimation of quantiles from original and obfuscated data.

No.	Original	Uniform	Laplace
"0.1"	580.8	578.394	555.663
"0.2"	612.8	622.305	596.067
"0.3"	645.2	650.741	633.059
"0.4"	675.6	673.521	664.011
"0.5"	700	695.237	693.693
"0.6"	727	720.346	734.52
"0.7"	750	762.636	770.46
"0.8"	786	831.202	809.933
"0.9"	826.6	888.999	879.513

estimates for both Laplace and Uniform Errors. However, we would prefer the use of Laplace over Uniform Error, since the Uniform has a serious drawback, as explained in the next paragraph.

In the case of Uniform Error, the estimate of $G(x)$ is given by the expression

$$\hat{G}(x) = \frac{a}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\phi\left(x, Z_j + \left(m + \frac{1}{2}\right)a, b\right)$$

which is nondecreasing if,

$$\hat{g}(x) = \frac{a}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\phi'\left(x, Z_j + \left(m + \frac{1}{2}\right)a, b\right) \geq 0,$$

that is if,

$$-\frac{c}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\left(x - Z_j - \left(m + \frac{1}{2}\right)a\right)e^{-\frac{\left(x-Z_j-\left(m+\frac{1}{2}\right)a\right)^2}{2b^2}} \geq 0$$

where c is a positive constant.

However, this term may become negative for certain cases. $\hat{G}(x)$ can therefore decrease at times, which is not at all desirable, since it is an estimate of a cumulative distribution function. In our simulations, we found that this problem arose several times, while in case of Laplace Error, this problem never arose. However, theoretically Equation (5), resulting from Laplace noise distribution, could not be proven to have a nondecreasing distribution function, either.

We have currently checked results for Uniform and Laplace distributions. However, the choice of an optimal density function for obfuscation and estimation has not yet been properly defined. It would be a challenging problem to define the optimal criterion and find a density that is capable of satisfying the criterion. The same challenge applies to finding an optimal ε (as defined in subsec. 2.5) for a given data set $(X_1, X_2, \ldots, X_n)$.

As discussed in Section 3, the error in estimation is mainly a result of the bias of the estimate. Hence, an estimation of bias and its correction should lead to a better resolution of the problem.

Table 11. Showing estimated risk of disclosure at ten selected points for increasing error value and theoretically determined risk value.

	X_i										
d	814	750	764	574	614	669	616	674	714	740	True value
10	0.154	0.114	0.141	0.127	0.14	0.125	0.141	0.145	0.143	0.163	0.139
20	0.282	0.215	0.25	0.272	0.264	0.254	0.261	0.268	0.274	0.283	0.259
30	0.393	0.33	0.343	0.368	0.369	0.344	0.369	0.385	0.386	0.38	0.362
40	0.47	0.418	0.453	0.462	0.457	0.431	0.456	0.475	0.493	0.475	0.451
50	0.542	0.491	0.533	0.539	0.522	0.494	0.528	0.549	0.553	0.559	0.527
60	0.617	0.575	0.602	0.588	0.58	0.573	0.579	0.604	0.619	0.608	0.593
70	0.666	0.636	0.659	0.649	0.631	0.637	0.627	0.651	0.677	0.649	0.65
80	0.709	0.689	0.707	0.696	0.68	0.691	0.678	0.703	0.722	0.685	0.698
90	0.742	0.724	0.752	0.736	0.723	0.728	0.729	0.745	0.754	0.734	0.74
100	0.775	0.754	0.785	0.771	0.764	0.756	0.771	0.779	0.796	0.779	0.776

Moreover, as mentioned in Section 4, if the boundary values of the original data are known, the obfuscation in the boundary region degrades. There is no known solution to this problem.

Having obtained a quantile estimate, computation of a confidence interval for the unknown population quantile could be an interesting problem for future work.

However, the problem discussed can easily be applied to many real life problems. The technique used to solve the above problem can be applied to solve the equations for other error distributions too. Unlike the historical technique to solve such problems, as given in Fan (1991), this technique can be applied to cases where the characteristic function of the Error distribution may take nonpositive value in some regions over the real line.

Appendix

Proof of Lemma 2.1

Proof. Putting the form of $f(y)$ in Equation (3), we have

$$H(z) = \frac{1}{a}\int_{0}^{a} G(z-y)dy$$

Now differentiating with respect to z we have

$$h(z) = \frac{1}{a}\{G(z) - G(z-a)\},$$

which gives

$$G(z) = ah(z) + G(z-a).$$

Now, from this relation we have

$$G(z-a) = ah(z-a) + G(z-2a).$$

Inserting this in the expression for $G(z)$ we find

$$G(z) = ah(z) + ah(z-a) + G(z-2a)$$

Repeating this by putting the values of $G(z - ma)$ for $m = 1, 2, \ldots$ in a similar way, we arrive at the given result.

Proof of Lemma 2.2

Proof. The lemma is proved using the following result from Polyanin and Manzhirov (Polyanin and Manzhirov 2008).

Result: *Consider the equation* $\int_{-\infty}^{\infty} K(x-t)y(t)dt = f(x), -\infty < x < \infty$ *where* $y(\cdot)$ *is the unknown function to be determined. Suppose,*

(i) $f(x), y(x) \in L_2(-\infty, \infty)$
(ii) $K(x) \in L_1(-\infty, \infty)$

where the function space $L_k(S)$ for some set S and integer k, is the set of all real-valued functions $\left\{f : S \to R, \int_{-\infty}^{\infty} |f(x)|^k dx < \infty\right\}$.

Then, $y(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{f}(u)}{\tilde{K}(u)} e^{iux} du$, *where* $\tilde{f}$ *is the Fourier Transform of* f, $\tilde{K}$ *is the Fourier Transform of* K.

Now to apply the given result in our problem note that our equation is

$$H(z) = \int_{-\infty}^{\infty} \mathrm{G}(y)\phi_\sigma(z-y)dy = \int_{-\infty}^{\infty} \mathrm{G}(z-y)\phi_\sigma(y)dy$$

But $H(\cdot)$ and $G(\cdot)$ are not $L_2(-\infty, \infty)$. So taking the derivative w.r.t. z, we get

$$h(z) = \frac{d}{dz}\int_{-\infty}^{\infty} \mathrm{G}(z-y)\phi_\sigma(y)dy.$$

Now, since g($\cdot$) is bounded, for some real $0 < \mathrm{M} < \infty$, we have,

$$\frac{d}{dz}(G(z-y)\phi_\sigma(y)) = g(z-y)\phi_\sigma(y) \leqslant M\phi_\sigma(y)$$

Now $\int_{-\infty}^{\infty} M\phi_\sigma(y)dy = M < \infty$. Hence we can interchange the integration and differentiation sign which gives us,

$$h(z) = \int_{-\infty}^{\infty} g(z-y)\phi_\sigma(y)dy$$

Here, we have used the Leibniz rule for infinite range.

Now, since $g(\cdot)$ and $h(\cdot)$ are bounded by assumption (A1), they are L_2 $-$ *bounded* by Lemma 2.3 of the book "Deconvolution Problems in Non-Parametric Statistics" by Meister (2009). Also, $\phi_\sigma \in L_1(-\infty, \infty)$.

Hence, applying the last result, in our problem,

$$g(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{h}(k)}{\tilde{\phi}_\sigma(k)} e^{ikx} dk$$

But h is not known. So, we replace it by $\hat{h}$, the Kernel Density Estimate of h using standard Gaussian Kernel and bandwidth selected by Silverman's "Rule of Thumb". The general form of such kind of estimators with an arbitrary kernel function $K(\cdot)$ was discussed by Fan (1991) where the kernel estimators of mixture densities were studied along with their asymptotic properties. It is given by

$$\hat{h}(x) = \frac{1}{nb}\sum_{j=1}^{n} K\left(\frac{x - Z_j}{b}\right) \qquad (7)$$

where $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $b = 1.06\, n^{-\frac{1}{5}} A$ as defined in the statement of the Lemma. Plugging in, we get,

$$\tilde{\hat{h}}(k) = \int_{-\infty}^{\infty} \left\{ \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - Z_j}{b} \right) \right\} e^{-ikx} dx$$

$$= \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

Since

$$\frac{1}{b} \int_{-\infty}^{\infty} e^{-ikx} K\left(\frac{x - Z_j}{b} \right) dx = \frac{1}{\sqrt{2\pi} b} \int_{-\infty}^{\infty} e^{-ikx} e^{-\frac{(x - Z_j)^2}{2b^2}} dx,$$

which is the characteristic function of a normal random variable with mean Z_j and standard deviation b at the point $(-k)$ and we know that to be equal to $e^{-ikZ_j - \frac{k^2 b^2}{2}}$.

Also, note that

$$\tilde{\phi}_\sigma(k) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} e^{-ikx} dx$$

$$= e^{\frac{-k^2\sigma^2}{2}}$$

Therefore, $\frac{\tilde{h}(k)}{\tilde{\phi}_\sigma(k)} = \frac{\frac{1}{n}\sum_{j=1}^{n} e^{-ikx - \frac{k^2 b^2}{2}}}{e^{-\frac{k^2\sigma^2}{2}}} = \frac{1}{n}\sum_{j=1}^{n} e^{-ikx - \frac{k^2(b^2 - \sigma^2)}{2}} \in L_2(-\infty, \infty)$ if $b^2 - \sigma^2 > 0$, that is, $b > \sigma$

If $b > \sigma$, then,

$$\hat{g}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2(b^2 - \sigma^2)}{2}} e^{ixk} dk$$

$$= \frac{1}{\sqrt{2\pi} n \sqrt{\mathrm{b}^2 - \sigma^2}} \sum_{j=1}^{n} e^{-\frac{(x - Z_j)^2}{2(\mathrm{b}^2 - \sigma^2)}}.$$

where we have changed the order of summation and integration. This is nothing but the mean of n normal p.d.f.s with mean Z_j and variance $b^2 - \sigma^2$. Hence, we get the form given in Lemma 2.2.

Proof of Lemma 2.3

Proof: Proceeding in the same way as in Lemma 2.2, we have

$$\tilde{\hat{h}}(k) = \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

and the Fourier transform at point k of the Laplacian error density with scale parameter σ, denoted as $\tilde{\ell}_\sigma(k)$, is given by,

$$\tilde{\ell}_\sigma(k) = \int_{-\infty}^{\infty} \frac{1}{2\sigma} e^{-|x|/\sigma} e^{-ikx} dx = (1 + \sigma^2 k^2)^{-1}$$

Hence, the ratio becomes

$$\frac{\tilde{\hat{h}}(k)}{\tilde{\ell}_\sigma(k)} = \frac{1}{n} \sum_{j=1}^{n} (1 + \sigma^2 k^2) e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

This function is now in $L_2(-\infty, \infty)$ $\forall b, \sigma$. After taking the inverse Fourier transform, we have

$$\hat{g}(x) = \frac{1}{n} \sum_{j=1}^{n} I_j,$$

where

$$\begin{aligned} I_j &= \frac{1}{2\pi} \int_{-\infty}^{\infty} (1 + \mathrm{k}^2 \sigma^2) e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk \\ &= I_{1j} + I_{2j}, \end{aligned}$$

$$I_{1j} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk, \text{ and,}$$

$$I_{2j} = \frac{1}{2\pi} \int_{-\infty}^{\infty} k^2 \sigma^2 e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk.$$

Note that the integrand in I_{1j} is nothing but a constant multiple of the characteristic function of $N(0, 1/b)$ at $(x - Z_j)$ and hence, it can easily be shown that

$$I_{1j} = \phi(x, Z_j, b)$$

It should now be noted that

$$I_{2j} = \frac{\sigma^2}{2\pi} \int_{-\infty}^{\infty} k^2 e^{\frac{-k^2 b^2}{2}} \{\cos(k(x - Z_j)) + i\sin(k(x - Z_j))\} dk$$

Since the sine function is odd and the cosine function is even, we can write

$$I_{2j} = \frac{\sigma^2}{\pi} \int_{0}^{\infty} \cos(k(x - Z_j)) k^2 e^{\frac{-k^2 b^2}{2}} dk$$

Defining $c_j = \frac{\sqrt{2}}{b}(x - Z_j)$ and making a change of variables, we get the expression

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\int_0^\infty \cos\left(c_j\sqrt{y}\right)\sqrt{y}e^{-y}dy$$

Next, expanding $\cos\left(\mathrm{c_j}\sqrt{\mathrm{y}}\right)$ by a Taylor series and changing the order of summation and integration, we have

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2\mathrm{m}}}{(2m)!}\Gamma\left(m + \frac{3}{2}\right)$$

where $\Gamma(x)$ denotes the Gamma function evaluated at the point x. Using the properties of the Gamma function that $\Gamma(x + 1) = x\Gamma(x)$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ we can further calculate

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{(2m)!}\left(m + \frac{1}{2}\right)\left(m - \frac{1}{2}\right)\ldots\frac{1}{2}\mathit{\Gamma}\left(\frac{1}{2}\right)$$

$$= \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{(2m)!}\left(m + \frac{1}{2}\right)\left(m - \frac{1}{2}\right)\ldots\frac{1}{2}\mathit{\Gamma}\left(\frac{1}{2}\right)$$

$$= \frac{\sigma^2}{\sqrt{\pi}}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}m!}\frac{2m + 1}{2}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}\left\{2\sum_{m=1}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}(m - 1)!} + \sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}(m)!}\right\}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}\left\{2(-1)\left(\frac{c_j}{2}\right)^2 e^{-\left(\frac{c_j}{2}\right)^2} + e^{-\left(\frac{c_j}{2}\right)^2}\right\}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}e^{-(c_j/2)^2}[1 - 2(c_j/2)^2]$$

$$= \frac{\sigma^2}{b^2}\left[1 - \left(\frac{x - Z_j}{b}\right)^2\right]\phi(x, Z_j, b)$$

where we inserted the expression $c_j = \frac{\sqrt{2}}{b}(x - Z_j)$ in the last step. Thus, we can conclude that

$$\hat{g}(x) = \left(1 + \frac{\sigma^2}{b^2}\right)\left\{\frac{1}{n}\sum_{i=1}^{n}\phi(x, Z_j, b)\right\} - \frac{\sigma^2}{b^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x - Z_j}{b}\right)^2\phi(x, Z_j, b)$$

Hence, integrating $\hat{g}(u)$ over $(-\infty, x)$ we get Equation (5). Moreover, making a simple change of variable $\frac{u^2}{2} = y$ in the term

$$\int_{-\infty}^{\frac{x-Z_j}{b}} u^2 \phi(u) du$$

one can easily check whether it is equal to

$$0.5 + 0.5^{*} sign(x - Z_j) \mathcal{G}_{(3/2,1)} \left(\frac{x - Z_j}{b} \right)^2$$

as stated in Equation (6).

6. References

Fan, J. 1991. "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems." *The Annals of Statistics* 19(3): 1257–1272. Available at: http://www.jstor.org/stable/2241949 (accessed December 2017).

Fuller, W.A. 1993. "Masking Procedures for Microdata Disclosure Limitation." *Journal of Official Statistics* 9(3): 383–406. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/masking-procedures-for-microdata-disclosure-limitation.pdf (accessed December 2017).

Kim, H.J. and A.F. Karr. 2013. *The Effect of Statistical Disclosure Limitation on Parameter Estimation for a Finite Population*. NISS, October.

Meister, A. 2009. *Deconvolution Problems in Nonparametric Statistics*. Berlin Heidelberg: Springer Verlag.

Mukherjee, S. and G.T. Duncan. 1997. *Disclosure Limitation through Additive Noise Data Masking: Analysis of Skewed Sensitive Data. Disclosure Limitation through Additive Noise Data Masking: Analysis of Skewed Sensitive Data*. IEEE.

Polyanin, A.D. and A.V. Manzhirov. 2008. *Handbook of Integral Equations*. Chapman and Hall/CRC.

Poole, W.K. 1974. "Estimation of the Distribution Function of a Continuous Type Random Variable Through Randomized Response." *Journal of the American Statistical Association* 69(348): 1002–1005.

Sinha, B., T.K. Nayak, and L. Zayatz. 2011. "Privacy Protection and Quantile Estimation from Noise Multiplied Data." *Sankhya B* 73: 297–315. Doi: https://doi.org/10.1007/s13571-011-0030-z.

Zayatz, L., K.T. Nayak, and B.K. Sinha. 2011. "Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection." *Journal of Official Statistics* 27(2): 527–544. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee-5bf7be7fb3/statistical-properties-of-multiplicative-noise-masking-for-confidentiality-protection.pdf (accessed December 2017).

Received March 2016
Revised May 2017
Accepted October 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 695–720, http://dx.doi.org/10.2478/JOS-2018-0033

Accounting for Spatial Variation of Land Prices in Hedonic Imputation House Price Indices: a Semi-Parametric Approach

Yunlong Gong[1,2] *and Jan de Haan*[2,3]

Location is capitalized into the price of the land the structure of a property is built on, and land prices can be expected to vary significantly across space. We account for spatial variation of land prices in hedonic house price models using geospatial data and a semi-parametric method known as mixed geographically weighted regression. To measure the impact on aggregate price change, quality-adjusted (hedonic imputation) house price indices are constructed for a small city in the Netherlands and compared to price indices based on more restrictive models, using postcode dummy variables, or no location information at all. We find that, while taking spatial variation of land prices into account improves the model performance, the Fisher house price indices based on the different hedonic models are almost identical. The land and structures price indices, on the other hand, are sensitive to the treatment of location.

Key words: Geospatial information; hedonic modeling; land and structure prices; mixed geographically weighted regression; residential property

JEL Classification: C14; C33; C43; E31; R31.

1. Introduction

The construction of house price indices is difficult because houses are traded infrequently and because properties are unique in terms of their location and structural characteristics. Hedonic regression and repeat sales methods both deal with these problems. The repeat sales method controls for location and unchanged structural characteristics as the prices of the 'same' properties are tracked over time (in a regression framework). However, this method suffers from several problems. For example, since only houses that are sold at least twice in the data set are used, it ignores single sales and is prone to sample selection bias. Also, the repeat sales method cannot provide information on the shadow prices of the property characteristics and thus does not allow the estimation of, for example, price

[1] Department of Land Resource Management, China University of Mining and Technology, Daxue Road 1, 221116 Xuzhou, China. Email: ylgong@cumt.edu.cn
[2] OTB-Research for the Built Environment, Delft University of Technology, Juliannalaan 134, 2628 BL Delft, The Netherlands.
[3] IT and Methodology Division, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands. Email: j.dehaan@cbs.nl

Acknowledgments: The authors would like to thank participants at the Economic Measurement Group Workshop, December 2014, Sydney, and at the Second International Conference of the Society for Economic Measurement, July 2015, Paris, for helpful comments on an earlier version of the article. Thanks are also given to two anonymous referees and the editor for critical review and constructive comments.

indices of the land the structure sits on. Given these problems with repeat sales methods, we focus on hedonic regression methods.

The hedonic regression method has its limitations as well. A general problem in the context of housing is omitted variables bias; it is not possible to include all the structural characteristics into the model, even if data on these characteristics were available (which is usually not the case). In addition, the true relationship between housing characteristics and house prices is unknown. The treatment of location is an important issue. One may for instance include locational variables in the model, such as distance to the city center and amenities. However, this is a rather data intensive method, and listing all the nodes of interest within the area is virtually impossible. Instead, researchers often include dummy variables at some aggregate level, such as postcode areas, to approximate the location effects. This is obviously a crude approach and could potentially lead to "location biases". In this article, we focus on the use of geospatial data, that is information on the exact location of properties in terms of geographic coordinates, to measure the effect of location.

Not properly accounting for location is likely to result in spatial correlation of house prices, which will impact on the precision of parameter estimates in hedonic models. Spatial correlation can be modeled in various ways, for instance via spatial lags or spatial errors, where a spatial weight matrix is designed to relate the feature of a point in space to the features of neighboring points. Such spatial econometric methods have been applied in time dummy hedonic models to estimate house price indices (Hill et al. 2009; Dorsey et al. 2010). Spatial error modelling has also been combined with state-space house price models which allow the parameters to follow a stochastic process along the time dimension; the price index can then be constructed through imputations (Rambaldi and Rao 2011, 2013). Others have directly extended the spatial filter by including time so that both spatial and temporal correlations are accounted for; these spatiotemporal autoregressive (STAR) models can generate a price index surface (Pace et al. 1998; Tu et al. 2004; Sun et al. 2005).

A disadvantage of the above methods is that the value of location and land is not explicitly modeled. For some purposes, like taxation and national accounting, being able to decompose the property value into land and structures values would be quite useful (Diewert et al. 2015; Rambaldi et al. 2015). In the present article, we attain this decomposition using a simplified version of the so-called builder's model (Diewert et al. 2011, 2015). We further assume that the value of location is capitalized into the price of land but not into the price of structures so that land prices are expected to vary across space whereas the price of structures is 'fixed'. The spatial variation of land prices is estimated by Geographically Weighted Regression (GWR), a nonparametric method proposed by Brunsdon et al. (1996) and Fotheringham et al. (1998b). Combining the land and structures components, we form a semi-parametric house price model and estimate it by Mixed Geographically Weighted Regression (MGWR). The (annual) house price index and its land and structures components are subsequently constructed in an imputation framework.

Our article tries to fill a gap in the *Handbook on Residential Property Price Indices* (Eurostat et al. 2013) in which the use of geospatial data to estimate hedonic house price models is not well covered. Geospatial data has been used before to estimate house price indices using a semi-parametric method. Clapp (2004), for example, estimated the value of location and overall property price change by Local Polynomial Regression (LPR). Our work differs from Clapp's approach in a number of ways. The most important difference is that we incorporate the value of location into land prices and hence are able to construct a land price index, whereas Clapp treats it as an additive term to house value and thus cannot distinguish between land and structures values.

The article proceeds as follows. Section 2 outlines some basic ideas about the hedonic house price model that decomposes the property value into land and structures values and about the inclusion of additional structural characteristics into the model. Section 3 explains how we treat location; the GWR and MGWR approaches will be discussed in detail. Section 4 describes how we calculate the hedonic imputation indices. Section 5 presents empirical evidence for the Dutch city of "A" and discusses the results. Section 6 concludes and identifies some potential improvements.

2. A Simplification of the 'Builder's Model'

2.1. Some Basic Ideas

Our starting point is the 'builder's model' proposed by Diewert et al. (2011, 2015). It is assumed that the value of a property i in period t, p_i^t, can be split into the value of the land $\left(\alpha^t z_{iL}^t\right)$, the value of the structure $\left(\beta^t z_{iS}^t\right)$ and a random error term u_i^t with zero mean:

$$p_i^t = \alpha^t z_{iL}^t + \beta^t z_{iS}^t + u_i^t. \tag{1}$$

The land and structure values are assumed to be proportional to the plot size z_{iL}^t and the size of living space z_{iS}^t, respectively. The shadow prices of land and structures in (1), α^t and β^t, are the same for all properties, irrespective of their location. In Section 3 we relax this assumption and allow for spatial variation in the price of land.

When applying Model (1) to the data of a sample S^t of properties sold in period t, a few problems arise. First, the model has no intercept term, which hampers the interpretation of R^2 and the use of standard tests in Ordinary Least Squares (OLS) regression. Second, a high degree of collinearity between land size and structure size can be expected, so that α^t and β^t will be estimated with low precision. To resolve these drawbacks, Equation (1) is divided by structure size z_{iS}^t, giving

$$p_i^{t^*} = \alpha^t r_i^t + \beta^t + \varepsilon_i^t, \tag{2}$$

where $p_i^{t^*} = p_i^t / z_{iS}^t$ is the *normalized property price*, that is, the value of the property per square meter of living space, $r_i^t = z_{iL}^t / z_{iS}^t$ is the ratio of plot size to structure size, and $\varepsilon_i^t = u_i^t / z_{iS}^t$. The model now has an intercept term and a single explanatory variable. In what follows, we focus on this normalized model.

2.2. Adding Structures Characteristics

Models (1) and (2) only incorporate structure size and plot size, which may lead to omitted variable bias. Here we discuss the inclusion of additional characteristics for the structures by linearizing the method proposed by Diewert et al. (2015).

We first consider the age effect and assume a straight-line depreciation model. The adjusted value of the structure is $\beta^t\left(1-\delta^t a_i^t\right)z_{iS}^t$, where δ^t is the depreciation rate and a_i^t is age of the structure. It is assumed that structure age is available in the data set as an ordinal (categorical) rather than continuous variable. Using multiplicative dummy variables D_{ia}^t that take on the value 1 if in period t property i belongs to age category a ($a=1,\ldots,A$) and 0 otherwise, and after reparameterizing to eliminate the term $\beta^t z_{iS}^t$, the adjusted value of structure can be expressed as $\sum_{a=1}^{A}\gamma_a^t D_{ia}^t z_{iS}^t$, where γ_a^t represents the unit price of a structure belonging to age category a. While using discrete age may be somewhat problematic, it introduces some flexibility in that age dummies will not only reflect depreciation of structure but also capture vintage effect.

When incorporating another attribute, such as the number of rooms, the new value of the structures becomes $\beta^t\left(1-\delta^t a_i^t\right)\left(1+\mu^t z_{iM}^t\right)z_{iS}^t$, where μ^t is the parameter for the number of rooms z_{iM}^t. Using dummies D_{iM}^t for the number of rooms ($m=1,\ldots,M$), and reparameterizing again, the new adjusted value of structure becomes $\sum_{a=1}^{A}\gamma_a^t D_{ia}^t z_{iS}^t+\sum_{m=1}^{M}\lambda_m^t D_{im}^t z_{iS}^t+\sum_{a=1}^{A}\sum_{m=1}^{M}\eta_{am}^t D_{ia}^t D_{im}^t z_{iS}^t$. To save degrees of freedom, we ignore the second-order interaction terms $D_{ia}^t D_{im}^t$ and obtain the *normalized* model

$$p_i^{t^*}=\theta^t+\alpha^t r_i^t+\sum_{a=1}^{A-1}\gamma_a^t D_{ia}^t+\sum_{m=1}^{M-1}\lambda_m^t D_{im}^t+\varepsilon_i^t. \tag{3}$$

In this model, an intercept term θ^t is included by excluding dummy variables for age class A and category M. For a property belonging to age class a ($a=1,\ldots,A-1$) and category m ($m=1,\ldots,M-1$) for number of rooms, the unit price of structures equals $\theta^t+\gamma_a^t+\lambda_m^t$. Additional categorical variables for the structures can be incorporated in a similar way.

3. Land and Spatial Heterogeneity

3.1. Location and the Price of Land

It is widely accepted that the value of location is capitalized into the price of land. In most empirical studies it is assumed that the price of land varies across postcode areas but is the same within each postcode area. An example is Diewert and Shimizu (2013) who estimated the 'builder's model' for Tokyo. Applying the same strategy to postcode dummy variables D_{ik} as was used in Subsection 2.2 for adding structure characteristics, an improved version of Model (3) for the *normalized* property price is

$$p_i^{t^*}=\theta^t+\sum_{k=1}^{K}\alpha_k^t D_{ik}r_i^t+\sum_{a=1}^{A-1}\gamma_a^t D_{ia}^t+\sum_{m=1}^{M-1}\lambda_m^t D_{im}^t+\varepsilon_i^t. \tag{4}$$

Each postcode area now has its own land price α_k^t. This might be still too crude, however, depending of course on the level of detail of the postcode system. A more general version of Model (4) is found by assuming that the price of land can differ at the individual property level, that is, at the micro location. We denote the property-specific land price by α_i^t, yielding

$$p_i^{t^*} = \theta^t + \alpha_i^t r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{m=1}^{M-1} \lambda_m^t D_{im}^t + \varepsilon_i^t. \tag{5}$$

This model obviously cannot be estimated by standard regression techniques. In Subsection 3.2 below, we discuss a semi-parametric approach that enables us to estimate Model (5).

3.2. Mixed Geographically Weighted Regression

The parameters for the structures characteristics (θ^t, γ_a^t, and λ_m^t) in Model (5) are constant across space, whereas the land price $\left(\alpha_i^t\right)$ differs between properties. In other words, we account for spatial heterogeneity, or spatial nonstationarity as it is often referred to Brunsdon et al. (1996), of the price of land. One method that deals with spatial heterogeneity of parameters is the 'expansion method' (Casetti 1972; Jones and Casetti 1992). In our case, the price of land would be viewed as an unknown function of the property's location in terms of latitude x_i and longitude y_i or a similar geographic coordinate system. This function can then be approximated using a Taylor-series expansion of some order; typically, second-order approximations are applied. Although the expansion method makes use of geospatial data, it is basically parametric because it calibrates a prespecified parametric model for the trend of land prices across space (Fotheringham et al. 1998a).

In this article, we adopt a truly nonparametric approach, namely *Geographically Weighted Regression* (GWR), to dealing with spatial heterogeneity of parameters (Brunsdon et al. 1996; Fotheringham et al. 1998b). Let us for a moment ignore the structures characteristics to explain how the property-based land prices can be obtained. Defining $\alpha_i = \alpha(x_i, y_i)$ and using matrix notation, Model (5) without structures characteristics can be written as

$$\mathbf{p}^* = \mathbf{r} \circ \boldsymbol{\alpha} + \boldsymbol{\eta}, \tag{6}$$

where $\mathbf{p}^* = (p_1^*, p_2^*, \cdots, p_n^*)^T$, $\mathbf{r} = (r_1, r_2, \cdots, r_n)^T$, $\boldsymbol{\alpha} = (\alpha(x_1, y_1), \alpha(x_2, y_2), \ldots, \alpha(x_n, y_n))^T$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)^T$, and $\circ$ is an operator that multiplies each element of $\boldsymbol{\alpha}$ by the corresponding element of $\mathbf{r}$. We have dropped the superscript t for convenience; it should be clear that we estimate models for each time period separately. In Model (6), the land price at point i is a realization of the continuous function $\alpha(x,y)$ at that point.

Model (6) can be estimated using a moving kernel window approach, which is essentially a form of Weighted Least Squares (WLS) regression. To obtain an estimate for the price of land $\alpha(x_i, y_i)$ for property i, a WLS regression is run on a subset of properties close to i on the premise that a property j which is closer to property i has a bigger influence in the estimation of $\alpha(x_i, y_i)$. That is

$$\alpha(x_i, y_i) = (\mathbf{r}^T \mathbf{w}(x_i, y_i) \mathbf{r})^{-1} \mathbf{r}^T \mathbf{w}(x_i, y_i) \mathbf{p}^*, \tag{7}$$

where $\mathbf{w}(x_i, y_i) = \mathrm{diag}[w_1(x_i, y_i), w_2(x_i, y_i), \ldots, w_n(x_i, y_i)]$ is an n by n spatial weighting matrix. In this way, we are able to estimate land prices not only for observed properties, but also for any imaginary location within the study area, enabling us to plot a continuous surface of land prices. The predicted values of the house prices are

$$\hat{\mathbf{p}}^* = \mathbf{r} \circ \boldsymbol{\alpha} = \mathbf{s}\mathbf{p}^*, \tag{8}$$

where the so-called hat matrix $\mathbf{s}$ is given by

$$\mathbf{s} = \begin{bmatrix} r_1\left(\mathbf{r}^T\mathbf{w}(x_1, y_1)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_1, y_1) \\ r_2\left(\mathbf{r}^T\mathbf{w}(x_2, y_2)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_2, y_2) \\ \vdots \\ r_n\left(\mathbf{r}^T\mathbf{w}(x_n, y_n)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_n, y_n) \end{bmatrix}.$$

The weights w_{ij} $(i \neq j)$ should follow a monotonic decreasing function of distance between (x_i, y_i) and (x_j, y_j). There is a range of possible functional forms from which we have chosen the frequently-used *bi-square* function

$$w_{ij} = \begin{cases} \left(1 - d_{ij}^2/h^2\right)^2 & \text{if } d_{ij} < h \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where h denotes the bandwidth. The choice of bandwidth involves a trade-off between bias and variance. A larger bandwidth generates an estimate with larger bias but smaller variance whereas a smaller bandwidth produces an estimate with smaller bias but larger variance. The usual solution is to select the optimal bandwidth by minimizing the cross-validation (CV) statistic

$$CV(h) = \sum_{i=1}^{n} \left[p_i^* - \hat{p}_{\neq i}^*(h)\right]^2, \tag{10}$$

where $\hat{p}_{\neq i}(h)$ is the predicted price of property i where the observation for i has been omitted from the calibration process.

The above nonparametric GWR approach to dealing with spatial heterogeneity of land prices has to be extended by including structures characteristics with spatially fixed parameters, as shown in Model (5). This leads to a specific instance of the semi-parametric Mixed GWR (MGWR) approach discussed by Brunsdon et al. (1999), where some parameters are spatially fixed and the remaining parameters are allowed to vary across space. The estimation of the MGWR model is more complicated than that of the GWR model. To outline the estimation procedure, we write Model (5) in matrix notation as

$$\mathbf{p}^* = \mathbf{r} \circ \boldsymbol{\alpha} + \mathbf{D}_S\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{p}^*$, $\mathbf{r}$ and $\boldsymbol{\alpha}$ have the same meaning as in Equation (6), $\mathbf{D}_S$ is the matrix of structures characteristics included in Model (5), given by

$$\mathbf{D}_S = \begin{bmatrix} 1 & D^a_{11} & \cdots & D^a_{1,A-1} & D^m_{11} & \cdots & D^m_{1,M-1} \\ 1 & D^a_{21} & \cdots & D^a_{2,A-1} & D^m_{21} & \cdots & D^m_{2,M-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & D^a_{n1} & \cdots & D^a_{n,A-1} & D^m_{n1} & \cdots & D^m_{n,M-1} \end{bmatrix},$$

and $\boldsymbol{\beta} = (\theta, \gamma_1, \cdots, \gamma_{A-1}, \lambda_1, \cdots, \lambda_{M-1})^T$ is the vector of coefficients relating to $\mathbf{D}_S$ to be estimated.

We use the estimation method proposed by Fotheringham et al. (2002), which is less computationally intensive than the method described in Brunsdon et al. (1999). If the parameters $\boldsymbol{\beta}$ were known, the GWR approach (7) could be used to estimate $\boldsymbol{\alpha}$ by regressing $\mathbf{r}$ on $\mathbf{p}^* - \mathbf{D}_S\boldsymbol{\beta}$. Similarly, OLS estimates of $\boldsymbol{\beta}$ could be obtained by regressing $\mathbf{D}_S$ on $\mathbf{p}^* - \mathbf{r} \circ \boldsymbol{\alpha}$ if the property-based parameters $\boldsymbol{\alpha}$ were known. In practice, a four-step estimation procedure is followed; for details, see Fotheringham et al. (2002), Mei et al. (2006) and Geniaux and Napoléone (2008). This four-step procedure involves:

(1) regressing each column of $\mathbf{D}_S$ against $\mathbf{r}$ using the GWR approach described by (7) and then computing the residuals $\mathbf{Q} = (\mathbf{I} - \mathbf{s})\mathbf{D}_S$,
(2) regressing the dependent variable $\mathbf{p}^*$ against $\mathbf{r}$ using the GWR approach (7) and then computing the residuals $\mathbf{R} = (\mathbf{I} - \mathbf{s})\mathbf{p}^*$,
(3) regressing the residuals $\mathbf{R}$ against the residuals $\mathbf{Q}$ using OLS in order to obtain the estimates $\hat{\boldsymbol{\beta}} = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{R}$,
(4) subtracting $\mathbf{D}_S\hat{\boldsymbol{\beta}}$ from $\mathbf{p}^*$ and regressing this part against $\mathbf{r}$ using GWR approach in (7) to obtain estimates $\hat{\alpha}(x_i, y_i) = [\mathbf{r}^T\mathbf{w}(x_i, y_i)\mathbf{r}]^{-1}\mathbf{r}^T\mathbf{w}(x_i, y_i)(\mathbf{p}^* - \mathbf{D}_S\hat{\boldsymbol{\beta}})$.

The predicted values for the property prices in Equation (11) can be expressed as

$$\hat{\mathbf{p}}^* = \mathbf{s}(\mathbf{p}^* - \mathbf{D}_S\hat{\boldsymbol{\beta}}) + \mathbf{D}_S\hat{\boldsymbol{\beta}} = \mathbf{L}\mathbf{p}^*, \tag{12}$$

with $\mathbf{L} = \mathbf{s} + (\mathbf{I} - \mathbf{s})\mathbf{D}_S\left[\mathbf{D}_S^T(\mathbf{I} - \mathbf{s})^T(\mathbf{I} - \mathbf{s})\mathbf{D}_S\right]^{-1}\mathbf{D}_S^T(\mathbf{I} - \mathbf{s})^T(\mathbf{I} - \mathbf{s})$, which is the hat matrix for Equation (11).

As discussed above, the parameter estimates and the predicted property prices depend on the choice of weights, hence on the choice of bandwidth h. The optimal value for h is determined by minimizing the CV statistic given by (10). In the case of MGWR, the CV statistic is equivalent to (Mei et al. 2006)

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_i^* - \hat{p}_i^*(h)}{1 - l_{ii}(h)}\right]^2, \tag{13}$$

where $\hat{p}_i^*(h)$ is the predicted price for property i and $l_{ii}(h)$ is the ith diagonal element of matrix $\mathbf{L}$ in Equation (12).

4. Hedonic Imputation Price Indices

This section addresses the issue of estimating quality-adjusted property price indices. Suppose sample data is available for periods $t = 0, \ldots, T$, where 0 is the base period (the starting period of the time series we want to construct), and suppose Model (5) has been estimated separately for each period. The predicted property prices are given by $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \left[\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^t\right] z_{iS}^t$. For short, we write the predicted unit price of structures, $\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^t$, as $\hat{\beta}_i^t$ and the predicted overall property price as $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t$ $(t = 0, \ldots, T)$.

We denote the sample of properties sold in the base period by S^0. The hedonic imputation Laspeyres property price index going from period 0 to period t is defined by

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \hat{p}_i^{t(0)}}{\sum_{i \in S^0} \hat{p}_i^0}, \tag{14}$$

Equation (14) may need some explanation. All quantities are equal to 1, reflecting the fact that each property is considered unique. The index is not affected by compositional change because it is based on a single sample. Most, if not all, of the properties sold in period 0 are not resold in period t, and the ‘missing prices’ have to be imputed by $\hat{p}_i^{t(0)}$. We also replaced the observed base period prices p_i^0 by the predicted values $\hat{p}_i^0$, a method known as *double imputation*. Hill and Melser (2008) discussed different types of hedonic imputation methods in the context of housing. For a general discussion of the difference between hedonic imputation indices and time dummy indices, see Diewert et al. (2009) and de Haan (2010).

The $\hat{p}_i^{t(0)}$ are estimated period t constant-quality property prices, that is, estimates of the prices that would prevail in period t for properties sold in period 0 if the properties’ price-determining characteristics were equal to those of the base period, which serves to adjust for quality changes of the individual properties. These constant-quality prices are estimated by $\hat{p}_i^{t(0)} = \hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0$, where $\hat{\beta}_i^{t(0)} = \hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^0 + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^0$ denotes the estimated constant-quality price of structures.

Substitution of $\hat{p}_i^0 = \hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0$ and $\hat{p}_i^{t(0)} = \hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0$ into (14) yields

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \left[\hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0\right]}{\sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]} = \hat{s}_L^0 \frac{\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0}{\sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0} + \hat{s}_S^0 \frac{\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0}{\sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0}, \tag{15}$$

where $\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0 / \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0$ is a price index of land and $\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0 / \sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0$ is a price index of structures. Equation (15) decomposes the overall house price index into structures and land components; the weights $\hat{s}_L^0 = \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0 / \sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]$ and $\hat{s}_S^0 = \sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0 / \sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]$ are estimated shares of land and structures in the total value of property sales in period 0. The double imputation method ensures that the weights sum to unity.

An alternative to the Laspeyres index is the hedonic double imputation Paasche price index, defined on the sample S^t of properties sold in period t $(t = 1, \dots, T)$:

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \hat{p}^t_i}{\sum_{i \in S^t} \hat{p}^{0(t)}_i}. \tag{16}$$

The imputed constant-quality prices $\hat{p}^{0(t)}_i$ are estimates of the prices that would prevail in period 0 if the property characteristics were those of period t, which are estimated as $\hat{p}^{0(t)}_i = \hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}$, where $\hat{\beta}^{0(t)}_i = \hat{\theta}^0 + \sum_{a=1}^{A-1} \hat{\gamma}^0_a D^t_{ia} + \sum_{m=1}^{M-1} \hat{\lambda}^0_m D^t_{im}$ denotes the period 0 constant-quality price of structures. By substituting the constant-quality prices and the predicted prices $\hat{p}^t_i = \hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS}$ into (16), the hedonic imputation Paasche index can be written as

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \left[\hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS}\right]}{\sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]} = \hat{s}^{t(0)}_L \frac{\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL}}{\sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}} + \hat{s}^{t(0)}_S \frac{\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS}}{\sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}}, \tag{17}$$

where $\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL} / \sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}$ and $\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS} / \sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}$ are Paasche price indices of land and structures, which are weighted by $\hat{s}^{t(0)}_L = \sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL} / \sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]$ and $\hat{s}^{t(0)}_S = \sum_{i \in S^t} \hat{\beta}^0_i z^t_{iS} / \sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]$. The weights are now of a hybrid nature and reflect the shares of land and structures in the estimated total value of property sales in period t, evaluated at base period prices.

A drawback of the above indices is that they are based on the sample of either the base period or the comparison period t, but not on both samples. When constructing an index going from 0 to t, the sales in both periods should ideally be taken into account in a symmetric fashion. The double imputation Fisher price index

$$P^{0t}_{Fisher} = \left[P^{0t}_{Laspeyres} \times P^{0t}_{Paasche}\right]^{\frac{1}{2}} \tag{18}$$

does so by taking the geometric mean of the Laspeyres and Paasche price indices. The Fisher index formula is not consistent in aggregation, which means that decomposing the Fisher property index into structures and land components like Equation (15) and (17) is not possible. In other words, the Fisher property index can only be derived directly from house price relatives, but not from aggregating the Fisher structures index and land index, whereas the Laspeyres and Paasche indices can be obtained in both ways.

Double imputation Laspeyres, Paasche, and Fisher property price indices and the land price indices based on the more restrictive hedonic Models (4) or (3) are found by replacing $\hat{\alpha}^0_i$ and $\hat{\alpha}^t_i$ in (15) and (17) by the corresponding postcode-specific estimates $\hat{\alpha}^0_k$ and $\hat{\alpha}^t_k$ or the city-wide estimates $\hat{\alpha}^0$ and $\hat{\alpha}^t$. In the latter case, the estimated land price index of course equals $\hat{\alpha}^t / \hat{\alpha}^0$, irrespective of the index number formula used.

5. Empirical Evidence

5.1. The Data Set

The data set we utilize was provided by the Dutch Association of Real Estate Agents. It contains residential property sales for a small city (population is around 60,000) in the northeastern part of the Netherlands, the city of "A", and covers the first quarter of 1998 to the fourth quarter of 2007. Statistics Netherlands has geocoded the data. We excluded sales of condominiums and apartments as the treatment of land deserves special attention in this case. The resulting total number of sales in the data set during the ten-year period is 6,058, representing approximately 75 per cent of all residential property transactions in "A".

Our data set contains information on time of sale, transaction price, a range of structures characteristics, and land characteristics. We included only three structures characteristics in our models, that is, usable floor space, type of house and building period. Note that, because a sample period of ten years is relatively short and building period is available in decades only, we decided to use building period in the models rather than approximate age of the structures (from building period in decades and time of sale). For land, we used plot size and postcode or latitude/longitude. We deleted 43 observations with missing values or prices below EUR 10,000, properties with more than ten rooms and those with ratios of plot size to structure size (usable floor space) larger than ten as well as transactions in rural areas. Finally, we removed 32 outliers or influential observations detected by Cook's distance and were left with 5,983 observations during the sample period.

Table A1 in the Appendix reports summary statistics by year for the numerical variables. The average transaction price and the price per square meter of floor space increased significantly from 1998 to 2007. Average land size and usable floor space were quite stable over time. The urban area of the city of "A" seems to have expanded along the east-west axis; the standard deviation of the x coordinate in later years is generally much larger than that in earlier years.

5.2. Estimation Results for Hedonic Models

Given the small size of the city of "A" and the resulting low number of observations, we decided to use annual rather than biannual or quarterly data. We estimated three normalized hedonic models: Model (3), which does not include location and has a fixed land price across the city (denoted by FLP), Model (4) with nine postcode dummy variables, hence with postcode-varying land prices (PCLP), and Model (5) with location-varying land prices (LLP).

The FLP and PCLP models were estimated by OLS, while the MGWR approach described in Subsection 3.2 was used to estimate the LLP model. When applying the MGWR approach, a key point is the selection of the bandwidth in Equation (9) to decide which neighboring transactions will be used in the estimation of the land price for a specific property. Given that the transactions in our data set are not evenly distributed across space, using transactions within a certain distance may not be good practice: properties located in the densely-populated area will have many neighbors while other properties will have only few. We therefore constructed the weighting scheme using the *adaptive bi-square* function where the bandwidth relates to a fixed number N of nearest neighbors that are used in the estimation process. When computing the weights given by Equation (9), h equals the distance to the Nth nearest neighbor and changes with the target properties. In practice, the choice of N nearest neighbors is equivalent to the choice of window size, that is the fraction of the sample used. To find the optimal value, we varied the window size from ten per cent to 95 per cent using a five per cent interval and selected the size that yielded the lowest CV score as given by Equation (13). Each annual sample then has a unique optimal window size. The CV scores indicated that a ten per cent window size was preferred for most of the years, except for 1999, 2000, and 2002, with an optimal size of 15 per cent, and for 2003, with an optimal size of 30 per cent. However, for the construction of price indices we prefer the same window size for all years, in particular because the number of sales is almost evenly spread across the whole period. So we chose a window size of ten per cent for each year, leading to 60 nearest neighbors that were used in the estimation of the LLP models.

As an illustration, Table 1 contains the 2007 parameter estimates for the structures characteristics. Almost all of the estimates differ significantly from zero at the one per cent level. To some extent they vary across the different models. For example, the FLP intercept term is relatively high compared to the PCLP and LLP intercepts. Since dummy variables for houses built after 2000 and for detached houses were not included, the intercept measures the price in euros of structures per square meter of living space for detached houses built after 2000. In accordance with a priori expectations, detached dwellings are more expensive than other types of houses. For all models, there is a clear tendency for the structures to become less expensive as they are older.

Table 2 contains summary statistics for the estimated price per square meter of land from the three models. The three average land price series exhibit a similar pattern over time, which differs substantially from the changes in the average transaction price of the properties (see Table A1 in the Appendix). After a sharp increase in 1999, the estimated average land price fluctuated during a couple of years, experienced a dramatic drop in 2003, and then increased again.

As mentioned earlier, a virtue of MGWR approach is that it allows us to plot a continuous map with estimated prices of land per square meter. To produce the map, we first divided the city of “A” into 50 (meters) × 50 (meters) grids and retrieved the coordinates of each cell, and then estimated the unit land price of each grid based on their coordinates. For the year 2007, such a map is depicted in Figure 1, where the land prices were rescaled to the range [0, 1]. The postcode areas are indicated as well. While the spatial pattern in Figure 1 is largely consistent with the pattern found using the PCLP model (shown in Figure A1 in the appendix), the land prices estimates from the LLP model do vary within some of the postcode areas. This suggests that the use of postcode dummies, as in the PCLP model, is a rather crude strategy to incorporate spatial variation of land prices.

To formally compare the performance of the three hedonic models, two statistics were calculated, the Corrected Akaike Information Criterion (AICc) and the Root Mean Square Error (RMSE). The AICc takes into account the trade-off between goodness of fit and degrees of freedom. The AICc expressions for the FLP and PCLP models can be found in

Table 1. Parameter estimates for structures characteristics, 2007.

	FLP	PCLP	LLP
Intercept	1480.70*** (46.93)	1405.41*** (53.71)	1395.76*** (57.51)
Building period: 1960–1970	−370.48*** (25.94)	−389.50*** (36.67)	−398.40*** (41.75)
Building period: 1971–1980	−311.17*** (23.36)	−261.50*** (33.96)	−323.50*** (41.69)
Building period: 1981–1990	−232.93*** (23.37)	−173.08*** (32.59)	−226.14*** (42.87)
Building period: 1991–2000	−58.64*** (21.64)	−49.34* (26.55)	−115.13*** (37.26)
Terrace	−285.65*** (35.17)	−264.34*** (35.24)	−187.28*** (37.32)
Corner	−281.36*** (31.77)	−274.54*** (31.18)	−192.85*** (34.07)
Semidetached	−122.89** (47.96)	−149.50*** (47.57)	−96.93** (48.73)
Duplex	−151.08*** (30.60)	−147.24*** (30.17)	−104.56*** (31.03)

Notes: Model FLP and PCLP are estimated by OLS, while model LLP is estimated using the MGWR approach. Standard errors are reported in parentheses; ***, ** and * denote significance at the 1%, 5% and 10% level, respectively.

Table 2. Summary statistics for estimated land prices.

		PCLP		LLP			
	FLP	Mean	S.D.	Max	Median	Mean	S.D.
1998	116.80	131.50	31.14	231.03	122.66	125.49	28.66
1999	154.64	178.50	34.85	223.66	174.07	167.77	30.39
2000	239.77	239.41	36.24	319.32	251.34	241.83	44.27
2001	214.54	235.58	47.59	295.01	229.52	226.70	48.77
2002	234.77	245.11	38.41	323.63	255.05	242.23	40.89
2003	166.07	185.11	44.23	248.23	179.93	172.26	44.55
2004	186.40	197.19	29.75	254.20	197.70	195.41	33.78
2005	226.13	224.11	36.55	299.74	214.19	205.89	35.17
2006	202.84	195.77	30.85	274.24	207.43	201.27	32.05
2007	214.87	236.73	27.96	286.91	235.07	229.25	30.99

Notes: For FLP, the land price estimates are reported. For PCLP, the columns show the weighted mean and standard deviation of the estimated land prices for 9 postcode areas where the weights are equal to the share of transactions within each postcode area. For LLP, the columns provide summary statistics for the land price estimates of all transacted properties.

Hurvich and Tsai (1989); for the LLP model, it is defined by

$$AICc = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left(\frac{n + tr(\mathbf{L})}{n - 2 - tr(\mathbf{L})}\right),$$

where $\hat{\sigma}$ is the estimated standard deviation of the error term and $tr(\mathbf{L})$ the trace of the hat matrix described in Subsection 3.2. The RMSE measures the variability of the absolute prediction errors of the models and is given by

$$RMSE = \frac{1}{n}\sqrt{\sum_i (p_i - \hat{p}_i)^2}.$$

Table 3 shows the AICc and RMSE and their differences for the three models. A rule of thumb states that if the difference in the AICc for two models is larger than three, a significant difference exists in their performance (Fotheringham et al. 2002). It can be seen that the PCLP model performs much better than the FLP model in all years, as we would expect, and in turn that the LLP model outperforms the PCLP model (except for 2003, when the difference is insignificant). The same ranking is found if the RMSE is used to assess the various models. These results confirm the earlier finding that land prices vary across space, both between and within postcode areas.

Although LLP is obviously better suited to model the variation of land prices and to predict property prices, the PCLP model does a good job too. In several years, for example in 1998, 1999 and 2003, the inclusion of postcode dummy variables accounts for the major part of the variance in overall property prices, almost as much as the LLP model does. This does not come as a surprise though, given that the MGWR approach used for estimating the land price of a particular property in the LLP model utilizes the information of neighboring properties, most of which are likely to be located in the same postcode area.

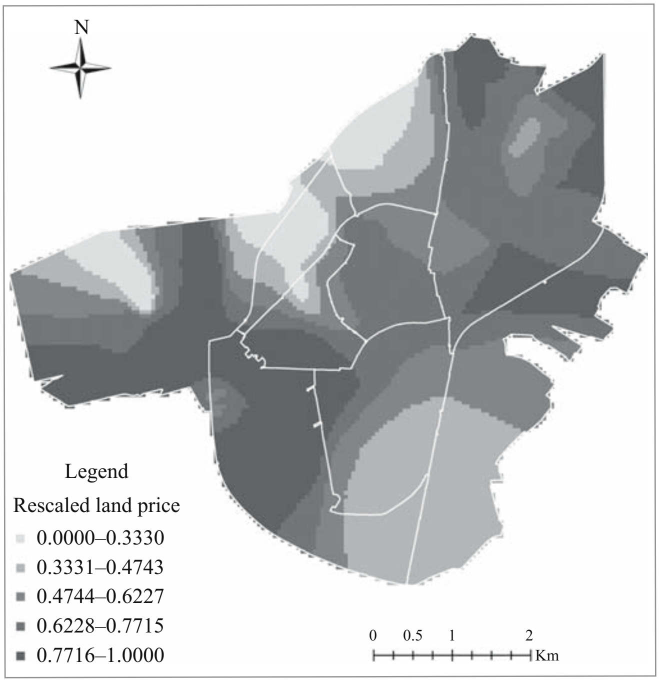


Fig. 1. Price of land per square meter, 2007.

5.3. Hedonic Imputation Price Indices

Changes in average property prices, and changes in their land and structure components, are affected by compositional change in the traded properties. Our hedonic house price indices and the land and structures components control for this. We estimated chained rather than direct price indices because the value shares of the land and structures will then be updated at an annual frequency. A drawback of chaining is that the resulting indices cannot be exactly decomposed because they are not consistent in aggregation.

In Figures 2–4, the estimated double imputation hedonic Laspeyres, Paasche, and Fisher price indices for the overall property are plotted, based on the three models (FLP, PCLP, and LLP). A comparison of Figures 2 and 3 shows that, for each model, the chained Laspeyres index sits above the Paasche index, as expected. The Laspeyres and Paasche indices based on PCLP and LLP are very similar; for the Laspeyres index, the difference can even hardly be noticed. This result is in accordance with our previous finding that the PCLP model captures the spatial variation of land prices reasonably well.

Not using location information at all does make a difference, at least for the Laspeyres and Paasche house price indices. The FLP-based Laspeyres and Paasche indices seem to be biased downwards and upwards, respectively. However, the biases almost cancel out in the Fisher indices, as can be seen in Figure 4: the FLP-based Fisher index is very similar to the PCLP-based and LLP-based Fisher indices. In other words, the hedonic imputation Fisher house price index is insensitive to the treatment of location in the hedonic model, which is a surprising result.

Table 3. Model comparison.

	FLP		PCLP				LLP					
	AICc	RMSE	AICc	ΔAIC_{PF}	RMSE	$\Delta RMSE_{PF}$	AICc	ΔAIC_{LP}	ΔAIC_{LF}	RMSE	$\Delta RMSE_{LP}$	$\Delta RMSE_{LF}$
1998	6487.71	91.32	6372.45	−115.26	80.89	−10.43	6366.21	−6.24	−121.50	77.85	−3.04	−13.47
1999	7056.56	146.82	6990.52	−66.04	136.14	−10.68	6982.93	−7.59	−73.63	131.28	−4.86	−15.54
2000	7216.89	151.11	7164.26	−52.63	142.00	−9.11	7127.51	−36.75	−89.38	133.30	−8.70	−17.81
2001	7380.41	147.00	7294.10	−86.31	134.36	−12.64	7279.66	−14.44	−100.75	128.87	−5.49	−18.13
2002	7718.63	152.44	7643.97	−74.66	141.19	−11.25	7632.34	−11.63	−86.29	135.62	−5.57	−16.82
2003	7769.06	159.02	7702.07	−66.99	148.23	−10.79	7701.91	−0.16	−67.15	143.93	−4.30	−15.09
2004	7968.62	159.66	7947.61	−21.01	154.80	−4.86	7927.92	−19.69	−40.70	147.91	−6.89	−11.75
2005	8060.84	161.52	7993.88	−66.96	150.93	−10.59	7984.11	−9.77	−76.73	145.10	−5.83	−16.42
2006	8597.81	175.46	8565.36	−32.45	168.94	−6.52	8517.73	−47.63	−80.08	157.67	−11.27	−17.79
2007	9006.25	177.18	8960.58	−45.67	169.24	−7.94	8929.11	−31.47	−77.14	159.98	−9.26	−17.20

Note: ΔAIC_{PF} is equal to AICc for PCLP minus AICc for FLP; ΔAIC_{LP} and ΔAIC_{LF} are equal to AICc for LLP minus AICc for PCLP and FLP, respectively; $\Delta RMSE_{PF}$, $\Delta RMSE_{LP}$ and $\Delta RMSE_{LF}$ have a similar meaning.

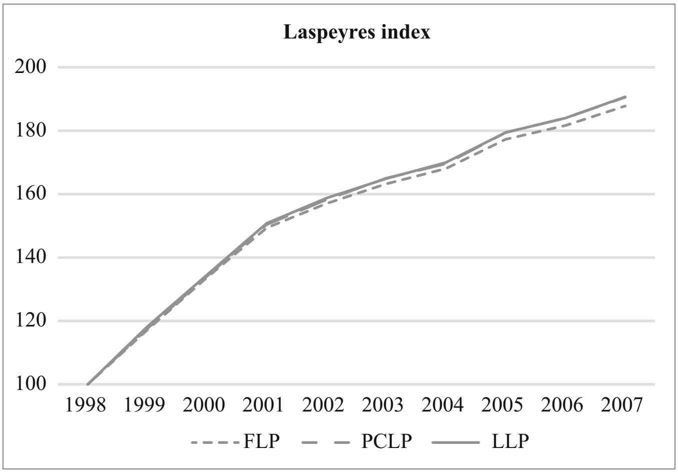


Fig. 2. Hedonic imputation Laspeyres house price index.

Figure 5 plots the Fisher price indices for land. The PCLP- and LLP-based indices, which explicitly account for location, are similar, though the LLP-based index is less volatile, at least during 2003–2007. The FLP-based index seems to be significantly upward biased. For example, between 1999 and 2000 as well as between 2003 and 2005, the FLP-based index rises much faster than the other two indices. A possible explanation is the following. Suppose specific locational attributes improved over time or that consumers' preferences changed towards locations with specific characteristics. This will have caused land prices in some areas to appreciate significantly relative to other areas. If, as in the FLP model, such heterogeneity is not accounted for, bias in the average estimated land price is likely to occur. The treatment of location in the FLP model may not only have produced biased levels of land prices, it might easily have led to a biased trend as well.

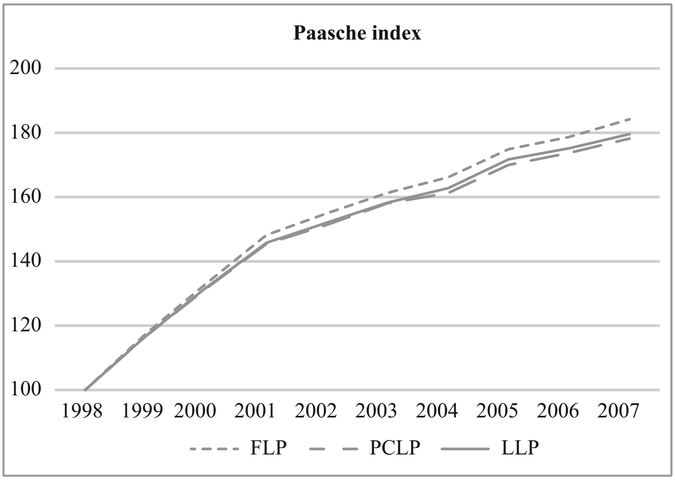


Fig. 3. Hedonic imputation Paasche house price index.

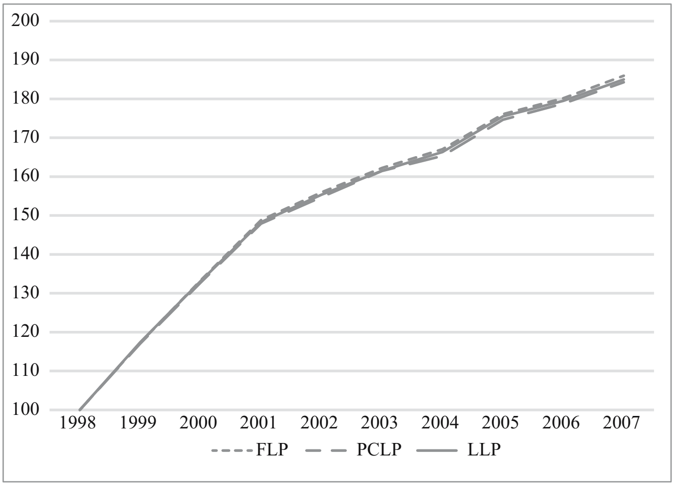


Fig. 4. Hedonic imputation Fisher house price index.

Figure 6 shows the Fisher price indices for structures based on the three models. Again, the PCLP-based and LLP-based indices are similar. The FLP-based index sits below these two indices, which is not surprising given the above results for land. Since the hedonic model in this paper leaves out many structural characteristics, which may be correlated with location, the decomposition of house price index is not strictly orthogonal. In this sense, upward bias in estimated land prices using the FLP model is therefore likely to result in downward bias in structures prices.

Figure 7 shows the LLP-based value share estimates for both structures and land. Prior to 2003, these shares are quite volatile, but from 2003 on they remain fairly constant. The average estimated shares for structures and land across the entire sample period are 0.67 and 0.33. The FLP- and PCLP-based shares exhibit similar patterns and levels; the value

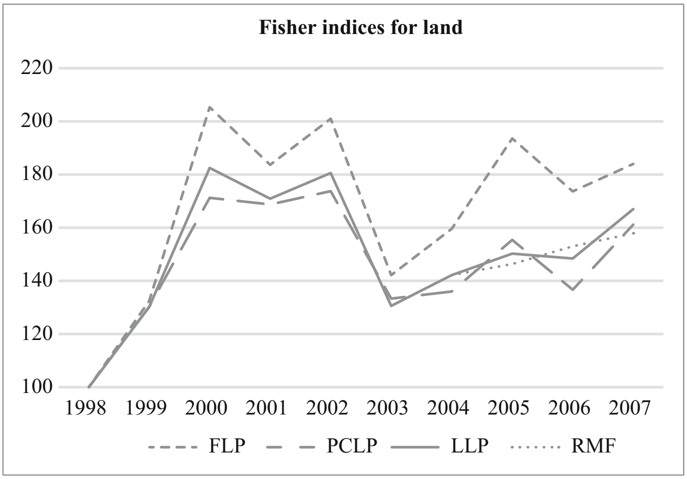


Fig. 5. Hedonic imputation Fisher price indices for land.

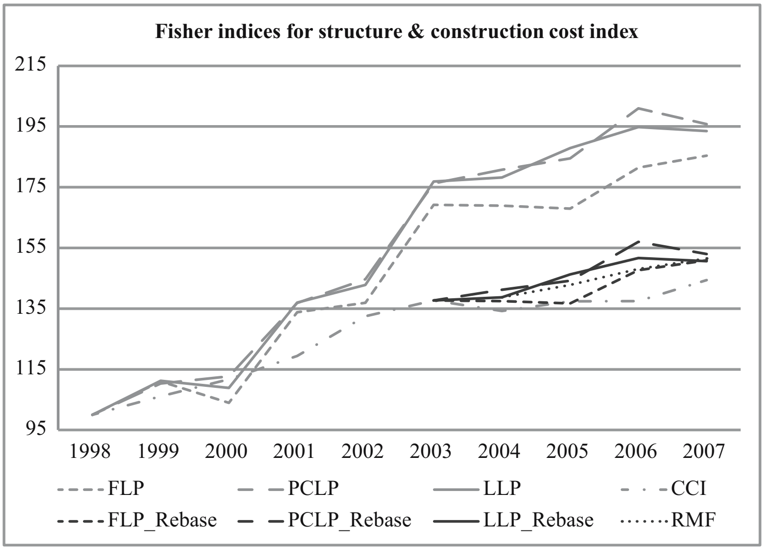


Fig. 6. Hedonic imputation Fisher price indices for structures and official construction cost index.

shares for structures are 0.68 and 0.66, respectively, hence for land 0.32 and 0.34. Given that the estimated value share of structures is twice as large as that of land, overall house price indices are affected most by changes in structures prices. Yet, combining Figures 4, 5, 6, and 7 suggests that the increase in house prices between 1998 and 2001 was driven mainly by the increase of land prices: both the (average) price of land and its value share show a sharp increase.

5.4. Discussion

Figures 5, 6, and 7 raise a number of issues. The first issue is the volatility of the land and structures price indices. Volatile series can be expected with sparse data (without

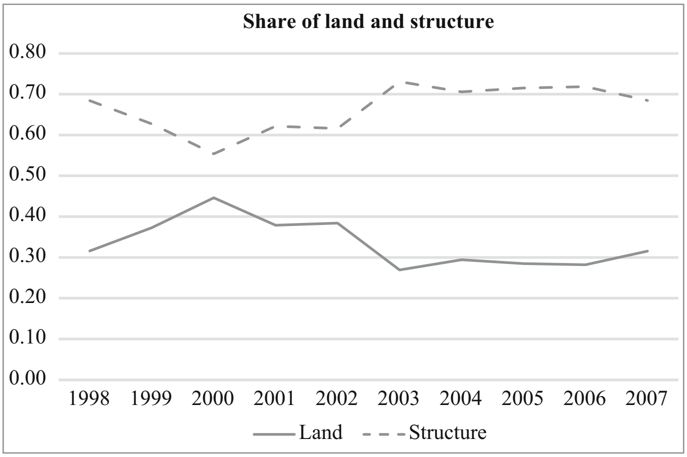


Fig. 7. Estimated value shares of land and structures, LLP model.

smoothing). Another potential cause is multicollinearity. Diewert et al. (2015) found that multicollinearity between land and structure size led to price changes for land and structures which consistently had opposite signs. To deal with this type of multicollinearity, they incorporated exogenous information in the hedonic models; see also (e.g., Diewert et al. 2009; Diewert and Shimizu 2013; Francke and van de Minne 2017). More specifically, their (final) models did not endogenously determine a price index of structures but used the published construction cost index as the measure of structures price change. We did not follow their approach for two reasons: an endogenously estimated trend in the price of structures does not necessarily have to be equal to that of construction costs, and multicollinearity does not seem to be the most important issue.

In Figure 8, the LLP-based Fisher price indices for land and structures from Figures 5 and 6 are copied. In some years, for example in 2003 when the land price index suddenly falls and starts to sit below the structures price index, the price changes for land and structures have opposite signs, but in other years the price changes are in the same direction. The variance inflation factor (VIF) for the ratio of plot size to structure size did not point to significant multicollinearity either. Further, there is a considerable amount of variation in these ratios in our data set; see Table A1. We therefore suspect that multicollinearity is not the main issue.

The second issue is whether the trends of the (Fisher) price indices for land and structures are plausible. For land, this can hardly be checked since information on the price change of land covering our sample period is not available for the Netherlands. Rambaldi et al. (2015), using an unobserved component approach, estimated an endogenous monthly land price index for the city of "A" from August 2003 to June 2008, denoted by RMF index. We converted their series into an annual series by averaging the monthly indices, rebased the resulting index to 2004, and then spliced it on to the LLP land price index for 2004 (see Figure 5). Our LLP hedonic land price index in 2005, 2006, and 2007 is very similar to the RMF index, which is reassuring, except that the latter index is smoother.

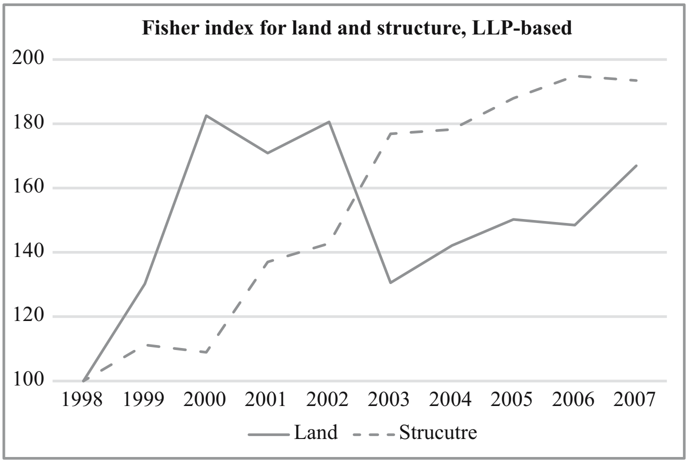


Fig. 8. Chained Fisher price indices for land and structures, LLP model.

For structures we use the nationwide construction cost index (CCI) for new dwellings published by Statistics Netherlands as a benchmark. This price index, rebased to 1998=100, is shown in Figure 6 as well. Our hedonic structures price indices appear to rise much faster than the construction cost index. As mentioned above, a construction cost index does not necessarily have to coincide with an implicit price index for structures derived from a hedonic model. In a competitive market, where developers also have sufficient time to meet demand, construction cost is believed to be equal to the market value of the structure (Davis and Heathcote 2007; Davis and Palumbo 2008). However, in reality the market is characterized by restrictions on new construction and high costs of replacing old structures by new ones. In this case, a markup on construction costs can be expected. During a housing boom, like our study period, the mark up may well be growing over time. Kuminoff and Pope (2013), who estimated land values for US metropolitan areas using a hedonic approach, indeed found that in some (though not all) areas the increase in the market value of the structures exceeded the increase in replacement costs in the booming period.

Omitted variables bias, resulting in quality-change bias, may have played a role as well. We included only a few structures characteristics in the hedonic models. Unless they would be highly collinear with included variables, adding characteristics will lead to better quality adjustment for structures and lower the hedonic price indices for structures if the average quality of structures improved over time.

Importantly, the major part of the differences between our hedonic indices and the construction cost index stems from a big increase in our indices in 2003; as of 2003, the deviation is relatively small. We reproduced the RMF price index for structures estimated by Rambaldi et al. (2015) in Figure 6 and, as was the case for land, their index is very similar to our LLP structures price index in 2005, 2006, and 2007.

The sudden increase in estimated structures prices and drop in estimated land prices in 2003 are worth examining in more detail. At first glance, sample selection bias might matter, for example if the spatial distribution of transacted properties in 2003 was very different from that in other years, or if unique properties, like properties with a very large of plot size to structure size ratios, were transacted in 2003. However, after a careful check of the data, we exclude this possibility. It could be that the 2003 results are "real" in the sense that a shock affected households' decision-making in the Dutch housing market or perhaps in the local market of "A". This is quite plausible given that the house price appreciation rate suddenly dropped from above ten per cent to around four per cent at the time around 2002 or 2003. But it is not clear to us what that shock might have been.

The third issue concerns the low share of land in the value of properties sold, which was estimated at roughly one third across the sample period. Rambaldi et al. (2015) estimated the land value share for the city of "A" during the period 2003–2008 between 0.30 and 0.40. van de Minne and Francke (2012) estimated the share of land for properties (excluding apartments/condominiums) sold during 2003–2010 in the Dutch city of 's Hertogenbosch at 0.39 on average. In a follow-up study (Francke and van de Minne 2017), where they made a distinction between the part of the land plot that the structure sits on and the part used as gardens, the estimate was almost 0.50. It is not unreasonable to find that the value share of land for the city of "A" is lower than that for 's Hertogenbosch. The city of "A" lies in a less prosperous part of the Netherlands with fewer amenities, and we expect this to have a downward effect on the price of land but not on the price of structures, hence on the value share of land.

De Groot et al. (2015), who also used hedonic modeling to decompose property values into land and structures components, estimated the price of land for most Dutch cities, though unfortunately not for "A". They found substantial cross-city differences. For example, the price per square meter of land in 2005 was estimated at EUR 717 for the capital city of Amsterdam, EUR 308 for 's-Hertogenbosch, and EUR 184 for Leeuwarden. Like "A", Leeuwarden is a city in the northeastern part of the Netherlands but bigger. In light of their findings, our MGWR estimates of the average price of land for the city of "A", EUR 206 in 2005 (Table 2), and the value share of land are not surprisingly low after all.

6. Summary and Conclusions

Hedonic house price models used for constructing house price indices usually do not explicitly model the value of land. In the present article, we assumed that the value of location is capitalized into land and attempted to account for spatial variation of land prices in the construction of hedonic imputation house price indices. We linearized the 'builder's model' proposed by Diewert et al. (2015), allowed the price of land to vary across individual properties, and estimated the model for the normalized property price (the property price per square meter of living space) by MGWR, a semi-parametric method, on annual data for the Dutch city of "A". We then constructed chained imputation Laspeyres, Paasche and Fisher indices, and compared these indices with price indices based on more restrictive models, that is a model where land prices vary across postcode areas and a model with no variation in land prices, both estimated by OLS.

The Fisher house price indices were quite insensitive to the choice of model, but the Laspeyres and Paasche indices for the 'fixed' land price model differed from those for the models where location was explicitly included. The use of postcode area dummy variables produced price indices very similar to indices obtained by MGWR. Hill and Scholz (2017), who treated location as a 'separate characteristic' in their hedonic models in that they estimated property-specific shift terms for the overall property price, also concluded that the use of geocoded information did not significantly improve hedonic imputation house price indices compared to indices based on models with postcode dummy variables. This result is reassuring for statistical agencies that do not have the expertise or resources to apply more sophisticated methods. It should be noted that the similarity between PCLP-based and LLP-based house price indices could also be due to the small size and homogeneity of the city "A" where relatively little variation of land prices can be expected.

Apart from being able to capture spatial variation of land prices at the property level, the MGWR-based LLP model has two additional advantages. A potential problem with the PCLP model is that if a large number of postcode areas are distinguished, observations in some areas may not be available, leading to difficulties in the construction of hedonic imputation price indices. The LLP model deals with this problem by using data of the nearest neighbors which are not necessarily confined to a particular postcode area. Most importantly, The LLP model can generate a continuous map of land prices for a city, which will be more informative than a discrete map that only shows differences between postcode areas.

For some purposes, separate price indices for land and structures are needed. As was demonstrated already by Diewert et al. (2015), the decomposition into land and structures using hedonic modeling is not straightforward and raises several statistical and functional form issues. First, our LLP-based price indices of land and structures for the city of "A" are a bit volatile, compared to indices produced by smoothing methods such as the unobserved component approach (Rambaldi et al. 2015). The volatility may be due to sparse data and also to multicollinearity (though we believe the latter is less important). Second, the structures price index increases faster than the official construction cost index, perhaps due a failure to fully control for changes in structures characteristics. Third, the estimated large drop in land prices and increase in structures prices in 2003 seems a bit unusual. While these results could be caused by methodological issues, they could also reflect the impact of a housing market shock which affected households' preferences. Finally, at first glance, the estimated value share of land seems to be rather low. The above-mentioned issues may have played a role here, but the low land share could also be a real phenomenon: households may not value a square meter of land in the city of "A" as much as they do in more prosperous cities with more and better amenities. In future work it would be useful to re-examine our models and compare the results for the city of "A" with those for bigger cities in the western part of the Netherlands, like Amsterdam, Rotterdam or The Hague. Having more observations might also enable us to estimate biannual or even quarterly price indices.

We did not address functional form problems. The original 'builder's model' is nonlinear, in particular due to the treatment of net depreciation. We linearized the model, which basically means we ignored interaction terms, and replaced age by building period in the empirical estimation. Another potential type of misspecification arises from the linear relation between land price and plot size in our models. As Diewert et al. (2015), Francke and van de Minne (2017) and others have argued, the marginal price of land tends to decrease with plot size. Diewert et al. (2015) accounted for this form of nonlinearity by using linear splines. In future work we may modify our normalized models by using linear splines as well and estimate different parameters for the plot size to structure size ratio for different categories of lot size or by explicitly specifying some nonlinear function of this ratio. Furthermore, it would be useful to explicitly allow for net depreciation, as in the original models.

Appendix

Table A1. Summary statistics by year.

	Total	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
# of obs.	5983	545	549	559	574	597	597	612	618	651	681
Transaction price (EUR)											
Mean	157073.87	95124.15	117936.77	131907.96	144672.16	151363.75	162956.98	174998.71	180882.00	191491.09	198546.51
S.D.	72782.29	40240.34	53569.32	54793.53	58064.72	53220.31	63278.10	82975.61	68777.60	76120.61	83639.92
Standardized price (EUR/m^2)											
Mean	1232.38	742.30	930.70	1039.71	1168.13	1240.63	1287.24	1353.89	1420.07	1469.62	1518.50
S.D.	374.83	206.31	273.33	279.98	293.14	285.56	285.87	296.73	294.31	321.20	348.89
Lot size (m^2)											
Mean	251.57	234.08	259.73	242.23	239.68	239.20	250.46	261.38	248.93	263.15	270.98
S.D.	148.16	135.05	169.59	132.98	120.00	115.39	145.76	163.19	136.00	149.26	187.52
Floor space (m^2)											
Mean	125.87	126.00	125.42	126.48	123.34	122.05	125.29	126.57	125.89	128.52	128.39
S.D.	30.61	23.59	31.99	31.97	29.59	28.16	29.87	36.90	30.29	31.14	30.09
Ratio of lot size to floor space											
Mean	1.96	1.81	2.04	1.89	1.93	1.97	1.96	2.01	1.93	2.01	2.04
S.D.	0.82	0.77	0.99	0.72	0.72	0.80	0.84	0.80	0.72	0.78	0.95
x-coordinate											
Mean	233733.81	233972.85	234200.97	234180.34	233948.97	234007.39	233624.00	233480.63	233519.69	233222.34	233385.19
S.D.	1796.35	1453.72	1427.35	1551.87	1716.67	1713.60	1794.99	1984.82	1927.09	1918.80	1948.29
y-coordinate											
Mean	558597.10	558739.46	558805.54	558830.14	558660.23	558721.99	558522.02	558397.61	558549.11	558429.21	558410.25
S.D.	1414.88	1436.14	1463.14	1428.62	1424.92	1410.80	1451.63	1413.94	1354.34	1322.63	1381.24

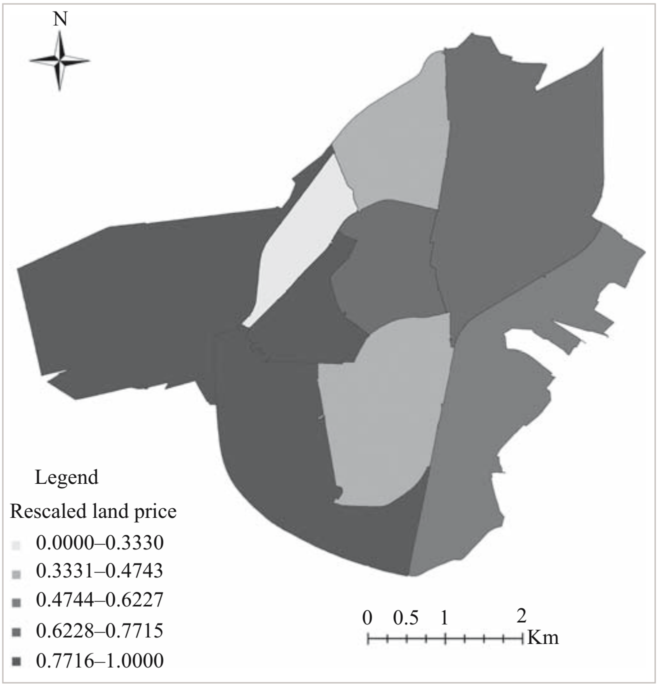


Fig. A1. Price of land per square meter, 2007, PCLP model.

7. References

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28: 281–298. DOI: http://dx.doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1999. "Some Notes on Parametric Significance Tests for Geographically Weighted Regression." *Journal of Regional Science* 39: 497–524. DOI: http://dx.doi.org/10.1111/0022-4146.00146.

Casetti, E. 1972. "Generating Models by the Expansion Method: Applications to Geographical Research." *Geographical Analysis* 4: 81–91. DOI: http://dx.doi.org/10.1111/j.1538-4632.1972.tb00458.x.

Clapp, J.M. 2004. "A Semiparametric Method for Estimating Local House Price Indices." *Real Estate Economics* 32: 127–160. DOI: http://dx.doi.org/10.1111/j.1080-8620.2004.00086.x.

Davis, M.A. and J. Heathcote. 2007. "The Price and Quantity of Residential Land in the United States." *Journal of Monetary Economics* 54: 2595–2620. DOI: https://doi.org/10.1016/j.jmoneco.2007.06.023.

Davis, M.A. and M.G. Palumbo. 2008. "The Price of Residential Land in Large US Cities." *Journal of Urban Economics* 63: 352–384. DOI: https://doi.org/10.1016/j.jue.2007.02.003.

de Groot, H.L.F., G. Marlet, C. Teulings, and W. Vermeulen. 2015. *Cities and the Urban Land Premium*. Cheltenham: Eaward Elgar.

de Haan, J. 2010. “Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and ‘Re-Pricing’ Methods.” *Jahrbücher für Nationalökonomie und Statistik* 230: 772–791.

Diewert, W.E., J. de Haan, and R. Hendriks. 2011. “The Decomposition of a House Price Index into Land and Structures Components: A Hedonic Regression Approach.” *The Valuation Journal* 6: 58–105.

Diewert, W.E., J. de Haan, and R. Hendriks. 2015. “Hedonic Regressions and the Decomposition of a House Price Index into Land and Structure Components.” *Econometric Reviews* 34: 106–126. DOI: http://dx.doi.org/10.1080/07474938.2014.944791.

Diewert, W.E., S. Heravi, and M. Silver. 2009. “Hedonic Imputation Versus Time Dummy Hedonic Indexes.” In *Price Index Concepts and Measurement*, edited by W.E. Diewert, J.S. Greenlees, and C.R. Hulten, 161–196. Chicago: University of Chicago Press.

Diewert, W.E. and C. Shimizu. 2013. *Residential Property Price Indexes for Tokyo*. Vancouver: The University of British Columbia (UBC Discussion Paper Series No. 13-07).

Dorsey, R.E., H. Hu, W.J. Mayer, and H. Wang. 2010. “Hedonic Versus Repeat-Sales Housing Price Indexes for Measuring the Recent Boom-Bust Cycle.” *Journal of Housing Economics* 19: 75–93. DOI: https://doi.org/10.1016/j.jhe.2010.04.001.

Eurostat, ILO, IMF, OECD, UNECE, and World Bank. 2013. *Handbook on Residential Property Price Indices*. Luxemburg: Publications Office of the European Union.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton. 1998a. “Scale Issues and Geographically Weighted Regression.” In *Modelling Scale in Geographical Information Science*, edited by N.J. Tate and P.M. Atkinson, 123–140. Chichester: Wiley.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons.

Fotheringham, A.S., M.E. Charlton, and C. Brunsdon. 1998b. “Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis.” *Environment and Planning A* 30: 1905–1927. DOI: https://doi.org/10.1068/a301905.

Francke, M.K. and A.M. van de Minne. 2017. “Land, Structure and Depreciation.” *Real Estate Economics* 45: 415–451. DOI: http://dx.doi.org/10.1111/1540-6229.12146.

Geniaux, G. and C. Napoléone. 2008. “Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditve Models.” In *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*, edited by A. Baranzini, J. Ramirez, C. Schaerer, and P. Thalmann, 101–127. New York: Springer.

Hill, R.J. and D. Melser. 2008. “Hedonic Imputation and the Price Index Problem: An Application to Housing.” *Economic Inquiry* 46: 593–609. DOI: http://dx.doi.org/10.1111/j.1465-7295.2007.00110.x.

Hill, R.J., D. Melser, and I. Syed. 2009. “Measuring a Boom and Bust: The Sydney Housing Market 2001 – 2006.” *Journal of Housing Economics* 18: 193–205. DOI: https://doi.org/10.1016/j.jhe.2009.07.010.

Hill, R.J. and M. Scholz. 2017. “Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines.” *Review of Income and Wealth* Forthcoming. DOI: http://dx.doi.org/10.1111/roiw.12303.

Hurvich, C.M. and C.L. Tsai. 1989. “Regression and Time Series Model Selection in Small Samples.” *Biometrika* 76: 297–307. DOI: https://doi.org/10.1093/biomet/76.2.297.

Jones, J.P. and E. Casetti. 1992. *Applications of the Expansion Method*. London: Routledge.

Kuminoff, N.V. and J.C. Pope. 2013. “The Value of Residential Land and Structures During the Great Housig Boom and Bust.” *Land Economics* 89: 1–29. DOI: https://doi.org/10.3368/le.89.1.1.

Mei, C., N. Wang, and W. Zhang. 2006. “Testing the Importance of the Explanatory Variables in a Mixed Geographically Weighted Regression Model.” *Environment and Planning A* 38: 587–598. DOI: https://doi.org/10.1068/a3768.

Pace, R.K., R. Barry, J.M. Clapp, and M. Rodriquez. 1998. “Spatiotemporal Autoregressive Models of Neighborhood Effects.” *Journal of Real Estate Finance and Economics* 17: 15–33. DOI: https://doi.org/10.1023/A:1007799028599.

Rambaldi, A.N., R.R.J. McAllister, and C.S. Fletcher. 2015. *Decoupling Land Values in Residential Property Prices: Smoothing Methods for Hedonic Imputed Price Indices*. Queensland: University of Queensland (School of Economics Discussion Paper Series No: 549).

Rambaldi, A.N. and D.S.P. Rao. 2011. *Hedonic Predicted House Price Indices Using Time-Varying Hedonic Models with Spatial Autocorrelation*. Queensland: University of Queensland (School of Economics Discussion Paper Series No: 432).

Rambaldi, A.N. and D.S.P. Rao. 2013. *Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes*. Queensland: University of Queensland (CEPA Working Paper Series No: WP03/2013).

Sun, H., Y. Tu, and S.-M. Yu. 2005. “A Spatio-Temporal Autoregressive Model for Multi-Unit Residential Market Analysis.” *The Journal of Real Estate Finance and Economics* 31: 155–187. DOI: https://doi.org/10.1007/s11146-005-1370-0.

Tu, Y., S.-M. Yu, and H. Sun. 2004. “Transaction-Based Office Price Indexes: A Spatiotemporal Modeling Approach.” *Real Estate Economics* 32: 297–328. DOI: http://dx.doi.org/10.1111/j.1080-8620.2004.00093.x.

van de Minne, A.M. and M.K. Francke. 2012. “De waardebepaling van grond en opstal [The determination of the value of land and structures].” *Real Estate Research Quarterly* 11: 14–24.

Received December 2016
Revised July 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 721–752, http://dx.doi.org/10.2478/JOS-2018-0034

Accounting for Complex Sampling in Survey Estimation: A Review of Current Software Tools

Brady T. West[1], *Joseph W. Sakshaug*[2], *and Guy Alain S. Aurelien*[3]

In this article, we review current state-of-the art software enabling statisticians to apply design-based, model-based, and so-called "hybrid" approaches to the analysis of complex sample survey data. We present brief overviews of the similarities and differences between these alternative approaches, and then focus on software tools that are presently available for implementing each approach. We conclude with a summary of directions for future software development in this area.

Key words: Complex sample survey data; statistical software; design-based analysis; model-based analysis; multilevel modeling.

1. Introduction

Secondary analysis of survey data arising from complex sample designs is a ubiquitous research methodology in many applied fields. The "complex" terminology refers to features of sample designs that deviate from a design featuring simple random sampling with replacement, which in a finite population sampling framework is in accord with the theoretical notion of independent and identically distributed data. These complex design features, which generally include unequal probabilities of selection into the sample, cluster sampling, and stratification of the target population prior to sampling (Heeringa et al. 2017), need to be accounted for by secondary data analysts and applied statisticians who have many tools at their disposal for analyzing these types of data sets. A failure to account for these design features in analysis can lead to substantially biased inferences (e.g., Skinner et al. 1989; West et al. 2016; Heeringa et al. 2017). Over a period of more than 80 years, many different methods have been proposed by statisticians and survey methodologists for correctly accounting for these sample design features when performing survey data analysis.

The variety of approaches discussed and proposed in the survey statistics literature can generally be grouped into two main categories: *design-based analysis*, where the

[1] Survey Research Center, University of Michigan-Ann Arbor, 4118 Institute for Social Research, 426 Thompson Street Ann Arbor, MI, 48106, U.S.A. Email: bwest@umich.edu
[2] Institute for Employment Research, Regensburger Strasse 104, Nuremberg, 90478, Germany. Email: joe.sakshaug@iab.de
[3] Walter R. McDonald & Associates, 12300 Twinbrook Pkwy, Suite310, Rockville, MD 20852, U.S.A. Email: alainshamir@gmail.com

Acknowledgments: This work was supported by a grant from the U.S. National Science Foundation (NCSE-1424081).

randomized selection mechanism underlying the probability sampling governs all subsequent inference, and *model-based analysis*, where all inference depends on probability models posited by the analyst (Hansen et al. 1983). More recently (e.g., Little 2015), statisticians have advocated "hybrid" approaches that combine optimal properties of model-based and design-based approaches. A statistician responsible for analyzing survey data therefore needs to select one or more of these approaches to employ, depending on the objectives of a researcher's study and the parameters of scientific interest. And, once an approach has been selected, the statistician needs to identify software that implements the selected approach. In the present article, we aim to provide statisticians and survey researchers with an up-to-date review of state-of-the-art statistical software capable of implementing each of these different approaches, depending on the specific analysis of interest.

When thinking about these alternative approaches and the software tools implementing them, one needs to consider the objectives of a given analysis of survey data. Is one merely interested in generating descriptive inferences (means, proportions, totals, etc.), or is one also interested in more "analytic" objectives (regression coefficients, odds ratios, etc.)? The identification of appropriate software requires a cross-classification of "objective" (descriptive vs. analytic) and "approach" (design-based vs. model-based); see Table 1. We note that so-called "hybrid" approaches to analytic studies combine features of both design-based and model-based approaches. In the discussion moving forward, we assume that a formal probability sampling plan has been used to select a given sample from a finite population, and that the analyst is weighing different analysis approaches with this sample in hand. We do not consider software for analyzing data from non-probability samples, which are currently receiving a great deal of research attention (e.g., Baker et al. 2013; Elliott and Valliant 2017), in this article.

This article reviews state-of-the-art software tools in each of the five domains indicated in Table 1. Modern survey statisticians need to speak multiple computing languages in general, understanding the pros and cons of each, and effectively communicate software alternatives for clients who desire to analyze survey data. Not all software packages share the same capabilities for analyzing complex sample survey data, and we aim to review the state of the art in this regard. The article is structured as follows. In each of Sections 2 through 6, we first present a brief overview of one of the five approaches in Table 1, and then review current software tools that are available for implementing that particular approach. We then conclude in Section 7 with a summary of important directions for future software development in this area.

Table 1. Five possible combinations of research objectives and analysis approaches, to guide a review of current software for the analysis of complex sample survey data.

	Design-based approaches	Model-based approaches
Descriptive objectives	1	2
Analytic objectives	3	4
	"Hybrid" approaches (5)	

2. Descriptive Objectives: Design-Based Approaches

2.1. Weighted Estimation

When analysts employ design-based approaches to the descriptive analysis of survey data, their analytic objectives generally involve design-unbiased estimation (i.e., estimation that is unbiased with respect to the probability sample design used) of simple descriptive parameters characterizing a finite target population, such as means, proportions, totals, percentiles, and row or column percentages in contingency tables. These approaches generally feature weighted estimation of the parameters of interest, in addition to design-unbiased, nonparametric estimation of sampling variance for the weighted estimates and design-adjusted tests of associations between variables (e.g., Rao and Scott 1984). These approaches are quite popular among nonstatisticians because they are widely implemented in different statistical software packages, and they yield robust population inferences that do not require parametric assumptions regarding the variables of interest.

In general, the respondent weights computed by organizations collecting and producing survey data account for three key aspects of the sample design and the data collection: 1) unequal probabilities of selection into the sample for different population elements, 2) adjustment for nonresponse during data collection, and 3) calibration of the (possibly adjusted) respondent weights to known population totals (Kish 1965; Kalton and Flores-Cervantes 2003; Lohr 2009; Valliant et al. 2013; Lavallee and Beaumont 2016; Heeringa et al. 2017; Haziza and Beaumont 2017). The first element of a respondent weight is generally referred to as a *design weight*. The design weight for a given sampled unit is defined as the inverse of the probability of inclusion for that unit in a given sample, and these design weights can be computed for *all* sampled units in a probability sample (where every population element has a known nonzero probability of inclusion), including respondents and nonrespondents. Inference in design-based approaches is driven by these probabilities of selection, and these components of the weight ensure that estimates computed using the weights appropriately reflect the probability of selection for a given case from a specified target population. Under an extremely unusual scenario where 100 percent of the sampled population units respond to a survey request, one could compute population estimates of target parameters that are unbiased with respect to the sample design using this single design weight.

Unfortunately, not all sampled population units will respond to a survey request. If nonresponding units differ systematically from responding units in terms of key features of interest, nonresponse bias in estimates computed using design-based approaches may result. For this reason, the design weights are often adjusted to account for differential nonresponse among different population subgroups, treating the probability of responding as an additional stochastic stage of sample selection (Cassel et al. 1983; Särndal and Swensson 1987; Ekholm and Laaksonen 1991), and multiplying the design weights for responding units by the inverse of their response probability. Because these probabilities of response are not known in practice, they need to be estimated. Given auxiliary data for respondents and nonrespondents that are generally predictive of both the probability of responding and key survey variables (Lessler and Kalsbeek 1992; Bethlehem 2002; Kalton and Flores-Cervantes 2003; Little and Vartivarian 2005; Beaumont 2005; Groves 2006;

Kreuter et al. 2010), the literature provides extensive guidance on optimal methods for estimating these response probabilities and using them to adjust the design weights for nonresponse (Little 1986; Ekholm and Laaksonen 1991; Eltinge and Yansaneh 1997; Grau et al. 2006; Wun et al. 2007; Haziza and Beaumont 2007; West 2009; Kott 2012; Valliant et al. 2013; Brick 2013; Flores-Cervantes and Brick 2016).

The next (and generally final) step in computing adjusted design weights is to calibrate the (nonresponse adjusted) weights for responding units to sum to known population control totals, ensuring sound population representation in terms of the marginal distributions of (generally sociodemographic) population characteristics. There is a vast literature on this topic (Deville and Särndal 1992; Lundström and Särndal 1999; Rao 2005; Kott 2006; Kim and Park 2010; Kott 2011), and one can use a variety of approaches to perform calibration adjustments in practice, including poststratification (Holt and Smith 1979), raking (Oh and Scheuren 1983; Deville et al. 1993), and generalized regression estimation (Valliant et al. 2000), which can utilize population information for both continuous and categorical variables. Kott and Liao (2012) outline a calibration procedure implemented in the WTADJUST procedure of the SUDAAN software that builds on the developments in prior calibration literature to provide "double protection" against misspecification of either a substantive model or a response model (based on the auxiliary variables used in the calibration adjustment) when using calibration for nonresponse adjustment.

The WesVar software produced by Westat (https://www.westat.com/our-work/information-systems/wesvar®-support), the `calibrate()` function in the R `survey` package (Lumley 2010), the `ipfraking` user-written package in Stata (Kolenikov 2014), the `sreweight` user-written command in Stata (Pacifico 2014), and the CALMAR 2 software developed by Le Guennec and Sautory (2002) are also capable of computing calibration adjustments to design weights based on the methods described above, given population information on the chosen auxiliary variables (see http://vesselinov.com/CalmarEngDoc.pdf for more details on the various calibration options in the CALMAR 2 software). The final calibrated weights may then be trimmed to minimize the impact of weight variance on the precision of weighted survey estimates (Potter 1990; Elliott and Little 2000; Kalton and Flores-Cervantes 2003; Beaumont 2008; see also Asparouhov and Muthén 2007 for optimal weight trimming approaches using the Mplus software). The weights that result from this process then need to be input by analysts into software procedures enabling design-unbiased point estimation of population parameters (see Subsection 2.4).

We note that the final overall respondent weights that result from this three-step process are essentially adjusted versions of the design weights, but software procedures enabling design-based analysis treat these final respondent weights as if they were design weights that are "known" with certainty. Because *estimates* of response propensity are often used to adjust the design weights for unit nonresponse, this uncertainty in the final respondent weights should be accounted for in variance estimation. This is best handled using replication techniques, as outlined in Subsection 2.2 below, where the adjustment process (based on estimates) can be repeated for each replicate sample, and the variance in the adjustments across replicates is incorporated into the final variance estimates (Valliant 2004).

2.2. Variance Estimation

Taylor Series Linearization (TSL) is a design-based variance estimation technique that is widely implemented in different statistical software procedures and often serves as a default variance estimation procedure in these procedures when applying design-based approaches to complex samples. The basic idea behind TSL is to use a Taylor series expansion to approximate a non-linear estimator (e.g., a ratio mean, a ratio estimator of total, a regression parameter, a correlation coefficient) using a *linear* function of estimated sample totals. Once the nonlinear estimator is "linearized," then unbiased, design-based variance estimation formulae reflecting the complex sampling features (stratification, cluster sampling, weighting) can be applied to estimate the variance of the linear function of sample totals. The variance of the linearized estimator is estimated within each stratum (if applicable), and the stratum variances are combined to produce the total variance of the estimator. Wolter (2007, Chapter 6) reviews the TSL literature and provides technical details.

There are two important issues that analysts need to handle carefully when employing TSL for design-based variance estimation: subpopulation analysis, and "singleton" sampling clusters. First, considering subpopulation analysis, complex sample designs often employ the sampling of clusters of population elements within sampling strata for reasons of cost efficiency. The clusters sampled at the first stage of random selection (possibly within strata) are often referred to as primary sampling units, or PSUs, and these could be geographic areas in area probability samples, naturally occurring groups of population elements (e.g., colleges), or individual sampled elements if no cluster sampling is employed (software enabling design-based analysis will estimate variances under this assumption if no cluster ID variables are indicated). When analyzing subpopulations (e.g., elderly males) and using TSL for variance estimation, analysts need to explicitly form binary variables indicating which sampled cases fall into the subpopulation of interest, and use these indicators for variance estimation (which is often facilitated by "subpopulation" options in the different software procedures, e.g., the `subpop()` option in Stata). This approach enables PSUs with no sample from the subpopulation to still be accounted for in the variance estimation (in that they contribute totals of zero for the variables of interest), rather than being removed entirely. The physical removal of entire PSUs due to the deletion of cases that do not belong to a subpopulation can lead to scenarios where sampling strata only have a single PSU present, preventing variance estimation within that stratum when using TSL (more on this below). See West et al. (2008) and Heeringa et al. (2017) for more on this TSL-specific issue, which becomes irrelevant when using *replication methods* for variance estimation (as clusters with no subpopulation sample simply do not contribute to replicate estimates).

Second, considering the "singleton" sampling cluster issue, some PSUs may also be selected with *certainty*, meaning (in a design-based setting) that they would be included in every possible hypothetical sample that might be selected; that is, they have a probability of inclusion of one. When employing TSL for variance estimation, there need to be at least two PSUs present within a sampling stratum to estimate the contribution of that stratum to the overall sampling variance, and certainty PSUs often define their own stratum (e.g., the city of New York in the United States). Data producers can facilitate variance estimation using TSL by dividing the sampled elements in a certainty PSU into multiple *random* *groups* (Wolter 2007), and providing codes for these "pseudo clusters" in a public-use data set. If a data user for some reason encounters a stratum with only a single PSU code present in such a data set, most design-based software will provide some form of ad-hoc solution for estimating the contribution of that stratum to the overall sampling variance. For example, Stata provides users with several choices via the `singleunit()` option in the `svyset` command, which is used to identify PSUs for variance estimation; SUDAAN provides the MISSUNIT option (see http://sudaansupport.rti.org/sudaan/page.cfm/Theory); and the `survey` package in R provides the user with a variety of global options (see http://faculty.washington.edu/tlumley/old-survey/exmample-lonely.html for examples).

Replication methods represent a second nonparametric design-based approach to estimating the variance of a weighted estimate. In general, these methods involve dividing the full sample into various subsamples, calculating an estimate of the parameter of interest within each subsample, and calculating the variation among the subsample estimates to estimate the variance of the full sample estimate. These methods can be implemented in various forms, including the random groups method (RGM), Jackknife repeated replication (JRR), balanced repeated replication (BRR), bootstrapping, and various modifications of these methods (Wolter 2007; Shao and Tu 1995). A key advantage of these replication methods is that they do not require the linearization of a nonlinear estimator (Krewski and Rao 1981), and can generally be applied to many different forms of estimators. These methods also enable survey organizations to disseminate public-use survey data sets including (adjusted) weights for each of the replicate samples in lieu of stratum and PSU codes, minimizing the risk of identifying survey respondents within small PSUs. This requires the data user to employ variance estimation software that supports the specific type of replication weighting scheme used by the survey organization, and nearly all major statistical software packages with procedures enabling variance estimation for complex samples currently enable the use of these "replicate weights" (e.g., SAS, Stata, R). At present, the bootstrapping approach can be applied to complex samples in Stata (Kolenikov 2010), R (Lumley 2010), Mplus (Asparouhov and Muthén 2010), WesVar, SAS, and SUDAAN (Gagne et al. 2014).

So how does a survey statistician choose which variance estimation procedure to use when employing a particular software procedure for design-based descriptive analysis? Numerous studies have compared the performance of these alternative variance estimation methods under different complex sample designs. These include Kish and Frankel (1968, 1970, 1974), Frankel (1971), Bean (1975), Campbell and Meyer (1978), Lemeshow and Levy (1978), Shah et al. (1977), Rao and Wu (1985, 1987, 1988), Kovar et al. (1988), Judkins (1990), Shao and Sitter (1996), Korn and Graubard (1999), Canty and Davison (1999), Rao and Shao (1999), Shao (2003), and Heeringa et al. (2017). These studies have consistently demonstrated that for many common types of survey estimates (e.g., means, proportions, regression coefficients), all methods perform well and differences between the methods are *negligible*. Exceptions include small samples, where linearization can be unstable and perform worse than replication methods, and quantiles, where alternative forms of linearization are needed given that quantiles cannot generally be approximated using smooth functions of population totals or means (Woodruff 1952; Francisco and Fuller 1991; Sitter and Wu 2001). Many of the studies above demonstrate that BRR and the bootstrap perform well for medians and functions of quantiles. In addition, linearization methods covering all possible nonlinear estimators (e.g., correlation coefficients) and complex sample designs may not be readily programmed in all software packages.

2.3. *Calculation of Degrees of Freedom for Confidence Intervals*

Analysts often desire to form confidence intervals for population parameters when applying design-based methods to complex samples. These intervals, which under an assumption of large-sample normality of the sampling distribution for the weighted estimator rely on a critical t-value, also require specification of the appropriate degrees of freedom for the critical t-value. At present, most statistical software computes these degrees of freedom based on the aforementioned assumption of large-sample normality, setting the degrees of freedom equal to the number of clusters used for variance estimation minus the number of strata (Heeringa et al. 2017). While this approach makes intuitive sense, given that design-based variance estimates are driven by between-cluster variance within strata and the standard deviation of the sampling distribution is estimated rather than known, it is heavily dependent on the aforementioned assumption and can be severely limited in certain cases (Valliant and Rust 2010). Valliant and Rust (2010) propose an alternative estimator of the degrees of freedom for the critical t-value and show that it leads to improved coverage in some cases, but more work in this area, including sensitivity analyses, is certainly needed. Furthermore, the alternative estimator proposed by Valliant and Rust has yet to make its way into any statistical software.

Dean and Pagano (2015) provide a recent review of several different methods for computing confidence intervals for estimated proportions in the descriptive context, with and without adjustment for the degrees of freedom according to a complex sample design. Via simulation, these authors found support for use of the logit, Wilson, Jeffreys, and Agresti-Coull intervals (Agresti and Coull 1998) in complex samples, especially when proportions are very small or very large. Some of these methods (e.g., the logit approach) are readily implemented in existing software (e.g., the `svy: tab` command in Stata). While the other methods may not be as widely implemented, these authors provide clear guidance on their computation in practice.

2.4. *Software*

We now consider state-of-the-art statistical software that is currently available for implementing the design-based descriptive estimation and inference approaches outlined above when analyzing complex sample survey data. Table 2 provides a list of presently available software procedures and profiles their capabilities, in particular considering 1) percentile estimation, 2) variance estimation options, and 3) subpopulation analysis. All procedures in Table 2 enable appropriate weighted estimation of various descriptive parameters. A key take-away message from Table 2 is that weighted estimation of percentiles, combined with design-based variance estimation for the weighted estimates based on the aforementioned approaches, is not yet widely implemented across the different software packages. Aside from this, most software packages offer similar capabilities for design-based descriptive analyses of complex sample survey data. Examples of the use of syntax for many of these procedures can be found at http://www.isr.umich.edu/src/smp/asda.

Table 2. Capability profile of current statistical software enabling design-based descriptive estimation based on complex sample survey data (all procedures enable weighted estimation).

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
SAS/STAT (V9.4)										
SURVEYMEANS	Y	Y	Y	Y	Y	Y	Y[1]	Y	Y	Y[2]
SURVEYFREQ	N	Y	Y	Y	Y	Y	Y[1]	Y	Y	Y
IBM SPSS Statistics: Complex Samples Module (V25)										
CSDESCRIPTIVES	N	Y	N	Y	N	N	N	N	Y	N
CSTABULATE	N	Y	N	Y	N	N	N	N	Y	Y
Stata (V15+)										
`svy: mean`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y[4]
`svy: prop`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y[4]
`svy: tab`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y
R: survey package[5]										
`svymean()`	N	Y	Y	Y	Y	Y	Y	Y[7]	N	N
`svyby()`	Y[6]	Y	Y	Y	Y	Y	Y	Y[7]	Y	N
`svytable()`	N	Y	Y	Y	Y	Y	Y	Y[7]	Y	Y
`svyquantile()`	Y	Y	Y	Y	Y	Y	Y	Y[7]	N	N
`svycontrast()`	Y	Y	Y	Y	Y	Y	Y	Y[7]	N	Y[8]
WesVar										
Mean	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Median	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Quantile	Y	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Ratio	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Totals	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y

Table 2. Continued.

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
Variance	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
CV	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Skewness	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Kurtosis	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
SUDAAN										
DESCRIPT	Y	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
TABULATE	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
RATIOS	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
PROC CROSSTAB	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
IVEware[10]										
%DESCRIBE	N	N	N	Y	Y	N	N	N	Y	Y
VPLX[11]										
Summary statistics	N	N	Y	Y	Y	Y	N	Y	Y	Y
Epi Info CSAMPLE[12]										
Summary statistics	N	N	N	Y	N	N	N	N	Y	Y
AM Software[13]										
Summary statistics	Y	N	N	Y	Y	Y	Y	N	N	N

Table 2. Continued.

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
Bascula 4[14]										
Summary statistics	N	N	Y	Y	N	Y	N	Y	N	N
CLUSTERS[15]										
Summary statistics	N	N	N	Y	N	N	N	N	Y	Y
Generalized estimation system[16]										
Summary statistics	N	N	N	Y	Y	N	N	Y	N	N
PCCARP[17]										
Summary statistics	Y	Y	N	Y	N	N	N	Y	Y	Y

[1] See https://support.sas.com/documentation/onlinedoc/stat/143/surveymeans.pdf for general discussion.

[2] Available in the newest version of SURVEYMEANS in SAS/STAT 14.2: see http://support.sas.com/kb/34/607.html.

[3] Via the `poststrata()` and `postweight()` options in the `svyset` command.

[4] Via the `lincom` post-estimation command.

[5] See http://r-survey.r-forge.r-project.org/survey/ for extensive annotated examples of the use of these procedures in R.

[6] Via inclusion of the `svyquantile()` function.

[7] When the `postStratify()` function has been used to update the `svydesign()` object.

[8] Given a `svyby()` object; covariances of subgroup estimates are not computed and accounted for in comparisons when using linearization.

[9] See https://www.researchgate.net/publication/255643504_Using_bootstrap_weights_with_Wes_Var_and_SUDAAN for discussion.

[10] http://www.iveware.org

[11] https://www.census.gov/sdms/www/vwelcome.html (still active)

[12] http://www.cdc.gov/epiinfo/index.html

[13] http://am.air.org/

[14] http://www.hcp.med.harvard.edu/statistics/survey-soft/bascula.html (A European product, no longer active)

[15] http://www.hcp.med.harvard.edu/statistics/survey-soft/clusters.html (A European product, still active)

[16] http://www.hcp.med.harvard.edu/statistics/survey-soft/genest.html (A product of Statistics Canada, still active)

[17] http://www.hcp.med.harvard.edu/statistics/survey-soft/pccarp.html (no longer active)

3. Descriptive Objectives: Model-Based Approaches

3.1. Overview

While descriptive inferences based on complex sample survey data sets generally tend to arise from design-based approaches (Little 2004), model-based approaches to descriptive inference have their relative merits as well. Design-based approaches can be heavily affected by *non-sampling errors*, such as unit nonresponse, given that they are governed by knowledge of sampling probabilities for all cases included in a sample. Unlike design-based approaches, which are based on the notion of random sampling from a finite population, strictly model-based approaches assume that some superpopulation model exists, from which the finite populations in the design-based setting are actually sampled. Interest lies in unbiased estimation of the parameters of that superpopulation model. Model-based approaches to making descriptive inference generally involve the specification of a probability model for a variable (or variables) of interest (where the variable for which descriptive inference is desired is a dependent variable), estimation of the descriptive parameters of interest (e.g., means) defined by the model, and estimation of the variance of that estimate with respect to the specified model (Binder and Roberts 2003).

One can also employ model-based prediction approaches when making descriptive inferences about finite populations. In this case, various auxiliary predictors available for the larger population (usually in aggregate form) may be included in the specification of the probability model for the variable of interest. In the case of complex sampling, these auxiliary predictors can and should generally include some function of the probability of selection, in addition to stratum identifiers (if these design features are relevant and informative about the variable of interest; Hansen et al. 1983; Little 2004). In these cases, predictions are computed on the variable of interest for nonsampled cases or nonrespondents, using the auxiliary information and parameter estimates in the specified model, and estimates are computed by combining the observed sample data on the dependent variable and the model-based predictions for nonsampled or nonresponding cases (Valliant et al. 2000). Variances of the resulting estimates are then computed with respect to the properties of the model used. Predictions for the nonsampled cases and measures of uncertainty for the descriptive parameter of interest may also be computed based on Bayesian methods (Little 2003), where informative design features should again be included in the specification of the model (likelihood) for the available data.

Särndal et al. (1992) describe an alternative approach that combines elements of design-based inference and model-based inference known as *model-assisted* inference, where design-based estimates of descriptive parameters (e.g., totals) are adjusted given known auxiliary variables for the entire population and their relationships with the variable of interest, and variances of the adjusted estimates are computed with respect to the randomization distribution (as in design-based inference). The generalized regression (GREG) estimator is a popular example of the model-assisted approach to making descriptive inferences from complex sample survey data. Valliant et al. (2000) provide a comprehensive theoretical overview of related model-based prediction approaches to the descriptive analysis of survey data.

Best practices in this area generally focus on how probabilities of selection/weights should be accounted for in the models, how to make the most efficient use of auxiliary information available for a finite population, and how to handle nonresponse. For example, Elliott and Little (2000) discuss efficient model-based Bayesian approaches to accommodating sampling weights in descriptive inference when large or highly-variable survey weights may cause design-based approaches to become very inefficient. Little (2004) presented a model-based approach to descriptive inference combining precision weighting and probability weighting in a Bayesian framework. This was further expanded on in Little (2012), who advocated "Calibrated Bayes" (CB) as a framework for survey inference. The basic idea behind CB is to use a Bayesian model-based approach to produce inferences that have good design-based properties. The CB approach is intended to combine the strengths of both design-based and model-based perspectives by explicitly accounting for survey design information in the model and using only weak prior distributions that allow the observed data to dominate the inference. Inferences are calibrated in the sense that they produce posterior credibility intervals that correspond to their nominal design-based coverage in repeated sampling (Little 2006, 2011, 2012, 2015). The incorporation of all key survey design features in the model is paramount to this approach to minimize the effect of model misspecification.

Peress (2010) discussed the use of selection models in a model-based approach to account for nonignorable nonresponse as a part of the modeling process in estimating a proportion. Using a related approach, Barnighausen et al. (2011) applied a Heckman-type bivariate probit selection model in estimating HIV prevalence estimates that adjusted for nonignorable nonresponse based on a set of selection variables correlated with survey participation. More recently, West and McCabe (2017) demonstrated how this approach can be implemented using the Stata software to make descriptive inferences in a longitudinal context, where nonignorable attrition may be occurring in the future waves of a panel survey.

3.2. *Software*

Regarding available software for implementing these model-based approaches to descriptive inference in surveys, there are not nearly as many "canned" software procedures implementing these approaches as there are for design-based approaches, meaning that statisticians would generally need to write code implementing these approaches for nonstatistical clients. For example, Zheng and Little (2003), Little and Zheng (2007), and Zangeneh and Little (2015) demonstrate the improvements in estimates of population totals when using a penalized spline regression model over a design-based Horvitz-Thompson approach when the sizes of nonsampled units are either known or unknown, and error variances in the model of interest may be heteroscedastic. Zangeneh and Little (2015) have developed R code implementing their proposed approach (available from the authors upon request).

Chapter 8 of Lunn et al. (2012) discusses how the BUGS software (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) can be used to generate predictions for nonsampled cases using a Bayesian approach, where again any complex sampling features would need to be accounted for in the model specification (Little 2004). More recently, the Stan software (http://mc-stan.org/) has become a popular alternative for similar types of Bayesian approaches, and this software can be readily used in R and Stata (among other platforms). Interested readers can see http://rpubs.com/corey_sparks/157901 for an example of a model-based descriptive analysis of survey data using a Bayesian approach in R (calling the Stan software). Valliant et al. (2000) provided a comprehensive library of S-plus code for implementing various model-based and model-assisted approaches, and these functions can generally be adapted in the R software with ease (examples are available upon request from the first author; see also Valliant et al. (2013) for additional examples). In general, we recommend that statisticians compare standard errors for descriptive estimates computed using design-based and model-based approaches, and determine whether efficiency gains are possible when employing the model-based approaches discussed in this section.

In more recent work, Si et al. (2015) presented model-based Bayesian methodology for making robust finite population inferences about means or proportions of interest when only final survey weights (and no stratum or cluster codes) are available for survey respondents. This model-based approach simultaneously predicts the distribution of the final survey weights among nonsampled cases in the population of interest and the values of the survey variable of interest for these cases (as a function of the weights), enabling simulations of the full population means or proportions based on posterior distributions for these descriptive parameters (given the sampled cases and their data). These authors demonstrated the advantages of this approach for the efficiency of descriptive finite population estimates (for both full populations and subpopulations), and implemented this approach in the Stan software (see http://www.isr.umich.edu/src/smp/asda for example Stan code).

4. Analytic Objectives: Design-Based Approaches

4.1. Overview

Design-based approaches that utilize (adjusted) design weights to fit regression models to complex sample survey data are in common use (e.g., DuMouchel and Duncan 1983; Pfefferman 1993; Pfeffermann and Sverchkov 2009; Pfefferman 2011; Lumley and Scott 2017). In the simple case of estimating the parameters of a specified linear regression model, the standard ordinary least squares (OLS) approach can be modified by incorporating the final respondent weight into the objective function that minimizes the finite population residual sum of squares. This weighted least squares (WLS) approach provides a closed-form, model-unbiased estimator for the regression parameters that also serves as a pseudo-maximum likelihood estimator for the regression parameters in the finite population (Binder 1981, 1983; Pfeffermann 2011, Section 3.4). Lohr (2014) describes how to estimate design effects in this context reflecting complex sampling features.

For generalized linear models featuring nonlinear relationships between the predictors and the expectation of the dependent variable of interest (e.g., logistic regression models), closed-form solutions do not exist for estimation of model parameters. Furthermore, “standard” maximum likelihood estimation is not possible with complex sample designs because the assumption of independent observations is violated by the stratification and cluster sampling inherent to complex samples (Archer et al. 2007). Binder (1981, 1983) proposed a pseudo-maximum likelihood estimation (PMLE) framework for fitting generalized linear models to complex sample survey data. The basic idea of the PMLE method is to estimate model parameters by replacing finite population likelihood estimating equations with design-unbiased, weighted estimating equations for the responding units. Positive evaluations of the PMLE method and its properties can be found in several studies (Binder 1983; Chambless and Boyle 1985; Roberts et al. 1987; Morel 1989; Skinner et al. 1989; Nordberg 1989; Pfeffermann 1993; Godambe and Thompson 2009). PMLE is now the standard method implemented in many software procedures for fitting generalized linear models to complex sample survey data.

Binder (1981, 1983) also proposed a general method for linearized variance estimation for pseudo-maximum likelihood estimates of regression coefficients that is implemented as the default method in many statistical software packages, and replication methods generally work equally well in many regression settings, as noted earlier. Hypothesis tests for regression parameters based on complex sample survey data are carried out by using the design-based estimates of the variances and covariances and applying commonly used test statistics, such as Student's *t* and the Wald chi-square or Wald F-test. Rao and Scott (1981, 1984, 1987) proposed a modified Wald chi-square statistic for survey data that accounts for complex sample design features. This procedure is implemented in many statistical software packages that support the analysis of complex sample survey data (e.g., the `svy: tab` command and the `test` post-estimation command in Stata).

4.2. *Should Survey Weights Even Be Used to Fit Models?*

The use of (adjusted) design weights to fit regression models has some limitations which have provoked controversy among statisticians (see Pfeffermann 1993; Gelman 2007; Pfefferman 2011; or Bollen et al. 2016 for reviews of the general issues). Design-based estimation strategies utilizing probability-weighted estimators generally yield larger variances than model-based estimation strategies (Korn and Graubard 1999). This loss in efficiency is more notable for small sample sizes and cases where there is large variation in the survey weights. For this reason, one best practice in this area is to examine the sensitivity of the regression results by comparing weighted and unweighted analyses, which is quite easy to do using current software. If these analyses yield notable differences, then this may indicate model misspecification and the weighted estimates should be reported to ensure that they are unbiased with respect to the sample design used. A review of formal tests for differences between weighted and unweighted regression analyses can be found in Bollen et al. (2016). It is also worth noting that the use of probability weighting in the analysis of complex sample survey data is not customary in some disciplines (e.g., economics) which favor the flexibility of explicitly featuring the relevant design variables as part of the model-building process.

4.3. *Software for Model Fitting*

Table 3 presents a summary of available software procedures for fitting regression models to complex sample survey data using design-based approaches. We emphasize software

Table 3. Current software procedures enabling design-based estimation of various regression models.

	Regression modeling options										Goodness of fit tests	
Software	Linear	Binary logistic	Ordinal logistic	Multinomial logistic	Poisson regression	Negative binomial regression	Probit	Cloglog	Survival (Cox) models	Quantile regression	Archer and Lemeshow test	Design-based model diagnostics?
Stata (V15+)												
`svy: regress`[2]	Y	N	N	N	N	N	N	N	N	N[1]	N	N
`svy: logit`	N	Y	N	N	N	N	N	N	N	N	Y[4]	N
`svy: mlogit`	N	N	N	Y	N	N	N	N	N	N	N	N
`svy: ologit`	N	N	Y	N	N	N	N	N	N	N	N	N
`svy: poisson`	N	N	N	N	Y[3]	N	N	N	N	N	N	N
`svy: nbreg`	N	N	N	N	N	Y[3]	N	N	N	N	N	N
`svy: stcox`	N	N	N	N	N	N	N	N	Y	N	N	N
`svy: probit`	N	N	N	N	N	N	Y	N	N	N	N	N
`svy: cloglog`	N	N	N	N	N	N	N	Y	N	N	N	N
SAS (V9.4)												
SURVEYREG	Y	N	N	N	N	N	N	N	N	N	N	N
SURVEYLOGISTIC	N	Y	Y	Y	N	N	Y	Y	N	N	N	N
SURVEYPHREG	N	N	N	N	N	N	N	N	Y	N	N	N
IBM SPSS Statistics: complex camples module (V25)												
CSLOGISTIC	N	Y	Y	Y	N	N	Y	Y	N	N	N	N
CSORDINAL	N	N	Y	N	N	N	Y	Y	N	N	N	Y
CSGLM	Y	N	N	N	N	N	N	N	N	N	N	N
CSCOXREG	N	N	N	N	N	N	N	N	Y	N	N	N
R survey package												
`svyglm()`	Y	Y	Y	Y	Y	Y	N	N	N	N	N	Y[5]
`rq()`	N	N	N	N	N	N	N	N	N	Y[6]	N	N
`svycoxph()`	N	N	N	N	N	N	N	N	Y	N	N	N

Table 3. Continued.

	Regression modeling options									Goodness of fit tests	
Software	Linear	Binary logistic	Ordinal logistic	Multinomial logistic	Poisson regression	Negative binomial regression	Probit	Cloglog	Survival (Cox) models	Quantile regression	Archer and Lemeshow test	Design-based model diagnostics?
Wesvar												
Linear regression	Y	N	N	N	N	N	N	N	N	N	N	N
Logistic regression	N	Y	N	N	N	N	N	N	N	N	N	N
Multinomial regression	N	N	N	Y	N	N	N	N	N	N	N	N
SUDAAN												
REGRESS	Y	N	N	N	N	N	N	N	N	N	N	N
LOGISTIC	N	Y	Y	Y	N	N	N	N	N	N	N	N
MULTILOG	N	N	Y	Y	N	N	N	N	N	N	N	N
LOGLINK	N	N	N	N	Y	N	N	N	N	N	N	N
SURVIVAL	N	N	N	N	N	N	N	N	Y	N	N	N
IVEware												
%REGRESS	Y	Y	Y	Y	Y	N	N	N	Y	N	N	N
Epi info CSAMPLE												
Regress	Y	N	N	N	N	N	N	N	N	N	N	N
Logistic	N	Y	N	N	N	N	N	N	N	N	N	N
AM software												
Regression	Y	Y	Y	Y	N	N	Y	N	N	N	N	N

[1] A full list of possible models that can be fitted in the design-based framework using Stata 15+ (including specialized modeling options, such as structural equation models, instrumental variables regression, finite mixture models, among others) can be found at https://www.stata.com/manuals/svy.pdf.

[2] Quantile regression models can be fitted using design-based approaches in Stata, given replicate weights and the user-written `bs4rw` command.

[3] Procedures for fitting zero-inflated versions of these models, including `svy: zip` and `svy: zinb`, are also available.

[4] Implemented in the post-estimation command `estat gof`, which can be executed after running `svy: logit`.

[5] Currently available in a working R package `svydiags`; see Heeringa et al. (2017).

[6] Possible when using the `rq()` function from the `quantreg` package in combination with replicate weights in the `survey` package; see http://www.isr.umich.edu/src/smp/asda/Additional%20R%20Examples%20bootstrapping%20with%20quantile%20regression.pdf for details.

procedures in general-purpose statistical software packages, but other stand-alone software tools primarily focused on modeling, such as Mplus (see http://www.statmodel.com/resrchpap.shtml for examples) and Latent GOLD (see the *Advanced/Syntax add-on* at https://www.statisticalinnovations.com/latent-gold-5-1/), can also easily fit common regression models using design-based approaches.

Readily apparent from Table 3 are the following take-away points: 1) different packages currently vary in terms of the different types of regression models that can be fitted using design-based methods; 2) design-based post-estimation goodness-of-fit tests are currently only implemented for logistic regression modeling in Stata; 3) design-based quantile regression is only implemented in the R `survey` package and Stata add-on commands at present; and 4) model diagnostics using design-based approaches are currently only available for linear regression models in a working package in R (see http://www.isr.umich.edu/src/smp/asda/svydiags-manual.pdf for details). Examples of the use of syntax for many of these procedures can be found at http://www.isr.umich.edu/src/smp/asda.

4.4. Software for Model Evaluation and Selection

Numerous design-adjusted model evaluation tools have been developed to evaluate the fits of regression models based on complex sample survey data. However, the implementation of some of these tools in popular statistical software packages is not yet widespread. A modified version of the R^2 statistic is often available for linear regression models, which estimates the “weighted” proportion of explained variance in the dependent variable after controlling for the independent variables. Residual diagnostics for complex samples more generally is an active area of research. Li and Valliant (2015) document the latest advances in this area and review their implementation in R; a working package for R entitled `svydiags` is available from these authors upon request, and examples of the use of this package are provided in Heeringa et al. (2017). Liao and Valliant (2012a, 2012b) developed collinearity diagnostics for identifying excessively high correlations between independent variables that explicitly account for complex sampling features; however, these diagnostics have not yet made their way into popular statistical software packages. Li and Valliant (2009, 2011a, 2011b) and Ryan et al. (2015) have proposed methods for identifying influential data points in linear and logistic regression analyses based on complex sample survey data, and these also need software development.

Model selection methods for complex samples have also seen recent development. For instance, Lumley and Scott (2015) developed survey analogues of the popular AIC and BIC information criteria for regression models fitted using pseudo-maximum likelihood estimation methods. These methods have been implemented in the R `survey` package. Archer et al. (2007) demonstrate that standard goodness-of-fit tests are not suitable for complex sample survey data and propose alternative tests that account for complex design features, including an F-test which is a survey analogue to the Hosmer-Lemeshow chi-square test for logistic regression. Heeringa et al. (2017) provide Wald tests for comparing nested regression models, following from Hosmer et al. (2013), who note that the standard likelihood ratio chi-square test is inappropriate for complex sample survey data due to the violation of key assumptions about the likelihood function that underlie the test. This issue is addressed further by Lumley and Scott (2013, 2014), who developed partial likelihood ratio tests for Cox regression models in the survival analysis context and adapted the Rao and Scott (1984) chi-square test to the case of design-based likelihood ratio tests in arbitrary regression models fitted to survey data. These approaches are also currently implemented in the `survey` package in R.

4.5. *Software for Structural Equation Modeling and Classification Trees*

There has also been some work on accounting for complex sample design features in structural equation and latent variable models (e.g., Muthén and Satorra 1995; Kaplan and Ferguson 1999; Stapleton 2002, 2006). These design-based approaches to fitting structural equation models are currently implemented in the Mplus software, the `lavaan.survey` package of R (Oberski 2014), the LISREL software, the `svy: sem` and `svy: gsem` commands of Stata, the Latent GOLD software, and PROC LCA (a user-written add-on for SAS). Some analysts of survey data may also be interested in building classification or regression trees for generating finite population predictions, and the recently-developed `rpms` package in R (Toth 2017) enables analysts to apply regression trees to complex sample survey data. A more general summary of additional specialized procedures for fitting models to survey data in R using design-based approaches can be found at https://cran.r-project.org/web/views/OfficialStatistics.html.

5. Analytic Objectives: Model-Based Approaches

5.1. *Overview*

Model-based approaches for analyzing survey data given analytic objectives vary. These approaches are typically implemented under a population- or sample-based modeling perspective. Under the population modeling perspective, all population units (including nonsampled units) are included in the analysis model, whereas under the sample-based perspective, only the sampled (or responding) units are analyzed. Under the population modeling perspective, one possible approach is to include all design variables and relevant interaction terms as covariates in the analysis model and effectively integrate these variables out (Pfeffermann 2011), leaving only the covariates of substantive interest. Implementing such an approach can be difficult for secondary analysts, because design variables for the entire population are typically not made available to secondary data users.

Model-based approaches for imputing both the design and substantive variables for the non-observed portion of the population have been proposed (Feder 2011; Si et al. 2015), though issues arise when the sample selection is dependent on the substantive variables of interest – a situation realized in a non-ignorable sampling setting (Pfeffermann and Sikov 2011). A further complication, noted by Pfefferman (2011), is that modeling the relationship between the design and substantive variables can be quite cumbersome, and integrating the design variables out of the model may result in an analysis model that does not reproduce the target model of substantive interest. Pfeffermann (2011) addresses this issue by demonstrating that the analysis model can be estimated without integrating the design variables out of the model. When not all design variables are available to the analyst for the entire population, then the sample weight is sometimes used as a proxy for the design variables (DuMouchel and Duncan 1983; Rubin 1985; Chambers et al. 1998; Wu and Fuller 2006). However, this still requires that the sample inclusion probabilities be made available to the secondary data analyst for the entire population, which may not be possible due to confidentiality or other data restrictions.

In contrast, approaches based on the sample modeling perspective need only make use of the design variables known for the sampled (or responding) units. Model-based methods employing maximum likelihood techniques estimate the unknown population parameters using the likelihood of the joint distribution of the design variables and sample covariates (Gelman et al. 2003; Little 2004). Alternative full-likelihood methods, which utilize the Missing Information Principle (Orchard and Woodbury 1972), have been explored in different contexts (Breckling et al. 1994; Chambers et al. 1998; Chambers and Skinner 2003, Chapter 2). Empirical likelihood methods have also been considered for complex samples (Hartley and Rao 1968; Owen 2001). These methods, while generally more computationally intensive than design-based approaches, can produce much more efficient estimates of regression parameters with improved coverage properties (see Pfeffermann et al. 2006 for an illustration).

5.2. *Software*

Given the considerations outlined in Subsection 5.1, model-based approaches to analytic objectives can make use of existing software procedures for fitting regression models. There is no need to use specialized software for design-based survey analysis to fit these models. The important aspect of implementing these procedures is making sure that the design features have been carefully accounted for in the design matrices of the specified models.

6. Analytic Objectives: “Hybrid” Approaches

6.1. *Overview*

So-called “hybrid” approaches to regression modeling of complex sample survey data employ multilevel models, and are distinguished by the explicit desire of the researcher to make finite population inferences about the components of variance in dependent variables of interest attributable to the different stages of a multi-stage sample design. The theory and methods for incorporating survey weights into pseudo-maximum likelihood estimation of the fixed effect and covariance parameters defining a multilevel model were initially described by Pfeffermann et al. (1998). These methods were later expanded on and evaluated via simulation by Kovacevic and Rai (2003), Grilli and Pratesi (2004), Asparouhov (2006), Rabe-Hesketh and Skrondal (2006), Carle (2009), and Pfeffermann (2011). Skinner and Holmes (2003) and Heeringa et al. (2017) have elaborated on the appropriate use of survey weights when fitting multilevel models to *longitudinal* survey data.

These methods for computing weighted estimates of the parameters in multilevel regression models all require the following: 1) *conditional* weights at lower levels of the data hierarchy (e.g., students within schools), which indicate inverses of the probability of selection *conditional* on a given higher-level unit (e.g., school) being sampled, and 2) unit-level weights at the highest level of the data hierarchy (e.g., counties), representing inverses of the probabilities of selection for the highest-level sampling unit. Pfeffermann et al. (1998) and Rabe- Hesketh and Skrondal (2006) clearly describe how the likelihood functions used to estimate these models are partitioned in a way that requires this combination of conditional and unconditional weights for unbiased estimation of the model parameters. More recently, Stapleton and Kang (2016) have described how to estimate design effects in this context, representing the effects of complex sampling features on the variance of estimated parameters in multilevel models.

This requirement that the conditional lower-level weights and unconditional higher-level weights be available for estimation has limited these approaches from gaining traction outside of the survey statistics literature (see West et al. 2015 for a recent case study), given the need for public-use data files to include these "specialized" weights for users, which could introduce disclosure risk concerns. The final respondent weights provided in a public-use survey data set typically represent inverses of the products of the probabilities of selection at *all* stages of a complex sample design; computation of the conditional weights at lower levels requires dividing the final weights by the higher-level sampling weights to determine the inverse of the conditional sampling probability required for estimation. The computation of these weights therefore represents an additional burden that survey organizations would need to take on for users interested in these "hybrid" approaches. Chantala et al. (2011) provide important practical guidance and software tools to assist with this process.

The conditional weights that are specific to each lower-level unit also need to be *scaled* or *normalized* across all higher-level units, to reduce the varying magnitudes of these weights across the higher-level units. This weight scaling is important because it minimizes the bias in parameter estimates based on the models (Pfeffermann et al. 1998). Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), and Carle (2009) describe alternative methods for performing weight scaling (e.g., normalizing the lower-level weights by dividing all of the weights in a higher-level unit by their average, so that they sum to the sample size within that unit). The literature to date has not demonstrated that one weight scaling method is superior over another; there has, however, been consistent agreement that weight scaling needs to be done to minimize bias, especially in the case of generalized linear regression models (e.g., multilevel logistic regression models; Rabe-Hesketh and Skrondal 2006). Weight scaling represents an additional data processing step that may not be "automatic" in the software that is presently available for these "hybrid" approaches (e.g., the `mixed` command in Stata); see Rabe-Hesketh and Skrondal (2006) for worked examples.

6.2. *Software*

At present, these approaches for weighted estimation of multilevel models are not widely implemented across statistical software packages. This kind of implementation will be especially important for these "hybrid" model-based approaches to gain traction among nonstatisticians. Software packages and specific procedures capable of implementing these "hybrid" approaches for both linear and generalized linear regression models include Stata (Version 15.1+), SAS (PROC GLIMMIX, SAS/STAT Version 13.1+; Zhu 2014), HLM (Version 7.01+), MLwiN (Version 2.35+; see http://www.bristol.ac.uk/cmm/software/

mlwin), Mplus (Version 7.4+; see http://www.statmodel.com), and the `gllamm` command for Stata (www.gllamm.org). The online documentation for each of these packages provides worked examples of implementing these "hybrid" approaches (e.g., type "`help mixed#sampling`" in the Stata Viewer). Importantly, all of these tools are capable of implementing either model-based approaches or "hybrid" approaches, depending on how the survey weights are used.

7. Directions for Future Software Development

First, considering design-based approaches, additional software options enabling variance estimation for quantile estimates are still needed, where BRR and bootstrap methods have been shown to produce the best confidence interval coverage (Kovar et al. 1988). Techniques for accounting for complex sample design features when evaluating the goodness-of-fit of various regression models in the design-based framework (e.g., Archer et al. 2007) also need theoretical and computational development. Furthermore, methods for assessing regression diagnostics need further theoretical development (especially for generalized linear models), and state-of-the-art diagnostic methods for linear regression models need to be more widely incorporated in survey analysis software. Finally, more research needs to consider whether there are better approaches to estimating the design-based degrees of freedom associated with a given variance estimate when forming confidence intervals, and implementation of alternative approaches (e.g., Valliant and Rust 2010) in existing software is still needed.

Second, considering model-based and "hybrid" approaches, the literature currently lacks a coherent theoretical framework enabling hypothesis testing for the variance components in a multilevel model estimated using pseudo-maximum likelihood estimation (see Zhang and Lin 2008 for a review of these methods). The recent work by Lumley and Scott (2015) needs to be adapted to these types of tests based on multilevel models estimated using sampling weights. Also important in this area will be the development of diagnostics for fitted multilevel models (Claeskens 2013) that recognize complex sampling features. Finally, there is still work to be done in assessing optimal approaches for fitting multilevel models to longitudinal survey data (Thompson 2015); for example, should time-varying weights be computed to adjust for differential attrition at different waves? Or should only cases with complete data be analyzed when fitting the multilevel models (Heeringa et al. 2017, Chapter 11)? Empirical and theoretical developments in this area will be important moving forward.

Finally, this review has not touched on statistical analysis approaches involving item-missing survey data, and how complex sampling features should be accounted for in this context. Briefly, initial work in this area suggested that models for imputing item-missing values should include the complex sample design features as covariates, similar to some of the model-based approaches discussed above (Reiter et al. 2006). More recently, methods have been developed for simulating synthetic populations, given the complex sampling features available for a sample, and then imputing missing values using straightforward methods in these simulated populations prior to making population inferences (Zhou et al. 2016b; Zhou et al. 2016c; see also Zhou et al. 2016a, for example R code). Alongside these developments using model-based imputation methods, Kim and Fuller (2004) and Kim and Shao (2014) have developed fractional hot-deck imputation techniques for complex sample survey data sets that may offer efficiency advantages over other competing imputation approaches. These approaches have been implemented in the SURVEYIMPUTE procedure of the SAS software (Version 9.4). Future research should consider the competing benefits and costs of the simulation-based imputation approaches and the fractional imputation approaches in terms of computational costs and the efficiency of the finite population estimates produced.

8. References

Agresti, A. and B. A. Coull. 1998. “Approximate Is Better than ‘Exact’ for Interval Estimation of Binomial Proportions.” *American Statistician* 52: 119–126. Doi: https://doi.org/10.1080/00031305.1998.10480550.

Archer, K.J., S. Lemeshow, and D.W. Hosmer. 2007. “Goodness-of-fit Tests for Logistic Regression Models When Data are Collected using a Complex Sampling Design.” *Computational Statistics and Data Analysis* 51: 4450–4464. Doi: https://doi.org/10.1016/j.csda.2006.07.006.

Asparouhov, T. 2006. “General Multi-level Modeling with Sampling Weights.” *Communications in Statistics—Theory and Methods* 35: 439–460. Doi: https://doi.org/10.1080/03610920500476598.

Asparouhov, T. and B. Muthén. 2007. “Testing for Informative Weights and Weights Trimming in Multivariate Modelling with Survey Data.” In Proceedings of the Survey Research Methods Section of the American Statistical Association, 2007, Salt Lake City, Utah, 3394–3399. Available at: https://www.statmodel.com/download/JSM2007000745.pdf (accessed April 14, 2017).

Asparouhov, T. and B. Muthén. 2010. “Resampling Methods in Mplus for Complex Survey Data.” *Mplus Technical Report, May 4, 2010*. Available at: https://www.statmodel.com/download/Resampling_Methods5.pdf (Accessed October 10, 2016).

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, and R. Tourangeau. 2013. “Summary Report of the AAPOR Task Force on Non-probability Sampling.” *Journal of Survey Statistics and Methodology* 1: 90–143. Doi: https://doi.org/10.1093/jssam/smt008.

Barnighausen, T., J. Bor, S. Wandira-Kazibwe, and D. Canning. 2011. “Correcting HIV Prevalence Estimates for Survey Nonparticipation using Heckman-type Selection Models.” *Epidemiology* 22: 27–35. Doi: 10.1097/EDE.0b013e3181ffa201.

Bean, J.A. 1975. “Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution.” In *Vital and Health Statistics: Series 2, Data Evaluation and Methods Research* 65: i–iv.

Beaumont, J.F. 2005. “On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment.” *Survey Methodology* 31: 227–231.

Beaumont, J.F. 2008. “A New Approach to Weighting and Inference in Sample Surveys.” *Biometrika* 95: 539–553. Doi: https://doi.org/10.1093/biomet/asn028.

Bethlehem, J.G. 2002. “Weighting Nonresponse Adjustments Based on Auxiliary Information.” In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 275–288. New York: Wiley.

Binder, D.A. 1981. “On the Variances of Asymptotically Normal Estimators for Complex Surveys.” *Survey Methodology* 7: 157–170.

Binder, D.A. 1983. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51: 279–292. Doi: 10.2307/1402588.

Binder, D.A. and G.R. Roberts. 2003. “Design-based and Model-based Methods for Estimating Model Parameters.” In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner, 29–48. Chichester, West Sussex: Wiley.

Bollen, K.A., P.P. Biemer, A.F. Karr, S. Tueller, and M.E. Berzofsky. 2016. “Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis.” *Annual Review of Statistics and Its Application* 3: 375–392. Doi: https://doi.org/10.1146/annurev-statistics-011516-012958.

Breckling, J.U., R.L. Chambers, A.H. Dorfman, S.M. Tam, and A.H. Welsh. 1994. “Maximum Likelihood Inference from Sample Survey Data.” *International Statistical Review* 62: 349–363. Doi: 10.2307/1403766.

Brick, J.M. 2013. “Unit Nonresponse and Weighting Adjustments: A Critical Review.” *Journal of Official Statistics* 29: 329–353. Doi: https://doi.org/10.2478/jos-2013-0026.

Campbell, C. and M. Meyer. 1978. “Some Properties of T Confidence Intervals for Survey Data.” In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 437–442. Available at: https://ww2.amstat.org/sections/srms/Proceedings/papers/1978_089.pdf (accessed April 14, 2017).

Canty, A.J. and A.C. Davison. 1999. “Resampling-based Variance Estimation for Labour Force Surveys.” *The Statistician* 48: 379–391. Doi: 10.1111/1467-9884.00196.

Carle, A.C. 2009. “Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations.” *BMC Medical Research Methodology* 9(49). Doi: https://doi.org/10.1186/1471-2288-9-49.

Cassel, C., C.-E. Särndal, and J. Wretman. 1983. “Some Uses of Statistical Models in Connection with the Nonresponse Problem.” In *Incomplete Data in Sample Surveys*, edited by W.G. Madow and I. Olkin, 143–160. New York: Academic Press.

Chambers, R.L., A.H. Dorfman, and S. Wang. 1998. “Limited Information Likelihood Analysis of Survey Data.” *Journal of the Royal Statistical Society (Series B)* 60: 397–411. Doi: 10.1111/1467-9868.00132.

Chambers, R.L. and C.J. Skinner (Editors). 2003. *Analysis of Survey Data*. New York: John Wiley and Sons.

Chambless, L.E. and K.E. Boyle. 1985. “Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models.” *Communications in Statistics-Theory and Methods* 14: 1377–1392. Doi: https://doi.org/10.1080/03610928508828982.

Chantala, K., D. Blanchette, and C.M. Suchindran. 2011. “Software to Compute Sampling Weights for Multilevel Analysis.” Technical Report, Carolina Population Center, UNC at Chapter Hill. Available at http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights (accessed January 30, 2018).

Claeskens, G. 2013. “Lack of Fit, Graphics, and Multilevel Model Diagnostics.” In *The SAGE Handbook of Multilevel Modeling*, edited by M.A. Scott, J.S. Simonoff, and B.D. Marx, 425–444. Los Angeles: SAGE Publications.

Dean, N. and M. Pagano. 2015. “Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys.” *Journal of Survey Statistics and Methodology* 3: 484–503. Doi: https://doi.org/10.1093/jssam/smv024.

Deville, J.C. and C.-E. Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87: 376–382. Doi: https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.C., C.-E. Särndal, and O. Sautory. 1993. “Generalized Raking Procedures in Survey Sampling.” *Journal of the American Statistical Association* 88: 1013–1020. Doi: https://doi.org/10.1080/01621459.1993.10476369.

DuMouchel, W.H. and G.J. Duncan. 1983. “Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples.” *Journal of the American Statistical Association* 78: 535–543. Doi: https://doi.org/10.1080/01621459.1983.10478006.

Ekholm, A. and S. Laaksonen. 1991. “Weighting Via Response Modeling in the Finnish Household Budget Survey.” *Journal of Official Statistics* 7: 325–337.

Elliott, M.R. and R.J. Little. 2000. “Model-based Alternatives to Trimming Survey Weights.” *Journal of Official Statistics* 16: 191–210.

Elliott, M.R. and R. Valliant. 2017. “Inference for Nonprobability Samples.” *Statistical Science* 32: 249–264. Doi: 10.1214/16-STS598.

Eltinge, J.L. and I.S. Yansaneh. 1997. “Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey.” *Survey Methodology* 23: 33–40.

Feder, M. 2011. “Fitting Regression Models to Complex Survey Data – Gelman’s Estimator Revisited.” In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland, August 2011. Available at: http://2011.isiproceedings.org/papers/950551.pdf (accessed January 30, 2018).

Flores-Cervantes, I. and J.M. Brick. 2016. “Nonresponse Adjustments with Misspecified Models in Stratified Designs.” *Survey Methodology* 42: 161–177.

Francisco, C.A. and W.A. Fuller. 1991. “Quantile Estimation with a Complex Survey Design.” *The Annals of Statistics* 19: 454–469. Doi: http://www.jstor.org/stable/2241867.

Frankel, M.R. 1971. *Inference from Survey Samples: an Empirical Investigation*. Institute for Social Research, University of Michigan, Ann Arbor, MI, USA.

Gagne, C., G. Roberts, and L.-A. Keown. 2014. “Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software.” *Statistics Canada: The Research Data Centres Information and Technical Bulletin, August 7, 2014*. Available at: http://www.statcan.gc.ca/pub/12-002-x/2014001/article/11901-eng.htm (accessed January 30, 2018).

Gelman, A. 2007. “Struggles with Survey Weighting and Regression Modeling.” *Statistical Science* 22: 153–164. Doi: 10.1214/088342306000000691.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. *Bayesian Data Analysis* (*2nd Edition*). Boca Raton, FL: Chapman & Hall/CRC.

Godambe, V.P. and M.E. Thompson. 2009. “Estimating Functions and Survey Sampling.” *Handbook of Statistics Vol 29B (Sample Surveys: Inference and Analysis)*: 83–101. Doi: https://doi.org/10.1016/S0169-7161(09)00226-0.

Grau, E., F. Potter, S. Williams, and N. Diaz-Tena. 2006. “Nonresponse Adjustment Using Logistic Regression: To Weight or Not to Weight?” In *Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria, VA, 2006*, 3073–3080. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.586.3263&rep=rep1&type=pdf (accessed January 30, 2018).

Grilli, L. and M. Pratesi. 2004. “Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs.” *Survey Methodology* 30: 93–104.

Groves, R.M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly* 70: 646–675. Doi: https://doi.org/10.1093/poq/nfl033.

Hansen, M.H., W.G. Madow, and B.J. Tepping. 1983. “An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys.” *Journal of the American Statistical Association* 78: 776–793. Doi: https://doi.org/10.1080/01621459.1983.10477018.

Hartley, H.O. and J.N.K. Rao. 1968. “A New Estimation Theory for Sample Surveys.” *Biometrika* 55: 547–557. Doi: https://doi.org/10.1093/biomet/55.3.547.

Haziza, D. and J.F. Beaumont. 2007. “On the Construction of Imputation Classes in Surveys.” *International Statistical Review* 75: 25–43. Doi: 10.1111/j.1751-5823.2006.00002.x.

Haziza, D. and J.F. Beaumont. 2017. “Construction of Weights in Surveys: A Review.” *Statistical Science* 32: 206–226. Doi: 10.1214/16-STS608.

Heeringa, S.G., B.T. West, and P.A. Berglund. 2017. *Applied Survey Data Analysis, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.

Holt, D. and T.M.F. Smith. 1979. “Post Stratification.” *Journal of the Royal Statistical Society, Series A (General)* 142: 33–46. Doi: http://www.jstor.org/stable/2344652.

Hosmer, D.W., S. Lemeshow, and X. Sturdivant. 2013. *Applied Logistic Regression, Third Edition*. New York, NY: Wiley.

Judkins, D.R. 1990. “Fay’s Method for Variance Estimation.” *Journal of Official Statistics* 6: 223–239.

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

Kaplan, D. and A.J. Ferguson. 1999. “On the Utilization of Sample Weights in Latent Variable Models.” *Structural Equation Modeling: A Multidisciplinary Journal* 6: 305–321. Doi: https://doi.org/10.1080/10705519909540138.

Kim, J.K. and W.A. Fuller. 2004. “Fractional Hot Deck Imputation.” *Biometrika* 91: 559–578. Doi: https://doi.org/10.1093/biomet/91.3.559.

Kim, J.K. and J. Shao. 2014. *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL: CRC Press.

Kim, J.K. and M. Park. 2010. “Calibration Estimation in Survey Sampling.” *International Statistical Review* 78: 21–39. Doi: 10.1111/j.1751-5823.2010.00099.x.

Kish, L. 1965. *Survey Sampling*. New York, NY: Wiley.

Kish, L. and M.R. Frankel. 1968. "Balanced Repeated Replication for Analytical Statistics." In *Proceedings of the Social Statistics Section of the American Statistical Association,* 1968, 2–10. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y1968/Balanced%20Repeated%20Replications%20For%20Analytical%20Statistics.pdf (accessed January 30, 2018).

Kish, L. and M.R. Frankel. 1970. "Balanced Repeated Replications for Standard Errors." *Journal of the American Statistical Association* 65: 1071–1094. Doi: https://doi.org/10.1080/01621459.1970.10481145.

Kish, L. and M.R. Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society. Series B (Methodological)* 36: 1–37. Doi: http://www.jstor.org/stable/2984767.

Kolenikov, S. 2014. "Calibrating Survey Data using Iterative Proportional Fitting (Raking)." *The Stata Journal* 14: 22–59.

Kolenikov, S. 2010. "Resampling Variance Estimation for Complex Survey Data." *Stata Journal* 10: 165–199.

Kott, P.S. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32: 133.

Kott, P.S. 2011. "A Nearly Pseudo-Optimal Method for Keeping Calibration Weights From Falling Below Unity In The Absence Of Nonresponse Or Frame Errors." *Pakistan Journal of Statistics* 27: 391–396.

Kott, P.S. 2012. "Why One Should Incorporate the Design Weights When Adjusting for Unit Nonresponse Using Response Homogeneity Groups." *Survey Methodology* 38: 95–99.

Kott, P.S. and D. Liao. 2012. "Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-Weighting Routine." *Survey Research Methods* 6: 105–111. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5076.

Korn, E.L. and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York, NY: Wiley.

Kovačević, M.S. and S.N. Rai. 2003. "A Pseudo Maximum Likelihood Approach to Multilevel Modelling of Survey Data." *Communications in Statistics-Theory and Methods* 32: 103–121. Doi: https://doi.org/10.1081/STA-120017802.

Kovar, J.G., J.N.K. Rao, and C.F.J. Wu. 1988. "Bootstrap and Other Methods to Measure Errors in Survey Estimates." *Canadian Journal of Statistics* 16: 25–45. Doi: 10.2307/3315214.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R.M. Groves, and T.E. Raghunathan. 2010. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173: 389–407. Doi: 10.1111/j.1467-985X.2009.00621.x.

Krewski, D. and J.N.K. Rao. 1981. "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods." *The Annals of Statistics* 9: 1010–1019. Doi: http://www.jstor.org/stable/2240615.

Lavallée, P. and J.F. Beaumont. 2016. "Weighting: Principles and Practicalities." In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T.W. Smith, and Y. Fu, 460–476. London: Sage.

Le Guennec, J., and O. Sautory. 2002. "CALMAR 2: Une nouvelle version de la macro Calmar de redressment d'echantillon par calage." Actes des Journeés de Méthodologie Statistique, INSEE, Paris. Available in French at: http://jms.insee.fr/files/documents/2002/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF (accessed January 30, 2018).

Lemeshow, S. and P. Levy. 1978. "Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Units per Stratum—A Comparison of Balanced Replication and Jackknife Techniques." *Journal of Statistical Computation and Simulation* 8: 191–205. Doi: https://doi.org/10.1080/00949657908810266.

Lessler, J.T. and W.D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. Wiley.

Li, J. and R. Valliant. 2009. "Survey Weighted Hat Matrix and Leverages." *Survey Methodology* 35: 15–24.

Li, J. and R. Valliant. 2011a. "Linear Regression Influence Diagnostics for Unclustered Survey Data." *Journal of Official Statistics* 27: 99–119.

Li, J. and R. Valliant. 2011b. "Detecting Groups of Influential Observations in Linear Regression using Survey Data: Adapting the Forward Search Method." *Pakistan Journal of Statistics* 27: 507–528.

Li, J. and R. Valliant. 2015. "Linear Regression Diagnostics in Cluster Samples." *Journal of Official Statistics* 31: 61–75. https://doi.org/10.1515/jos-2015-0003.

Liao, D. and R. Valliant. 2012a. "Variance Inflation Factors in the Analysis of Complex Survey Data." *Survey Methodology* 38: 53–62.

Liao, D. and R. Valliant. 2012b. "Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data." *Survey Methodology* 38: 189–202.

Little, R.J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54: 139–157. Doi: http://www.jstor.org/stable/1403140.

Little, R.J.A. 2003. "The Bayesian Approach to Sample Survey Inference." In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner, 49–57. Chichester, West Sussex: Wiley.

Little, R.J.A. 2004. "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99: 546–556. Doi: https://doi.org/10.1198/016214504000000467.

Little, R.J.A. 2006. "Calibrated Bayes: A Bayes/Frequentist Roadmap." *The American Statistician* 60: 213–223. Doi: https://doi.org/10.1198/000313006X117837.

Little, R.J.A. 2011. "Calibrated Bayes, for Statistics in General, and Missing Data in Particular." *Statistical Science* 26: 162–186. Doi: 10.1214/10-STS318.

Little, R.J.A. 2012. "Calibrated Bayes: An Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder)." *Journal of Official Statistics* 28: 309–372.

Little, R.J.A. 2015. "Calibrated Bayes, An Inferential Paradigm for Official Statistics in the Era of Big Data." *Statistical Journal of the IAOS* 31: 555–563. Doi: https://doi.org/10.3233/SJI-150944.

Little, R.J.A. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Little, R.J.A. and H. Zheng. 2007. “The Bayesian Approach to the Analysis of Finite Population Surveys.” *Bayesian Statistics* 8: 1–20.

Lohr, S. 2009. *Sampling: Design and Analysis*, *Second Edition*. Boston, MA: Cengage Learning.

Lohr, S. 2014. “Design Effects for a Regression Slope in a Cluster Sample.” *Journal of Survey Statistics and Methodology* 2: 97–125. Doi: https://doi.org/10.1093/jssam/smu003.

Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. New York, NY: Wiley.

Lumley, T. and A. Scott. 2013. “Partial Likelihood Ratio Tests for the Cox Model under Complex Sampling.” *Statistics in Medicine* 32: 110–123. Doi: https://doi.org/10.1002/sim.5492.

Lumley, T. and A. Scott. 2014. “Tests for Regression Models Fitted to Survey Data.” *Australian & New Zealand Journal of Statistics* 56: 1–14. Doi: https://doi.org/10.1111/anzs.12065.

Lumley, T. and A. Scott. 2015. “AIC and BIC for Modeling with Complex Survey Data.” *Journal of Survey Statistics and Methodology* 3: 1–18. Doi: https://doi.org/10.1093/jssam/smu021.

Lumley, T. and A. Scott. 2017. “Fitting Regression Models to Survey Data.” *Statistical Science* 32: 265–278. Doi: https://10.1214/16-STS605.

Lundström, S. and C.E. Särndal. 1999. “Calibration as a Standard Method for Treatment of Nonresponse.” *Journal of Official Statistics* 15: 305–327.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2012. *The BUGS book: A Practical Introduction to Bayesian Analysis*. CRC press.

Morel, G. 1989. “Logistic Regression under Complex Survey Designs.” *Survey Methodology* 15: 203–223.

Muthén, B.O. and A. Satorra. 1995. “Complex Sample Data in Structural Equation Modeling.” *Sociological Methodology* 25: 267–316. Doi: https://doi.org/10.2307/271070.

Nordberg, L. 1989. “Generalized Linear Modeling of Sample Survey Data.” *Journal of Official Statistics* 5: 223–239.

Oberski, D.L. 2014. “lavaan.survey: An R package for Complex Survey Analysis of Structural Equation Models.” *Journal of Statistical Software* 57: 1–27. Doi: https://doi.org/10.18637/jss.v057.i01.

Oh, H.L. and F.J. Scheuren. 1983. “Weighting Adjustment for Unit Nonresponse.” In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin, 143–184. New York: Academic Press.

Orchard, T. and M.A. Woodbury. 1972. “A Missing Information Principle: Theory and Applications.” Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, 697–715. University of California Press: Berkeley, CA. Available at: https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200514117 (accessed January 30, 2018).

Owen, A.B. 2001. *Empirical Likelihood*. New York: Chapman & Hall.

Pacifico, D. 2014. “sreweight: A Stata Command to Reweight Survey Data to External Totals.” *The Stata Journal* 14: 4–21.

Peress, M. 2010. "Correcting for Survey Nonresponse using Variable Response Propensity." *Journal of the American Statistical Association* 105: 1418–1430. Doi: https://doi.org/10.1198/jasa.2010.ap09485.

Pfeffermann, D. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61: 317–337. Doi: https://doi.org/10.2307/1403631.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why Model? Why Is It a Problem? How Can We Approach It?" *Survey Methodology* 37: 115–136.

Pfeffermann, D., F.A.D.S. Moura, and P.L.D.N. Silva. 2006. "Multi-level Modelling Under Informative Sampling." *Biometrika* 93: 943–959. Doi: https://doi.org/10.1093/biomet/93.4.943.

Pfeffermann, D. and A. Sikov. 2011. "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information." *Journal of Official Statistics* 27: 181–209.

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 (Statistical Methodology)* 60: 23–40. Doi: https://doi.org/10.1111/1467-9868.00106.

Pfeffermann, D. and M. Sverchkov. 2009. "Inference Under Informative Sampling." In *Handbook of Statistics – Sample Surveys: Inference and Analysis (Volume 29, Part B)*, edited by V.N. Gudivada, V.V. Raghavan, V. Govindaraju, and C.R. Rao, 455–487.

Potter, F.J. 1990. "A Study of Procedures to Identify and Trim Extreme Sampling Weights." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 225–230. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1990_034.pdf (accessed January 30, 2018).

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: https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J.N.K. 2005. "Interplay Between Sample Survey Theory and Practice: An Appraisal." *Survey Methodology* 31: 117–138.

Rao, J.N.K. and J. Shao. 1999. "Modified Balanced Repeated Replication for Complex Survey Data." *Biometrika* 86: 403–415. Doi: https://doi.org/10.1093/biomet/86.2.403.

Rao, J.N.K. and A.J. Scott. 1981. "The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-way Tables." *Journal of the American Statistical Association* 76: 221–230. Doi: https://doi.org/10.2307/2287815.

Rao, J.N.K. and A.J. Scott. 1984. "On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics* 12: 46–60. Doi: http://dx.doi.org/10.1214/aos/1176346391.

Rao, J.N.K. and A.J. Scott. 1987. "On Simple Adjustments to Chi-square Tests with Sample Survey Data." *The Annals of Statistics* 15: 385–397. Doi: https://doi.org/10.1214/aos/1176350273.

Rao, J.N.K. and C.F.J. Wu. 1985. "Inference from Stratified Samples: Second-order Analysis of Three Methods for Nonlinear Statistics." *Journal of the American Statistical Association* 80: 620–630. Doi: https://doi.org/10.2307/2288478.

Rao, J.N.K. and C.F.J. Wu. 1987. "Methods for Standard Errors and Confidence Intervals from Sample Survey Data: Some Recent Work." *Bulletin of the International Statistical Institute* 3: 5–21.

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83: 231–241. Doi: https://doi.org/10.2307/2288945.

Reiter, J.P., T.E. Raghunathan, and S.K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32: 143–149.

Roberts, G., J.N.K. Rao, and S. Kumar. 1987. "Logistic Regression Analysis of Sample Survey Data." *Biometrika* 74: 1–12. Doi: https://doi.org/10.2307/2336016.

Rubin, D.B. 1985. "The Use of Propensity Scores in Applied Bayesian Inference." In *Bayesian Statistics 2*, edited by J.M. Bernardo, M.H. Degroot, D.V. Lindley, and A.F.M. Smith, 463–472. Elsevier Science Publishers B.V.

Ryan, B.L., J. Koval, B. Corbett, A. Thind, M.K. Campbell, and M. Stewart. 2015. "Assessing the Impact of Potentially Influential Observations in Weighted Logistic Regression." *The Research and Data Centres Information and Technical Bulletin (Statistics Canada)* 7. Available at: http://www.statcan.gc.ca/pub/12-002-x/2015001/article/14147-eng.htm (accessed January 30, 2018).

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag Inc.

Särndal, C.E. and B. Swensson. 1987. "A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse." *International Statistical Review* 55: 279–294. Doi: https://doi.org/10.2307/1403406.

Shah, B.V., M.M. Holt, and R.E. Folsom. 1977. "Inference about Regression Models from Sample Survey Data." *Bulletin of the International Statistical Institute* 47: 43–57.

Shao, J. 2003. "Impact of the Bootstrap on Sample Surveys." *Statistical Science* 18: 191–198.

Shao, J. and R.R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91: 1278–1288. Doi: https://doi.org/10.2307/2291746.

Shao, J. and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Si, Y., N.S. Pillai, and A. Gelman. 2015. "Bayesian Nonparametric Weighted Sampling Inference." *Bayesian Analysis* 10: 605–625. Doi: http://dx.doi.org/10.1214/14-BA924.

Sitter, R.R. and C. Wu. 2001. "A Note on Woodruff Confidence Intervals for Quantiles." *Statistics & Probability Letters* 52: 353–358. Doi: https://doi.org/10.1016/S0167-7152(00)00207-8.

Skinner, C.J. and D.J. Holmes. 2003. "Random Effects Models for Longitudinal Survey Data." Chapter 14 in *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner. John Wiley and Sons.

Skinner, C.J., D. Holt, and T.F. Smith. 1989. *Analysis of Complex Surveys*. John Wiley & Sons.

Stapleton, L.M. 2002. "The Incorporation of Sample Weights into Multilevel Structural Equation Models." *Structural Equation Modeling* 9: 475–502. Doi: https://doi.org/10.1207/S15328007SEM0904_2.

Stapleton, L.M. 2006. "An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data." *Structural Equation Modeling* 13: 28–58. Doi: https://doi.org/10.1207/s15328007sem1301_2.

Stapleton, L.M. and Y. Kang. 2016. "Design Effects of Multilevel Estimates From National Probability Samples." *Sociological Methods & Research*, available at http://journals.sagepub.com/doi/abs/10.1177/0049124116630563 (accessed January 30, 2018). Doi: https://doi.org/10.1177/0049124116630563.

Thompson, M.E. 2015. "Using Longitudinal Complex Survey Data." *Annual Review of Statistics and Its Application* 2: 305–320. Doi: https://doi.org/10.1146/annurev-statistics-010814-020403.

Toth, D. 2017. "rpms: An R Package for Modeling Survey Data with Regression Trees." Available at https://cran.r-project.org/web/packages/rpms/vignettes/rpms_2017_02_10.pdf (accessed January 1, 2018).

Valliant, R. 2004. "The Effect of Multiple Weighting Steps on Variance Estimation." *Journal of Official Statistics* 20(1): 1–18.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: a Prediction Approach*. New York: Wiley.

Valliant, R. and K.F. Rust. 2010. "Degrees of Freedom Approximations and Rules-of-Thumb." *Journal of Official Statistics* 26: 585–602.

West, B.T. 2009. "A Simulation Study of Alternative Weighting Class Adjustments for Nonresponse when Estimating a Population Mean from Complex Sample Survey Data." In *Proceedings of the section on Survey Research Methods: Joint Statistical Meetings*, 4920–4933. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2009/Files/305394.pdf (accessed January 30, 2018).

West, B.T., L. Beer, W. Gremel, J. Weiser, C. Johnson, S. Garg, and J. Skarbinski. 2015. "Weighted Multilevel Models: A Case Study." *American Journal of Public Health* 105: 2214–2215. Doi: https://dx.doi.org/10.2105%2FAJPH.2015.302842.

West, B.T., P.A. Berglund, and S.G. Heeringa. 2008. "A Closer Examination of Subpopulation Analysis of Complex-Sample Survey Data." *The Stata Journal* 8: 520–531.

West, B.T. and S.E. McCabe. 2017. "Alternative Approaches to Assessing Nonresponse Bias in Longitudinal Survey Estimates: An Application to Substance Use Outcomes among Young Adults in the U.S." *American Journal of Epidemiology* 185: 591–600. Doi: https://doi.org/10.1093/aje/kww115.

West, B.T., J.W. Sakshaug, and G.A.S. Aurelien. 2016. "How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data?" *PLoS ONE* 11. Doi: https://doi.org/10.1371/journal.pone.0158120.

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

Woodruff, R.S. 1952. "Confidence Intervals for Medians and other Position Measures." *Journal of the American Statistical Association* 47: 635–646. Doi: https://doi.org/10.2307/2280781.

Wu, Y.Y. and W.A. Fuller. 2006. "Estimation of Regression Coefficients with Unequal Probability Samples." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, 3892–3899. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2006/Files/JSM2006-000807.pdf (accessed January 30, 2018).

Wun, L.M., T.M. Ezzati-Rice, N. Diaz-Tena, and J. Greenblatt. 2007. "On Modeling Response Propensity for Dwelling Unit (DU) Level Non-response Adjustment in the Medical Expenditure Panel Survey (MEPS)." *Statistics in Medicine* 26: 1875–1884. Doi: https://doi.org/10.1002/sim.2809.

Zangeneh, S.Z. and R.J. Little. 2015. "Bayesian Inference for the Finite Population Total from a Heteroscedastic Probability Proportional to Size Sample." *Journal of Survey Statistics and Methodology* 3: 162–192. Doi: https://doi.org/10.1093/jssam/smv002.

Zhang, D. and X. Lin. 2008. "Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics." In *Random Effect and Latent Variable Model Selection*, edited by D.B. Dunson. Springer Lecture Notes in Statistics, 192.

Zheng, H. and R.J. Little. 2003. "Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples." *Journal of Official Statistics* 19: 99–117.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016a. "Synthetic Multiple-Imputation Procedure for Multistage Complex Samples." *Journal of Official Statistics* 32: 231–256. Doi: https://doi.org/10.1515/JOS-2016-0011.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016b. "Multiple Imputation in Two-Stage Cluster Samples Using the Weighted Finite Population Bayesian Boostrap." *Journal of Survey Statistics and Methodology* 4: 139–170. Doi: https://doi.org/10.1093/jssam/smv031.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016c. "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation." *Biometrics* 72: 242–252. Doi: https://10.1111/biom.12413.

Zhu, M. 2014. "Analyzing Multilevel Models with the GLIMMIX Procedure." Paper SAS026-2014. Cary, NC: SAS Institute, Inc.

Received July 2017
Revised December 2017
Accepted February 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 753–784, http://dx.doi.org/10.2478/JOS-2018-0035

Generalized Method of Moments Estimators for Multiple Treatment Effects Using Observational Data from Complex Surveys

Bin Liu[1], *Cindy Long Yu*[2], *Michael Joseph Price*[2], *and Yan Jiang*[3]

In this article, we consider a generalized method moments (GMM) estimator to estimate treatment effects defined through estimation equations using an observational data set from a complex survey. We demonstrate that the proposed estimator, which incorporates both sampling probabilities and semiparametrically estimated self-selection probabilities, gives consistent estimates of treatment effects. The asymptotic normality of the proposed estimator is established in the finite population framework, and its variance estimation is discussed. In simulations, we evaluate our proposed estimator and its variance estimator based on the asymptotic distribution. We also apply the method to estimate the effects of different choices of health insurance types on healthcare spending using data from the Chinese General Social Survey. The results from our simulations and the empirical study show that ignoring the sampling design weights might lead to misleading conclusions.

Key words: Observational data; propensity score; semiparametric; treatment effects; two-phase sampling design.

1. Introduction

Observational data from a complex survey has increasingly become useful for causal inference because they can provide timely results with low cost. Survey data contains information on the treatment selections, which enables us to estimate the effects of treatments that cannot feasibly be evaluated with a randomized trial. In a survey, a treatment can be broadly defined as one of the survey questions, for example whether or not an individual has quit smoking, how often an individual does a physical exam, or what types of health insurance an individual has chosen. We can use the existing survey data to estimate effects of those treatments on health care spending, even if we cannot randomize the health behavior or the health insurance enrollment of an individual. Also because a well-designed survey sample is often a good representative of the target population, the treatment effect results can be generalized to the target population level if the survey weights are appropriately incorporated. Propensity score methods are well-established statistical methods to remove treatment selection bias in observational studies if the

[1] Ant Financial, Hangzhou, China. Email: lb88701@alibaba-inc.com
[2] Iowa State University – Department of Statistics, Ames, Iowa 50011, United States. Emails: cindyyu@iastate.edu and michael.price@pioneer.com
[3] Renmin University of China - School of Statistics and The Center for Applied Statistics, Beijing, China. Email: jiangyan@ruc.edu.cn

selection probability model is correctly specified (Rosenbaum and Rubin 1983). Many observational data sets have multiple treatment options. In order to handle the complexity in multiple treatment groups, theoretical results support using the inverse of the estimated treatment selection probabilities as weights to adjust for selection bias and attain asymptotic efficiency (Hahn 1998; Hirano et al. 2003; Cattaneo 2010). This kind of estimator is called inverse probability weighted (IPW) estimator, and the estimated selection probabilities are called propensity scores. We also consider IPW estimators in this article to address the potential confounding in observational studies. However, it is very common that people ignore survey weights in observational data when using the IPW estimators yet claim that the estimated treatment effects are generalizable to the target population, causing misleading guidance in causal inference. Failure to properly account for the complex survey design may lead to biased treatment effect estimates and incorrect variance estimation.

Several authors have emphasized the importance of incorporating survey weights in their IPW estimators, for example DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015). The general idea is to multiply the inverse of the estimated propensity scores by the sampling design weights. However most of the papers, except for Ashmead (2014), do not provide theoretical justification for such survey adjusted estimators, and variance estimation is seldom discussed. Yu et al. (2013) proposes a semiparametric two-phase regression estimator to estimate marginal mean treatment effects in observational data sets from complex survey designs. This article considers a more general set up in which parameters of interest are defined through estimation equations, and uses the generalized method of moments (GMM) for parameter estimation. Similarly to Yu et al. (2013), this article draws a connection between the two-phase sampling in survey statistics and the estimation of treatment effects from an observational database. The observational data set, denoted as A_1 (with size n), is considered as a first-phase sample from a finite population, according to a known sampling probability π_{1i} for subject i. The second-phase sampling is a partitioning of the first-phase sample (observational data set) into mutually exclusive and self-selected treatment groups, $A_{21}, \ldots, A_{2G}$, where G is the number of treatments. This partitioning in the second-phase can be viewed as a multinomial sampling in survey statistics, and its self-selection probabilities π_{2ig} for subject i into group g ($g = 1, \ldots, G$) can be estimated using the semiparametric approach in Cattaneo (2010).

Our article differs from DuGoff et al. (2014), Zanutto (2006), Ashmead (2014) and Ridgeway et al. (2015) in the following ways. (i) Their papers consider two treatments, while our article deals with multi-level treatment selection. (ii) In their work, the propensity scores are estimated using a parametric linear logistic regression, while our propensity scores, that is π_{2ig} in our situation, are estimated through a semiparametric approach. Thus, our approach should be more robust to the misspecification of the selection probability model. (iii) In their work, the parameters of interest are treatment means. We are interested in estimating treatment specific parameters defined through estimation equations. In addition to providing generality, defining parameters through estimation equations can facilitate variance estimation. For example, if a parameter is a function of means, such as correlation or domain mean (see more details in Subsection 2.1), the variance estimation of GMM estimators for such parameter scan be easily calculated through the sandwich formula associated with the asymptotic variance for a GMM estimator. Ashmead (2014) also utilizes estimation equations in their weighting estimator.

This article also differs from Yu et al. (2013) in the following aspects. We extend Yu et al. (2013), which only focuses on estimating marginal treatment means, to estimate parameters defined through estimation Equations (see $\hat{\boldsymbol{\theta}}_g^{(1)}$ in Subsection 2.3). This article also proposes the second estimator to gain efficiency by incorporating the first phase and second phase means of covariates into the estimation equations (see $\hat{\boldsymbol{\theta}}_g^{(2)}$ in Subsection 2.3). This is similar to the effect of calibrating the second phase means of covariates to their first phase means seen in the optimal two-phase regression estimator discussed in Fuller (2009). Additionally, Yu et al. (2013) assumes sample missing at random (SMAR), which is commonly used in literature, while this article considers population missing at random (PMAR), the framework proposed in Berg et al. (2016) (see more details in Subsection 2.1). It makes sense to use PMAR assumption in the context of casual inference study using observation dataset. We discuss situations when PMAR holds but SMAR fails, and argue that when it happens survey weights should be included in the estimation of π_{2ig}, that is the propensity scores.

We provide theoretical justification for our estimator in a combined framework of a finite population and a superpopulation, and also propose variance estimators. We demonstrate the validity of our estimator through simulation studies, and show that the estimator that ignores the design weights might be subject to biases. We also explore the feasibility of our method using data from the Chinese General Social Survey to estimate the effects of different choices of health insurance types on health care spending. The article is organized as below. Section 2 introduces the framework and the proposed estimators. Section 3 presents an asymptotic normality and variance estimation. Simulation studies and an empirical study are reported in Sections 4 and 5 respectively. Section 6 concludes. Appendix collects the conditions and a sketch of the proof for the main theorem in the article.

2. Proposed Estimators

In this section, we introduce our estimators. Subsection 2.1 discusses the basic set-up, Subsection 2.2 introduces the semiparametric approach for estimating the self-selection probabilities, and Subsection 2.3 proposes the estimators.

2.1. Basic Setup

Let U be a finite population with size N containing $(\mathbf{Y}_i, Z_i)$, where $i = 1, \ldots, N$ indexes a subject, Z_i is a covariate variable, and $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{iG}]^T$ is a vector of potential outcomes for G different treatments depending on covariate Z_i. Let δ_{1i} be the sampling indicator from the survey design, defined by $\delta_{1i} = 1$ if unit i is selected into A_1 and zero otherwise. Let π_{1i} and π_{1ij} be the first and second order inclusion probabilities of the sampling design, defined as,

$$[\pi_{1i}, \pi_{1ij}] = [Prob(\delta_{1i} = 1), Prob(\delta_{1i} = 1, \delta_{1j} = 1)].$$

We assume the sampling weights are appropriately adjusted for any nonresponse. If the weights are adjusted due to nonresponse, the method can be used but with awareness of that the variation from estimating $\hat{\pi}_{1i}$ is not accounted for. Let δ_{2ig} $(g = 1, \ldots, G)$ be the self-selection indicator of subject i selecting treatment g, defined by $\delta_{2ig} = 1$ if unit i selects treatment g and zero otherwise. The self-selection process leads to the partitioning in the second phase. Assume conditioning on a covariate X_i, the self-selection indicators $\boldsymbol{\delta}_{2i} = \left[\delta_{2i1}, \ldots, \delta_{2iG}\right]$ follow a multinomial distribution with probabilities,

$$\pi_{2ig} = Prob(\delta_{2ig} = 1|X_i), \quad \text{for} \quad g = 1, \ldots, G, \tag{1}$$

that is for any subject i,

$$\boldsymbol{\delta}_{2i} = \left[\delta_{2i1}, \ldots, \delta_{2iG}\right] \sim multinomial\left(1; \pi_{2i1}, \ldots, \pi_{2iG}\right),$$

where $\sum_{g=1}^{G} \pi_{2ig} = 1$ for any i, and $\boldsymbol{\delta}_{2i}$ is independent of $\boldsymbol{\delta}_{2j}$ for any subjects $i \neq j$. Here covariates Z_i and X_i can be totally different, or can have overlap. We use separate notations in order to emphasize that the outcome response variables $\mathbf{Y}_i$ and the self-selection indicators $\boldsymbol{\delta}_{2i}$ can depend on different sets of covariates. We discuss how to identify Z_i and X_i practically in Section 4. Both Z_i and X_i have compact supports and are observed in A_1. They are written to be univariate forms in order to reduce notation burden. It is straightforward to extend to multivariate covariates, which are considered in the simulation studies and the empirical study of this article. We suppose that $\left(\mathbf{Y}_i, \delta_{1i}, \delta_{2i}, X_i, Z_i\right); i = 1, \ldots, N$ are identically independently distributed (i.i.d.) generated from a superpopulation ξ.

In the context of simple random sampling, a common missing at random (MAR) assumption is $\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|(X_i, Z_i)$. With this MAR assumption, the selection bias can be removed by applying the propensity score method (Rosenbaum and Rubin 1983; Hirano et al. 2003). However, in the context of a complex survey, unequal probabilities of sampling can complicate the relationship between $\mathbf{Y}_i$, (X_i, Z_i), $\boldsymbol{\delta}_{2i}$ and the sample inclusion indicator δ_{1i}. Even if

$$\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|(X_i, Z_i), \tag{2}$$

holds for a specific superpopulation model,

$$\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|\{(X_i, Z_i), \delta_{1i} = 1\}, \tag{3}$$

may not hold. Following Berg et al. (2016), we call Assumption (2) population missing at random (PMAR), and Assumption (3) sample missing at random (SMAR) to emphasize it depends on the realized sample (that is conditional on $\delta_{1i} = 1$). The SMAR has been used previously (Pfefferman 2011 and Little 1982). However, it is natural to consider PMAR in our context because the mechanisms underlying the selection propensity are conceptualized as inherent characteristics of the subjects in the population. For example, whether or not a person decides to stop smoking heavily depends on this person's perseverance and personality type; what types of insurance a person has chosen depends on the nature of this person's work. In these examples, the self-selection probabilities depend on subjects' inherent characteristics that have nothing to do with whether or not the subjects were selected into the survey that was typically designed for other general purposes. Berg et al. (2016) also provides examples of situations in which PMAR may be considered reasonable. They argue that if both PMAR and SMAR hold, weights are not needed in their imputation model; however if PMAR holds but SMAR fails, it is necessary to include weights to produce consistent estimators. A situation in which PMAR holds while SMAR does not can arise if a design variable omitted from the first phase sample is related to both the sampling inclusion probabilities and the response variable. An example of such a design variable is location in a situation where design strata are functions of location, the location is correlated with the response variable, but the specific location is masked from the analyst because of concerns associated with confidentiality. Using Lemma 1 of Berg et al. (2016), we identify the following two conditions of the sampling and the self-selection mechanisms for which PMAR implies SMAR: (1) $\delta_{1i} \perp \mathbf{Y}_i | (X_i, Z_i), \boldsymbol{\delta}_{2i}$; or (2) $\boldsymbol{\delta}_{2i} \perp \left(\mathbf{Y}_i, \delta_{1i}\right) | \left(X_i, Z_i\right)$. The first condition states that the sampling mechanism is noninformative given covariates (X_i, Z_i) within all the second phase self-selected groups A_{2g}. The second condition states that the self-selection mechanism is independent of either $\mathbf{Y}_i$ or sample inclusion given $\left(X_i, Z_i\right)$. Like Berg et al. (2016), we suggest to include survey weights into the estimation of the self-selection probabilities π_{2ig} when SMAR fails (see Subsection 2.2). In our simulation studies, we consider both noninformative sampling (Condition (1) above holds), and informative sampling (Condition (1) above fails).

The true parameter of interest, $\boldsymbol{\theta}_g^0$ $(g = 1, \ldots, G)$, is a d_θ-dimensional vector satisfying,

$$E\left[\mathbf{m}_g\left(Y_{ig}, Z_i; \boldsymbol{\theta}_g\right)\right] = 0, \tag{4}$$

in the superpopulation, where $\mathbf{m}_g(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$, hereafter denoted as $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ to save space, is an r-dimensional function with $r \geq d_\theta$. Sometimes in addition to treatment marginal means, people might be interested in estimating treatment correlations or treatment domain means. For example in our empirical study, it is interesting to understand whether the correlations between annual medical expenditure and age (or household income) differ significantly across different health insurance type groups; or whether the means of annual medical expenditure for very sick people (domain means) are significantly different across health insurance type groups. The parameter defined through Equation (4) includes treatment correlations and treatment domain means as special cases. More specifically, if the parameter of interest is $\boldsymbol{\theta}_g^0 = \left[P_g, \mu_g, \sigma_g^2, R_g\right]^T$, where $P_g = Prob(Y_{ig} \leq C)$ for some C, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_i)$, then the estimation equation can be defined as,

$$\begin{aligned}\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = \Big[&1_{Y_{ig} \leq C} - P_g, Y_{ig} - \mu_g, (Y_{ig} - \mu_g)^2 - \sigma_g^2, (Y_{ig} - \mu_g)(Z_i - \mu_z) \\ &-R_g\sqrt{\sigma_g^2}\sqrt{\sigma_z^2}, Z_i - \mu_z, (Z_i - \mu_z)^2 - \sigma_z^2\Big]^T.\end{aligned} \tag{5}$$

If the parameter of interest is a treatment specific domain mean, $\boldsymbol{\theta}_g^0 = E(Y_{ig} | Z_i \leq C)$, then the estimation equation can be written as,

$$\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = [Y_{ig} 1_{Z_i \leq C} - \boldsymbol{\theta}_g P_z, 1_{Z_i \leq C} - P_z]^T. \tag{6}$$

Here in both examples, μ_z, σ_z^2 or P_z are all nuisance parameters.

2.2. *Semiparametric Estimation of π_{2ig}*

Because of the difficulty in specifying a parametric form for π_{2ig} and the constraint, $\sum_{g=1}^{G} \pi_{2ig} = 1$, we adopt the semiparametric method in Cattaneo (2010) to estimate π_{2ig}.

Let $\{r_k(X_i)\}_{k=1}^{\infty}$ be a sequence of known approximating functions, and assume that the generalized logit of π_{2ig} can be approximated by $R_K(X_i)^T \gamma_{g,K}$ for $K = 1, 2, \ldots,$ where $R_K(X_i) = [r_1(X_i), r_2(X_i), \ldots, r_K(X_i)]^T$ and $\gamma_{g,K}$ is a vector of the real-valued coefficients of $R_K(X_i)$ for the g-th treatment selection. Let an estimator of the $K \times G$ matrix $\gamma_K = \left[\gamma_{1,K}, \gamma_{2,K}, \ldots, \gamma_{G,K}\right]$ be,

$$\hat{\gamma}_K = [\hat{\gamma}_{1,K}, \hat{\gamma}_{2,K}, \ldots, \hat{\gamma}_{G,K}] = \underset{\gamma_K | \gamma_{1,K} = \mathbf{0}_K}{argmax} \sum_{i \in A_1} b_i w_{1i} \sum_{g=1}^{G} \delta_{2ig} log \left[\frac{e^{R_K(X_i)^T \gamma_{g,K}}}{\sum_{g=1}^{G} e^{R_K(X_i)^T \gamma_{g,K}}} \right], \quad (7)$$

where $w_{1i} = \pi_{1i}^{-1}$, and $\mathbf{0}_K$ represents a $K \times 1$ zero vector used to constrain the sum $\sum_{g=1}^{G} \hat{\pi}_{2ig} = 1$. The estimated self-selection probabilities are

$$\begin{aligned} \hat{\pi}_{2ig} &= \frac{e^{R_K(X_i)^T \hat{\gamma}_{g,K}}}{1 + \sum_{g=2}^{G} e^{R_K(X_i)^T \hat{\gamma}_{g,K}}} \quad \text{for } \mathrm{g} = 2, 3, \ldots, \mathrm{G} \\ &= \left(1 + \sum_{g=2}^{G} e^{R_K(X_i)^T \hat{\gamma}_{g,K}}\right)^{-1} \quad \text{for } \mathrm{g} = 1. \end{aligned} \quad (8)$$

This solution is that of multinomial logistic regression where the probability for each g is approximated using a linear combination of the series of the approximating functions $R_K(X_i)$. Condition B in the Appendix specifies assumptions about $R_K(X_i)$, π_{2ig} and K to ensure $\hat{\pi}_{2ig}$ converges to π_{2ig} fast enough. Examples of $R_K(X_i)$ include a cubic polynomial basis, $R_K(X_i) = \left[1, X_i, X_i^2, X_i^3\right]^T$, or a quadratic spline basis with q knots $R_K(X_i) = \left[1, X_i, X_i^2, (X_i - \kappa_1)_+^2, \ldots, (X_i - \kappa_q)_+^2\right]^T$ where $(t)_+ = t$ if $t > 0$ and 0 otherwise, and $\kappa_1, \ldots, \kappa_q$ are knots in the compact support of X_i.

The b_i in Equation (7) is a user-specified constant that represents the properties of the sampling and the self-selecting mechanism. As discussed in Subsection 2.1, PMAR assumption does not necessarily imply SMAR assumption. If one believes SMAR assumption holds, then one can set $b_i = w_{1i}^{-1}$, which leads to unweighted estimation of $\hat{\pi}_{2ig}$. If SMAR is not satisfied, the unweighted estimator may lead to bias, and setting $b_i = 1$ is one way to attain an approximately unbiased estimator, see Berg et al. (2016) for further discussion of the choice of b_i. If it is difficult to verify SMAR assumption, we suggest to use the conservative choice of $b_i = 1$, which leads to consistent estimators under PMAR without requiring SMAR.

2.3. Proposed Estimators

Since the true parameter of interest $\boldsymbol{\theta}_g^0$ is defined through an estimation equation in (4), the GMM method with propensity scores is used for estimation. It is common that people simply ignore the sampling design weights in the first-phaseand calculate a naive estimator as,

$$\hat{\boldsymbol{\theta}}_g^{nw} = \underset{\boldsymbol{\theta}_g}{\arg\min} \left[\bar{\mathbf{m}}_g^{nw}(\theta_g)\right]^T \left[\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g)\right], \quad (9)$$

where

$$\bar{\mathbf{m}}_g^{nw}\left(\boldsymbol{\theta}_g\right) = \frac{1}{n}\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)}{\hat{\pi}_{2ig}}. \tag{10}$$

Here the superscript 'nw' means no weight. The estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ ignores the sampling weights by applying equal weights to the estimation equations in (10). Although it uses the propensity score $\hat{\pi}_{2ig}$ to adjust for selection biases in the second-phase, it does not account for the survey design in the first-phase, which might lead to biases and incorrect variance estimation when estimating the treatment effect parameters on the population level. This is demonstrated in the simulation studies of Section 4. Both Ridgeway et al. (2015) and Yu et al. (2013) analytically quantify biases caused by ignoring the survey weights in complex survey.

In order to obtain a consistent estimator for $\boldsymbol{\theta}_g^0$, the first-phase survey weights need to be included into the estimation equation. We propose the following GMM estimator,

$$\hat{\boldsymbol{\theta}}_g^{(1)} = \underset{\boldsymbol{\theta}_g}{\arg\min}\ [\bar{\mathbf{m}}_{2\pi g}(\theta_g)]^T[\bar{\mathbf{m}}_{2\pi g}(\theta_g)], \tag{11}$$

where

$$\bar{\mathbf{m}}_{2\pi g}(\theta_g) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\hat{\pi}_{2ig}}. \tag{12}$$

In order to improve efficiency, one can incorporate the information from covariate Z_i that is potentially correlated with the outcome responses into the estimation equations. We propose the second GMM estimator as,

$$\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right) = \underset{(\boldsymbol{\theta}_g,\mu_z)}{\arg\min}\, [\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]^T\hat{\boldsymbol{\Sigma}}_{Hg}^{-1}(\boldsymbol{\theta}_g, \mu_z)[\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)], \tag{13}$$

where

$$\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) = [\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g),\ \ \bar{z}_{2\pi g}(\mu_z),\ \ \bar{z}_{1\pi}(\mu_z)]^T, \tag{14}$$

$$\bar{z}_{2\pi g}(\mu_z) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}\frac{Z_i - \mu_z}{\hat{\pi}_{2ig}} \quad \text{and} \quad \bar{z}_{1\pi}(\mu_z)^T = \frac{1}{N}\sum_{i\in A_1} w_{1i}(Z_i - \mu_z). \tag{15}$$

$\hat{\mu}_z$ is an estimator for the nuisance parameter $\mu_z^0 = E(Z_i)$ and $\hat{\boldsymbol{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$ is the variance estimator of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$, which depends on the joint inclusion probabilities and is defined in (36) of Subsection 3.2. The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (13) is connected to a two phase sampling extension of the design unbiased difference estimator proposed by Särndal et al. (1992) and Breidt et al. (2005) when $\bar{\mathbf{m}}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$.

Remark 1: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $X_i = Z_i$, the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (11) is asymptotically equivalent to the regression estimator proposed in Yu et al. (2013).

Remark 2: The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (13) is more efficient than the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (11). The supplemental file provides a sketch of proof to show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ is the most efficient estimator among the class of estimators $\hat{\boldsymbol{\theta}}_g^{a}$ that use any fixed positive definite matrix $\mathbf{A}$ in the quadratic form minimization, that is $\hat{\boldsymbol{\theta}}_g^{a}$ is defined as

$$\left(\hat{\boldsymbol{\theta}}_g^{a}, \hat{\mu}_z^{a}\right) = \underset{(\boldsymbol{\theta}_g, \mu_z)}{\arg\min}\, [\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]^T \mathbf{A}^{-1}[\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]. \tag{16}$$

If the matrix is an identity matrix, then $\hat{\boldsymbol{\theta}}_g^{a}$ obtained in (16) is equivalent to $\hat{\boldsymbol{\theta}}_g^{(1)}$. Therefore $\hat{\boldsymbol{\theta}}_g^{(1)}$ is expected to be less efficient than $\hat{\boldsymbol{\theta}}_g^{(2)}$, which has been confirmed by the simulation studies in Section 4.

Remark 3: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ corresponds to the optimal two phase regression estimator discussed in Fuller (2009) (Theory 2.2.4). The optimality in Fuller (2009) is in terms of achieving the minimum variance for the limiting distribution of design consistent estimators of the form, $\bar{Y}_{2p,reg} = \bar{Y}_{2\pi} + \left(\bar{Z}_{1\pi} - \bar{Z}_{2\pi}\right)\hat{\beta}$, where $[\bar{Y}_{2\pi}, \bar{Z}_{2\pi}] = \left(\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_2}\left(\pi_{1i}^{-1}\pi_{2i}^{-1}\right)\left[Y_i, Z_i\right]$, $\bar{Z}_{1\pi} = \left(\sum_{i\in A_a} \pi_{1i}^{-1}\right)^{-1} \sum_{i\in A_a} \pi_{1i}^{-1} Z_i$, and π_{1i} (or A_1) and π_{2i} (or A_2) are the first phase and the second phase sampling probabilities (or samples). The efficiency gain of $\bar{Y}_{2p,reg}$ over $\bar{Y}_{2\pi}$ is similar to the effect of calibrating the second phase covariate mean $\bar{Z}_{2\pi g}$ to its first phase mean $\bar{Z}_{1\pi}$.

Remark 4: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $Z_i \equiv 1$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ coincides analytically with the weighting estimator discussed in Ashmead (2014) except that the propensity scores in Ashmead (2014) are estimated using a parametric logistic regression.

Remark 5: When the population mean of Z_i is available, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be easily extended to incorporate this additional information. For example, this case can occur when there are some demographic variables available on the population level. The extended estimator can be obtained by adding one more moment $\bar{z}_N(\mu_z) = N^{-1}\sum_{i\in U}(Z_i - \mu_z)$ into the $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (14). Efficiency gain should be expected since this estimator uses more information on the population level. By viewing the problem as a two-phase sampling problem, the method can be readily extended to multiple sampling phases. This extension is useful because the database A_1 can come from alarger sample within the database. This case covers the common situations where detailed treatment and outcome data is available for only a subsample of the data such as a subsample with medical chart adjudication of claims records or a subsample constructed by merging multiple sources of claims records and electronic medical records.

3. Asymptotic Normality and Variance Estimation

Since $\hat{\boldsymbol{\theta}}_g^{(1)}$ can be written as a special case of $\hat{\boldsymbol{\theta}}_g^{(2)}$, in Subsection 3.1 we derive the asymptotic normal distribution for $\hat{\boldsymbol{\theta}}_g^{(2)}$ only, and in Subsection 3.2 provide a linearized variance estimator for $\hat{\boldsymbol{\theta}}_g^{(2)}$. Subsection 3.3 gives a replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$.

3.1. Asymptotic Normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$

The asymptotic normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$ is established in Theorem 1 by combining two randomizations from the finite population level and the superpopulation level. For the finite population level, we consider a sequence of samples and finite populations indexed by N, where the sample size $n \longrightarrow \infty$ as $N \longrightarrow \infty$ (Isaki and Fuller 1982). To define the regularity conditions, we introduce the notation $\mathcal{F}_N$ to represent an element of the sequence of finite population with size N. To distinguish between the two randomizations, we use the notation "$|\mathcal{F}_N$" to indicate that the reference distribution is with respect to repeated sampling conditional on the finite population size N. For example, $E(\cdot|\mathcal{F}_N)$ and $V(\cdot|\mathcal{F}_N)$ denote the conditional mean and variance with respect to the randomization generated from repeated sampling from $\mathcal{F}_N$. And we use $E_\xi(\cdot)$, $Var_\xi(\cdot)$ and $Cov_\xi(\cdot,\cdot)$ to denote mean, variance and covariance with respect to the randomization from the superpopulation ξ. The proof of Theorem 1 uses a result given in Theorem 1.3.6 of Fuller (2009) that shows how to combine two asymptotic normalities from the finite population and the superpoulation levels. Because of the importance of this theorem to our results, we state this theorem as Fact 1:

Fact 1 (Theorem 1.3.6 of Fuller 2009): Suppose θ_0 is a true parameter on a superpopulation level, θ_N is its analogous part on a finite population level, and $\hat{\theta}$ is an estimator of θ_0 calculated from a sample. If $(\hat{\theta} - \theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, V_{11})$ almost surely (a.s.) and $(\theta_N - \theta_0) \xrightarrow{\mathcal{L}} N(0, V_{22})$, then, $(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N\ (0, V_{11} + V_{22})$. Here $(\hat{\theta} - \theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, V_{11})$ *a.s.* means that $\hat{\theta} - \theta_N$ converges in a distribution to a random variable with the distribution of $N(0, V_{11})$ almost surely with respect to the process of repeated sampling from the sequence of finite populations as $N \longrightarrow \infty$. V_{11} is the asymptotic variance of $\hat{\theta}$ on the finite population level, while V_{22} is the asymptotic variance of θ_N on the superpopulaton level.

The key step in our proof of Theorem 1 is to obtain an asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$,

$$\begin{aligned}\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) &= \frac{1}{N}\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} \\ &= \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\sum_{i\in U}\frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} E_\xi(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i) + o_p(n^{-1/2}).\end{aligned} \tag{17}$$

Define

$$\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z) = [m_{ig}(\boldsymbol{\theta}_g), Z_i - \mu_z]^T, \tag{18}$$

and similary we can show an asymptotic equivalent form of $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g, \mu_z)$ as,

$$\begin{aligned}\frac{1}{N}\sum_{i\in A_{2g}} \frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}\hat{\pi}_{2ig}} &= \frac{1}{N}\sum_{i\in U}\frac{\delta_{1i}\delta_{2ig}\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\sum_{i\in U}\frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} \\ &\quad \times E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)|X_i) + o_p(n^{-1/2}) \\ &= \frac{1}{N}\sum_{i\in A_1}\frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}} + o_p(n^{-1/2}),\end{aligned} \tag{19}$$

where

$$\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)=\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)\frac{\delta_{2ig}}{\pi_{2ig}}+\left(1-\frac{\delta_{2ig}}{\pi_{2ig}}\right)\boldsymbol{\mu}_{Hg}(X_i;\boldsymbol{\theta}_g,\mu_z),\ \text{ and}$$
$$\boldsymbol{\mu}_{Hg}(X_i,\boldsymbol{\theta}_g)=E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)|X_i). \tag{20}$$

Thus we can write $\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)$ in (14) as,

$$\begin{aligned}\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)&=\left[\frac{1}{N}\sum_{i\in A_{2g}}\frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}\hat{\pi}_{2ig}},\frac{1}{N}\sum_{i\in A_1}\frac{Z_i-\mu_z}{\pi_{1i}}\right]^T\\&=\left[\frac{1}{N}\sum_{i\in A_1}\frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}},\frac{1}{N}\sum_{i\in A_1}\frac{Z_i-\mu_z}{\pi_{1i}}\right]^T+o_p(n^{-1/2}).\end{aligned} \tag{21}$$

Then the large sample theory for $\hat{\boldsymbol{\theta}}_g^{(2)}$ is derived based on the asymptotic form of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)$ in Equation (21). We now state Theorem 1:

Theorem 1: Under the regularity conditions in the Appendix, for any $g=1,\ldots G$,

$$\sqrt{n}\left(\begin{bmatrix}\hat{\boldsymbol{\theta}}_g^{(2)}\\ \hat{\mu}_z\end{bmatrix}-\begin{bmatrix}\boldsymbol{\theta}_g^0\\ \mu_z^0\end{bmatrix}\right)\xrightarrow{\mathcal{L}}N\left(\mathbf{0},V_g\left(\boldsymbol{\theta}_g^0,\mu_z^0\right)\right),$$

where

$$V_g(\boldsymbol{\theta}_g,\mu_z)=\left[\Gamma_g^T(\boldsymbol{\theta}_g)\Sigma_{Hg}^{-1}(\boldsymbol{\theta}_g,\mu_z)\Gamma_g^T(\boldsymbol{\theta}_g)\right]^{-1}, \tag{22}$$

$$\Gamma_g(\boldsymbol{\theta}_g)=\left[E_\xi\left[\frac{\partial\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\partial\boldsymbol{\theta}_g}\right]\quad E_\xi\left[\frac{\partial\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\partial\mu_z}\right];\quad\mathbf{0}\quad-1\right], \tag{23}$$

$$\text{and}\quad\boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g,\mu_z)=\left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g,\mu_z)\quad\boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g,\mu_z);\quad\boldsymbol{\Sigma}_{12}^T(\boldsymbol{\theta}_g,\mu_z)\quad\boldsymbol{\Sigma}_{22}(\mu_z)\right]. \tag{24}$$

Here the notation $[\mathbf{a_{11}},\mathbf{a_{12}};\mathbf{a_{21}},\mathbf{a_{22}}]$ represents a 2 × 2 block matrix with blocks $\mathbf{a_{ij}}$. The term $\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g,\mu_z)$ in Equation (24) is related to the asymptotic variance of the first element in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g,\mu_z)=\lim_{N\to\infty}V_{\eta g,N}(\boldsymbol{\theta}_g,\mu_z)+\frac{n}{N}Var_\xi(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)), \tag{25}$$

$$\text{where}\quad V_{\eta g,N}(\boldsymbol{\theta}_g,\mu_z)=nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)\boldsymbol{\eta}_{jg}^T(\boldsymbol{\theta}_g,\mu_z). \tag{26}$$

The term $\boldsymbol{\Sigma}_{22}(\mu_z)$ in Equation (24) is related to the asymptotic variance of the second element in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{22}(\mu_z)=\lim_{N\to\infty}V_{z,N}(\mu_z)+\frac{n}{N}Var_\xi(Z_i), \tag{27}$$

$$\text{where} \quad V_{z,N}(\mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}(Z_i - \mu_z)(Z_j - \mu_z). \tag{28}$$

The term $\boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (24) is related to the asymptotic covariance between the two elements in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z) = \lim_{N\to\infty} C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N} Cov_\xi(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i), \tag{29}$$

$$\text{where} \quad C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j - \mu_z). \tag{30}$$

Equation (25) is connected to Fact 1 stated above, where its first term is $nV\left(N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)|\mathcal{F}_N\right)$ on the finite population corresponding to V_{11} in Fact 1, and its second term is $nV_\xi\left(N^{-1}\sum_{i\in U}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)\right)$ on the superpopulation level corresponding to V_{22} in Fact 1. The limit sign in the first term of Equation (25) indicates this is the limit with respect to the process of repeated sampling from a sequence of finite population as $N \to \infty$. Similar connections can be seen in Equations (27) and (29). The proof of Theorem 1 uses results from Pakes and Pollard (1989) (Theorems 3.2 and 3.3) which provides a general central limit theorem for estimators defined by minimization of the length of a vector valued random criterion function. The justification of Theorem 1 takes into account the finite population asymptotic framework and the semiparametric estimation of $\hat{\pi}_{2ig}$. The asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ in (17) is analytically similar to the mathematical forms of the doubly robust (DR) estimators when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, see Kim and Haziza (2014), Haziza and Rao (2006), Tan (2006), and Robins et al. (2007). One difference is that the consistency of the DR estimators requires one of the response model and the outcome model to be correctly specified, while our estimators estimate both the self-selection probabilities π_{2ig} and the outcome model semiparametrically. The regularity conditions on the sample design and tuning parameters for the semiparametric estimation are provided in the Appendix, and an outline of the proof for Theorem 1 can be found in Appendix A.

3.2. *Variance Estimation Based on the Asymptotic Normality*

We use the asymptotic variance $V_g(\boldsymbol{\theta}_g^0, \mu_z^0)$ in (22) to estimate the variance of $\hat{\boldsymbol{\theta}}_g^{(2)}$. To estimate $\boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$, an estimator of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ is obtained by,

$$\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z) = \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)\frac{\delta_{2ig}}{\hat{\pi}_{2ig}} + \left(1 - \frac{\delta_{2ig}}{\hat{\pi}_{2ig}}\right)\hat{\boldsymbol{\mu}}_{Hg}\left(X_i; \boldsymbol{\theta}_g, \mu_z\right), \tag{31}$$

where $\boldsymbol{\mu}_{Hg}(X_i, \boldsymbol{\theta}_g)$ is also estimated semiparametrically using the same bases $R_K(X_i)$, that is

$$\hat{\boldsymbol{\mu}}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z) = \hat{\beta}_g^T(\boldsymbol{\theta}_g, \mu_z)R_K(X_i), \quad \text{and} \tag{32}$$

$$\hat{\beta}_g(\boldsymbol{\theta}_g, \mu_z) = \left(\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)R_K(X_i)^T\right)^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)\mathbf{H}_{ig}^T(\boldsymbol{\theta}_g, \mu_z). \tag{33}$$

An estimator of $V_g\left(\boldsymbol{\theta}_g^0, \mu_z^0\right)$ is calculated as follows,

$$\hat{V}_g\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right) = \left[\hat{\Gamma}_g^T\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)\hat{\boldsymbol{\Sigma}}_{Hg}^{-1}\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right)\hat{\Gamma}_g^T\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)\right]^{-1}, \tag{34}$$

where

$$\hat{\Gamma}_g(\boldsymbol{\theta}_g) = \frac{1}{N}\left[\sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}\left(\boldsymbol{\theta}_g, \mu_z\right)}{\partial \boldsymbol{\theta}_g}\sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}\left(\boldsymbol{\theta}_g, \mu_z\right)}{\partial \mu_z}; \quad 0 \quad -1\right], \tag{35}$$

$$\text{and} \quad \hat{\boldsymbol{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z) = \left[\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) \quad \hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z); \quad \hat{\boldsymbol{\Sigma}}_{12}^T(\boldsymbol{\theta}_g, \mu_z) \quad \hat{\boldsymbol{\Sigma}}_{22}(\mu_z)\right]. \tag{36}$$

The term $\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) = \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\widehat{Var}_{\xi}\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right), \tag{37}$$

$$\text{where} \quad \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{jg}^T(\boldsymbol{\theta}_g, \mu_z), \tag{38}$$

and

$$\widehat{Var}_{\xi}\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{ig}^T(\boldsymbol{\theta}_g, \mu_z)$$

$$-\frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]^T. \tag{39}$$

The term $\hat{\boldsymbol{\Sigma}}_{22}(\mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{22}(\mu_z) = \hat{V}_{z,N}(\mu_z) + \frac{n}{N}\widehat{\mathrm{Var}}_{\xi}(Z_i), \tag{40}$$

$$\text{where} \quad \hat{V}_{z,N}(\mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}(Z_i - \mu_z)(Z_j - \mu_z), \quad \text{and} \tag{41}$$

$$\widehat{\mathrm{Var}}_{\xi}(Z_i) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)^2 - \frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)\right]^T, \tag{42}$$

The term $\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z) = \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\widehat{\mathrm{Cov}}_{\xi}\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i\right), \tag{43}$$

$$\text{where } \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j - \mu_z), \tag{44}$$

and

$$\widehat{\mathrm{Cov}}_{\xi}\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i\right) = \frac{1}{N}\sum_{i \in A_1} \pi_{1i}^{-1} \hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_i - \mu_z) - \frac{1}{N^2}\left[\sum_{i \in A_1} \pi_{1i}^{-1} \hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]\left[\sum_{i \in A_1} \pi_{1i}^{-1}(Z_i - \mu_z)\right]. \tag{45}$$

To construct a joint estimator for $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G]^T$, one can simply stack $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in the quadratic form of Equation (13). Define $\mathbf{H}_i(\boldsymbol{\theta}, \mu_z)$ as the stacked vector of $\mathbf{H}_{ig}(\boldsymbol{\theta}_g, u_z)'s$ in Equation (18) and $\boldsymbol{\eta}_i(\boldsymbol{\theta}, \mu_z)$ as the stacked vector of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)'s$ in Equation (20). The asymptotic theory and the variance estimator for $\hat{\boldsymbol{\theta}}^{(2)}$ can be derived by simply replacing $\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ by $\mathbf{H}_i(\boldsymbol{\theta}, \mu_z)$ and $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ by $\boldsymbol{\eta}_i(\boldsymbol{\theta}, \mu_z)$. Then we can obtain an inference for the treatment effects or any linear combination of treatment parameters, $\boldsymbol{\lambda}^T\boldsymbol{\theta}$.

3.3. Replication Variance Estimation

In surveys conducted on land, for example surveys about natural resources (soil, forest, water, etc.), non-responses hardly occur. However, in surveys with high non-response rates, such as almost all surveys conducted on people, the joint inclusion probabilities are typically not available because sampling weights have to be appropriately adjusted for nonresponse. After such adjustments, the joint inclusion probabilities change and are hard to be derived. In practice, a set of replicate weights are often provided instead, because (1) design weights are often adjusted due to nonresponse issues and a set of replicate weights are provided to account for the weight adjustment; (2) sometimes a few design variables are masked from users to keep confidentiality. An example of such design variable is location which is used for defining design strata in a study, but the specific location is omitted from the analyst because of concerns associated with confidentiality. In this subsection, we show how to use the replicate weights to construct a Jackknife variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$. Note that $\hat{\boldsymbol{\theta}}_g^{(2)}$ depends on the joint inclusion probabilities π_{1ij} which are typically not available when replicate weights are provided. We propose to use the Jackknife (JK) variance estimator for a two-phase sampling design discussed in Fuller (2009) and Kim et al. (2006). Assume that there is a replicate variance estimator that gives a consistent estimator for the variance of the total estimator based on the first-phase sample. We write the replication variance estimator as, $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1) = \sum_{b=1}^{B} c_b\left(\hat{\boldsymbol{\theta}}_1^{[b]} - \hat{\boldsymbol{\theta}}_1\right)(\hat{\boldsymbol{\theta}}_1^{[b]} - \hat{\boldsymbol{\theta}}_1)^T$, where B is the number of replicates, $\hat{\boldsymbol{\theta}}_1 = \sum_{i \in A_1} w_{1i}X_i$ is the total estimator of variable X using the first-phase sample, $\hat{\boldsymbol{\theta}}_1^{[b]} = \sum_{i \in A_1} w_{1i}^{[b]}X_i$ is the estimated total for the b^{th}replicate, $w_{1i}^{[b]}$ is the b^{th}replicate weights in the first-phase, and c_b is a factor associated with replicate b such that $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_1$. Suppose the second-phase total estimator is, $\hat{\boldsymbol{\theta}}_2 = \sum_{i \in A_2} w_{1i}\pi_{2i|1i}^{-1}X_i$, where $\pi_{2i|1i}$ is the conditional probability of selecting i for the phase 2 sample given that i is in the phase 1 sample, and A_2 is the phase 2 sample. Define the b^{th}replicate of $\hat{\boldsymbol{\theta}}_2$ as, $\hat{\boldsymbol{\theta}}_2^{[b]} = \sum_{i \in A_2} w_{1i}^{[b]}\pi_{2i|1i}^{-1}X_i$. A Jackknife variance estimator for $\hat{\theta}_2$

can be calculated as, $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2) = \sum_{b=1}^{B} c_b \left(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2\right)^T$. Kim et al. (2006) showed that $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_2$.

Following the idea of Fuller (2009 Subsection 4.4), let b be the index for the deleted Jackknife groups and the corresponding replicate version of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ be,

$$\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i \in A_{2g}} w_{1i}^{[b]} \left(\hat{\pi}_{2ig}^{[b]}\right)^{-1} \mathbf{m}_{ig}(\boldsymbol{\theta}_g), \tag{46}$$

where $\hat{\pi}_{2ig}^{[b]}$ is obtained by replacing w_{1i} by $w_{1i}^{[b]}$ in Equation (7). Then the replicate estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is,

$$\hat{\boldsymbol{\theta}}_g^{(1)[b]} = \arg\min_{\boldsymbol{\theta}_g} \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right]^T \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right], \tag{47}$$

and the replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is calcualted as,

$$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right) = \sum_{b=1}^{B} c_b \left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)\left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)^T. \tag{48}$$

Examples of $w_{1i}^{[b]}$ and c_b for a variety of designs are given in Särndal et al. (1992). For example, if the first-phase sample is drawn from a multi-stage cluster design, the Jackknife technique is usually applied at the primary sampling unit (PSU) levels. Assuming there are B PSUs and S_b is the b^{th} PSU deleted in the b^{th} replicate sample, the b^{th} replicate weight for the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \dfrac{B}{B-1} w_{1i} & \text{if} \quad i \notin S_b \end{cases}, \tag{49}$$

and $c_b = B^{-1}(B-1)$. As mentioned in Särndal et al. (1992), for stratified sampling designs, $w_{1i}^{[b]}$ and c_b need to be defined with care. We discuss this situation in Section 5 of the empirical study. If the first phase replicate weights are provided in practice, one can directly use them as $w_{1i}^{[b]}$. One thing to note is that Kim et al. (2006) assume π_{2ig} are known in their two phase replication variance estimator. The consistency theorem in Kim et al. (2006) needs to be modified to account for the variation from estimating $\hat{\pi}_{2ig}$ in our JK variance estimator, which can be our future study.

4. Simulation Study

In this section, we evaluate the performance of our estimators and variance estimators under four different simulation set-ups. We consider three treatment levels, and a population size of $N = 10{,}000$ and an expected sample size of $n = 1{,}000$. We generate i.i.d. realizations, $(\mathbf{Y}_i, \delta_{1i}, \boldsymbol{\delta}_{2i}, X_i, Z_i)$; $i = 1, \ldots, N$, according to the following super-population set-ups.

(1) Covariates: simulate covariates $\mathbf{Z}_i = [Z_{1i}, Z_{2i}]$ where $Z_{1i} \sim N(2,1)$ and $Z_{2i} \sim N(10,1)$, and $\mathbf{X}_i = [X_{1i}, X_{2i}]$ where $X_{1i} = Z_{1i}$ and $X_{2i} \sim N(0.5, 0.3^2)$.

(2) Potential response outcomes: the superpopulation model for potential outcomes is $Y_{ig} = \mu_g(\mathbf{Z}_i) + \sigma_g(\mathbf{Z}_i)\varepsilon_{ig}$, where

$$\mu_g(\mathbf{Z}_i) = \beta_{g0} + \beta_{g1}(Z_{1i} - 0.5) + \beta_{g2}(Z_{1i} - 0.5)^2 + \beta_{g3}Z_{2i},$$

$\epsilon_{ig} \sim N(0,1)$, $\sigma_g(\mathbf{Z}_i) = |\mu_g(\mathbf{Z}_i)|$, and $[\beta_{g0}, \beta_{g1}, \beta_{g2}, \beta_{g3}]$ equals to [5, 4, 2, 1] for $g = 1$, [0, 1, 0, 0] for $g = 2$, and $[-5, -4, -2, -0.5]$ for $g = 3$.

(3) First phase sampling: we consider two sampling designs, non-informative stratification sampling and informative Poisson sampling.

- Stratification (STS): population units are sorted by values of Z_{1i}, and then the population is divided into two subpopulations U_1 and U_2 with equal sizes. Simple random sampling is used to draw 80 percent of the sample from U_1 and 20 percent from U_2. For units in stratum s ($s = 1$ or 2), $\pi_{1i} = N_s^{-1}n_s$ and $\pi_{1ij} = \{N_s(N_s - 1)\}^{-1}n_s(n_s - 1)$, where n_s and N_s are the sample size and the population size in stratum s. The joint inclusion probability for two units in different strata is the product of their first order inclusion probabilities.
- Informative Poisson (Informative): the first-phase sample design is Poisson sampling with selection probability,

$$\pi_{1i} = \frac{exp\left(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}_i\|\right)}{1 + exp\left(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}_i\|\right)},$$

where $\|\mathbf{Y}_i\| = \sqrt{Y_{i1}^2 + Y_{i2}^2 + Y_{i3}^2}$. Modeling π_{1i} as a function of $\mathbf{Y}_i$ is a common way (i.e., Pfeffermann and Sverchkov 1999) to represent joint dependence of $\mathbf{Y}_i$ and π_{1i} on a design variable that is not contained in (X_i, Z_i). In this specification, we assume $\|\mathbf{Y}_i\|$ is known at the design stage of the survey, but is unavailable at the analysis stage.

(4) Second phase self-selection probability models: we consider two models for π_{2ig}.

- Logit Linear (LogitLinear):

$$\pi_{2ig} = \frac{\exp\left(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i}\right)}{\sum_{g=1}^{G}\exp\left(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i}\right)},$$

where $[\phi_{g0}, \phi_{g1}, \phi_{g2}]$ equals to $[-0.5, 0, 0]$ for $g = 1$, $[0.3, -0.3, -0.3]$ for $g = 2$, and $[0, -0.5, 0.5]$ for $g = 3$.

- Jump (JUMP):

$$\begin{aligned}[\pi_{2i1}, \pi_{2i2}, \pi_{2i3}] &= [0.90, 0.05, 0.05] \quad \text{if} \quad X_{1i} + X_{2i} \geq 3 \\ &= [1/3, 1/3, 1/3] \quad \text{if} \quad 2 \leq X_{1i} + X_{2i} < 3 \\ &= [0.05, 0.05, 0.90] \quad \text{if} \quad X_{1i} + X_{2i} < 2.\end{aligned}$$

The JUMP model violates the differentiability assumption of π_{2ig} in Condition B(2) in the Appendix. It is deliberately included in the simulation to see if our semiparametric approach can estimate a nonsmooth multiple treatment selection probabilities well.

For each $i \in U$, $\boldsymbol{\delta}_{2i}$ is simulated from $multinomial(1; \pi_{2i1}, \pi_{2i2}, \pi_{2i3})$. For $i \neq j$, $\pi_{1ij} = \pi_{1i}\pi_{1j}$. For STS design which is noninformative, SMAR holds and we set $b_i = w_{1i}^{-1}$ in Equation (7) to estimate $\hat{\pi}_{2ig}$. For Informative design, SMAR fails and we use $b_i = 1$ in Equation (7) to estimate $\hat{\pi}_{2ig}$.

We first simulate a finite population with size N from the superpopulation and then use indicators generated in (3) and (4) to obtain the first and second phase samples. We repeat the process to produce 1,000 MC samples. We are interested in estimating five parameters for each group, $\boldsymbol{\theta}_g = \left[P_g, \mu_g, \sigma_g^2, R_g, D_g\right]$, where $P_g = Prob(Y_{ig} \leq 0)$, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_{2i})$, and $D_g = E[E(Y_{ig}|Z_{1i} \leq 0.65)]$. The corresponding estimation equations $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ can be found in Equations (5) and (6). For each MC sample, we calculate the following four estimators:

- $\hat{\boldsymbol{\theta}}_g^{(1)}$: the estimator defined in (11). When $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, $\hat{\boldsymbol{\theta}}_g^{(1)}$ corresponds to the estimator in Yu et al. (2013) asymptotically.
- $\hat{\boldsymbol{\theta}}_g^{(2)}$: the estimator defined in (13).
- $\hat{\boldsymbol{\theta}}_g^{nw}$: the estimator defined in (9), and is included to see what happens when the survey weights are ignored in analyses.
- $\hat{\boldsymbol{\theta}}_g^{p}$: the estimator calculated the same way as $\hat{\boldsymbol{\theta}}_g^{(1)}$, except that $\hat{\pi}_{2ig}$ are estimated using a parametric multinomial regression. This estimator is introduced in order to have plausible comparisons in context of three treatments between our estimators and others that use parametric logistic regression to estimate propensity scores, see DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015).

We use a cubic spline base of X_{1i} for $R_K(X_{1i})$, as suggested by Breidt et al. (2005) which mentions that setting the degree of the spline equal to three is a popular choice in practice. Condition 4(B) in the Appendix gives a practical guidance for the choice of K, the number of knots in the spline. Condition 4(B) requires $K = O(n^{\nu})$, where ν has an upper bound $\nu \leq (4\eta + 2)^{-1}$ with $\eta = 1/2$ for spline bases. In our simulation studies, the sample size $n = 1{,}000$, suggesting $n^u = 5.6$. The choices of $K = 5, 4, 3, 2$ are tried and the corresponding $\hat{\pi}_{2ig}$ curves are plotted. It is found that there is not noticeable change in the $\hat{\pi}_{2ig}$ curves until K decreases to 2. So $K = 3$ is used and the locations of the three knots correspond to the 25*th*, 50*th*, and 75*th* quantiles of observed X_{1i}'s. A cubic spline base for $R_K(X_{2i})$ is constructed the same way. And the semiparametric bases are $R_K(\mathbf{X}_i) = \left[R_K^T(X_{1i}), R_K^T(X_{2i})\right]^T$.

If the dimension of $(\mathbf{X}_i, \mathbf{Z}_i)$ is big, in practice we suggest to run a multinomial regression using δ_{2i} on $(\mathbf{X}_i, \mathbf{Z}_i)$ to select covariates that are most significant, and then use them for estimation of $\hat{\pi}_{2ig}$. When using $\hat{\boldsymbol{\theta}}_g^{(2)}$, one can run a multiple linear regression of Y_{ig} on $(\mathbf{X}_i, \mathbf{Z}_i)$ in A_{2g} to identify covariates that are most useful for explaining the outcome Y_{ig}, and then add their first and second phase means in the estimation equations. It is not impossible to obtain a very small $\hat{\pi}_{2ig}$ computationally, which leads to extreme weights. A solution is to truncate such $\hat{\pi}_{2ig}$'s to a small constant L (which is set to be 0.0001 in our study), then adjust the truncated $\hat{\pi}_{2ig}$ by calibrating the second phase mean of U_i to its first phase mean, that is $\tilde{\pi}_{2ig} = F_g \hat{\pi}_{2ig}^t$ where $F_g = \left(\sum_{i \in A_1} w_{1i} U_i\right)^{-1} \sum_{i \in A_{2g}} w_{1i} \left(\hat{\pi}_{2ig}^t\right)^{-1} U_i$, and $\hat{\pi}_{2ig}^t$ is the truncated propensity score which equals to L if $\hat{\pi}_{2ig} < L$, otherwise remains unchanged. Here the variable U_i can be an important covariate chosen by users, or a weighted mean of $(\mathbf{X}_i, \mathbf{Z}_i)$ where weights indicate importance of the covariates. We use the average of the covariate $\mathbf{X}_i$ as U_i in both of the simulation studies and the empirical study.

Figures 1–4 show side-by-side boxplots of MC estimates of the four estimators for all treatment effects. Each figure represents one of four simulation setups: (STS-LogitLinear), (STS-JUMP), (Informative-LogitLinear), and (Informative-JUMP). In each subplot, the first two boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, and the third and fourth boxplots are for $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively. When comparing our estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{nw}$, $\hat{\boldsymbol{\theta}}_g^{nw}$ is highly biased in most of parameters and scenarios, due to ignoring the survey weights. The variances of $\hat{\boldsymbol{\theta}}_g^{nw}$ in general are smaller than those of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, which is expected especially when the survey weights are very different from each other. The coefficient of variation (CV) of the weights for the STS design is 0.75, and the CV of weights for the Informative design is 4.77. When comparing our estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{p}$, biases of $\hat{\boldsymbol{\theta}}_g^{p}$ are comparable to those of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ for the LogitLinear model because in this scenario $\hat{\theta}_g^{p}$ correctly assumes a parametric model for π_{2ig}. However, in the situation of JUMP models, $\hat{\boldsymbol{\theta}}_g^{p}$ has larger biases than $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ because π_{2ig} is misspecified parametrically. When comparing $\hat{\boldsymbol{\theta}}_g^{(1)}$ with $\hat{\boldsymbol{\theta}}_g^{(2)}$, both of their biases are comparable in all scenarios. However, the plots show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ consistently has smaller variances than $\hat{\boldsymbol{\theta}}_g^{(1)}$. The variance reduction of $\hat{\boldsymbol{\theta}}_g^{(2)}$ over $\hat{\boldsymbol{\theta}}_g^{(1)}$ indicates that efficiency gain occurs after adding the first and second phase means of covariates to the estimation equations, which confirms Remark 2. Additionally, it is promising to see that both $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ have relatively small biases even if the JUMP model fails to satisfy the differentiability assumption in the theory, indicating our semiparametric approach of estimating $\hat{\pi}_{2ig}$ works well for the nonsmooth function considered. We also tabulate the MC results into four tables for readers who prefer to see numbers rather than Figures (see Supplemental file, Tables 1–4).

Tables 1–2 contain the coverage probabilities of the 95 percent confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(2)}$ based on its asymptotic normality and its linearized variance estimator in Subsection 3.2, and the coverage probabilities of the 95 percent confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ based on the JK approach discussed in Subsection 3.3. The replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{nw}$ is calculated by replacing w_{1i} by N/n in Equation (49). This gives inappropriate variance estimation for $\hat{\boldsymbol{\theta}}_g^{nw}$ under an unequal probability sampling, but mimics what people do when they ignore survey weights. To create the JK replicates, we delete one unit at a time and set $B = 1{,}000$. The coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{(2)}$ using the linearized variance estimator seem to work well, except for the marginal mean μ_g under (STS-LogitLinear) and the marginal proportion P_g under (STS-JUMP). The rest of coverage probabilities are reasonably close to the nominal size 95 percent. The JK variance estimator of $\hat{\boldsymbol{\theta}}_g^{(1)}$ gives very good coverage probabilities. However the coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{nw}$ using the JK variance estimation are far away from the nominal size, especially under the Informative-JUMP model where the coverage probabilities are severely underestimated. Those under-coverages are due to the biases in $\hat{\boldsymbol{\theta}}_g^{nw}$, or inappropriate variance estimation, or both.

Our simulation studies demonstrate the validity of our estimators and variance estimators.

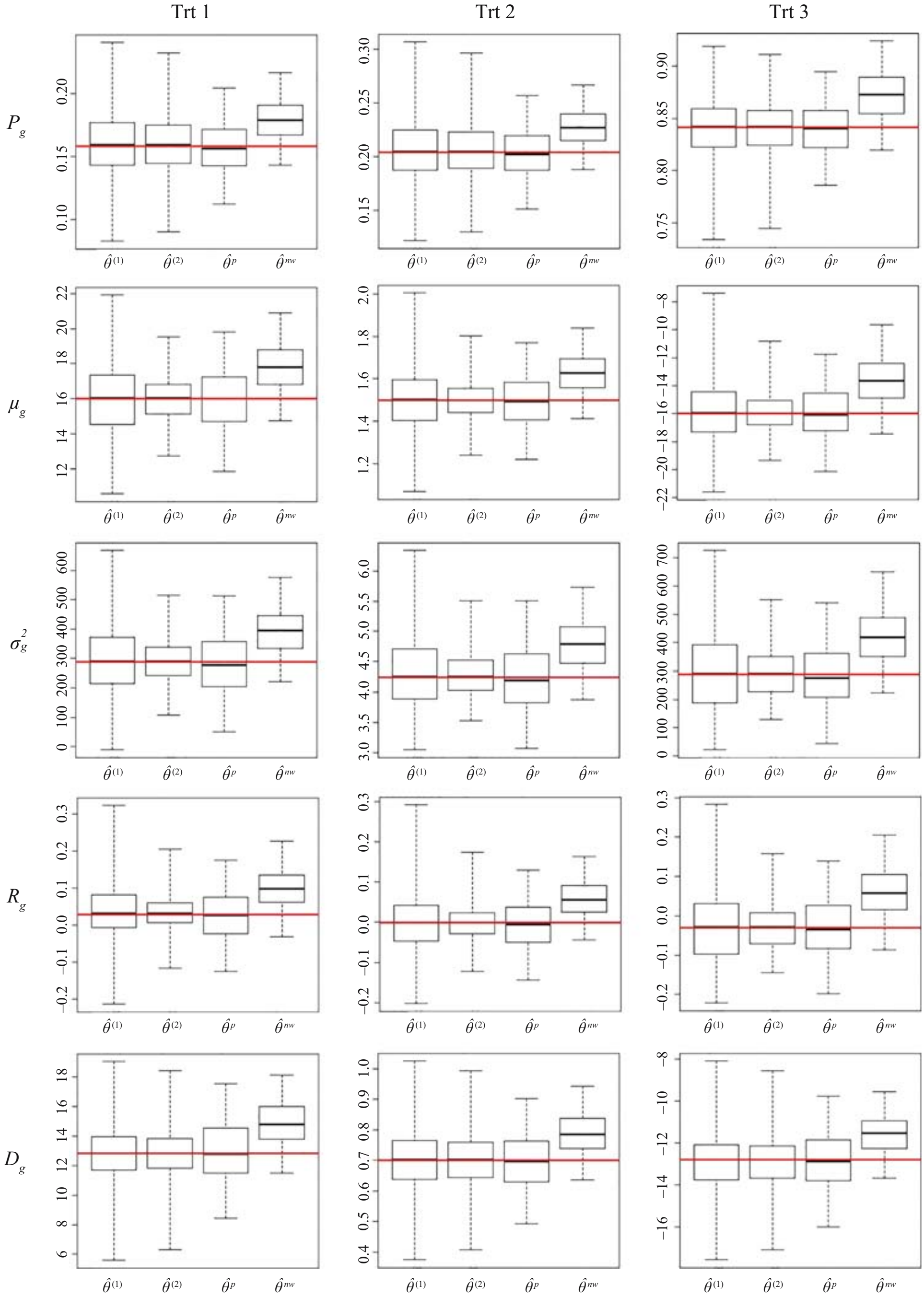


Fig. 1. **STS-LogitLinear***: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

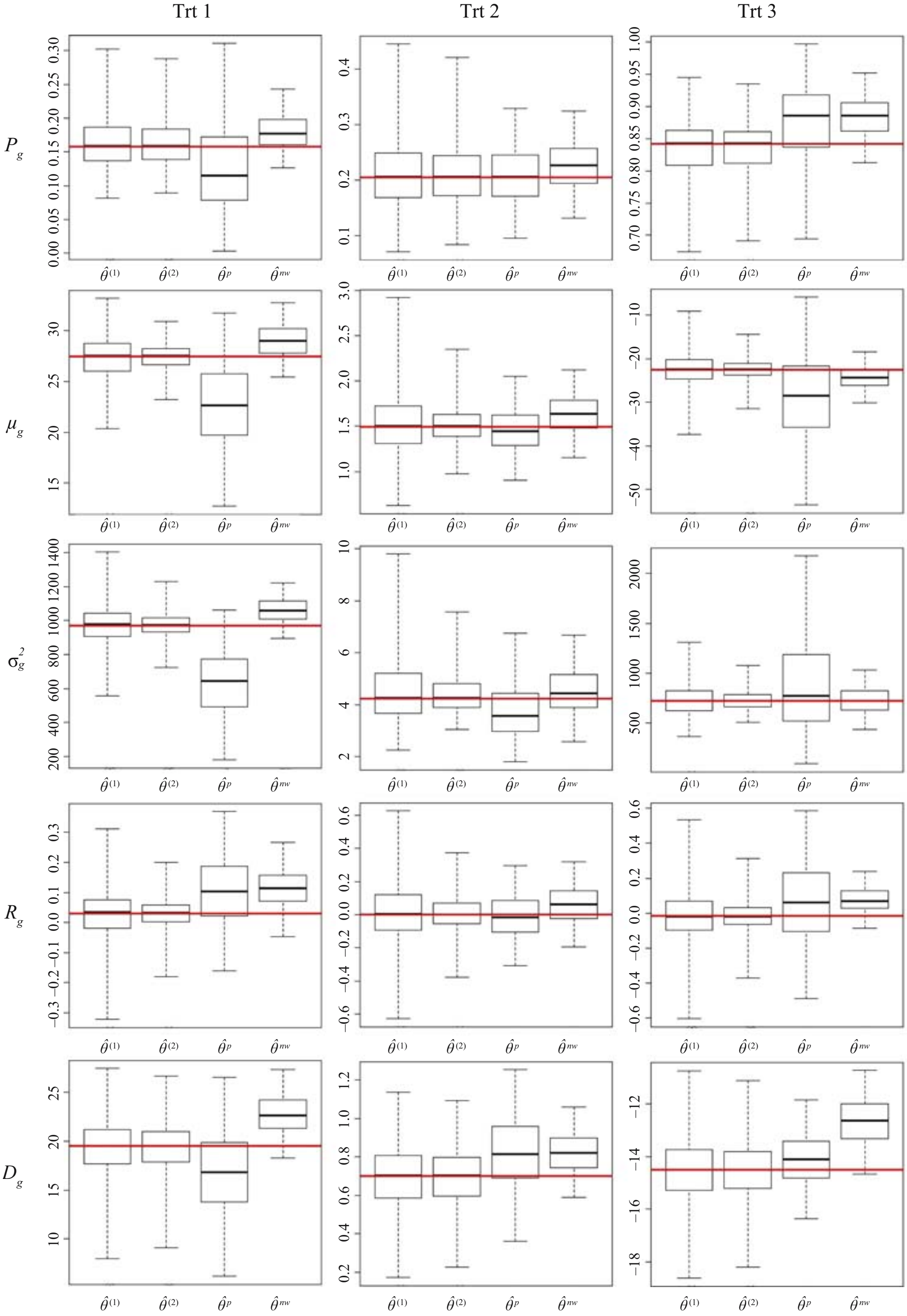


Fig. 2. **STS-JUMP***: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.*

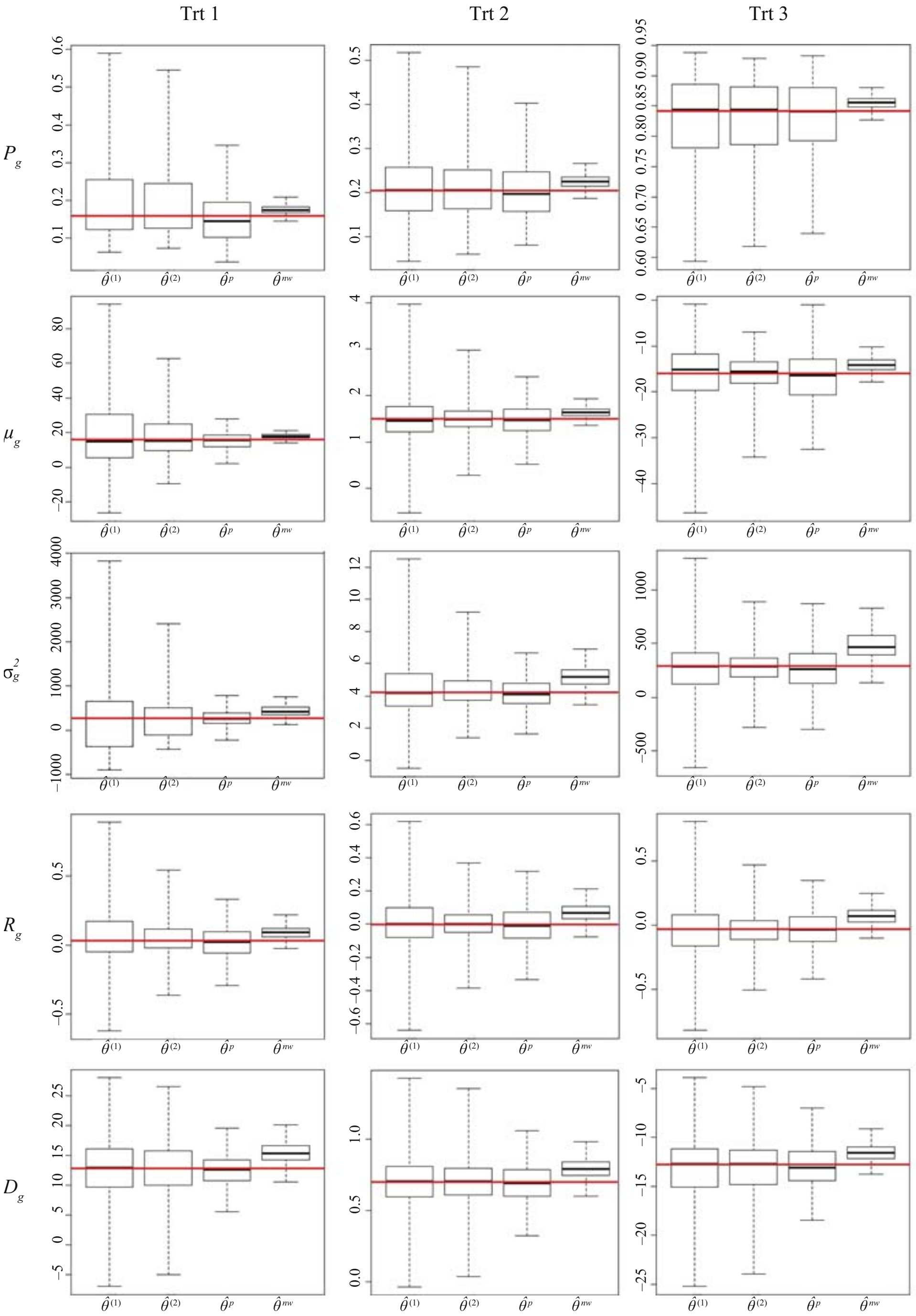


Fig. 3. ***Informative-LogitLinear****: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

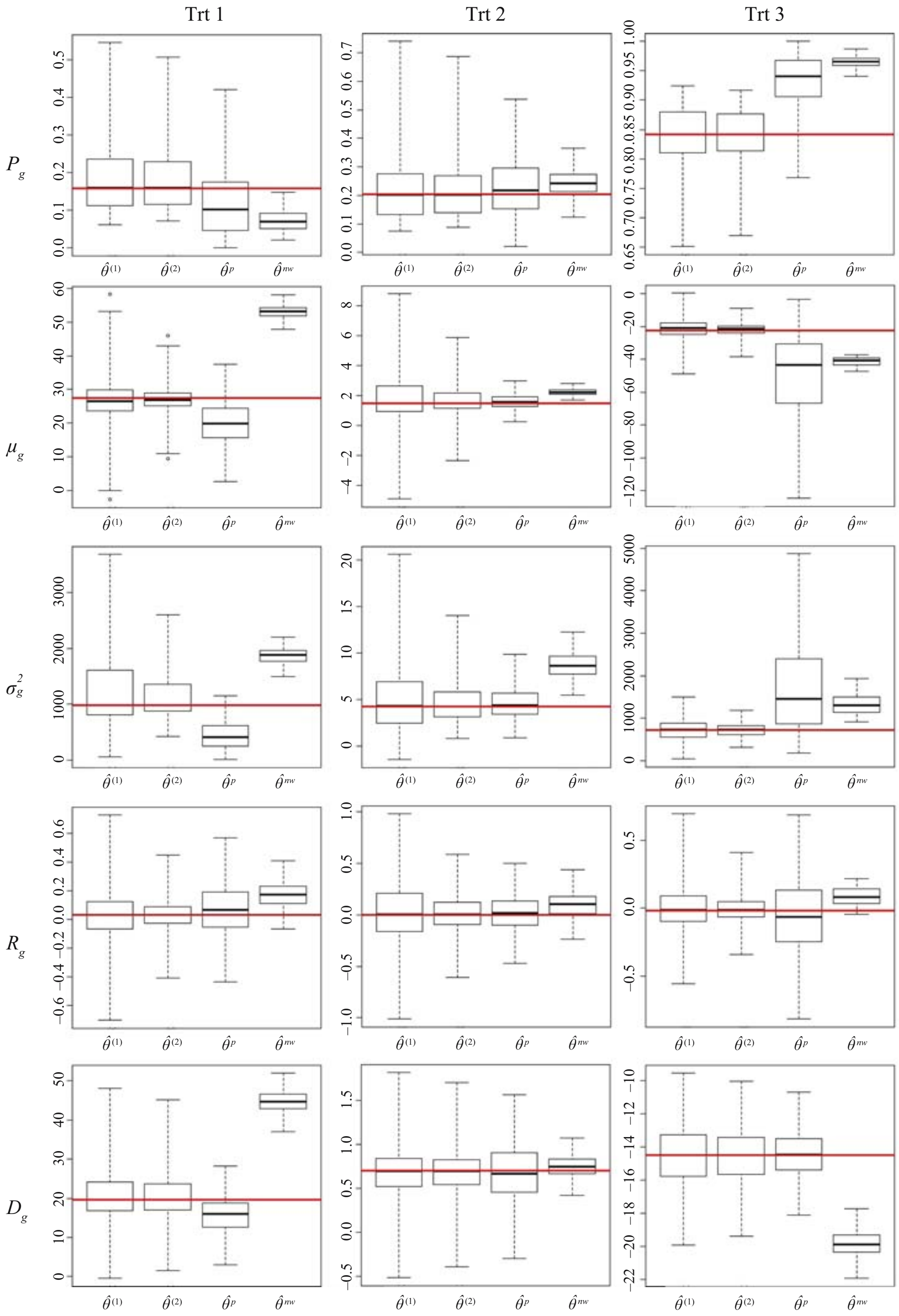


Fig. 4. ***Informative-JUMP****: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively in order. The horizontal line is located at the value of the true treatment effect.*

Table 1. ***Stratification:*** *The coverage probabilities of the 95 percent constructed intervals for the five estimated parameters using the linearized variance estimator* $\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(2)}$*, and the Jackknife variance estimators* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$ *and* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively.*

(a) STS-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	93.6	95.2	59.8	94.3	94.0	57.9	92.4	94.1	56.2
μ_g	95.4	95.5	58.7	95.2	95.1	57.9	88.1	94.1	62.3
σ_g^2	92.5	94.7	61.7	94.4	94.9	60.4	92.2	94.3	58.2
R_g	94.2	94.7	57.6	92.1	95.1	60.3	95.9	94.3	59.1
D_g	92.4	94.8	56.7	95.1	95.1	58.8	92.6	95.9	62.2

(b) STS-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	89.2	92.8	71.9	95.2	92.3	80.4	92.5	95.4	50.0
μ_g	92.2	95.3	73.0	95.6	93.3	76.8	94.5	95.3	79.3
σ_g^2	94.2	93.0	56.5	93.3	96.6	83.9	95.3	96.1	86.0
R_g	95.3	94.6	60.7	93.8	95.2	81.0	92.3	95.8	61.1
D_g	92.9	95.0	50.0	96.6	93.2	61.7	94.1	96.6	28.2

Table 2. ***Informative:*** *The coverage probabilities of the 95 percent constructed intervals for the five estimated parameters using the linearized variance estimator* $\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(2)}$*, and the Jackknife variance estimators* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$ *and* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(I)}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively.*

(a) Informative-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	94.3	95.2	58.9	94.1	96.4	49.3	95.4	94.7	49.2
μ_g	95.2	96.4	50.6	92.5	95.4	46.2	95.1	95.5	49.3
σ_g^2	92.0	94.1	46.2	95.4	95.1	50.6	93.7	96.2	44.4
R_g	94.9	96.0	45.1	90.6	94.8	42.8	93.8	95.4	38.3
D_g	93.4	96.2	59.3	93.8	95.7	48.1	93.1	94.9	45.7

(b) Informative-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	94.4	96.3	6.6	97.7	94.7	72.4	92.9	96.9	0.0
μ_g	92.5	97.2	0.0	91.1	97.1	2.0	92.2	93.9	0.0
σ_g^2	94.9	97.0	0.0	95.5	92.3	2.0	94.6	96.5	20.6
R_g	92.7	93.2	41.6	95.2	95.0	71.4	94.3	95.4	48.4
D_g	92.2	95.0	0.0	91.0	94.3	81.9	95.0	96.0	0.0

Table 3. ***Empirical study with weights in estimation of*** $\hat{\pi}_{2ig}$: *The treatment effect estimates using estimators* $\hat{\boldsymbol{\theta}}_g^{nw}$ *and* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *defined in Subsection 2.3. The parameter of interests are* $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ *and* $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ *where* I_{di} *is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95 percent confidence intervals are in brackets.*

(a) Treatment mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1349.57 (215.90) [926.40 1772.74]	309.408 (28.23) [254.07 364.74]	−1040.165 (698.47) [−2409.17 328.84]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1210.57 (353.50) [517.71 1903.44]	−21.45 (29.17) [−78.61779 35.71]	−1232.03 (56.56) [−1342.88 −1121.18]

(b) Treatment domain mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	3214.18 (32.22) [3151.03 3277.34]	811.56 (38.69) [735.73 887.39]	−2402.62 (46.48) [−2493.73 −2311.52]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3320.93 (9.97) [3301.39 3340.47]	4.49 (2.69) [−0.77 9.76]	−3316.43 (240.85) [−3788.50 −2844.37]

5. Empirical Study

In this section, we investigate the feasibility of our method in estimating the mean annual medical expenditures under different choices of health insurance types in China. We use the data from the Chinese General Social Survey (CGSS) conducted by the National Survey Research Center at the Renming University of China in 2010. The population consisted of all Chinese adults (18+) in mainland China. A sample of 12,000 adults was drawn for the base questionnaire and a subsample of 4,000 adults was drawn for the health care questionnaire. Data were collected by in-person interviews. The sample for the CGSS survey was selected using a multi-stage cluster sampling design. In the first stage, the primary sampling units (PSUs) were districts which were divided into two strata. Stratum 1 contained 67 districts in five major cities (Shanghai, Beijing, Guangzhou, Shenzhen and Tianjin), and Stratum 2 contained 2,795 districts in the rest of the area of China. In both strata, a probability proportional to size (PPS) design with the resident population size as the size variable was used to select the PSUs (40 PSUs were selected in Stratum 1, and 100 PSUs were selected in Stratum 2). In the second stage, the secondary sampling units (SSUs) were communities. A PPS design with resident population size as the size variable was used to select 2 SSUs within each selected PSU in Stratum 1 and 4 SSUs within each selected PSU in Stratum 2. In the third stage, the ultimate sampling units (USUs) were households. In each selected SSU, 25 households were drawn by a systematic sampling method. Then a respondent was selected randomly within each household. Totally 12,000 households responded to the base questionnaire. Then every third household respondent in each SSU was selected to answer the health care questionnaire. The subsample of 4,000 was used in our investigation.

Table 4. ***Empirical study without weights in estimation of*** $\hat{\pi}_{2ig}$: *The treatment effect estimates using estimators* $\hat{\boldsymbol{\theta}}_g^{nw}$ *and* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *defined in Subsection 2.3. The parameter of interests are* $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ *and* $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ *where* I_{di} *is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95 percent confidence intervals are in brackets.*

(a) Treatment mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public − Private	Public − No insurance	Private − No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1301.04 (150.81) [1005.45 1596.63]	298.02 (42.79) [214.15 381.89]	− 1003.02 (169.31) [− 1334.87 − 671.17]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1205.295 (259.68) [696.32 1714.27]	− 13.23 (55.84) [− 122.68 96.22]	− 1218.52 (260.12) [− 1728.36 − 708.68]

(b) Treatment domain mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public − Private	Public − No insurance	Private − No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	2519.35 (239.67) [2049.60 2989.10]	829.45 (87.41) [658.13 1000.77]	− 1689.90 (257.46) [− 2194.52 − 1185.28]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3207.10 (17.14) [3173.51 3240.69]	4.092 (4.30) [− 4.34 12.52]	− 2343.00 (180.83) [− 2697.43 − 1988.57]

The response variable in our study is the annual medical expenditure. The treatment variable is the health insurance type (public health insurance, private health insurance, and no health insurance). Public health insurance is sponsored by Chinese government and is the main health insurance type in China. Six relevant covariates are chosen from the health care questionnaire in our study: age, household register (urban, rural, other), annual household income, physical condition (healthy, just so-so/or a little sick, sick, very sick), chronic disease (yes, no), and treatment to illness (self-treatment, go to hospital, no treatment). Due to some nonresponse units, the final data had a sample size of 3,866. The data weights were adjusted to deal with the nonresponse issue.

We are interested in estimating the following parameters, $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ and $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ where I_{di} is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. When estimating $\hat{\pi}_{2ig}$, we use $b_i = 1$ in Equation (7) to obtain conservative estimates since it is difficult to verify SMAR assumption. For comparison, we also report the results using $b_i = w_{1i}^{-1}$ in Equation (7).

Estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ are calculated and the Jackknife variance estimator discussed in Subsection 3.3 is used to calculate their standard errors. $\hat{\boldsymbol{\theta}}_g^{(2)}$ is not included into the empirical study becauseπ_{1ij} are not available. Since the design is a stratified multi-stage cluster design, we use the districts (PSUs) in different strata as the deleted Jackknife groups S_b. The Jackknife variance estimator is,

$$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right) = \sum_{h=1}^{2} \frac{B_h - 1}{B_h} \sum_{b=1}^{B_h} \left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)\left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)^T, \qquad (50)$$

where $\hat{\boldsymbol{\theta}}_g^{(1)[b]}$ is the minimizer of Equation (47) and the replicate weight in the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) \neq h(b) \\ \dfrac{B_h}{B_h - 1}\pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) = h(b). \end{cases} \tag{51}$$

Here $h(i)$ is the stratum where unit i belongs to, $h(b)$ is the stratum where the b^{th} deleted group S_b belongs to, and $[B_1, B_2] = [40, 100]$. The replicate estimator $\hat{\boldsymbol{\theta}}_g^{nw[b]}$ for the estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ without survey weights and the variance estimator $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ can be obtained in the same way by simply replacing π_{1i} by nN^{-1} in (51). A spline base of degree 2 with 8 equally spaced knots in the data range is constructed for the two continuous variables (age and annual household income). Dummy variables are created for the remaining categorical variables and added to the model.

Table 3 and 4 contain the estimated treatment mean effects and estimated treatment domain mean effects for physical condition, along with standard errors (in parentheses) and 95 perecnt confidence intervals (in brackets), for $b_i = 1$ and $b_i = w_{1i}^{-1}$ cases respectively. The treatment effect estimates in Table 3(a) indicate that, when the data weights are neglected, the estimated mean medical expenditure of the public health insurance group is not significantly different from that of the no health insurance group. However, when the data weights are incorporated, the public health group is found to spend significantly more on the medical expenses than the no health insurance group. This makes sense because people who have no health insurance might be reluctant to spend money to see doctors. This trend is also seen in the domain treatment effects estimates in Table 3(b). In addition, when the data weights are neglected for the treatment mean effect estimates, the estimated mean medical expenditure of the private health insurance group is significantly different from that of the no insurance group, while incorporating the data weights finds these estimated means not significantly different. Table 4 gives the same story as Table 3 when comparing the public health insurance group versus the private health insurance group, and comparing the public health insurance group versus the no health insurance group. However, when comparing the private health group with the no insurance group, Table 4 reports significant difference in the treatment mean effect for both estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$. Note that the standard errors of the unweighted estimator are not consistently smaller than those of the weighted estimator because the variation of weights in the real data is small (the CV = 0.45).

This study demonstrates that our method is feasible in real data application and suggests that ignoring the weights of an observational data might lead to a misleading conclusion.

6. Conclusions

In this article, we consider a GMM estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ to estimate treatment effects defined through an estimation equation in an observational data set that is a sample drawn by a complex survey design. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ include both the first-phase sampling probabilities and the estimated second-phase selection probabilities to remove the biases due to ignoring unequal sampling design in the first-phase and the selection biases in the second-phase. The self-selection probabilities are estimated using a semiparametric approach in Cattaneo (2010) to deal with the situation with multiple treatments. Our simulation studies demonstrate that neglecting the first-phase design and handling only treatment selection could lead to erroneous treatment effect estimation. The proposed estimator is designed to handle multiple treatments and do not require strong model assumption of the selection probability as in a fully parametric solution. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be readily extended to multiple sampling phases as well when the data set is a subsample of a larger survey sample.

Appendix

The notation of $|\cdot|$ represents the norm of a matrix, defined as $|A| = \sqrt{trace(A'A)}$ and the notation of $||\cdot||$ denotes the sup-norm in all arguments for functions.

We first give regular conditions on the sample designs in both phases. The following notations, I_i, π_i and π_{ij}, denote the sampling indicator, the first and second inclusion probabilities either for the first-phase design or for the second-phase design. For example, $I_i = \delta_{1i}$ or $I_i = \delta_{2ig}$ for any g, and $\pi_i = \pi_{1i}$ or $\pi_i = \pi_{2ig}$ for any g, depending on whether the design if the first-phase design or the second-phase design.

Condition A:

(1) Any variable v_i such that $E[|v_i|^{2+\delta}] < \infty$, where $\delta > 0$, satisfies $\sqrt{n}(\bar{v}_{HT} - \bar{v}_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, V_\infty)$ *a.s.*, where $(\bar{v}_{HT}, \bar{v}_N) = N^{-1}\sum_{i=1}^N (\pi^{-1}v_iI_i, v_i)$, $V_\infty = lim_{N\to\infty}V_N$, and $V_N = nV(\bar{v}_{HT}|\mathcal{F}_N)$ is the conditional variance of the Horvitz-Thompson estimator (Horvitz and Thompson 1952), $\bar{v}_{HT}$, given $\mathcal{F}_N$.

(2) $nN^{-1} \to f_\infty \in [0, 1]$.

(3) There exist constant C_1, C_2 and C_3 such that $0 < C_1 \le nN^{-1}\pi_i^{-1} < \infty$, and $\left|n(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1}\right| \le C_3 < \infty$ *a.s.*

Condition A(1) and A(2) are regular conditions assumed for a survey design in a finite population framework. Condition A(3) is used in Fuller (2009). The part of condition A(3) related to the joint selection probabilities is used in the proofs to bound sums of covariance induced by the sample design. Condition A(3) holds for simple random sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)(N-1)^{-1}N - 1$, and for Poisson sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = 0$, and can hold for cluster sampling and stratified sampling. Fuller (2009) explains that a designer has the control to ensure condition A(3). Note that for the second-phase design in our situation, $(\pi_{2ij,g} - \pi_{2ig}\pi_{2jg})\pi_{2ig}^{-1}\pi_{2jg}^{-1} = 0$ for any g because our second-phase design is a multinomial extension of Poisson sampling.

Next we give regular conditions on the tuning parameters of the semiparametric basis. For simplicity, we consider the special case of power series and spline series.

Condition B:

(1) The smallest eigenvalue of $E[R_K(X_i)R_K(X_i)']$ is bounded away from zero uniformly in K.

(2) There exists a sequence of constant $\zeta(K)$ such that $\|R_K(X_i)\| \leq \zeta(K)$ for $K \longrightarrow \infty$ and $\zeta(K)K^{1/2}n^{-1/2} \longrightarrow 0$.

(3) For all g, $\pi_{2ig}(X_i)$ and $\boldsymbol{\mu}_{mg}(X_i, \boldsymbol{\theta}_g) = E[\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i]$ are s-time differentiable with $sd_x^{-1} \geq 5\eta/2 + 1/2$, where d_x is the dimension of X_i, and $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.

(4) $K = O(n^{\nu})$ with $4sd_x^{-1} - 6\eta \geq \nu^{-1} \geq 4\eta + 2$, where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.

Condition B(1) and B(2) are standard assumptions and are automatically satisfied in the case of power series or spline series. Condition B(3) and B(4) describe the minimum smoothness required as a function of the dimension of X and the choice of basis, and the relationship between the sample size and the number of bases. Under B(3) and B(4), by Lorentz (1986), there exists a K-vector $\gamma^*_{g,K}$ for any g such that

$$\left\| log\left(\frac{\pi_{2ig}(X)}{1 - \sum_{g=2}^{G} \pi_{2ig}(X)}\right) - R_K^T(X)\gamma^*_{g,K}\right\| = O\left(K^{-\frac{s}{\nu}}\right), \tag{52}$$

where $R_K^T(X)\gamma^*_{g,K}$ is the best L_∞ approximation for the logarithm of the odds ratio of treatment g to the base treatment. The property (52) is used to derive the convergence rate of $\hat{\pi}_{2ig}$ to π_{2ig} as follows,

$$\|\hat{\pi}_{2ig} - \pi_{2ig}\| = O_p(\xi(K)K^{1/2}n^{-1/2} + \xi(K)K^{1/2}K^{-s/d_x}) = o_p(1). \tag{53}$$

For details, see Theorem B-1 of Cattaneo (2010).

Next we give regular conditions on the estimation equation function $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$.

Condition C:

(1) $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ is differentiable with respect to $\boldsymbol{\theta}_g$.

(2) Both $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ and its first derivative with respect to $\boldsymbol{\theta}_g$ have bounded $2 + \delta$ moments. More specifically, $E[|h(Y_i, Z_i; \boldsymbol{\theta})|^{2+\delta}] < M$, where $h\left(Y_i, Z_i; \boldsymbol{\theta}\right)$ denote an element of $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ or an element of its first derivative with respect to $\boldsymbol{\theta}_g$.

(3) $\Gamma_g\left(\boldsymbol{\theta}_g^0\right)$ is full rank.

(4) Assume that $\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})$ converges to 0 uniformly in $\boldsymbol{\theta}$, where $\bar{h}_{HT}(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} h_i(Y_i, Z_i; \boldsymbol{\theta})$, $\bar{h}_N(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} h_i(Y_i, Z_i; \boldsymbol{\theta})$, and $h_i(Y_i, Z_i; \boldsymbol{\theta})$ has the same interpretation as in condition C(2) above. This condition means that for all $\epsilon > 0$, there exists a $\delta > 0$ such that $Pro(|\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})| > \epsilon) < \delta$, for all N greater than some value M, and for all $\boldsymbol{\theta}$.

A: Proof of Theorem 1

The proof of Theorem 1 proceeds in two steps. The first step is to show that the asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g\right)$,

$$\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i \in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\frac{\delta_{1i}(\delta_{2ig} - \pi_{2ig})}{\pi_{1i}\pi_{2ig}}\boldsymbol{\mu}_{mg}(X_i; \boldsymbol{\theta}_g) + o_p(n^{-1/2}), \tag{A.1}$$

where $\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)=E_\xi((\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i)$. In order to show (A.1), we first decompose $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ into

$$\begin{aligned}\frac{1}{N}\sum_{i\in A_{2g}}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}}&=\frac{1}{N}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}}-\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}+\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig})\right\}\\&+\frac{1}{N}\sum_{i\in A_1}\left\{-\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig})+\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig})\right\}\\&+\frac{1}{N}\sum_{i\in A_1}\left\{-\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig})+\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\}\\&+\frac{1}{N}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}-\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\}.\end{aligned}\tag{A.2}$$

By the result in (53), the first three terms in (A.2) can be shown to have order $o_p(n^{-1/2})$ asymptotically, which leads to Equation (A.1). Similar arguments can be used to show $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g)=\frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)+o_p(n^{-1/2})$. The justification of those orders follows Cattaneo (2010), and we refer readers to Cattaneo (2010) for details.

The second step is to show the following two conditions of Pakes and Pollard (1989) hold: (1) $sup_{\boldsymbol{\theta}_g\in\Theta}|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))|=o_p(1)$, and (2) for every sequence of real numbers $\delta_n\rightarrow 0$, $sup_{|\boldsymbol{\theta}_g-\boldsymbol{\theta}_g^0|\le\delta_n}\left|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))-\bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g^0\right)\right|=o_p(n^{-1/2})$. By Equation (A.1), we can show that

$$\begin{aligned}E(\bar{\mathbf{m}}_{2\pi g}-E(\mathbf{m}_g(\theta_g))^2&=E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}}-\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}-E\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\right)\right)^2\\&\quad+o\left(n^{-1/2}\right)\\&\le 2T_{1N}+2T_{2N}+o\left(n^{-1/2}\right),\end{aligned}\tag{A.3}$$

where $T_{1N}=E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}}-E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))\right)^2$ and $T_{2N}=E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}\right)^2$. It is easy to show $T_{1N}=O(N^{-1})$ and $T_{2N}=O(N^{-1})$. Then we have $E(\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_g(\boldsymbol{\theta}_g)))^2=O\left(\frac{1}{N}\right)\Rightarrow\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_g(\boldsymbol{\theta}_g))=o_p(1)$. Condition (1) of Pakes and Pollard (1989) holds. Similarly, we can show that $sup_{(\boldsymbol{\theta}_g,\mu_z)}|\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g,\mu_z)-E(\boldsymbol{H}_{ig}(\boldsymbol{\theta}_g,\mu_z))|=o_p(1)$.

By Equation (A.1), we can also show that $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_g(\boldsymbol{\theta}_g))-\bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g^0\right)=T_{3N}-T_{4N}+o_p\left(n^{-1/2}\right)$, where $T_{3N}=\frac{1}{N}\sum_{i\in U}\frac{\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}}-E\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)$ and $T_{4N}=\frac{1}{N}\sum_{i\in U}\frac{E\left[\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)|X\right](\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}$. When $\left|\boldsymbol{\theta}_g-\boldsymbol{\theta}_g^0\right|\le\delta_n$, we have

$$E\left(T_{3N}^2\right)=\frac{1}{N}Var\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)$$
$$+E\left[\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\frac{\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)}{\pi_{1i}}\frac{\mathbf{m}_{jg}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{jg}\left(\boldsymbol{\theta}_g^0\right)}{\pi_{1j}}\right]$$
$$+E\left[\frac{2}{N^2}\sum_{i\in U}\left(\frac{1}{\pi_{2ig}}-1\right)\frac{\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)^2}{\pi_{1i}}\right]\leq\frac{1}{N}O\left(\delta_n^2\right)=o\left(\frac{1}{N}\right) \tag{A.4}$$

$$E\left(T_{4N}^2\right)\leq E\left[\frac{1}{N^2}\sum_{i\in U}E\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\middle|X\right)^2\right]$$
$$\leq E\frac{1}{N}E\left[\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)^2\middle|X\right]\leq\frac{1}{N}O\left(\left|\theta_g-\theta_g^0\right|^2\right)=o\left(\frac{1}{N}\right). \tag{A.5}$$

Then we have $T_{3N}=o_p(n^{-1/2})$ and $T_{4N}=o_p(n^{-1/2})$ when $|\boldsymbol{\theta}-\boldsymbol{\theta}^0|\leq\delta_n$, thus Condition (2) of Pakes and Pollard (1989) is verified. Similarly, we can show that for every sequence of real numbers $\delta_n\longrightarrow 0$,

$$\sup_{\left\|\begin{bmatrix}\boldsymbol{\theta}_g\\ \mu_z\end{bmatrix}-\begin{bmatrix}\boldsymbol{\theta}_g^0\\ \mu_z^0\end{bmatrix}\right\|\leq\delta_n}\left|\bar{\mathbf{H}}_{2\pi g}\left(\boldsymbol{\theta}_g,\mu_z\right)-E\left(\mathbf{H}_{ig}\left(\boldsymbol{\theta}_g\right)\right)-\bar{\mathbf{H}}_{2\pi g}\left(\boldsymbol{\theta}_g^0,\mu_z^0\right)\right|=o_p(n^{-1/2}). \tag{A.6}$$

For a vector $c=[c_1,c_2]^T$, we know $|c|\leq\sqrt{2}(|c_1|+|c_2|)$. Therefore, Condition (1) and (2) of Pakes and Pollard (1989) in terms of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)$ can be verified. The details of the proof can be obtained upon request.

7. References

Ashmead, R. 2014. "Propensity Score Methods for Estimating Causal Effects from Complex Survey Data." Ph.D. Dissertation, Ohio State University. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1417616653.

Berg, E., J.K. Kim, and C. Sinner. 2016. "Imputation under Informative Sampling." *Journal of Survey Statistics and Methodology* 4: 436–462. Doi: 10.1093/jssam/smw032.

Breidt, F.J., G. Claeskens, and J.D. Opsomer. 2005. "Model-Assisted Estimation for Complex Surveys Using Penalised Splines." *Biometrika* 92(4): 831–846. Doi: 10.1093/biomet/92.4.831.

Cattaneo, M.D. 2010. "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability." *Journal of Econometrics* 155(2): 138–154. Doi: 10.1016/j.jeconom.2009.09.023.

DuGoff, E., M. Schuler, and E. Stuart. 2014. "Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys." *Health Services Research* 49(1): 284–303. Doi: 10.1111/1475-6773.12090.

Fuller, W.A. 2009. *Sampling Statistics*, Vol. 56, John Wiley and Sons. Doi: 10.1002/9780470523551.

Hahn, J. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66(2): 315–331. Doi: 10.2307/2998560.

Haziza, D. and J.N.K. Rao. 2006. "A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data." *Survey Methodology* 32(1): 53. Doi: 12-001-X20060019257.

Hirano, K., G. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1161–1189. Doi: 10.1111/1468-0262.00442.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. Doi: 10.1080/01621459.1952.10483446.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. Doi: 10.1080/01621459.1982.10477770.

Kim, J.K. and D. Haziza. 2014. "Doubly Robust Inference with Missing Data in Survey Sampling." *Statistica Sinica* 24: 375–394. Doi: 10.5705/ss.2012.005.

Kim, J.K. A. Navarro, and W. Fuller. 2006. "Replication Variance Estimation for Two-Phase Stratified Sampling." *Journal of the American Statistical Association* 101: 312–320. Doi: 10.1198/016214505000000763.

Little, R.J.A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77: 237–250. Doi: 10.1080/01621459.1982.10477792.

Lorentz, G. 1986. *Approximating of Functions*. New York: Chelsea Publishing Company. Doi: 10.1112/jlms/s1-43.1.570b.

Pakes, A. and D. Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57: 1027–1057. Doi: 10.2307/1913622.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why Model? Why is it a Problem? How Can we Approach it?" *Survey Methodology* 37: 115–136. Retrieved from https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11602-eng.pdf?st=XWOwbI5k.

Pfeffermann, D. and M. Sverchkov. 1999. "Parametric and Semiparametric Estimation of Regression Models Fitted to Survey Data." *Sankhya B* 61: 166–186. Retrieved from http://www.jstor.org/stable/25053074.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable." *Statistical Science* 22(4): 544–559. Doi: 10.1214/07-STS227D.

Rosenbaum, P.R. and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. Doi: 10.1093/biomet/70.1.41.

Ridgeway, G., S.A. Kovalchik, B.A. Griffin, and M.U. Kabeto. 2015. “Propensity Score Analysis with Survey Weighted Data.” *Journal of Causal Inference* 3(2): 237–249. Doi: 10.1515/jci-2014-0039.

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer. Doi: 10.1007/978-1-4612-4378-6.

Tan, Z. 2006. “Regression and Weighting Methods for Causal Inference Using Instrumental Variables.” *Journal of the American Statistical Association* 101: 1607–1618. Doi: 10.1198/016214505000001366.

Tan, Z. 2008. “Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting.” *Biometrika* 94: 122. Doi: 10.1093/biomet/asq035.

Yu, C., J. Legg, and B. Liu. 2013. “Estimating Multiple Treatment Effects Using Two-phase Semiparametric Regression Estimators.” *Electronic Journal of Statistics* 7(2013): 2737–2761. Doi: 10.1214/13-EJS856.

Zanutto, E. 2006. “A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data.” *Journal of Data Science* 4: 67–91. Retrieved from http://www.jds-online.com/v4-1.

Received June 2016
Revised October 2017
Accepted November 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 785–786, http://dx.doi.org/10.2478/JOS-2018-0036

Book Review

William Cecere[1]

Jeremy Dawson. *Analysing Quantitative Survey Data*, 2017; Sage Publications Ltd. ISBN 978-1-4739-0751-5, 88 pp, USD 40.

This book targets non-experts conducting a survey and subsequent analysis for the first time. It is part of a larger series of books called *Mastering Business Research Methods.* This series is for Master's level students doing research for a dissertation in the business and management disciplines. The focus of this edition is primarily on reliability and validity of survey items under the framework of classical test theory. The examples in the book are given using primarily SPSS software. I found this edition easy to understand and apply for someone new to surveys.

The first chapter "An Introduction to Classical Test Theory and Quantitative Survey Data" gives a motivation for using surveys due to their flexibility with types of data collected. Six different types of questionnaire data are identified, grouped into categorical and numeric major types, in addition to a discussion of Likert-scale data. Classical test theory is introduced along with the assumptions surrounding it. This provides a more formal outline to the later discussion of reliability and validity testing. The author classifies three types of analysis using survey data: data tidying (including reliability and validity checks), descriptive analysis, and inference.

The second chapter approaches the framing of the concepts of reliability and validity with reference to classical test theory. The author lays out the philosophical underpinnings of positivism and interpretivism. Positivism is a philosophy holding that every rationally justifiable assertion can be scientifically verified or is capable of logical or mathematical proof. Interpretivism is a counter to positivism, claiming that the social realm requires a different epistemology for study. The author describes how interpretivism is appropriate for survey topics since the meaning of the questions depends on the respondents' understanding of it. This is a nice contrast between the social science and natural science philosophical platforms. The remainder of this chapter covers many examples of the types and aspects of validity and reliability analysis, hitting on many common sources of survey error. The chapter concludes with a discussion of the advantages of using multi-item scales.

Chapter 3 discusses steps that an analyst will need to go through to get multi-item scale variables ready for analysis. A nice flow-chart of the steps for analysis is shown to illustrate. The author provides detailed examples and tips for how to enter, code, and do basic summary statistics for survey data in SPSS. These will be especially helpful to a new

[1] Westat Inc, 1600 Research Boulevard, Rockville, Maryland, 20850-3195, U.S.A. Email: WilliamCecere@westat.com

practitioner of survey analysis. Factor analysis is introduced as a tool for establishing the validity of multi-item scales. Exploratory and confirmatory factor analysis are explained and a decision process is given to help determine how to properly use each. How to create ‘scale-scores’ using SPSS is then displayed and the chapter closes with a table of standard statistical tests for examining relationships within survey data.

The real depth necessary for a practitioner to apply validity and reliability analysis are provided in Chapter 4. The instructions and recommendations Dawson provides are geared towards management students using Likert scale data. Issues such as extracting factors, determining the number of factors, which variables should be excluded, and factor interpretation are covered using SPSS output for exploratory factor analysis. It is mentioned that confirmatory factor analysis is a special case of structural equation modeling and is not covered in SPSS. Several options are listed and the examples are shown using the software Amos. Useful diagrams and screenshots are provided to help the user construct and interpret the models. Finally, basic techniques for reliability analysis using SPSS are displayed.

Chapter 5 walks through three case study examples to illustrate the use of the techniques laid out in the previous chapter. The first case study looks at perception of job role and its impact on employee performance as an example of exploratory factor analysis. The second example describes a study that uses confirmatory factor analysis to validate the existence of five levels of affective well-being. The author cites a paper showing two studies where reliability analysis is used to examine factors regarding attitudes towards affirmative action. This chapter provides a good structure for how these methods can be incorporated into a study to help answer a wider set of questions.

Chapter 6 examines the role and assumptions of each method outlined when analyzing quantitative survey data. The author attempts to assess other options compared to classical test theory. Opinions on reliability and a comparison of principal component analysis versus factor analysis are also provided. A helpful table is given weighing the positives and negatives of different software options for performing confirmatory factor analysis. Since this book does not cover the theory or background to the methods used, this is a helpful perspective to give first-time users so that they might know of some limitations and alternatives.

This book covers what the author calls data tidying and focuses primarily on methods of assessing reliability and validity under the classical test theory framework. The title suggests a broader scope than the book covers. Perhaps a title hinting at data tidying would have been more appropriate. The nature of the text is quite hands-on with many tips and examples for beginners but a more advanced survey practitioner will require additional resources. Although simple in scope, Dawson provides a valuable resource to his target audience of Master students using survey data for the first time.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 787–789, http://dx.doi.org/10.2478/JOS-2018-0037

Book Review

Stephanie Coffey[1]

Barry Schouten, James Wagner, and Andy Peytchev. *Adaptive Survey Design*. 2017 Boca Raton: CRC Press, ISBN 978-1-4987-6787-3, 252 pp, USD 89.95.

Adaptive survey design (ASD) is earning increased attention as a framework for maintaining or improving survey quality. Noncontact, nonresponse, and the cost of carrying out data collection are all increasing at varying rates. At the same time, the increased computerization of survey operations, along with the increased processing power of computers, means it is possible to generate, aggregate, and process more paradata and survey data, helping survey methodologists, managers, and statisticians understand characteristics of data collection at a more detailed level. *Adaptive Survey Design*, the timely new book by Barry Schouten, James Wagner, and Andy Peytchev, places itself squarely in this environment, providing both motivation for and detailed guidance on implementing ASDs to improve survey outcomes.

This is the first published book addressing the developing field of ASD, and as a result, the authors cover a significant amount of material across five major sections. Section I, *Introduction to Adaptive Survey Design*, lays the foundation for the rest of the book through the introduction of several concepts that return throughout the text. Standard survey methodology topics including survey costs, survey errors, and the variability of survey implementations are tied together to motivate the need for the flexibility to adapt data collection protocols in order to improve survey outcomes. The authors also thoughtfully discuss the nomenclature of ASDs, responsive designs, and their interaction. Their definitions and context provide clear boundaries for what the authors will discuss throughout the book. This is useful for survey practitioners, whether new or familiar with the material, as what qualifies as an ASD is not always consistent in the working literature. This section ends with the introduction of several case studies that return throughout the book to illustrate concepts.

Section II, *Preparing an Adaptive Survey Design*, discusses three components required before implementing an ASD: strata, design features, and models for nonresponse. First, this section covers the process of stratification – stressing that strata should be based on covariates related to survey variables and likelihood of response to different data collection protocols. Data collection features, such as incentives, mode of contact, or case prioritization, are tailored to specific strata and are discussed second. The hope is that the application of particular data collection features to specific strata will result in superior survey outcomes to those obtained through a traditional, non-adaptive designs. Last, this

[1] U.S. Census Bureau, Center for Adaptive Design, 4600 Silver Hill Road, Washington, DC, 20233. U.S.A. Email: stephanie.coffey@census.gov

section addresses statistical models that may be used during an ASD. Two main categories of models are discussed – those that can be used to support changes in data collection procedures, and those that are used to monitor the potential for nonresponse bias. The authors acknowledge that the quality of the strata and models for nonresponse are directly tied to the predictive power of the available auxiliary data, which can vary considerably. Short discussions of potential sources of additional data (such as commercial data or paradata) and methods for stratification (such as response propensity variation, influence on estimates, or machine learning algorithms) point the reader in helpful directions, while making it clear that preparing for an ASD will have some unique elements from survey to survey.

Section III, *Implementing an Adaptive Survey Design*, discusses other aspects of implementing ASDs, and opens with a discussion of costs and logistics. The authors note that, in order to assess cost-quality tradeoffs, survey practitioners need cost models, but they are often very difficult to estimate. The authors help the reader conceptualize a cost model designed for an ASD, and then illustrate the estimation of those model parameters through regression. They also mention the use of expert opinion in constructing cost parameters. This section also discusses the optimization of adaptive survey design, and distinguishes between trial-and-error and numerical optimization, either through mathematical optimization or simulation. Each of these methods have strengths and weaknesses, and the authors suggest using them to validate one another, to the extent possible. Lastly, this section addresses the robustness of ASDs. Both Section II and Section III discuss the estimation of various design parameters, including response to various data collection features, and the related costs of those features. Here, the authors address how inaccuracy in those design parameters can impact the success of ASDs.

Section IV, *Advanced Features of Adaptive Survey Design*, includes the most statistical content of the book, and addresses two main topics. The first chapter reviews the more common indicators of nonresponse bias used in the literature, and classifies them into two types – those that rely only on covariates and a survey response indicator, and those that additionally rely on response data. Again, this requires consideration of the quality of the available covariates or auxiliary data. The second chapter addresses the "during or after" argument – that is, is it worth it to undertake the statistical and operational complexities to design and execute an ASD, or can the same reductions in nonresponse bias be attained through nonresponse adjustments using available covariates? The authors discuss some theoretical evidence of the potential for ASDs to reduce nonresponse bias, even after nonresponse adjustment, and the conditions required for bias reduction. Illustrative examples are provided using the introduced case studies and are particularly helpful here.

In Section V, *The Future of Adaptive Survey Design*, the authors propose a research agenda to further ASD, and itemize nine areas in three categories requiring further experience or research. The first category focuses on proving the utility of ASDs through accumulating evidence of success of ASDs across a range of designs. The second category includes topics related to the implementation of ASDs, including the need to explore statistical methods to inform adaptation (such as Bayesian models for incorporating information) and the need for flexible survey software that can accommodate more complex adaptation. The last category focuses on methodological advances for furthering ASD, including designing paradata to meet the needs of ASD, the optimization of decision making through numerical methods, and ASD's ability to address sources of error beyond nonresponse and cost. This section also offers a more expanded discussion of ASD for reducing measurement error, in particular. The authors make it clear through this agenda that there are open questions for exploration throughout the survey lifecycle.

The book is a success due to its accessibility and applicability. *Adaptive Survey Design* is written to appeal to a broad range of survey practitioners: methodologists, managers, and statisticians. The only real prerequisite is an understanding of the survey lifecycle process – how surveys are designed, conducted, processed, and analyzed. This is by design – the authors are clear that implementing an ASD *should* involve individuals throughout the survey process, as their knowledge and involvement is key to implementing adaptation successfully. At the same time, the authors provide mathematical detail for those interested, and clearly identify gaps in the existing research and unanswered questions for survey practitioners to consider.

Beyond clear organization and communication, what sets this book apart is the inclusion of case studies from the authors' own experiences with adaptive design. The examples include a random-digit-dial telephone survey, a multimode survey that can be linked to administrative data, and an in-person interview made up of a screener and a personal interview. By including such varied examples, the content and examples in the book are applicable to a wide range of data collection designs. Many aspects of surveys will evolve in the future, from cost and nonresponse patterns to available auxiliary data and design features. However, the underlying concepts of adapting data collection that are detailed in this book will continue to be valuable.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 791–793, http://dx.doi.org/10.2478/JOS-2018-0038

Book Review

Hanyu Sun[1]

Jelke Bethlehem. *Understanding Public Opinion Polls. 2018 Boca Raton*, FL: Taylor & Francis Group. ISBN-13: 978-1-4987-6974-7, 286 pp, USD 47.96.

With the increasing use of public opinion polling to collect information on almost every aspect of modern life, it becomes more and more difficult for people to separate the good polls from the bad ones and to determine whether the outcomes of a poll are valid or not. An easy-to-follow guidance is much needed. This book "Understanding Public Opinion Polls" by Dr. Jelke Bethlehem is written for professionals who have no prior training in relevant fields but need to judge the outcome of polls such as journalists, politicians, and decision makers among other non-experts in polling. The book, in the author's own words, is a "not-so-technical introduction" to public opinion polling. It can be viewed as providing a checklist to evaluate the quality of different polls or a how-to cookbook for conducting polls with appropriate methodologies that result in valid outcomes. The book covers all elements of conducting a poll, including the definition of a target population, the questionnaire design, the mode choice, sample design, weighting, data analysis, and the publication of the results. It is not uncommon for an "all-inclusive" book on survey statistics and methodology to be "too heavy" for entry-level readers. This book achieves a balance by including examples in each chapter to make it easier for readers to digest the relevant information on statistics and methodology.

Chapter 1 of the book provides an overview of the book's content. Chapter 2 gives a brief history of polls starting from their origin in ancient Greece to the development of modern sampling theory, the emergence of public opinion polls, and finally the rise of online polls. The remainder of the chapters can be grouped into three parts. Part One includes Chapter 3 to Chapter 7 and each covers a key component of a poll. Part Two focuses on two specific types of polls that are of the most interest to readers: online polls in Chapter 8 and election polls in Chapter 9. Part Three of the book is about data analysis and publication. The book concludes with Chapter 12 that summarizes all chapters and provides a checklist of poll quality for readers.

Part One includes five chapters with each covering a vital component of a poll. Chapter 3 provides an overview of the questionnaire design covering topics such as the types of survey questions, the question order effects, and questionnaire pretesting methods. The author recommends ignoring the poll outcomes if the instrument is poorly designed. Chapter 4 describes four major modes of data collection (i.e., face-to-face, telephone, mail, and online polls). The author provides pros and cons for each mode in terms of data quality and costs. After reading this chapter, it should become clear to readers that some

[1] Westat Inc, 1600 Research Blvd. Rockville Maryland 20850-3195, U.S.A. Email: hanyusun@westat.com

tradeoff has to be made when selecting the mode of data collection. Chapter 5 covers sampling as another key component of a poll. Throughout the book, the author emphasizes the importance of selecting a random sample using probability based methods so that unbiased estimates of population characteristics can be computed and the precision of the estimates can be determined. The author elaborates more on this topic in Chapter 5. Instead of reviewing the theory of sampling – which can be "too heavy" for readers without relevant background—the author provides step-by-step instructions on how to draw a random sample using devices such as the random number generator or a spreadsheet. In this chapter, the author also reviews quota sampling and issues associated with self-selection to prepare readers for the upcoming review of online polls. Chapter 6 is on estimation. Again, the author uses easy-to-follow examples to explain concepts associated with estimates, estimators and the margin of error. With an example, the author demonstrates how to estimate a population mean and percentage as well as how to determine the sample size. The first four chapters of Part One cover what constitutes a good poll. In Chapter 7, the author changes gear to talk about why nonresponse occurs, its consequences, and how to use adjustment weighting to remove or reduce the bias.

In Part Two, the author applies the guidance that was provided in Part One to examine two types of polls that are of the most interest– online polls and election polls. In Chapter 8, the author discusses issues of online polls under the total survey error framework. Online polls have become more and more popular because they seem to be easy, cheap, and fast at collecting large amounts of data. However, they also suffer from serious methodological issues, such as undercoverage, self-selection bias, nonresponse, and measurement errors. Readers should have a more comprehensive view of online polls after reading this chapter. Next, the author changes subjects and describes election polls in Chapter 9. This includes both pre-election polls and exit polls. Pre-election polls are conducted before an election takes place whereas exit polls are conducted on the day of the election. The author describes the pros and cons for both types of polls. As in previous chapters, the author provides examples in Chapter 8 and Chapter 9 to help readers understand what is covered.

Part Three of the book is devoted to analysis and publication. Chapter 10 is about data analysis. It mainly focuses on exploratory analysis such as how to examine the distribution of a single variable, how to examine the relationship between two variables, and how to present data in graphs such as bar charts, boxplots, and scatterplots. The author uses a small data set and the open source software R to illustrate the type of analysis covered in the chapter. This is a cheap and simple approach for readers who do not have access to expensive statistical software. Chapter 11 describes how to produce a research report—the key components of a research report and how to present the poll outcomes using graphs in a meaningful way. The book concludes with Chapter 12, summarizing all topics covered in previous chapters and providing a checklist for readers to determine the quality of a poll. If readers are running short of time, they can go to Chapter 12, particularly the checklist (i.e., Table 12.1), to guide their evaluation of a poll's quality.

In summary, this is a comprehensive how-to guide for readers who need to judge the outcomes of polls and who want to conduct polls but don't have the necessary training in relevant fields. The book covers all the key components of a poll and provides ample examples to explain abstract concepts and procedures. The examples are not only relevant to what is being described but also will allow readers to connect them with the polls they see on a regular basis. The idea that readers will be able to evaluate the quality of a poll by marking yes or no for each item on the checklist from Chapter 12 is quite useful. If there are many “No” answers, it is clear that the outcomes of the poll cannot be taken seriously. But what if the answers to five or six of the nine checklist items are “Yes”, what shall the reader do? And does it matter which five or six items are marked as “Yes”? That is, are all the items equally important when judging the poll quality? Questions like these may motivate readers to dive deeper and learn more about statistics and survey research. Nevertheless, this book is a well-written introduction for anyone who is interested in public opinion polling.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 795–796, http://dx.doi.org/10.2478/JOS-2018-0039

Erratum
Optimal Stratification and Allocation for the June Agricultural Survey

Jonathan Lisic, Hejian Sang, Zhengyuan Zhu, and Stephanie Zimmer

Erratum concerning the article "Optimal Stratification and Allocation for the June Agricultural Survey" by Jonathan Lisic, Hejian Sang, Zhengyuan Zhu, and Stephanie Zimmer published in Journal of Official Statistics, Volume 34, Number 1, 2018, pages 121–148 (https://doi.org/10.1515/jos-2018-0007).

This article has an error and related omission in the literature review, as well as some errors in the specification of the simulation. Neither of these errors affect any results in the paper or conclusions drawn.

The error and related omission in the literature review occur on page 122, paragraph 2. The reference, Lavallée and Hidiroglou (1988) is incorrect and should be replaced with a reference to Hidiroglou (1986). Both papers are similar in that they provide methods to optimally stratify and allocate univariate populations into take-all, take-none, and take-some stratum under a coefficient of variation (CV) constraint. However, Lavallée and Hidiroglou (1988) improves on Hidiroglou (1986) by allowing for an arbitrary number of take-some strata. This important contribution should have been included on page 122 paragraph 2, revised below.

> One major advantage that a priori and conditional allocation designs have over optimal stratified designs is that they are easy to obtain. Optimal stratified designs require an exploration of a combinatorial space to find an optimal design. This is a non-trivial problem for even small population and sample sizes. A solution to the problem of finding a univariate optimal stratified design using Neyman allocation for a fixed sample size was proposed by Dalenius and Hodges (1959). This method is commonly known as the cum $\sqrt{f}$ method (Särndal et al. 1991, Section 3.7 and Horgan 2006). Similar methods such as Hidiroglou (1986) and the multivariate extensions in Benedetti et al. (2010) and Benedetti and Piersimoni (2012) provide optimal designs under CV constraints, but are restricted to no more than three strata. These strata include a census (take-all), a sampled (take-some), and an unsampled (take-none) stratum for cut-off sampling. Lavallée and Hidiroglou (1988) introduced a univariate method that allows for an arbitrary number or take-some strata for the univariate case. These approaches are designed for highly skewed populations, exploiting the similarity of the underlying population to a geometric progression (Gunning et al. 2004). Benedetti and Piersimoni (2012) introduced a method for stratification which uses multiple administrative variables. This method, which is motivated by the Lavallée and Hidiroglou method, partitions the population into two strata, one which is sampled and one, which is a take-all stratum.

The partitioning is determined such that the sample size is minimized for a target coefficient of variation of a response variable. In addition to allocations with goals of increasing precision, allocations also consider data collection costs and other practical constraints such as the method proposed by Valliant et al. (2014) to allocate sample in household surveys using Address-Based Sampling Frames and available commercial data.

The errors in the specification of the simulation occur in two areas. First, the sample size of the univariate homoscedastic case should of been $n = 70$ instead of $n = 23$. Second, on page 134 after Equation (11) z_i was missing the γ exponent in relation to v_i; v_i is correctly defined as $v_i = \bar{z}_i^{\gamma}$ in the homoscedastic case and $v_i = z_{1,i}^{\gamma}$ in the heteroscedastic case.

References

Benedetti, R., M. Bee, and G. Espa. 2010. "A Framework for Cut-Off Sampling in Business Survey Design." *Journal of Official Statistics* 26(4): 651–671.

Benedetti, R. and F. Piersimoni. 2012. "Multivariate Boundaries of a Self Representing Stratum of Large Units in Agricultural Survey Design." *Survey Research Methods* 6: 125–135.

Dalenius, T. and J.L.J. Hodges. 1959. "Minimum Variance Stratification." *Journal of the American Statistical Association* 54: 88–101.

Gunning, P., J. Horgan, and W. Yancey. 2004. "Geometric Stratification of Accounting Data." *Contaduría y Administración*.

Hidiroglou, M.A. 1986. "The Construction of a Self-Representing Stratum for Large Units in Survey Design." *The American Statistician* 40(1): 27–31.

Horgan, J.M. 2006. "Stratification of Skewed Populations: A Review." *International Statistical Review* 74: 67–76.

Lavallée, P. and M. Hidiroglou. 1988. "On the Stratification of Skewed Populations." *Survey Methodology* 14: 33–43.

Särndal, C.-E., B. Swensson, and J. Wretman. 1991. *Model Assisted Survey Sampling*. Springer.

Valliant, R., F. Hubbard, S. Lee, and C. Chang. 2014. "Efficient Use of Commercial Lists in US Household Sampling." *Journal of Survey Statistics and Methodology* 2(2): 182–209.