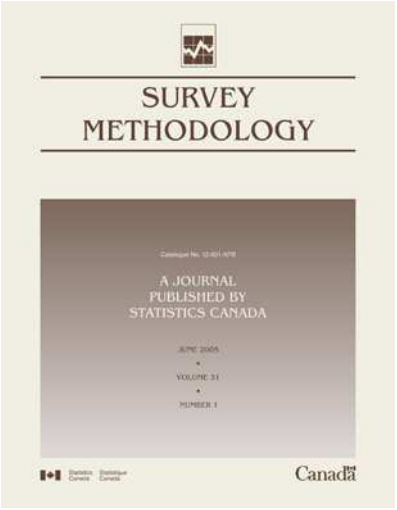
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website at www.statcan.ca.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Accessing and ordering information

This product, catalogue no. 12-001-XIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

This product, catalogue no. 12-001-XPE, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$15.00	CAN$30.00

All prices exclude sales taxes.

The printed version of this publication can be ordered by

- Phone (Canada and United States) 1 800 267-6677
- Fax (Canada and United States) 1 877 287-4369
- E-mail infostats@statcan.ca
- Mail Statistics Canada
  Finance Division
  R.H. Coats Bldg., 6th Floor
  100 Tunney's Pasture Driveway
  Ottawa (Ontario) K1A 0T6
- In person from authorised agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE, Vol. 32, no. 1
ISSN: 1492-0921

Catalogue no. 12-001-XPB, Vol. 32, no. 1
ISSN: 0714-0045

Frequency: Semi-Annual

Ottawa

La version française de cette publication est disponible sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

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

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Dedicated to the family of M.P. Singh: His wife, Savitri and his children, Mala, Mamta and Rahul


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology

A journal Published by Statistics Canada

Volume 32, Number 1, June 2006

Contents


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 1-2
Statistics Canada, Catalogue No. 12-001-XIE

In This Issue

This issue of the *Survey Methodology* journal opens with a special article to honour the memory of M.P. Singh, the founding Editor who led the journal for thirty years to its current stature as an internationally recognized source for new developments in survey methods and methods for production of official statistics. In this article many of M.P.'s closest colleagues and friends from over the years share their memories of him, and reflect on his career and his contributions.

In the first regular paper of this issue, Thompson discusses random walk designs for sampling from a networked population. He shows how this approach can lead to network samples where the inclusion probabilities can be estimated independently of how the initial sample of nodes is chosen, leading to valid design-based inference methods. Selection preference can be given to certain types of nodes or graph characteristics through choice of the random walk mechanism. He describes both uniform and targeted random walk designs, and presents some examples for illustration.

Durrant and Skinner consider the use of imputation and weighting to correct for measurement error in the estimation of a distribution function. They consider various nearest neighbor and hot-deck imputation methods, and propensity score weighting under different response models. They discuss the theoretical properties of these methods, and compare them via simulations to estimate the distribution of hourly pay based on United Kingdom Labour Force Survey data. They conclude that an approach based on fractional imputation seems best overall in terms of efficiency and robustness.

Harms and Duchesne look at the problem of estimation of quantiles using survey data. They calibrate an interpolated estimate of a distribution function to given quantiles of an auxiliary variable, and then invert the resulting calibrated interpolated estimator of the distribution function of the variable of interest. They compare their approach with other methods in a simulation study.

In their paper, Haziza and Rao propose a new regression imputation method that uses the response probabilities. The new method leads to valid estimators under either the nonresponse model approach or the imputation model approach. In the nonresponse model approach, the response mechanism is parametrically modelled and is not restricted to the uniform nonresponse model, while in the imputation model approach the variables of interest are modelled and nonresponse is assumed to be ignorable. The authors also provide estimators of the variance under their imputation method. Simulation results for both point and variance estimation are reported that show the good performance of the proposed regression imputation method.

The paper by Zanutto and Zaslavsky deals with the problem of estimation in the U.S. decennial census of population under sampling for nonresponse follow-up. Instead of trying to obtain information from all nonrespondents, a sample is drawn for follow-up, thus creating a small area estimation problem. The proposed strategy consists of predicting the number of nonrespondent households in different categories using a hierarchical loglinear model and then imputing detailed person and household information using donor imputation. The idea in the first step is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at higher levels of geography. The performance of the proposed model compares favourably to other models in a simulation study.

In Théberge's article, a new approach is proposed for 2006 Reverse Record Check (RRC) sample allocation for measuring census undercoverage and a part of overcoverage. RRC estimates are used together with census counts to produce population estimates which will be used in calculating Canadian federal government equalization payments to the provinces. The proposed approach will provide an allocation that achieves the proper balance between four objectives. It first consists of establishing a separate allocation for each objective. Then, by province, the maximum sample size is used for each allocation. Finally, the RRC's sub-provincial sample allocation is obtained by using calibration to smooth the stratum-level parameters.

In his paper Longford discusses how to design a survey when estimates are required for a number of small areas, with possibly different priorities for different small areas, by minimizing a weighted sum of expected variances. He first develops his ideas for direct estimation, and then extends to composite estimation that combines the direct estimator with a synthetic estimator. The approach is illustrated by the resulting sample allocations, under various assumptions, to cantons in a Swiss household survey.

You and Chapman propose a Hierarchical Bayes estimation approach for small area estimation when the sampling errors of direct estimators are estimated. They demonstrate the approach by producing small area estimates from two data sets and investigate the sensitivity of their approach to model assumptions.

Khoshgooyanfard and Monazzah compare synthetic, composite and Empirical Bayes small area estimation methods for producing intercensal estimates of unemployment rates for provinces in Iran. They find that both composite and Empirical Bayes approaches lead to satisfactory results.

The short note by Gabler, Häder and Lynn, the last paper in this issue, provides an interesting extension to the earlier paper by Gabler, Häder and Lahiri that appeared in Survey Methodology (1999). It offers a practical solution for obtaining design effects when different exclusive domains use different sample problems.

Finally, we note that *Survey Methodology* is now available on-line in a fully searchable pdf format. All articles published in the journal are now being made available free of charge directly on the Statistics Canada web site upon release. There are also plans to include past issues. All the articles from the latest seven issues have been posted, and work is in progress to add those of the previous 10 years. Printed copies of the journal will still be produced for subscribers. Older issues can be obtained upon request in paper or pdf scanned formats. The journal can be accessed from Statistics Canada's web site at www.statcan.ca/bsolc/english/bsolc?catno=12-001-X.

Harold Mantel, Deputy Editor

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 3-10
Statistics Canada, Catalogue No. 12-001-XIE

M.P. Singh Remembered

Introduction

Don Royce
Statistics Canada

In August of 2005 the world of survey methodology lost one of its leading figures with the death of Dr. M.P. Singh at the age of sixty-three, just a few months short of his planned retirement. M.P. and I had discussed his upcoming retirement only briefly, but it was intuitively clear to both of us that he would continue as the Editor of *Survey Methodology* even after he left Statistics Canada. *Survey Methodology* was a part of his life, and I was only too happy to offer M.P. the chance to work part-time from his family's home in Toronto so that he could continue to nurture the journal that he had led for over thirty years. Sadly, this arrangement never came to be realized.

In the series of articles that follow, many of M.P.'s closest colleagues and friends (the two are indistinguishable) recall M.P. Singh the statistician, editor, collaborator, leader, and human being. I am deeply indebted to Eric Rancourt of Statistics Canada for suggesting this series of articles, and to all of the authors who gave their time and talents to put into words their memories of M.P. Singh. Although words can never completely capture the essence of a person, the articles that follow do a marvellous job of describing the life of M.P. Singh and remind us of the legacy he left to all of us who were fortunate enough to know him. We hope M.P. would be pleased.

Some Reminiscences

J.N.K. Rao
Carleton University, Ottawa

I first met Mangala Prasad Singh (fondly known to many of us as M.P.) in 1968 while I was a visiting professor at the Indian Statistical Institute (ISI), Calcutta. M.P. was a Ph.D. student at the ISI working under the supervision of M.N. Murthy. While doing his Ph.D. he also worked in the National Sample Survey (NSS) of India. NSS was located in the ISI campus and M.P. worked under renowned survey statisticians at the NSS and ISI including P.C. Mahalanobis, D.B. Lahiri and M.N. Murthy. He received solid training in both design and theory of sample surveys. M.P. made good use of that sound training throughout his illustrious career by following the principles of efficient design subject to cost and operational considerations and insisting on sound theory before implementing new survey designs or redesigning continuing surveys such as the Canadian Labour Force Survey (LFS).

A major part of M.P. Singh's thesis was on the efficient use of auxiliary information. He studied the case of two auxiliary variables, one positively correlated and the other negatively correlated with the variable of interest, and developed ratio-cum-product estimators of totals. Murthy (1967) devoted a section in his well-known sampling book to ratio-cum-product estimators. M.P. published several papers on the efficient use of auxiliary information based on his thesis work: ratio-cum-product estimators (*Metrika* 1967; *Sankhyā* 1969), multivariate product estimation (*Journal of the Indian Society of Agricultural Statistics* 1967) and systematic sampling in ratio and product estimation (*Metrika* 1967). He also published an important paper in the *Annals of Statistics*, 1967 on the relative efficiency of two-phase sampling strategies under a super-population model. The first phase consisted of simple random sampling to collect data on an auxiliary variable x that was used at the second phase to select a PPS sample without replacement and collect data on the variable of interest, y.

M.P. was also dabbling with inferential issues in survey sampling at the time of my visit to ISI and he encountered technical problems in proving some admissibility results: An estimator is admissible in a class of unbiased estimators if no other estimator in the class is uniformly more efficient. Unfortunately, the criterion of admissibility is not sufficiently selective and as a result other admissibility related criteria for unique choice were proposed in the literature. I was also interested in inferential issues at that time and this led to our collaboration on admissibility related topics. The resulting work constituted a part of his Ph.D. thesis. We ultimately published a paper based on this work in the *Australian Journal of Statistics*, 1973 based on our 1969 ISI Technical Report. Our results demonstrated the practical irrelevance of a criterion called hyper-admissibility that leads to the Horvitz-Thompson (HT) estimator of total as the *unique* choice for *any* sampling design. Subsequently, D. Basu obtained similar results independently in his 1971 landmark paper on inferential issues, and his famous example of circus elephants put a stop to research on unrealistic criteria that lead to unique choice for any design.

M.P. also showed that hyper-admissibility when applied to variance estimation leads to a "bad" variance estimator as the unique choice.

Soon after joining Statistics Canada in 1970 as a Methodologist, M.P. was actively involved in the redesign of the LFS that led to several innovations. M.P. proposed the use of systematic PPS sampling without replacement with initial randomization for selecting the primary units from the non-self-representing units (NSRUs) and the random group method with one primary unit from each random group selected by PPS sampling from the self-representing units (SRUs). In the 1960s I studied those methods theoretically from the point of view of efficiency and variance estimation. M.P. on the other hand recognized their practical advantages in the context of LFS. Both systematic PPS sampling and random group method permitted sample expansion as well as easy rotation of sample primary units over time and the random group method enabled the adaptation of Keyfitz' ingenious method for changing out-dated size measures within each random group. A joint paper with Dick Platek on updating size measures was published in *Metrika*, 1975. The LFS group under the able guidance of M.P. made several methodological advances to improve the efficiency of the design as well as estimation. Given M.P. Singh's past interest in the effective use of auxiliary information, the LFS switched to generalized regression estimation to accommodate several post-stratification variables. The LFS group also was the first to recognize the merits of re-sampling variance estimation and the jackknife was adopted for variance estimation. More recently, regression composite estimation was introduced in the LFS under M.P. Singh's leadership, using a method suggested by Wayne Fuller and myself that is good for both change and level estimation. This method and an earlier method of Avi Singh fit in well with the existing LFS estimation system based on generalized regression. Three papers on regression composite estimation for LFS, including a joint paper of M.P. with Jack Gambino and Brian Kennedy, were published in the June 2001 issue of *Survey Methodology*.

M.P. also had a keen interest in small area estimation, dating back to 1976. His team made important methodological contributions to small area estimation. M.P. and his colleagues proposed simple synthetic estimators as well as a new estimator called the sample dependent estimator. The latter estimator is a simple composite estimator with weights designed to account for realized sample sizes smaller than expected sample sizes in the areas. Sample dependent estimators became quite popular and many agencies worldwide have used them. M.P. Singh's 1994 joint paper in *Survey Methodology* with Jack Gambino and Harold Mantel addresses several practical issues pertaining to small area estimation. I particularly like the section on design issues. It presents an excellent illustration of compromise sample allocation in the LFS to satisfy reliability requirements at the provincial level as well as sub-provincial level. A section in my 2003 Wiley book on Small Area Estimation is devoted to design issues largely based on the 1994 paper. M.P. played an active role in organizing a highly successful international conference on Small Area Estimation in 1985 and acted as co-editor of a 1987 book *Small Area Statistics* published by Wiley based on the invited papers presented at the conference.

M.P. thoroughly enjoyed working as Editor-in-Chief of *Survey Methodology*. He maintained close contact with his team of Associate Editors and introduced many innovative ideas including theme papers on both theory and practice and the Waksberg series of papers. The luncheon gatherings M.P. organized at the Annual Joint Statistical Meetings were always a big hit with the Associate Editors! As an Associate Editor located in Ottawa and consultant to Statistics Canada, I had many conversations with M.P. on matters related to the journal over the past 25 years. M.P. also played an active role in the Statistical Society of Canada (SSC) and he was instrumental in raising the profile of survey sampling at the SSC Annual Meetings.

M.P. was remarkably accurate in palm reading. In 1999 he read my palms and warned me of health problems. Indeed, I faced an unexpected health problem in 2001 due to complications from appendicitis. A few months before M.P.'s death, Avi Singh told me that M.P. had read his own palms and predicted recovery from his serious health problems. Both Avi and I were very confident that we would see M.P. back at work. However, it is a common belief in India that palmists reading their own hands cannot predict their futures accurately. Unfortunately, this belief proved to be true in this instance.

M.P. was truly a great friend of mine and I will miss him very much. It is fitting that his ashes were immersed in the sacred river Ganges in the holiest city for Hindus, Varanasi (also called Benares), where M.P. was born. His soul has gone to Heaven but his legacy will remain with us.

M.P. and his Research Days

T.J. Rao
Indian Statistical Institute, Kolkata

I had first met M.P. when he came to attend the Fourth Summer Course (Advanced) for Statisticians organized by the Research and Training School (RTS) of the Indian Statistical Institute (ISI) in May–June 1964 at the University of Kerala in the South Indian city of Trivandrum (now Thiruvanathapuram). This course was meant for research

scholars and junior faculty of ISI and other Universities. M.P. came from the Benares Hindu University (BHU) where he was a temporary lecturer. He obtained his Bachelor's degree in Statistics from the same University (BHU) and a Masters from University of Poona. I was among the research scholars that were selected from ISI for this course. We did not have much interaction during the course.

A little later, M.P. was offered a job in the Sampling Division of the National Sample Survey (NSS) Department, which at that time was part of ISI. Professors D.B. Lahiri, S. Rajarao and M.N. Murthy among others were already heading several divisions of NSS by then. Besides being occupied with the designing of the large scale sample surveys conducted by NSS, M.P. spent his spare time on research problems in sample surveys. Lahiri and Murthy encouraged methodological research in the NSS and had started a seminar series as well as release of technical reports similar to the RTS technical reports of ISI. M.P. and I spoke on our research on sampling problems in these seminars organized by NSS as well as RTS. Most of the work of M.P., which he made into technical reports of the NSS Series, got published later on in well known journals.

With his expertise in the NSS on multi purpose surveys, he got interested in the problems of utilization of auxiliary information in sample surveys. His early work related to ratio and product methods of estimation. M.P. successfully and intelligently considered the case of multiple auxiliary variables of which some are positively correlated and some negatively correlated with the study variable and used ratio estimators for the former and product estimators for the latter and produced the "ratio cum product estimator" (Singh 1967). This paper is often quoted and several scholars, especially from India, published extensions. Jointly with M.N. Murthy, he developed interesting concepts of admissibility of estimators (Murthy and Singh 1969). During the year 1968, Professor J.N.K. Rao visited ISI and we were very fortunate to have interaction with him.

M.P. was very much interested in attending conferences. He never missed any at his alma mater BHU nor the sessions of the Indian Science Congress. He took the task of writing his thesis very seriously and used to have discussions with Professors M.N. Murthy, J.N.K. Rao and D. Basu. He submitted his research work as a Thesis (Singh 1969) for the degree of Doctor of Philosophy (Ph.D.) of the Indian Statistical Institute in 1969 under the general guidance of M.N. Murthy. He left the NSS and ISI in 1970 to join Statistics Canada.

All the research scholars of ISI during 1965–70 and his colleagues at the NSS miss him very much.

References

Singh, M.P. (1967). Ratio cum Product Method of Estimation. *Metrika*, 12, 34-42.

Singh, M.P. (1969). *Some aspects of Estimation in Sampling from Finite Populations.* Ph.D. Thesis submitted to the Indian Statistical Institute.

Murthy, M.N., and Singh, M.P. (1969). On the Concepts of Best and Admissible Estimators in Sampling Theory**.** *Sankhyā*, 31, 343-354.

M.P. Singh

Nanjamma Chinnappa
Statistics Canada (retired)

While many know M.P. the statistician and of his achievements in statistics, I will try to write about M.P. the man.

I had not met M.P. until I came to Canada, although I had heard that he was the young man appointed in my position when I resigned my job at the National Sample Survey (NSS) department of the Indian Statistical Institute in Kolkata, India. I heard that when Dr. M.N. Murthy (then the head of the Methodology area in the NSS) sent me the draft of his book Sampling Theory and Methods for review, M.P. was the one who read my comments and discussed them with Dr. Murthy. Much later, when Dr. Murthy heard that I was hired by Statistics Canada, he gave me M.P.'s telephone number in Ottawa. So, when we arrived in Ottawa, I called M.P. from the hotel we were put up in and to my surprise he drove to the hotel on a cold, damp morning in late September and took me to Statistics Canada. That warm and friendly gesture brightened my day and my introduction to Statistics Canada.

M.P. hailed from the ancient city of Benares in India and it would appear that some of the qualities for which that city is famous had rubbed off on him. He was gentle, friendly to all, unflappable, resilient and wise. Many have told me how he was never too busy to listen to their problems and always helped with kind words and suggestions. Many young statisticians have benefited from his advice related to their research and career.

M.P. was fond of classical Indian music and dance. A family-oriented man, he was a pillar of strength for his wife and children during their times of need. At social gatherings he was full of fun and laughter. And when he first fell seriously ill some years ago he told me that it was his faith in God and in himself that helped him to recover. He will long be remembered, not only as a statistician of repute but as a good man who befriended and helped many.

A Career in Survey Methodology

Gordon Brackstone
Statistics Canada (retired)

M.P. Singh spent almost his whole career in the methodology area of Statistics Canada. He joined the organization in 1970, after obtaining a Ph.D. in survey sampling from the Indian Statistical Institute. At the time of his death he was Director of the Household Survey Methods Division in the Methodology Branch. His rise through the organization was steady rather than meteoric: he became a section Chief in 1973, an Assistant Director in 1982, and a Director in 1994. This steady progression mirrored his approach to survey methodology which valued thoroughness in research and testing to build firm foundations for implementation and further improvement.

Our careers at Statistics Canada coincided, give or take a year at either end, and intersected frequently, particularly from 1982 onwards. In the early 1980s when we felt the need to improve integration and oversight of Statistics Canada's methodology research work, there was little doubt in my mind who we would ask to head this effort and M.P. was duly appointed as the first Chair of the Methodology Research Committee. In this role until 1987 he initiated the planning processes and reporting requirements that, with further improvements from his successors, have governed the management of methodology research for two decades. It was during this same period that Statistics Canada's annual methodology symposia became established, with M.P. playing a key role in several of the earliest symposia (and many more subsequently).

In his long career at Statistics Canada, M.P. was involved in a broad range of methodological work, but his name will always be most closely associated with two projects: the design of the Canadian Labour Force Survey (LFS), and the Editorship of the journal, *Survey Methodology*.

The LFS provides the foundation for Statistics Canada's household survey program. Not only is it the source of monthly estimates of labour market conditions in Canada, but its frame is also the sampling basis for many other household surveys, including several longitudinal surveys introduced in the 1990s. Its efficient design is therefore crucial to the cost-effectiveness of Canada's social statistics program. First introduced in 1945, the LFS has typically undergone at least a sample redesign after each decennial Census. M.P. happened to join Statistics Canada just in time for the major post-1971 Census redesign. This redesign encompassed not only the sampling scheme, but also the questionnaire, the methods of collection, and the processing systems. Such a major redesign required extensive interdisciplinary project teamwork and M.P. became a key player in the methodological aspects of this redesign. His papers from that period focus on optimizing the multi-stage design and updating the sample. He was co-author of the official description of the methodology of the Canadian Labour Force Survey (Platek and Singh 1976).

Following this redesign pressure to produce labour market estimates for smaller regions increased. This lead him to develop methods for small area estimation from the LFS (Drew, Singh and Choudhry 1982). By the time of the post–1981 Census redesign, M.P. had become the Chair of the Redesign Committee responsible for oversight of the whole redesign. In addition to the usual sampling efficiency objectives, this redesign aimed to produce better sub-provincial data and to enhance the role of the LFS as a vehicle for conducting other household surveys. Naturally M.P. was again a principal author of the description of the new design (Statistics Canada 1990).

The efforts to make the LFS frame a basis for other household surveys were so successful that by the late 1990s a problem of overload had arisen. With the introduction of longitudinal surveys in addition to the regular survey program, concerns over the burden on the frame were increasing. In addition, the need for more targeted survey frames for certain sub-populations was being felt. M.P., set about finding alternative approaches, including approaches that would take advantage of the address register being developed for Census purposes. Some of these approaches were incorporated into the post-2001 redesign of the LFS that was just being introduced at the time of his death; some more ambitious ideas for a new frame for household surveys are still under consideration by his successors.

For more than 30 years M.P. guided methodological input to the LFS. His many papers, often co-authored with his staff, bear witness to his lasting imprint on the design of this flagship survey, and his guidance of many younger statisticians in the early stages of their careers.

Over this same period, M.P. also bore another heavy responsibility as Editor of *Survey Methodology*. The evolution of this journal from its inception in 1975 to its 25th Anniversary has been described by its founder, Richard Platek (1999), who had the foresight to appoint M.P. as its first Editor.

Under M.P.'s leadership the journal passed many milestones. In 1982 it became an official Statistics Canada publication – fully bilingual and priced. Authorship was expanded beyond Statistics Canada employees; a highly qualified panel of associate editors was recruited; theme issues were introduced, often attracting the best papers from a recent conference or symposium; the Editor's *In This Issue* feature was introduced to provide an overview of content; special 25th anniversary issues were published in 1999 – 2000, along with an index for Volumes 1–26. Over

this period, arrangements were negotiated, firstly with the International Association of Survey Statisticians and later with other statistical societies, to provide discounted subscriptions. More recently electronic versions of the journal have been made available.

Throughout these developments M.P. was at the helm, planning future issues, on guard for interesting research worthy of inclusion, encouraging potential authors, recruiting and pestering associate editors through the refereeing process, working with Statistics Canada's publication and marketing staffs to improve and promote the journal. On the journal's Management Board from 1987-2004, I witnessed first-hand and admired his enthusiasm and perseverance in the face of many difficulties. It was for him, I believe, a true labour of love.

These brief descriptions of just two of M.P.'s many contributions to Statistics Canada and the statistics profession cannot do full justice to his career. I hope they give an impression of an ever dependable professional who combined a deep understanding and research ability in statistical methods with an appreciation of the practical constraints of applying statistical methods to surveys. His style was based on reason and persistence, without bluster and shunning confrontation, coupled with an innate concern for the feelings of others. It was always a pleasure to work with M.P. and an honour to be associated with his accomplishments.

References

Drew, D., Singh, M.P. and Choudhry, H. (1982). Evaluation of Small Area Estimation Techniques. *Survey Methodology*, 8, 17-47.

Platek, R., and Singh, M.P. (1976). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue number 71-526.

Platek, R. (1999). Survey Methodology – The First 25 Years. *Survey Methodology*, 25, 109-111.

Statistics Canada (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistics Canada, Catalogue number 71-526.

In Memory of M.P. Singh

Fritz Scheuren
2005 President, American Statistical Association

In M.P. Singh last summer we lost an individual known throughout the whole statistical world as a scholar, a gentleman, and a doer. When I spoke about him at the fall 2005 Statistics Canada Methodology Symposium, it was from this perspective.

I will be brief however, providing only a sample of what could be said. Others are writing too. I will leave it to them to say more.

My memories of M.P. go back over 20 years. Exactly when we first met is now obscure to me but I have been one of his associate editors (AE's) at *Survey Methodology* for at least that long.

He used to like to have me look at papers on record linkage, sometimes sample weighting or estimation, and, less commonly, on missing data topics. His selections were ones I invariably learned from. By and large, after his initial screening, the incoming quality was excellent and, working under him my job was to make sure that the journal versions that eventually resulted were even better.

His editorship of *Survey Methodology* was challenging. The Journal had to have closely argued mathematical statistical formulations but these also had to be ones that could be put into practice. In other words the ideas had to be very good, as well as eminently useful. And they have consistently been both. No mean feat.

Many outstanding younger professionals, when they first submit a paper, demonstrate just one of these two attributes in their submissions, usually the mathematical side of their topic. For submissions that achieved at least one of these, my interpretation of the goal M.P. set for his AE's was to help authors, through the referee and AE comments, to achieve the second goal too. And what a journal he created with his vision!

By the way, he suggested that I might have tended to overdo my author support role but I think that secretly he was pleased with my approach of never giving up on what could become a great paper, if given patience. And there were several papers I handled that his patience was tried but eventually rewarded in the end.

M.P. had toughness, though, that complemented his unfailing gentleness. He firmly held all of us to high standards in guiding *Survey Methodology* with a sure hand. Even when his health began to fail, his spirit always remained visible.

The one word summary I used to characterize M.P. at the fall conference was to call him a "Mensch." Now this German word for "person" may be familiar to many of you in its Yiddish sense of a complete or whole human being. But frankly "Mensch" is really untranslatable. That is why it has stayed in Yiddish here (although I have not written in Hebrew letters, as would have been appropriate). Certainly no simple definition can do justice to either the word or the individual that M.P. was.

We all miss him greatly. He was a good friend, a loving family man, open to new ideas, careful in his advice about practice and rigorous in his thinking. M.P. will forever be a model of what it means to be a sampling statistician.

Some Recollections of M.P. Singh

David A. Binder
Statistics Canada (retired)

My memories of M.P. Singh over the many years that I knew him are all very fond. His strengths, both as an outstanding survey statistician, and as a kind and gentle person, were characteristics that were unmatched.

It was in the summer of 1970 when I first met M.P. Singh. I was working as a summer student in Agriculture Division at Statistics Canada. M.P. Singh and J.C. (John) Koop were the methodologists working with Agriculture Division at the time. I was sharing an office with Jack Graham who was on sabbatical leave from Carleton University. Jack's comment to me at that time was how fortunate Statistics Canada was to have M.P. and John there as survey methodologists, as they were two of the finest survey statisticians in the world. In fact, it was such outstanding talent at Statistics Canada that helped me decide that it would be a good place to start my career.

Most people knew M.P. through his dozens of published papers, his stewardship of the journal, *Survey Methodology*, and his interventions at statistical conferences. His publications included papers on household survey designs and redesigns, estimation (including composite estimation and domain estimation), small area estimation, and nonresponse adjustment. His insights into the many complexities of survey methods were often reflected by his questions and suggestions at conferences and meetings.

He also co-edited monographs on panel surveys (Kasprzyk *et al.* 1989) and on small area statistics (Platek *et al.* 1987), and he wrote a review article on *Survey Methodology* in the *Encyclopedia of Statistical Sciences* (Singh 1988).

As editor of *Survey Methodology* since its inception in 1975, M.P. oversaw the evolution of the journal from its beginnings as mainly a vehicle for staff at Statistics Canada to publish their research to a top international journal with regular contributions from around the world. *Survey Methodology* has been adopted by the Section on Survey Research Methods of the American Statistical Association, and by the International Association of Survey Statisticians as a publication for members of those organizations. This is a reflection of the many years of M.P.'s "labour of love" on the journal. His gentleness and kindness were even reflected in his encouraging remarks when writing a letter of rejection to authors!

Over the years M.P. was a leader in adapting to the changing technology for households surveys. He always pursued ways to improve data collection methods. He guided Statistics Canada though the world of face-face interviewing, into telephone interviewing, and computer-assisted methods. Most recently, he was keen to develop methods to improve efficiency by introducing the concept of a master sample for household survey designs at Statistics Canada, and he was instrumental in convincing managers from across Statistics Canada of the potential merits of this concept.

M.P. was a major influence at Statistics Canada to ensure the quality and the stature of research in Statistical Methods. The Bureau's accomplishments in this area have received recognition from around the world, and Statistics Canada is now often asked to participate in research activities, such as presenting invited papers at meetings, participating on panel discussions, and joining various advisory committees and panels. To help achieve this stature, the Methodology Research Committee was created in 1982-1983, with M.P. as its first chair. There he helped develop a research agenda and a strategic plan for the Methodology Branch. Although the research agenda has changed over time, the Methodology Research Program is still flourishing, thanks to the management structure and support that M.P. helped put in place.

Throughout my career at Statistics Canada, I was able to benefit greatly from M.P.'s presence. At management meetings and at meetings where he represented Methodology management, he always ensured that we kept our distinctiveness as methodologists, ensuring that decisions we took made sense for our group.

Even with all of M.P.'s accomplishments as a survey statistician, it was his character that I admired the most. His selfless compassion for others, no matter what their level of competence, was his greatest strength, in my opinion. I can recall one occasion when the two of us were interviewing a highly qualified candidate whom we brought to Ottawa from a fair distance away. However, after just a few minutes, it was clear that, in spite of this person's qualifycations, he was not suitable for a position in the Methodology Branch. Yet, M.P. managed to make the candidate feel comfortable, after having made a special trip to Ottawa for the interview, by discussing that which the candidate was most familiar with, even though M.P. also recognized that the candidate was unsuitable for the Branch.

M.P. always praised others when their accomplishments were noteworthy. This is one of the many reasons why he was endeared by so many, and why so many will miss him.

References

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (Ed.) (1989). *Panel Surveys*. New York: John Wiley & Sons, Inc.

Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Ed.) (1987). *Small Area Statistics*: *An International Symposium*. New York: John Wiley & Sons, Inc.

Singh, M.P. (1988). Encyclopedia of Statistical Sciences, Vol. 9, (Eds. D.L. Banks, Read, B. Campbell and S. Kotz), 109-110. New York: John Wiley & Sons, Inc.

Manager and Mentor

Jack Gambino
Statistics Canada

Others have written of M.P. Singh's important and varied contributions to the statistics profession and to Statistics Canada. I had the good fortune to work closely with M.P. for 17 years and got to know a side of him that only those who worked with him on a regular basis saw and appreciated. I saw M.P. in his role as editor of *Survey Methodology*, including his involvement in the day-to-day activities that led to each issue of the journal, in his role as manager, and in his role as supervisor and mentor.

In the 1980s, when I first joined Statistics Canada, it was impossible not to come across M.P. Singh. To me, for the first few years, he was the person who asked probing questions at each and every methodology seminar I attended. Much later, when we happened to sit on some of the same committees, I was always amazed when, during meetings, he would come up with good questions on topics that were clearly not on methodology turf. Invariably, his questions helped to clarify the issues, not only for methodologists, but for everyone in attendance. The lesson I learned from this was not to assume that I'm the only one who doesn't fully understand the topic under discussion.

M.P. the Editor: I first got to know M.P. personally when I joined his subdivision in 1988. He immediately recruited me as an assistant editor of *Survey Methodology*. This was standard practice for M.P. – when people with a strong technical background came into his sights, they became potential assistant editors for the journal. Those of us lucky enough to become assistant editors learned a great deal from the experience. As M.P. grew to trust our judgment over time, he relied increasingly on our views, for example, in dealing with a paper that had received conflicting referee reports.

M.P. the Manager: M.P.'s approach to assistant editors is illustrative of how he managed more generally. He let people prove themselves and, with rare exceptions, each employee's abilities grew in parallel with M.P.'s confidence in him or her. Many managers follow a specific management philosophy, sometimes jumping on whatever the latest management fad is. M.P. was not in that category. He was an intuitive manager and had a knack for spotting future "talent" early in their careers. He was also a non-authoritarian manager who encouraged his staff in their work. Although M.P. was an open, easygoing manager, he knew when to put his foot down, as many of us who worked with him found out the hard way, albeit on rare occasions.

M.P. was a strategic thinker who liked to discuss both statistical and management issues thoroughly. This sometimes led to long meetings where we were all expected to give our views. And just when we would think that an issue was settled, M.P. would throw in a new twist that got the discussion going again! The advantage of M.P.'s approach, of course, was that by the end of the meeting we all understood the ins and outs of the subject under discussion and almost always reached a consensus.

Throughout his career, M.P. took a strong interest in the development of researchers and the research function at Statistics Canada. He viewed an active research program as essential for the continued success of Statistics Canada. As a result, he worked to increase the professional visibility of researchers, and more generally of survey methodologists, within the Statistical Society of Canada and other organizations.

M.P. the supervisor and mentor: After working in M.P.'s area for a few years, I had the good fortune to report directly to him. Separating M.P. the supervisor from M.P. the mentor is impossible. He took a keen interest in his immediate employees' careers, giving them advice and steering them toward the right choices or, more importantly, steering them away from the wrong ones. What was interesting was the way he often did this. Rather than be direct, he would often lead the employee, in a near-Socratic way, to the realization that something was not such a good idea. Another technique was the "look" – anyone who got to know M.P. well learned to tell at a glance when M.P. thought an idea was particularly bad.

I learned a great deal about surveys from M.P. but more importantly, I think, I learned from him what makes a good manager, motivator and mentor. Thus I come to the realization that perhaps his greatest role was *M.P. the teacher*. Those of us who worked closely with M.P. over the years will continue to benefit from his example for the rest of our careers, and I expect we will pass on what we learned from him, filtered through our own unique experiences, to the next generation as well.

In his Own Words

Eric Rancourt
Statistics Canada

M.P. was a man of impressive personality. Many of his employees and colleagues did not have the chance to work closely with him, but for those who did, M.P. would reveal himself as a very comprehensive and human character. Below are a few quotes from him that others and I have collected. These words usually came to us at a comforting time and always made us come out of his office on a positive note.

- Don't bother setting up a meeting, my door is always open to discuss anything.
- It's good to have a pet project.
- We don't design surveys to calculate the variance.
- I'm sure it can be done.
- You're telling me that 2 out of 3 of your findings did not make it to the survey! Don't complain; if as much as 10% of your ideas get implemented, you'll have a great career!
- There is a sign by the highway that says 100 km/h; that doesn't mean you have to go to 100 km/h.
- Don't worry, there is still time.
- After all the efforts we make in designing surveys, what we remember and appreciate the most is not the methods or results; it is the people we worked with.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 11-24
Statistics Canada, Catalogue No. 12-001-XIE

Targeted Random Walk Designs

Steven K. Thompson [1]

Abstract

Hidden human populations, the Internet, and other networked structures conceptualized mathematically as graphs are inherently hard to sample by conventional means, and the most effective study designs usually involve procedures that select the sample by adaptively following links from one node to another. Sample data obtained in such studies are generally not representative at face value of the larger population of interest. However, a number of design and model based methods are now available for effective inference from such samples. The design based methods have the advantage that they do not depend on an assumed population model, but do depend for their validity on the design being implemented in a controlled and known way, which can be difficult or impossible in practice. The model based methods allow greater flexibly in the design, but depend on modeling of the population using stochastic graph models and also depend on the design being ignorable or of known form so that it can be included in the likelihood or Bayes equations. For both the design and the model based methods, the weak point often is the lack of control in how the initial sample is obtained, from which link-tracing commences. The designs described in this paper offer a third way, in which the sample selection probabilities become step by step less dependent on the initial sample selection. A Markov chain "random walk" model idealizes the natural design tendencies of a link-tracing selection sequence through a graph. This paper introduces uniform and targeted walk designs in which the random walk is nudged at each step to produce a design with the desired stationary probabilities. A sample is thus obtained that in important respects is representative at face value of the larger population of interest, or that requires only simple weighting factors to make it so.

Key Words: Adaptive sampling; Link-tracing designs; Markov chain Monte Carlo; Network sampling; Random walk; Respondent-driven sampling; Sampling in graphs; Sampling hidden population.

1. Introduction

Populations with linkage or network structure are conceptualized as graphs, with the nodes of the graph representing the units of the population and the edges or arcs of the graph representing the relationships or links between the units in the population. A central problem of studies in graph settings is that for many of the populations of interest it is difficult or impossible to obtain samples using conventional designs, and the samples obtained may be at face value highly unrepresentative of the larger population of interest. In practice, often the only practical methods of obtaining the sample involve following links from sample nodes to add more nodes and links to the sample. For example, in studies of hidden human populations such as injection drug users, sex workers, and others at risk for HIV/AIDS or hepatitis C, social links are followed from initially identified respondents to add more research participants to the sample. Similarly, in investigations of the characteristics of the Internet, the usual procedure is to obtain a sample of web sites by following links from initial sites to other sites.

Klovdahl (1989) used the term "random walk" to describe a procedure for obtaining a sample from a hidden population by asking a respondent to identify several contacts, one of whom is selected at random to be the next respondent, with the pattern continuing for a number of steps. Heckathorn (1997) described methods of "respondent-driven sampling" using procedures of this type. The motivation for using designs like this in practice is to penetrate deeper into the hidden population to obtain respondents who are more "representative" of the population than the more conspicuous initial respondents may be. In studies of the Internet, the parallel idea is that of the "random surfer", who selects a web page at random, clicks at random on one of the links on that page, thus moving to another page, and so on (Brin and Page 1998). The random walk design can be conceptualized as a Markov chain (Heckathorn 1997, 2002, Henzinger, Heydon, Mitzenmacher and Najork 2000, Salganik and Heckathorn 2004). In this paper some modifications of these Markov chain designs are described, with the object of obtaining stationary probabilities of equal or specified values in order to obtain simple estimates of characteristics of the population graph of interest.

Approaches to inference from samples in a graph setting include design-based, model-based, and combination methods. In the design based approach, all values of node and link variables in the graph are considered fixed or given, and inference is based on the design-induced probabilities involved in selecting the sample. In the model based approach, the population is itself viewed as a realization of a stochastic graph model, which provides the joint probability distribution of all the node and link variables. Previous design-based approaches include the methods of network or

1. Steven K. Thompson, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6. E-mail: thompson@stat.sfu.ca.

multiplicity sampling (Birnbaum and Sirken 1965), adaptive cluster sampling applied in a graph setting (Thompson and Collins 2002), and a few of the methods in the snowball sampling literature (Frank 1977, 1978, Frank and Snijders 1994). A method combining design and model based approaches is used in Felix-Medina and Thompson (2004) for studying a hidden population in which link-tracing follows from a probability survey sample from a frame that covers only part of the population.

The advantage of design-based methods is that populations such as socially networked hidden human populations are difficult to model realistically, and the design-based inference does not rely on modeling assumptions for properties such as unbiasedness and consistency of estimators. Design-based inference methods do rely on the design being implemented according to plan, however, and exact implementation of a given design may be a very great challenge in studies of hidden human populations. This was the motivation for the development of a range of model-based methods for inference from samples in graphs, including maximum likelihood and Bayes techniques (Thompson and Frank 2000, Chow and Thompson 2003). Assuming that the initial sample is "ignorable" in the likelihood sense (Rubin 1976), or that the design is of known form so that it can be included in the likelihood and Bayes equations, these methods work for a very wide range of link-tracing sampling procedures, including most variations of the snowball and network methods. In reality, however, the initial sample may be selected in a fashion that is anything but ignorable, with selection probabilities depending on node value, node degree, and other factors. The pervasive problem of initial sample selection in link-tracing studies has been remarked upon by Spreen (1992) among others.

The approach pursued in the present paper does not assume total control over all design possibilities, but rather seeks to work with the way samples naturally tend to get selected in networked populations, whether by ethnographers or other social scientists, members of the population themselves, or automated web crawlers. Starting with those natural selection processes, we introduce iterative modifications to obtain sampling procedures that step by step approach desired selection probabilities.

Although the underlying structure of the designs in this paper depends on Markov chains, the estimators and quantities of most interest to investigators may not in fact be Markovian. For example, while the sequence of selections of sample units may depend at each step only on the most recently selected unit, the sequence by which distinct units are added to the sample depends on all units selected thus far. For this reason, the properties of a number of alternative estimators with different designs are examined using simulation, by repeatedly selecting samples from stochastic graph realizations and from an empirical population from a study of a people at high risk for HIV/AIDS transmission.

Random walk designs are described in section 2. Uniform and targeted walk designs are introduced in sections 3 and 4 respectively. Examples are worked in section 5, including an illustrative example using as the population a realization of a stochastic graph model and an empirical example using data from a study of a population at high risk for HIV/AIDS.

2. Random Walk

The population of interest is a graph, given by a set of N nodes with labels $U = \{1, 2, \ldots, N\}$ and values $\mathbf{y} = \{y_1, \ldots, y_N\}$ and an $N \times N$ matrix $\mathbf{A}$ indicating relationships or links between nodes. An element a_{ij} of $\mathbf{A}$ is one if there is a link from node i to node j and zero otherwise. The diagonal elements a_{ii} are assumed to be zero. For node i, the row sum $a_{i\bullet}$ is the out-degree or number of nodes to which i has a link and the column sum $a_{\bullet i}$ is the in-degree or number of nodes which link to i. With an undirected graph, the matrix $\mathbf{A}$ is symmetric and the in-degree of any node equals its out-degree.

Let W_k denote the unit or node of the graph that is selected at the k^{th} wave. If i is the node selected at the k^{th} wave, then for wave $k+1$ one of the nodes linked from i is selected at random. Thus, $\{W_0, W_1, W_2, \ldots\}$ is a Markov chain with

$$P(W_{k+1} = j \mid W_k = i) = a_{ij} / a_{i\bullet}. \qquad (1)$$

Let $\mathbf{Q}$ denote the transition matrix of the chain with elements $q_{ij} = P(W_{k+1} = j \mid W_k = i)$. The chain is a random walk in that at each step, one of the neighboring states of the present state is selected at random.

If the graph consists of a single connected component, that is, if every node of the graph is reachable from every other node by some path, then the chain is irreducible and its stationary probabilities $(\pi_1, \ldots, \pi_N)$ satisfy $\pi_j = \sum \pi_i q_{ij}$ for $j = 1, \ldots, N$. In fact, with the simple random walk design in a connected undirected graph the stationary probabilities can be shown (Salganik and Heckathorn 2004) to be

$$\pi_j \propto a_{\bullet j}.$$

That is, for an undirected graph consisting of only one connected component, the long term selection frequency for any node is proportional to its in-degree, which, for a nondirected graph, equals the out-degree.

Suppose one wishes to estimate a characteristic of the population graph, such as the population mean of the node values $\mu_y = \sum_{i=1}^{N} y_i / N$ using data from a random walk

sample. The sample mean $\bar{y} = \sum_{i \in s} y_i$ is in general not unbiased because the value y_i of a node may be related to its degree and hence to its probability of being selected. However, one can obtain an approximately unbiased estimate by weighting each sample y–value by the reciprocal of its in-degree, assuming that that information is available from the data (Salganik and Heckathorn 2004).

2.1 Random Walk with Random Jumps

In a graph with separate components or with unconnected nodes, the simple random walk just described does not have the property that every node can be eventually reached from every other node. Without this property, the limiting distribution of the random walk is sensitive to the starting distribution, since the limiting probability for a node depends on the initial probability of starting in the component that contains that node. A modification of the design which overcomes this problem allows for a jump with small probability to a node at random from the whole graph. At each step, this random walk follows a randomly selected link with probability d and, with probability $1-d$, jumps to another node in the graph at random or with specified probability. In the Internet search literature, d is referred to as the "damping factor", since a value of d less than one damps the effect of the out-degree of a given node (Brin and Page 1998).

The transition probabilities for the random walk with jumps are given by

$$q_{ij} = \begin{cases} (1-d)/N + da_{ij}/a_{i\bullet} & \text{if} \quad a_{i\bullet} > 0 \\ 1/N & \text{if} \quad a_{i\bullet} = 0. \end{cases} \qquad (2)$$

With the small probability $1-d$ of a random jump at any step, the Markov chain walk can potentially reach any node in the graph from any other, so that the chain is irreducible. Further, the random jumps, which include the possibility of going to node i from node i, ensure that the chain is aperiodic so that the stationary probabilities are limiting probabilities. With $d < 1$ the stationary probability of node i is not a simple function of its own in-degree, but depends also on the stationary probabilities of the nodes that link to it.

More generally, the jumps can be made with any specified probabilities $\mathbf{p} = (p_1, \ldots, p_N)$ and the probability of a jump can depend on the current state, so that the transition probabilities are

$$q_{ij} = \begin{cases} (1-d_i)p_j + d_i a_{ij}/a_{i\bullet} & \text{if} \quad a_{i\bullet} > 0 \\ 1/N & \text{if} \quad a_{i\bullet} = 0. \end{cases}$$

Estimates which are approximately design-unbiased for population graph characteristics can be obtained by weighting sample values inversely proportional to the limiting Markov chain selection probabilities, but with the additional problem that these limiting probabilities are unknown and must be estimated from the sample data (see Henzinger *et al.* 2000 for an approach to this).

For the remainder of this paper, "random walk" or "ordinary random walk" will refer to the random walk with jumps unless it is specifically stated to be a random walk without the option of jumps.

3. Uniform Walk

In this section a modification of the random walk design is proposed which leads to uniform stationary probabilities $\boldsymbol{\pi} = (\pi, \ldots, \pi)$.

Consider first the case of the population graph consisting of only one connected component. Let $\mathbf{Q}$ be the transition matrix for the simple random walk with transition probabilities q_{ij} given by (1). Suppose that at step k the state of the process is i. A tentative selection is made using the transition probabilities in the i^{th} row of $\mathbf{Q}$. Suppose that the tentative selection is node j. If the out-degree $a_{j\bullet}$ of node j is less than the out-degree $a_{i\bullet}$ of node i, then the selection for the next wave is node j, that is, $W_{k+1} = j$. If, on the other hand, the out degree of node j is greater than the out degree of node i, then a uniform random number Z is selected from the unit interval. If $Z < a_{i\bullet}/a_{j\bullet}$, then $W_{k+1} = j$. Otherwise, $W_{k+1} = i$.

Using the Hastings-Metropolis method (Hastings 1970), the transition matrix for the modified walk in the connected graph is constructed with elements

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for} \quad i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{a_{i\bullet}}{a_{j\bullet}}, 1\right\}.$$

With a population graph containing separate components or isolated nodes, the random walk with jumps, having transition matrix $\mathbf{Q}$ given by (2), can be modified to give

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for} \quad i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{q_{ji}}{q_{ij}}, 1\right\}.$$

Thus, for two mutually connected nodes i and j, the acceptance probability for a transition from i to j is

$$\alpha_{ij} = \min\left\{\frac{(1-d)/N + d/a_{j\bullet}}{(1-d)/N + d/a_{i\bullet}}, 1\right\}.$$

For a transition from an isolated unit to one in a component larger than one node, the acceptance probability is $\alpha_{ij} = 1-d$. Other acceptance probabilities have $\alpha_{ij} = 1$. Note also that for a directed graph, the acceptance probability for following an asymmetric link would be zero.

The uniform walk is implemented, when the current state is i, by selecting a candidate next state, say j, using the transition probabilities in the i^{th} row of **Q**. A standard uniform random number Z is selected and, if $Z < \alpha_{ij}$, the next state is j, whereas otherwise the walk stays at i for one more step.

The quantity α_{ij} with the uniform walk designs depends on the known transition probabilities of the basic random walk, so does not require estimation for implementation.

4. Targeted Walk

The same approach can be used to construct a walk having any specified stationary probabilities, for example selecting nodes with high y values with higher probabilities or selecting nodes to have probabilities strictly proportional to degree, even when the graph contains separate connected components. Let $\pi_i(y)$ denote the desired stationary selection probability for the i^{th} node as a function of its y value. For example, in a study of a hidden human population at risk for HIV/AIDS, suppose it is desired to sample injection drug users $(y_i = 1)$ with twice the probability of noninjectors $(y_i = 0)$. The relevant transition probabilities for the value-targeted walk, using again the Hastings-Metropolis method, are

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for} \quad i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}.$$

Note that the basic transition probability is known, since it depends only on out-degree of observed nodes, the chosen probability d, and the specified ratio π_j / π_i.

For a walk in which the relative selection probability depends on y value, the ratio $\pi_j(y_j)/\pi(y_i)$ is specified and

$$\alpha_{ij} = \min\left\{\frac{\pi_j(y_j) q_{ji}}{\pi_i(y_i) q_{ij}}, 1\right\}.$$

As another example of a targeted walk, the target distribution could be to have nodes selected proportional to their out-degree, that is, the number of links out. Since the degree for an isolated node is zero, one possibility, referred to as the "degree + 1" targeted walk, simply adds one to each degree, so that $\pi_i \propto a_{i\bullet} + 1$ is the target selection probability.

A slightly different choice, referred to simply as the degree-targeted walk, adds one only to the degree of isolated nodes, so that $\pi_i \propto \max(a_{i\bullet}, 1)$. For a degree-targeted walk of this type, the acceptance probability for a transition between two mutually connected nodes is

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}(1-d)/N + 1}{a_{i\bullet}(1-d)/N + 1}, 1\right\}.$$

For a transition between an isolated node and one with positive degree, the probability is

$$\alpha_{ij} = \min(a_{j\bullet}(1-d), 1).$$

The transition probability between two nodes each having positive degree is

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}}{a_{i\bullet}}, 1\right\}.$$

In that case

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet} q_{ji}}{a_{i\bullet} q_{ij}}, 1\right\}.$$

Since isolated nodes, without any links to other nodes, have degree zero, to give them a positive selection probability their degree can arbitrarily be assigned the value "1" in the degree-targeted walk calculation, or the value 1 can be added to the degree of every node.

5. Nonreplacement Walk Designs

The limiting distribution results of the previous sections apply exactly to walk designs with replacements, so that the selection of nodes can proceed indefinitely through the finite population. Some of the estimators used in the examples to follow, are based however on the sequence of distinct units selected through that process. The sequence of distinct units, which in effect provides a walk sample without replacement, can add new units only until the number of distinct nodes in the sample equals that of the finite population, at which point the sample mean and the population mean coincide.

A different procedure for selecting a walk sample without replacement is to directly confine the selection of the next unit at any step from the set of units not already selected, as with the "self-avoiding random walk" (Lovász 1993). If a select-reject procedure is used as with the

targeted walks, the next selection is made from the set of units not having been tentatively selected at all, whether or not the unit was accepted.

6. Estimators Based on the Values of the Accepted Nodes

With a uniform random walk with replacement the draw-by-draw sample mean of the sequence of accepted values is asymptotically unbiased for the mean of the population, because the limiting selection probabilities are all equal. The draw-by-draw sample mean is the nominal mean including repeat values, so a node's value is weighted by the number of times it is selected. With a without-replacement design this same estimator is not precisely asymptotically unbiased because the limiting probabilities are not exactly equal. The standard variance estimator based on a within-walk sample variance is not unbiased because of the dependencies within walks. Variance estimators are examined empirically in the examples.

With a targeted walk in which the limiting probability π_i of node *i* is proportional to c_i, an asymptotically consistent estimator, based on the limiting probabilities, is provided by the generalized ratio estimator

$$\hat{\mu} = \frac{\sum_{s_a} y_i / c_i}{\sum_{s_a} 1/q_i}.$$

Note that the Horvitz-Thompson estimator can not be used because the proportionality constant in the inclusion probabilities is unknown, whereas in the generalized ratio estimator it cancels out. Again the limiting probabilities on which the estimator is based hold exactly for the with-replacement design. For the without-replacement variation, the estimator is examined empirically in the examples.

7. Examples

7.1 Realized Stochastic Graph

Figure 1 depicts first a small simulated population having 60 nodes. Nodes having value $y = 1$ are colored dark and nodes with value $y = 0$ are light. The entire realization is taken to be our population of interest. The model producing the realization is a stochastic block model in which the probability of a link between any two nodes depends on the values of the nodes. Links are more likely between nodes of the same type, and the dark nodes are more highly connected than the light nodes. For example, it may be of interest to estimate the proportion of positive nodes (that is, nodes with $y = 1$) in the graph. In the population graph, 24 of the 60 nodes are positive, so the true proportion is 0.4. To the right is shown the same graph but with node sizes proportional to the random walk limiting selection probabilities. Because of the higher linkage tendencies of the positive nodes, many of them have higher than average selection probabilities.

In the bottom row of Figure 1 a random walk and a uniform walk selected from the population are shown. Each starts from the same randomly selected node, labelled "1", and proceeds until five distinct nodes are selected. The arrows show the direction of following links and a jump to a new node selected at random from the graph is shown as a dotted line. Note that the random walk backtracks from the third selected node to the second one before following a new link to the fourth sample node. From the first sample node, the uniform walk passes up the higher-probability node selected by the random walk, accepting instead another of the nodes linked to it. Either of these walks can at any time take a random jump, though in the examples illustrated only the uniform walk happens to take one, in the transition from the third to the fourth sample node.

7.2 Empirical Population

Data from a study on the heterosexual transmission of HIV/AIDS in a high-risk population in Colorado Springs (Potterat *et al.* 1993, Rothenberg *et al.* 1995) are shown in Figures 2 and 3. The 595 people interviewed in the study population are represented by the nodes of the graph, and the reported sexual relationships between the respondents are shown as links between nodes. (Additional sexual links from any of the 595 to persons who were not subsequently interviewed are not shown.) The study population includes at-risk people including injecting drug users, sex workers, their sexual and drug-use partners and other close social contacts. The node variable depicted indicates sex work, with a positive value ($y = 1$) colored dark. Only sexual links are shown, though many coincide with the drug-related links. The largest sexually connected component of the graph contains 219 of the people. The next largest connected component contains 12 people, followed by a number of components of four, three and two people. The remaining nodes represent people without reported sexual contacts within the interviewed population.

The observed pattern of this population, with one connected component very much larger than the others, has been described by researchers as not atypical of studies of hidden, at-risk populations. We are using this population solely as an empirical population from which to select samples to compare sampling designs and estimators.

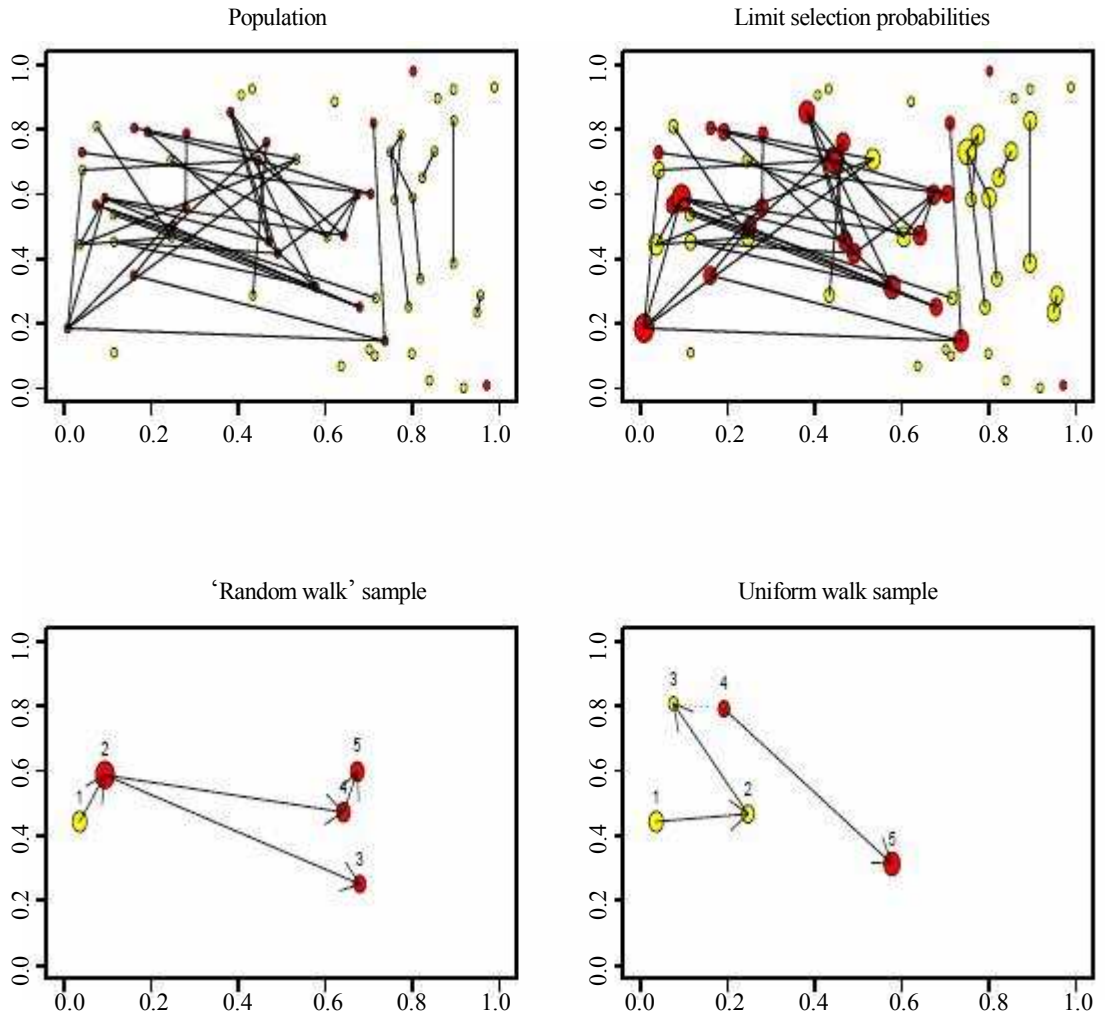


Figure 1. Top left: Population is realization of stochastic block graph model. Top right: The random walk limit probabilities of the nodes. Bottom left: Random walk of 5 steps. Bottom right: Uniform walk of 5 steps. Arbitrary axes scales are provided as a visual aid in identifying sample nodes with population nodes.

Figure 3 shows the same population with node size drawn proportional to random walk limiting selection probability.

Each plot of Figure 4 shows a cumulative sample mean of a single walk which is continued until 120 distinct nodes have been selected. The actual proportion of positive (1–valued) nodes in the empirical population (0.2235) is shown by the horizontal line in each plot.

In the top row of Figure 4, an ordinary random walk with a randomly selected starting node is shown. The left plot shows the cumulative sample mean of the distinct units. The right plot shows the same data but with the draw-by-draw sample mean, which includes repeat selections of the same node, so that each node value is weighted by the number of times that node was selected during the random walk.

In the bottom row of Figure 4 the same two types of sample mean are shown for a uniform walk that is continued until 120 distinct nodes are selected. Notice that, for the ordinary random walk, the sample mean wanders mainly above the actual mean, representing the positive bias resulting from the preferential selection of the more highly connected, high-risk people in the population. For the uniform walk, the sample mean wanders closer to the actual value, sometimes above and sometimes below. Each of these plots also gives indication of the autocorrelation present within a single Markov chain.

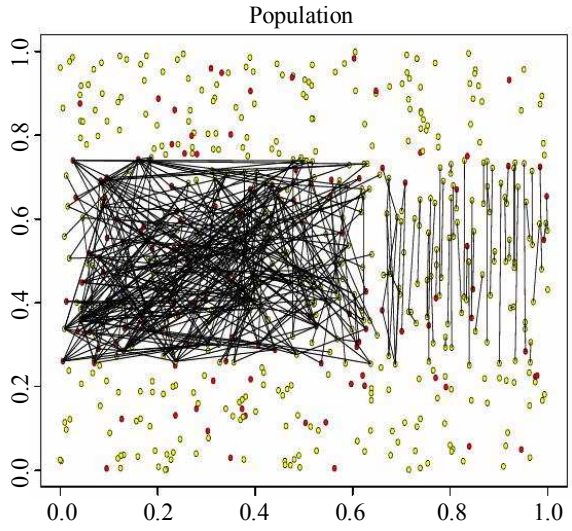


Figure 2. High-risk population in Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat *et al.* 1993, Rothenberg *et al.* 1995, and personal communications). Dark circles represent highest-risk individuals, in this case those who have exchanged sex for money. Links shown between individuals are sexual and drug injecting partnerships.

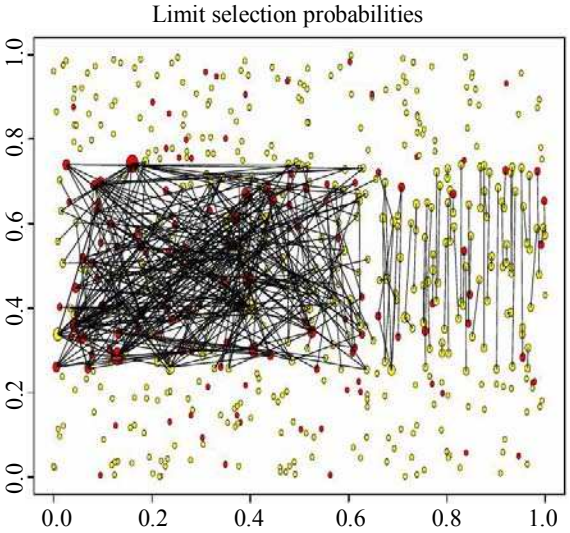


Figure 3. Limiting random walk selection probabilities for Colorado Springs population. Notice that in the real population many of the individuals with the highest-risk behavior also have high selection probabilities with the ordinary random walk, and so will tend to be overrepresented in a sample.

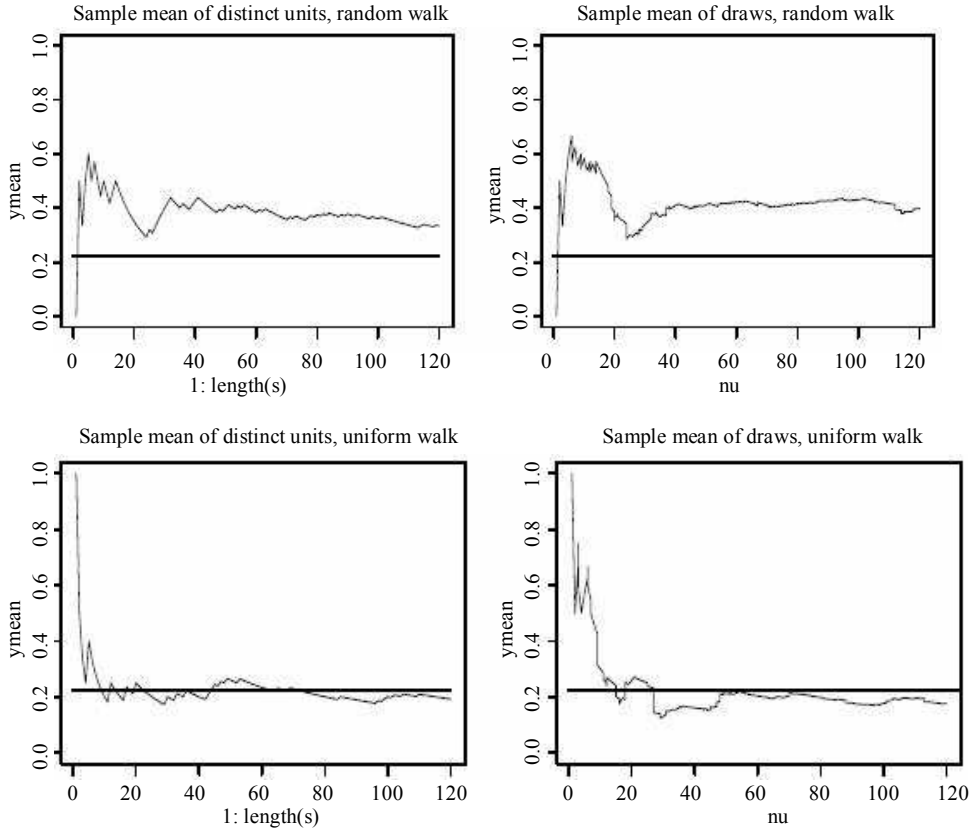


Figure 4. Sample paths of sample means for a single random walk of length 120 nodes. The top two plots are with an ordinary random walk, while the bottom two are with a uniform walk. Sample mean of the distinct units, up to the wave given by the x-axis, is plotted on the left. On the right is the sample mean of the nominal draws, so that node value is weighted by the number of times the node is selected.

The plots in Figure 5 show the expected node value as a walk progresses wave by wave, for different types of walks and with different initial distributions from which the first node is selected, for the empirical population with 595 nodes. Thus, for the k^{th} wave, the plots show $\mathrm{E}(Y_k)$, where Y_k is the value of the node selected at the k^{th} wave. The dashed line shows the actual mean for the Colorado Springs population (0.2235). The other three lines represent three different starting distributions. In all cases, the line that starts out the lowest is the uniform initial distribution, since the mean for the initial randomly selected node equals the mean for the population. The value-dependent initial distribution, in which positive nodes ($y = 1$) have twice the initial selection probability of zero nodes ($y = 0$), gives the expected value line that is in all cases mostly in the middle initially and shows the strongest tendency toward initial periodicity. The degree-based initial distribution, in which initial probability of selection for a node is proportional to its degree (plus one, since isolated nodes have zero degree), forms the top line in each of the plots.

The six plots in Figure 5 show the expected values for six different types of walks. For a random walk that follows links only, without the possibility of random jumps, the long term distribution is dependent on which component the walk starts in, which depends on the initial distribution. The three separate lines in the first figure reflect the sensitivity to the initial distribution. The random walk with jumps, on the other hand, enables any node to be reached from any other so that a limiting distribution is approached quite rapidly whatever the initial distribution. With the uniform random walk, the walk that starts with the uniform distribution stays in the uniform distribution wave after wave, and the walks that start with either of the unequal distributions depicted approach this distribution fairly rapidly. Each of the value-dependent and degree-dependent walks also approaches its limiting distribution fairly rapidly, with the expected node value considerably higher than the average node value in the population. The "degree + 1" walk approaches a distribution with selection probabilities proportional to one plus the degree for each node, while the "degree" walk has limiting probabilities proportional to the actual degree except that isolated nodes are assigned degree one.

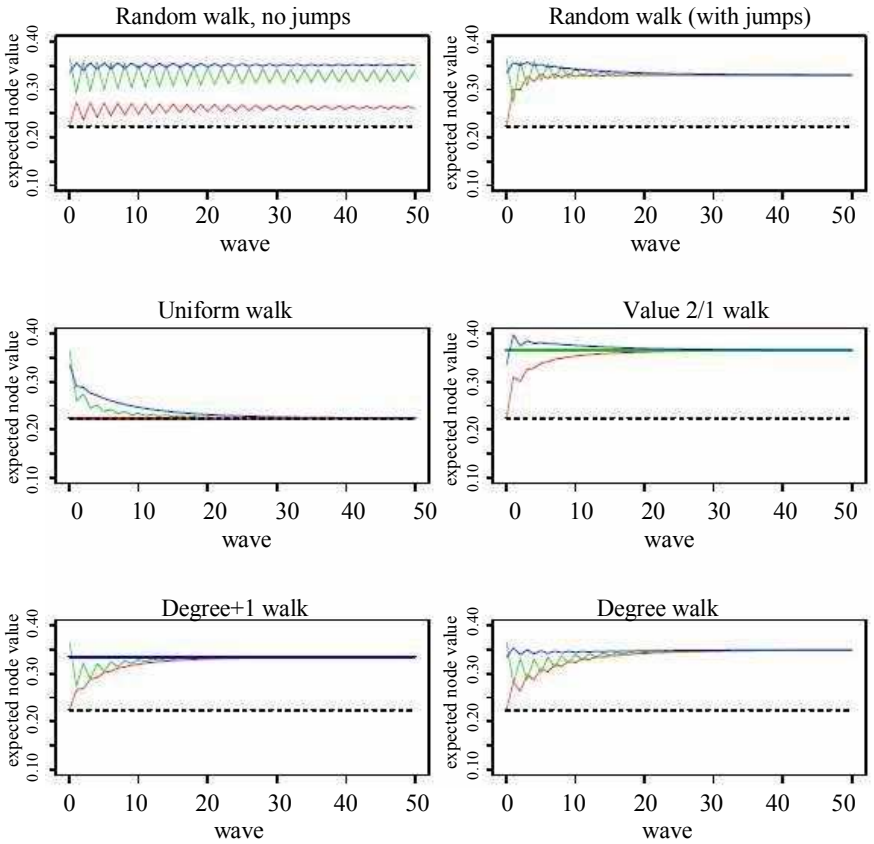


Figure 5. Expected value of node by wave for different walk designs with the Colorado Springs empirical population. Each plot shows one walk design. The dashed line is the actual mean. The other three lines show expected value for three different starting distributions. In each case the lower of the three lines starts with the uniform distribution, the middle line with the value 2/1 distribution, and the top line with the degree distribution.

Tables 1 and 2 show the calculated values of the expected value of y for the Colorado Springs study population for each type of walk, wave by wave, and with different starting distributions for the mode selections. Results for ordinary random walks are in Table 1 and for uniform walks are in Table 2. The expected values are shown for the initial selections, waves 1, 2, 3, 4, 5, 6, 8, 16, and 32, and for the limit as the number of waves approaches infinity. The three initial distributions, for the selection of the first node of a walk, are random, selection in which positive nodes have twice the probability of zero-valued nodes, and selection proportional to in-degree of each node plus one. Note that, with k independent walks of a given design, the expectations at wave j would apply to the sample mean of the k y–values at wave j from each of the walks.

Table 1
Random Walks: Expected Value of y for Waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and Infinite. Wave 0 is the Initial Selection. Three Different Initial Selection Probability Assumptions are Used: Initial Random Selection ($\pi_0 = 1/N$ for all Nodes), Nodes with Value $y = 1$ Have Twice the Selection Probability of Nodes with Value $y = 0 (\pi_0 \propto y+1)$, and Initial Selection Probability Proportional to in-Degree Plus One ($\pi_0 \propto a_{\bullet j}+1$). The Actual Mean of the Node Values for this Population is 0.2235294

wave	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0.2235294	0.3653846	0.3349894
1	0.2998771	0.2752690	0.3560839
2	0.3005446	0.3587093	0.3507451
3	0.3273606	0.3082865	0.3570490
4	0.3177081	0.3594697	0.3500041
5	0.3320705	0.3179675	0.3528395
6	0.3231213	0.3542086	0.3469835
8	0.3256034	0.3490933	0.3440449
16	0.3291087	0.3372548	0.3363884
32	0.3302606	0.3313908	0.3315119
∞	0.3303787	0.3303787	0.3303787

Table 2
Uniform Walks: Expected Value of y for Waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and Infinite, with Three Different Initial Selection Assumptions

wave	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0.2235294	0.3653846	0.3349894
1	0.2235294	0.2590239	0.2903147
2	0.2235294	0.2741356	0.2877974
3	0.2235294	0.2447258	0.2761270
4	0.2235294	0.2511473	0.2707929
5	0.2235294	0.2372440	0.2646280
6	0.2235294	0.2420866	0.2600923
8	0.2235294	0.2371714	0.2522952
16	0.2235294	0.2285370	0.2352150
32	0.2235294	0.2243635	0.2256228
∞	0.2235294	0.2235294	0.2235294

For the ordinary random walks, starting with the initial sample, the observed value is unbiased for the population value only for the initial selection, and thereafter the bias rapidly rises to its limiting value of 0.3303787–0.223594. With the initial samples biased toward the positive nodes, the bias changes less as the walk progresses.

For the uniform walk, an initial random selection coincides with the stationary distribution, so that the walk continues to be unbiased wave after wave. With the initial selection in which positive nodes have twice the selection probability of zero-valued nodes, the bias is greatly reduced with each of the first few waves and the selected node values approach their unbiased limiting state. With the initial selection proportional to in-degree plus one, the bias requires a few more waves to become small. The rapid initial approach of the expected value toward the limiting value suggests that it may be desirable to have an initial "burn in" period which is not used in the estimation part. Even a very short burn in of one to three waves could substantially reduce the bias of estimators based on short walks.

Figures 6–9 show the sampling distributions of sample means and weighted estimators for different walk designs with the Colorado Springs data set. Each histogram is based on 1,000 simulations of the sampling design applied to the empirical population. For the designs in Figures 6 and 7, each sample consists of 24 walks, each having length 5, that is, continuing until 5 distinct nodes are selected. Figure 5 shows the distributions of sample means for random walks (top row) and uniform walks (bottom row). The distribution of the mean of the 24 sample means of 5 distinct units is given on the left. On the right, the mean of the 24 draw-by-draw means, incorporating repeat selections, is given.

The actual proportion (0.2235) of the y values in the empirical population is indicated by the solid triangle, while the mean of the sampling distribution is indicated by the hollow triangle. The sample means for the random walks are biased upward, while the sample means for the uniform walk are nearly unbiased. Neither is precisely unbiased, because of the way the walk continues until a fixed number of distinct nodes is selected, instead of proceeding for a fixed number of waves.

Figure 7 shows the distribution of the generalized ratio estimator for the targeted walks having stationary probabilities related to node value and to degree (node degree plus one). For comparison purposes, each of these walks was started in its own stationary distribution, in effect giving the distributions of the estimators after "burn in". These estimators are not unbiased, since effective sample size is fixed, which affects the actual probabilities with with distinct nodes are selected in sequence, and because the denominator of the estimator is random, being the sum of the sample weights.

Figures 8 and 9 show the distributions of the same estimators and designs as in Figures 6 and 7, but with each sample consisting on one long walk of 120 distinct nodes.

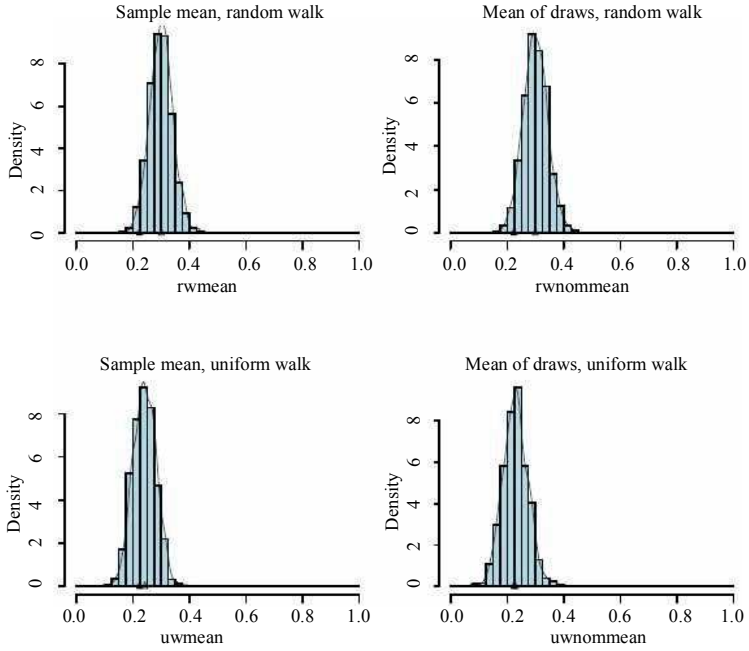


Figure 6. Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1,000.

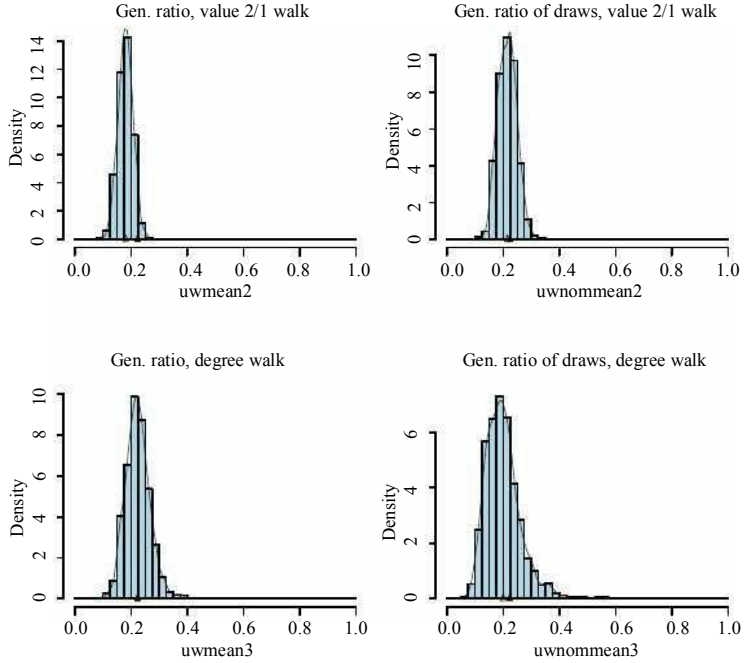


Figure 7. Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1,000.

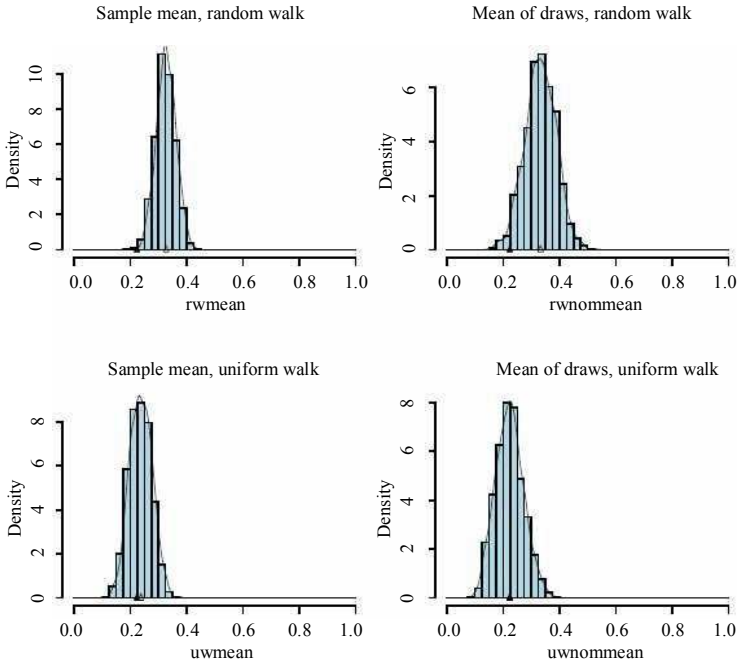


Figure 8. Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1,000.

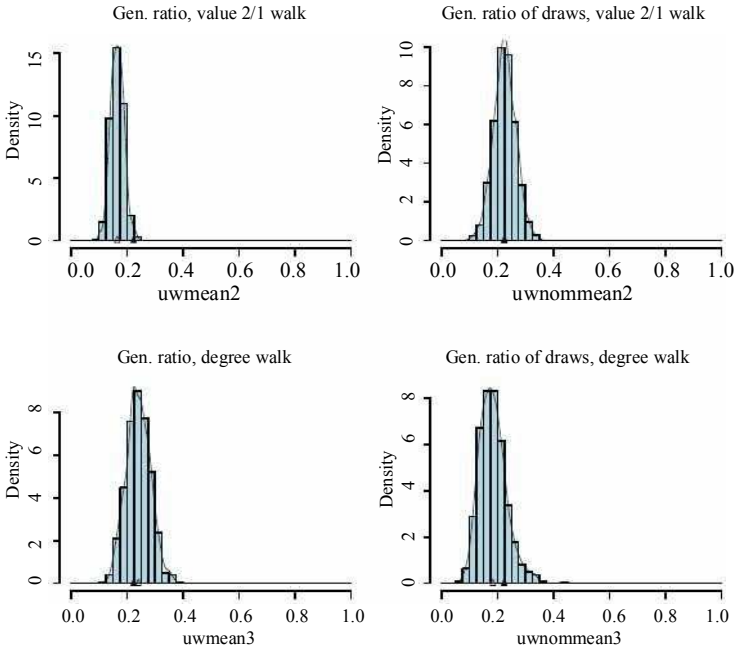


Figure 9. Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1,000.

Tables 3–6 summarize the expected values and mean square errors of the estimators with the various strategies, based on the 1,000 simulation runs with the Colorado Springs data set serving as the population.

Tables 7 and 8 give the variance and the expected values of between-walk sample variances, where available, and of within-walk sample variances for the uniform walk designs.

Table 3
Means and Mean Square Errors for Sample Means of Distinct Units and Draw-by-Draw Means for Random Walks and Uniforms Walks. The Design Uses 24 Walks Each Continuing Until 5 Distinct Nodes are Included

design:	random walk	random walk	uniform walk	uniform walk
estimator:	Sample mean	draw mean	sample mean	draw mean
mean	0.3008000	0.2994872	0.2423000	0.2289125
m.s.e.	0.007617465	0.007608868	0.002016378	0.001974826

Table 4
Means and Mean Square Errors for Weighted Means (Generalized Ratio Estimator), Using the Distinct Units in Each Walk or the Draw-by-Draw Selections for Value-Dependent Walks and Degree-Dependent Walks. The Design Uses 24 Walks Each Continuing Until 5 Distinct Nodes are Included

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1805114	0.2144555	0.2235257	0.1994530
m.s.e.	0.002546968	0.001195507	0.001807981	0.004382568

Table 5
Means and Mean Square Errors for Sample Means of Distinct Units and Draw-by-Draw Means for Random Walks and Uniform Walks. The Design Uses One Walk Continuing Until 120 Distinct Nodes are Included

design:	random walk	Random walk	uniform walk	uniform walk
estimator:	sample mean	draw mean	sample mean	draw mean
mean	0.3274083	0.3325171	0.2379333	0.2232534
m.s.e.	0.012004961	0.014902382	0.001777285	0.002442825

Table 6
Means and Mean Square Errors for Weighted Means (Generalized Ratio Estimator), Using the Distinct Units in Each Walk or the Draw-by-Draw Selections for Value-Dependent Walks and Degree-Dependent Walks. The Design Uses One Walk Continuing Until 120 Distinct Nodes are Included

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1652275	0.2254267	0.2404622	0.1835336
m.s.e.	0.003952703	0.001578039	0.002115518	0.003951540

Table 7
Variance of Estimators and Expected Values of Between-Walk and Within-Walk Sample Variances for the Uniform Random Walk, for the Design with 24 Walks of 5 Distinct Nodes Each

	estimator: sample mean	draw-by-draw mean
variance of estimator:	0.001665709	0.001947796
E (between-walk variance)	0.001584203	0.001919005
E (average within-walk variances)	0.001515521	0.001231983

Table 8
Variance of Estimators and Expected Values of Within-Walk Sample Variance for the Uniform Random Walk, for the Design with a Single Walk of 120 Distinct Nodes. (No Between-Walk Sample Variance is Available for this Design)

	estimator: sample mean	draw-by-draw mean
variance of estimator:	0.001571384	0.002445194
E (average within-walk variances)	0.001510515	0.001429126

Table 9
Acceptance Rates for the Uniform and Targeted Walks in the Empirical Population

design:	uniform walk	value walk	degree + 1 walk	degree walk
acceptance rate	0.62	0.60	0.85	0.88

8. Acceptance Rates

The principal advantages of the controlled Markov chain sampling designs, such as the uniform and targeted walks, are (1) they make the limiting selection probabilities known from the data so that they can be used in estimation; (2) the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected; (3) the estimates are design based and so certain of their key properties do not depend on assumptions, which might turn out to be incorrect, about the population graph itself; and (4) with increasing chain length, the expected values of estimates tend to move toward the corresponding graph quantities even when the initial selection distribution is different from the limiting one. Further, the uniform walk design produces a sample that, without weighting or analysis, is at face value "representative" in some respects of the larger population.

An important practical concern with the uniform and targeted walks is the acceptance rate, that is, the average probability a tentatively selected node is accepted. Tentatively selected nodes that are rejected do not contribute to the simple estimators. For a population such as the Internet, in which tentative selections and accept/reject decisions can be automated and made quickly, the acceptance rate may not be critical. Sampling simply continues until a suitable number of nodes are accepted. For studies of hidden human populations, sample sizes tend to be small. Members of the population are difficult to find and interviews may be time consuming. In some studies, however, the decision to accept or reject, based on a tentatively selected person's out degree, may be fairly quickly ascertained through a short screening interview. Even so, it is desirable to have a sampling method with as high an acceptance rate as possible.

The random walks have acceptance probability equal to one, but do not in general have known or controlled limiting probabilities. If one thinks of the underlying random walk as the natural, uncontrolled walk through a population, then a controlled walk having a limiting distribution close to the natural random walk of the population would be expected to have a higher acceptance rate than a controlled having a limiting distribution very different from the natural random walk. That is, a controlled walk with a stationary distribution not far from the underlying random walk distribution should require less modification through the rejection of tentatively selected nodes than one with stationary distribution far from the natural random walk tendencies.

As mentioned earlier, the stationary probabilities for an ordinary random walk in a nondirected graph with a single component are proportional to the degrees of the nodes. When there is more than one connected component, the random jump innovation is necessary to ensure that every node is reachable and to produce a single stationary distribution not dependent on the starting distribution, and the limiting probabilities are influenced by, but not strictly proportional to, the node degrees. Even with the random jump innovation and the induced acceptance probabilities, the targeted walks producing stationary probabilities proportional to node degrees may be the closer than the other controlled walks under consideration to the natural random walk distribution. Indeed, in Figure 5 it is evident that, for the empirical population, the equilibrium distribution of the expected node value for the degree + 1 walk is closer to the equilibrium for the random walk with jumps than is any of the other controlled designs studied.

For the empirical population from the HIV/AIDS heterosexual transmission study, the acceptance rates for the different designs are given in Table 9. For the uniform walk design, the acceptance rate was 62 percent. For the value walk, giving twice the limiting probability for the high risk as for the low risk people, the acceptance rate was 60 percent. For the degree walk, in which the limiting probability was proportional to the degree plus one, the acceptance rate was 85 percent. For the degree walk with ond added only for the degree of the isolated nodes, the acceptance rate was 88 percent.

9. Discussion

The uniform and target walk sampling designs serve to make the limiting selection probabilities known from the data so that they can be used in estimation. Further, the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected. Dependence on the initial selection, which may be uncontrolled, decreased step by step.

The estimators used in this paper with the uniform and targeted walk designs can be said to be design based. Even though the exact design based selection probabilities may be unknown if they are unknown in the initial selection, the stationary selection probabilities are used in the estimators. With increasing chain length, these probabilities become more accurate and the expected values of estimates move toward the corresponding graph quantities. The design based estimation methods have the advantage that certain of their properties, such as design unbiasedness or consistency, do not depend on model based assumptions that would possibly be incorrect. The design based estimates have the additional attractive quality that they are very simple and easy to understand and explain, and can even produce data that can be presented without analysis or interpretation as representative in important characteristics of the wider population of interest.

The use of Markov Chain Monte Carlo algorithms for data analysis with complicated models is common in statistics. The methods described here are unusual in that the Markov Chain methods are applied to real-world populations to actually obtain the data, with the result that the data thus obtained can be easily analyzed by hand. In fact, one could go a step farther and construct a complex Bayes stochastic graph model for the population, using Markov Chain Monte Carlo methods in the conventional fashion in analyzing the data as well as in their collection.

The uniform or targeted walk designs are useful to obtain samples of accepted nodes that have certain desirable properties in relation to the population, that provide very simple estimators of population quantities, or that could provide an initial sample for another design. It should be noted that nodes that were observed but then "rejected" under the design are actually still part of the data. Their values can still be incorporated into estimates if desired using the Rao-Blackwell method applied once the chain has reached approximate equilibrium, though the estimates then are computationally complex.

Another alternative is to use model based methods such as Bayes estimates. The model based methods require, in addition to adequate stochastic graph modeling of the population, an ignorable initial selection procedure, which is not in general satisfied with initial selections biased by node or degree values, or else adequate modeling of the nonignorable selection procedure as part of the likelihood. Targeted walk designs producing an asymptotic distribution unrelated to the nonignorable selection procedure and hence approximately unrelated to node or degree values outside of the sample could provide the initial selections for a sample with which model based inference methods could then be applied.

Acknowledgements

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102 and DMS-0406229), and the National Institutes of Health (R01-DA09872). I would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study.

References

Birnbaum, Z.W., and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Serie 2, No.11. Washington: Government Printing Office.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, Elsevier, 107-117.

Chow, M., and Thompson, S.K. (2003). Estimation with link-tracing sampling designs-a Bayesian approach. *Survey Methodology*, 29, 197-205.

Felix-Medina, M.H., and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.

Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.

Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Henzinger, M.R., Heydon, A., Mitzenmacher, M. and Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 295-308.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World*, (Ed. M. Kochen) Norwood, NJ: Ablex Publishing, 176-210.

Lovász, L. (1993). Random walks on graphs: A survey. In *Combinatorics, Paul Erdös is Eighty*, (Eds. D. Miklós, D. Sós and T. Szöni), János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks*, (Eds. R.H. Needle, S.G. Genser and R.T. Trotter) Drug Abuse, and HIV Transmission, NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Salganik, M.J., and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Methodologie Sociologique*, 36, 34-58.

Thompson, S.K., and Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. S*urvey Methodology*, 26, 87-98.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 25-36
Statistics Canada, Catalogue No. 12-001-XIE

Using Missing Data Methods to Correct for Measurement Error in a Distribution Function

Gabriele B. Durrant and Chris Skinner [1]

Abstract

This paper considers the use of imputation and weighting to correct for measurement error in the estimation of a distribution function. The paper is motivated by the problem of estimating the distribution of hourly pay in the United Kingdom, using data from the Labour Force Survey. Errors in measurement lead to bias and the aim is to use auxiliary data, measured accurately for a subsample, to correct for this bias. Alternative point estimators are considered, based upon a variety of imputation and weighting approaches, including fractional imputation, nearest neighbour imputation, predictive mean matching and propensity score weighting. Properties of these point estimators are then compared both theoretically and by simulation. A fractional predictive mean matching imputation approach is advocated. It performs similarly to propensity score weighting, but displays slight advantages of robustness and efficiency.

Key Words: Donor imputation; Fractional imputation; Hot deck imputation; Multiple imputation; Nearest neighbour imputation; Predictive mean matching; Propensity score weighting.

1. Introduction

Measurement error may lead to biased estimation of distribution functions (Fuller 1995). In this paper we consider approaches to correcting for this bias when, in addition to sample observations on the erroneously measured variable, values of the accurately measured variable are available for a subsample. When the subsample is selected using a randomised scheme, the set-up is an instance of the well-studied problem of double sampling (*e.g.*, Tenenbein 1970). In this case, unbiased estimates can be constructed from the subsample alone, but use of data on the correlated surrogate variable for the whole sample may improve efficiency. See, for example, Luo, Stokes and Sager (1998). In this paper we shall suppose that the subsample is not selected by a known randomised scheme, but rather by an unknown missing data mechanism. We shall just assume that the accurate variable is missing at random (MAR) (Little and Rubin 2002), conditional on variables measured on the whole sample. Some inference methods are available for this problem if we are willing to make strong parametric assumptions about the true distribution (*e.g.*, Buonaccorsi 1990) or about the measurement error model (*e.g.*, Luo *et al.* 1998). We shall not consider such methods further, however, since we suppose that we are dealing with an application where such assumptions are unrealistic. Instead, the novel feature of this paper is to view inference in this measurement error set-up as a missing data problem and to consider the application of imputation and weighting methods from the missing data literature. Our focus will be on the choice of such methods to improve point estimation of the distribution function, in terms of bias, efficiency and robustness to model assumptions. We shall only consider variance estimation briefly.

This paper is motivated by an application to the estimation of the distribution of hourly pay in the United Kingdom (UK), using data from the UK Labour Force Survey (LFS). In the LFS there are two ways of measuring hourly pay. The traditional method is to obtain information about earnings and hours worked and to derive a measure of hourly pay from this information. We refer to the variable derived in this way as the *derived hourly pay* variable. A more recent method of measuring hourly pay is to ask respondents directly about their hourly pay. We refer to the resulting measure of hourly pay as the *direct variable*. Skinner, Stuttard, Beissel-Durrant and Jenkins (2002) describe and provide empirical evidence of many sources of measurement error in the derived variable and conclude from their study that the direct variable measures hourly pay much more accurately than the derived variable. The problem with the direct variable is that it is missing for about 43% of all cases. The application is outlined in Section 8 and described in greater detail in Skinner *et al.* (2002), who also proposed the use of imputation to address the measurement error problem. This paper extends that work by considering a wider class of approaches to missing data and by comparing their properties both theoretically and via simulation. The imputation approach developed in this paper, which extends that considered by Skinner *et al.* (2002), has now been implemented by the UK Office for National Statistics as a new approach to producing low pay estimates.

The paper is structured as follows. The estimation problem is discussed in section 2. Imputation and weighting

1. Gabriele B. Durrant and Chris Skinner, University of Southampton, United Kingdom. E-mail: cjs@soton.ac.uk.

approaches are set out in sections 3 and 4 respectively and their properties are studied and compared theoretically in section 5 and via a simulation study in section 7. Variance estimation is considered briefly in section 6. Section 8 discusses the application of the methods to the LFS. Some concluding remarks are given in section 9.

2. The Estimation Problem

Let y_i be the (true) value of a variable of interest associated with unit i in a finite population U. The distribution function of the variable in U is:

$$F(y) = N^{-1} \sum_{i \in U} I(y_i \le y), \tag{1}$$

where $I(.)$ is the truth function ($I(E) = 1$ if E is true and $= 0$ otherwise) and y may take any specified value. Suppose that a survey is conducted on a sample $s \subset U$ and that the variable is measured as y_i^* for units $i \in s$. The difference between y_i^* and y_i represents measurement error. Suppose that the true value y_i is recorded for a subset of sample units and that we write $r_i = 1$ if y_i is recorded and $r_i = 0$ otherwise. Let x_i be a vector of auxiliary variables also recorded in the survey. Our data consist of values y_i^*, x_i and r_i for $i \in s$ and values y_i for $i \in s$ when $r_i = 1$. The problem is how to use these data to make inference about $F(y)$.

In the LFS application, the units are employees, s is the set of unit respondents in the LFS sample, y_i^* is the value of the derived hourly pay variable and y_i is the value of the direct variable for employee i. The value y_i is assumed equal to the true hourly pay.

The primary feature of this inference problem that concerns us is the missingness of y_i values and we consider two approaches to handle this missingness:

- imputation of y_i for units $i \in s$ where $r_i = 0$, using the values y_i^* and x_i as auxiliary information;
- weighting of an estimator based upon the responding subsample $s_1 = \{i \in s;\ r_i = 1\}$, in particular, the use of propensity score weighting (Little 1986).

These approaches to estimating $F(y)$ will be discussed in the following two sections.

Inference will be discussed under a model-based framework, in which it is assumed that the population values (y_i, y_i^*, x_i, r_i), $i \in U$, are independently and identically (IID) distributed and that sampling is ignorable, that is the distribution of (y_i, y_i^*, x_i, r_i) is the same whether or not $i \in s$. In section 8 we shall comment on how the methods developed under these assumptions may be adapted to handle the sampling design of the LFS and the use of weights to compensate for unit non-response in the survey.

3. Imputation Approaches

Suppose initially that it is possible to observe y_i for all $i \in s$. Then, under the assumptions given in the previous section,

$$\hat{F}(y) = n^{-1} \sum_{i=1}^{n} I(y_i < y) \tag{2}$$

would be an unbiased estimator of $F(y)$, in the sense that $E[\hat{F}(y) - F(y)] = 0$ for all y, where we write $s = \{1,...,n\}$ and the expectation is with respect to the model, conditional on the selected sample s. To address the problem that y_i is missing when $r_i = 0$, suppose that y_i is replaced in (2) by an imputed value y_i^I when $r_i = 0$ (and $i \in s$) and let $\tilde{y}_i = y_i$ if $r_i = 1$ and $\tilde{y}_i = y_i^I$ otherwise. The resulting estimator of $F(y)$ is

$$\tilde{F}(y) = n^{-1} \sum_{i=1}^{n} I(\tilde{y}_i < y). \tag{3}$$

A sufficient condition for $\tilde{F}(y)$ to be an unbiased estimator of $F(y)$ is that the conditional distribution of y_i^I given $r_i = 0$, denoted $f(y_i^I \mid r_i = 0)$, is the same as the conditional distribution $f(y_i \mid r_i = 0)$. However, since y_i is only observed when $r_i = 1$, the data provide no direct information about $f(y_i \mid r_i = 0)$ without further assumptions. We consider two possible assumptions.

Assumption (MAR): r_i and y_i are conditionally independent given y_i^* and x_i.

Assumption (Common Measurement Error Model): r_i and y_i^* are conditionally independent given y_i and x_i.

The first assumption is the standard one made when using imputation or weighting (Little and Rubin 2002) and is the one which we shall make. The second assumption is that the measurement error model, defined as the conditional distribution of y_i^* given y_i and x_i, is the same for respondents ($r_i = 1$) and nonrespondents ($r_i = 0$). We shall use the second assumption in the simulation study in section 7 to assess robustness of MAR-based procedures. Inference under the second assumption is more difficult, however, and appears to require stronger modelling assumptions about the distribution of y_i and x_i; we are considering this problem in other research and do not pursue this further in this paper. The plausibility of these two assumptions for the LFS application is discussed further in Skinner *et al.* (2002).

Under the MAR assumption we have $f(y_i \mid y_i^*,\ x_i,\ r_i = 0) = f(y_i \mid y_i^*,\ x_i,\ r_i = 1)$ and a sufficient condition for $\tilde{F}(Y)$ to estimate $F(Y)$ unbiasedly is that

$$f(y_i^I \mid y_i^*,\ x_i,\ r_i = 0) = f(y_i \mid y_i^*,\ x_i,\ r_i = 1). \tag{4}$$

We therefore consider an imputation approach where the conditional distribution of y given y^* and x is 'fitted' to the respondent $(r_i = 1)$ data and then the imputed values y_i^I are 'drawn from' this fitted distribution at the values y_i^* and x_i observed for the nonrespondents. Suppose that the conditional distribution $f(y_i \mid y_i^*, x_i, r_i = 1)$ may be represented by a parametric regression model:

$$g(y_i) = h(y_i^*, x_i; \beta) + e_i, \; E(e_i \mid y_i^*, x_i) = 0 \qquad (5)$$

where $g(.)$ and $h(.)$ are given functions and β is a vector of regression parameters. A point predictor of y_i, given an estimator $\hat{\beta}$ of β based on respondent data, is

$$\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]. \qquad (6)$$

Using $\hat{y}_i$ for imputation may, however, lead to serious underestimation of $F(y)$ for low values of y, since such simple regression imputation is expected to reduce the variation in $F(y)$ artificially (Little and Rubin 2002, page 64). This effect might be avoided by taking $y_i^I = g^{-1}[h(y_i^*, x_i; \hat{\beta}) + \hat{e}_i]$, where $\hat{e}_i$ is a randomly selected empirical residual (Little and Rubin 2002, page 65). Our experience is, however, that this approach fails to generate imputed values which reproduce the 'spiky' behaviour of hourly pay distributions in our application and may lead to bias around these spikes. We prefer therefore to restrict attention to donor imputation methods, which set $y_i^I = y_{d(i)}$ $(r_i = 0)$ for some donor respondent $j = d(i)$ for which $r_j = 1$. The imputed value from a donor will always be a genuine value and will respect the spiky behaviour in our application. The basic donor imputation method we consider is predictive mean matching (Little 1988), that is nearest neighbour imputation with respect to $\hat{y}_i$, defined by (6), *i.e.*,

$$\begin{aligned} &\text{impute } y_i \text{ by } y_{d(i)} \\ &\quad \text{satisfying } |\hat{y}_i - \hat{y}_{d(i)}| = \min_{j:\, r_j = 1} |\hat{y}_i - \hat{y}_j| \end{aligned} \qquad (7)$$

where $r_i = 0$ and $r_{d(i)} = 1$.

Corollary 2 of Theorem 1 of Chen and Shao (2000) then provides theoretical justification for the approximate unbiasedness of the resulting estimator $\tilde{F}(y)$ for $F(y)$, if the following four conditions hold: (i) y_i is missing at random (MAR) conditional on $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$, where $\beta = \text{plim}(\hat{\beta})$, (ii) the conditional expectation of y_i given z_i is monotonic and continuous in z_i, (iii) z_i and $E(y_i \mid z_i)$ have finite third moments and (iv) the probability of response given z is bounded above zero. These conditions seem plausible provided: the MAR assumption above holds; the distribution of y_i only depends on y_i^* and x_i via z_i; y_i^* is a reasonably good proxy for y_i. In addition, Chen and Shao's (2000) result needs to be adapted for the fact that the nearest neighbour is defined with respect to $\hat{\beta}$ whereas the above conditions are with respect to β. This adaptation seems plausible since, for a sufficiently large number of respondents, close neighbours with respect to $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]$ should also be close neighbours with respect to $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$.

There are thus theoretical grounds that nearest neighbour imputation with respect to $\hat{y}_i$ will lead to an approximately unbiased estimator of $F(y)$, subject to the MAR assumption and certain additional plausible conditions. It is also of interest to consider the efficiency of $\tilde{F}(y)$. The variance of $\tilde{F}(y)$ for nearest neighbour imputation may be inflated if certain donors may be used much more frequently than others. We consider a number of approaches to reducing this variance inflation effect.

First, we may restrict the number of times that respondents are used as donors by defining imputation classes by disjoint intervals of values of $\hat{y}_i$ and drawing donors for a recipient by simple random sampling from the class within which the recipient's value $\hat{y}_i$ falls. The smoothing will be greatest if we draw donors without replacement. We denote this hot deck method HDIWR or HDIWOR, depending on whether sampling is with or without replacement. A second approach is to undertake donor selection sequentially and to penalize the distance function employed for determining the nearest neighbour $d(i)$ as follows

$$|\hat{y}_i - \hat{y}_{d(i)}|(1 + \mu t_{d(i)}) = \min_{j:\, r_j = 1} \{|\hat{y}_i - \hat{y}_j|(1 + \mu t_j)\}, \qquad (8)$$

where $\mu \in \mathbb{R}^+$ is a penalty factor, t_j is the number of times the respondent j has already been used as a donor, $r_i = 0$ and $r_{d(i)} = 1$ (Kalton 1983). A third approach is to employ repeated imputed values $y_i^{I(m)}$, $m = 1, ..., M$, for each recipient $i \in s$ such that $r_i = 0$. The resulting estimator of $F(y)$ is $M^{-1}\sum_m \tilde{F}^{(m)}(y)$, the mean of the resulting estimators $\tilde{F}^{(m)}(y)$. We refer to the third approach as fractional imputation (Kalton and Kish 1984; Fay 1996) rather than multiple imputation (Rubin 1996), since we do not require the imputation method to be 'proper', that is to fulfil conditions which ensure that the multiple imputation variance estimator is consistent. We do not stipulate this requirement here because our primary objective is point estimation. In our use of fractional imputation we aim to select donors $d(i, m)$, $m = 1, ..., M$, each a close neighbour to i, so that $\tilde{F}^{(m)}(y)$ remains approximately unbiased for $F(y)$. We consider the following variations of this approach.

(i) The $M/2$ nearest neighbours above and below $\hat{y}_i$ are taken, for $M = 2$ or 10, denoted NN2 and NN10 respectively.

(ii) $M/2$ donors are selected by simple random sampling with replacement from the M respondents above and from the M respondents below $\hat{y}_i$, for $M = 2$ or 10, denoted NN2(4) and NN10(20) respectively.

(iii) $M = 10$ donors are selected by simple random sampling with or without replacement from the imputation classes referred to in the HDIWR and HDIWOR methods described above. We refer to these as the HDIWR10 and HDIWOR10 methods.

For comparison we also consider the Approximate Bayesian Bootstrap method of multiple imputation (Rubin and Schenker 1986), denoted ABB10, defined with respect to the imputation classes referred to in the HDIWR and HDIWOR methods.

4. Weighted Estimation

The estimator $\tilde{F}(y)$ implied by the different imputation approaches considered in the previous section may be expressed in weighted form as:

$$\tilde{F}(y) = \sum_{i \in s_1} w_i I(y_i < y) / \sum_{i \in s_1} w_i, \qquad (9)$$

where $s_1 = \{i \in s; r_i = 1\}$ is the set of respondents and $w_i = 1 + d_i / M$, where d_i is the total number of times that respondent i is used as a donor over the M repeated imputations. Note that $\sum_{s_1} w_i = n$. Another choice of weight would be to set w_i equal to the reciprocal of an estimated value of the propensity score, $\Pr(r_i = 1 \mid y_i^*, x_i)$ (Little 1986). This approach has been proposed for the hourly pay application by Dickens and Manning (2004). The propensity score might be estimated, for example, under a logistic regression model relating r_i to y_i^* and x_i. Under the MAR assumption, the resulting estimator $\tilde{F}(y)$ will be approximately unbiased assuming validity of the model for the conditional distribution $f(r_i \mid y_i^*, x_i)$ and some regularity conditions, such as those described in section 3 for the imputed estimator. Note that the need to model $f(r_i \mid y_i^*, x_i)$ replaces the need to model $f(y_i \mid y_i^*, x_i)$ in the imputation approach.

5. Properties of Imputation and Weighting Approaches

In this section we investigate and compare the theoretical properties of the imputation and propensity score weighting approaches introduced in the previous two sections under various simplifying assumptions. We fix y and set $u_i = I(y_i < y)$. Letting $N \to \infty$ we suppose that the parameter of interest is $\theta = E(u_i)$. We consider the imputation approach first and suppose that y_i depends upon y_i^* and x_i only via $z_i = g^{-1}[h(y_i^*, x_i; \beta]$ and that y_i is missing at random given z_i. Ignoring the difference between β and $\hat{\beta}$, assuming s_1 is large, we consider nearest neighbour imputation with respect to z_i. As in (9) the imputed estimator of θ may be expressed as

$$\hat{\theta}_{\mathrm{IMP}} = \sum_{i \in s_1} w_i u_i / \sum_{i \in s_1} w_i \qquad (10)$$

where $w_i = 1 + d_i / M$ (and $\sum_{s_1} w_i = n$). We write the corresponding expression for propensity score weighting as $\hat{\theta}_{\mathrm{PS}}$ with w_i replaced by $w_{\mathrm{PS}i}$. Let $z_{\mathrm{PS}i}$ be the scalar function of y_i^*, x_i upon which r_i depends and write:

$$\Pr(r_i = 1 \mid y_i^*, x_i) = \pi(z_{\mathrm{PS}i}). \qquad (11)$$

Just as we ignored the difference between β and $\hat{\beta}$, we initially ignore error in estimating $\pi(z_{\mathrm{PS}i})$ and write $w_{\mathrm{PS}i} = \pi(z_{\mathrm{PS}i})^{-1}$.

The imputation and propensity score weighting approaches may be expected to yield similar estimators if z_i and $z_{\mathrm{PS}i}$ are similar, that is they are close to deterministic functions of each other, and M is large. To see this, consider a simple example of the imputation approach, where the donor is drawn randomly from an imputation class c of close neighbours with respect to z_i, containing m_c respondents and $n_c - m_c$ nonrespondents, as described in section 3. In this case, w_i will approach $1 + (n_c - m_c)/m_c = n_c / m_c$ as $M \to \infty$ and this is the inverse of the response rate within the class (David, Little, Samuhel and Triest 1983). More generally, with the fractional nearest neighbour imputation approach considered in section 3, the weight $w_i = 1 + d_i / M$ may be interpreted as a local (with respect to z_i) nonparametric estimate of $\Pr(r_i = 1 \mid z_i)^{-1}$ despite the fact that imputation is based upon a model for y_i given z_i rather than r_i given z_i. Thus, the imputation approach may be expected to lead to similar estimation results to propensity score weighting if z_i and $z_{\mathrm{PS}i}$ are deterministic functions of each other. In general, however, this will not be the case. Since $\Pr(r_i = 1 \mid z_i)$ may be expressed as an average of $\Pr(r_i = 1 \mid y^*, x)$ across values of y^* and x for which $z = z_i$, we may interpret w_i as a smoothed version of $w_{\mathrm{PS}i}$ and may expect it to show less dispersion. This suggests that it may be possible to use imputation to improve upon the efficiency of estimates based on propensity score weighting, as also discussed by David *et al.* (1983) and Rubin (1996, section 4.6). To investigate this further, assuming MAR and the other assumptions in sections 3 and 4 upon which the approaches are based, both imputation and weighting approaches lead to approximately unbiased estimation of $F(y)$ and we may focus our comparison on relative efficiency.

It follows from equation (3.3) of Chen and Shao (2000) that the variance of $\hat{\theta}_{\mathrm{IMP}}$ may be approximated for large n by

$$\mathrm{var}(\hat{\theta}_{\mathrm{IMP}}) \approx n^{-2} E\left[\sum_{s_1} w_i^2 V(u_i \mid z_i)\right] + n^{-1} V[\psi(z_i)], \qquad (12)$$

where $\psi(z_i) = E(u_i \mid z_i)$ and any impact of estimating β is ignored. Note that Chen and Shao (2000) consider single imputation with $M = 1$ but their proof of this result carries through if $M > 1$. It is convenient to reexpress this result as

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-1}\sigma^2 + n^{-2}E\left[\sum_{s_1}(w_i^2 - w_i)V(u_i \mid z_i)\right], \quad (13)$$

using the identity

$$V[\psi(z_i)] = \sigma^2 - E[V(u_i \mid z_i)], \quad (14)$$

where $\sigma^2 = V(u_i)$ and a corollary of Chen and Shao's (2000) Theorem 1 that

$$E\left[n^{-1}\sum_{s_1} w_i V(u_i \mid z_i)\right] = E[V(u_i \mid z_i)] + o_p(n^{-1/2}). \quad (15)$$

Note that $w_i^2 - w_i = (d_i / M)(1 + d_i / M) \geq 0$. Expression (13) may be interpreted from both 'missing data' and 'measurement error' perspectives. From a missing data perspective, the first term in (13) is just the variance of $\hat{\theta}$ in the absence of missing data and the second term represents the inflation of this variance due to imputation error. From a measurement error perspective, we may consider limiting properties under 'small measurement error asymptotics' (Chesher 1991), that is where $y_i^* \to y_i$ and $V(u_i \mid z_i)$ approaches zero. In this case, the second term also approaches zero and $\hat{\theta}_{\text{IMP}}$ becomes 'fully efficient', *i.e.*, its variance approaches σ^2 / n.

Let us now consider propensity score weighting. We make the corresponding assumption that y_i is missing at random given $z_{\text{PS}i}$. Linearising the ratio in (9), with $w_{\text{PS}i}$ in place of w_i, using the fact that $E(\sum_{s_1} w_{\text{PS}i}) = n$ and initially ignoring the impact of estimating the propensity score we may write

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) &\approx n^{-2}\,\text{var}\left[\sum_{s_1} w_{\text{PS}i}(u_i - \theta)\right] \\ &= n^{-1}\,E[w_{\text{PS}i}(u_i - \theta)^2], \end{aligned} \quad (16)$$

which may be expressed alternatively as

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & n^{-2}E\left[\sum_{s_1} w_{\text{PS}i}^2 V(u_i \mid z_{\text{PS}i})\right] \\ & + n^{-1}E\{w_{\text{PS}i}[\psi(z_{\text{PS}i}) - \theta]^2\} \end{aligned} \quad (17)$$

To compare the efficiency of weighting and imputation it is convenient to use (14) and (15) (which hold also with $w_{\text{PS}i}$ in place of w_i) to obtain

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & n^{-1}\sigma^2 \\ & + n^{-2}E\left[\sum_{s_1}(w_{\text{PS}i}^2 - w_{\text{PS}i})V(u_i \mid z_{\text{PS}i})\right] \\ & + n^{-1}E\left\{\sum_{s_1}[w_{\text{PS}i} - 1][\psi(z_{\text{PS}i}) - \theta]^2\right\}. \end{aligned} \quad (18)$$

Note that, in comparison with (13), this involves a third term, which does not necessarily converge to zero as y_i^* approaches y_i and $V(u_i \mid z_{\text{PS}i}) \to 0$. Hence propensity score weighting does not become fully efficient as the measurement error disappears. The second term of (18) may also be expected to dominate the second term of (13) when $V(u_i \mid z_i)$ and $V(u_i \mid z_{\text{PS}i})$ are constant and equal, since, recalling that $\sum_{s_1} w_i = E(\sum_{s_1} w_{\text{PS}i}) = n$, these second terms are primarily determined by the variances of the weights w_i and $w_{\text{PS}i}$, and, provided M is sufficiently large, we may expect w_i to display less variation than $w_{\text{PS}i}$, as argued above.

The above discussion ignores the potential impact of estimating β or estimating a parameter vector α upon which the propensity score $\Pr(r_i = 1 \mid y_i^*, x_i)$ may be assumed to depend. Kim (2004) shows in fact that the estimation of α by its maximum likelihood estimator $\hat{\alpha}$ reduces the variance of $\hat{\theta}_{\text{PS}}$ as follows:

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & \text{var}(\tilde{\theta}_{\text{PS}}) \\ & - \text{cov}(\tilde{\theta}_{\text{PS}}, \hat{\alpha})\,\text{var}(\hat{\alpha})^{-1}\,\text{cov}(\hat{\alpha}, \tilde{\theta}_{\text{PS}}), \end{aligned} \quad (19)$$

where $\tilde{\theta}_{\text{PS}}$ is the estimator $\hat{\theta}_{\text{PS}}$ with the estimated propensity scores replaced by their true values and where the left hand sides of (16), (17) and (18) should now be $\text{var}(\tilde{\theta}_{\text{PS}})$. We conclude from this fact and the previous discussion that, in general, $\hat{\theta}_{\text{IMP}}$ is not necessarily more efficient than $\hat{\theta}_{\text{PS}}$ or vice versa and we look to the simulation study in section 7 for numerical evidence. However, our conclusion that $\hat{\theta}_{\text{IMP}}$ is more efficient as measurement error disappears and $y_i^* \to y_i$ remains valid even in the presence of estimation error in α and β, since the impact of estimation error in β will disappear in this case with $z_i \to y_i^*$ whereas the second term in (19) when added to expression (18) will not in general reduce $\text{var}(\hat{\theta}_{\text{PS}})$ to σ^2 / n in this case.

Let us finally consider the impact of departures from the MAR assumption. Under small measurement error asymptotics where $y_i^* \to y_i$ and $V(u_i \mid z_i) \to 0$ so $y_i^I \to y_i$, the imputation approach will provide consistent inference about θ even if the MAR assumption fails. This is not the case for the propensity score weighting approach. This suggests that the imputation approach may display more robustness to departures from the MAR assumption if the amount of measurement error is relatively small.

6. Variance Estimation

Although point estimation is the primary focus of this paper, we do now consider linearization variance estimation briefly. For propensity score weighting we refer to Kim (2004). For the single and fractional imputation methods in section 3 based upon nearest neighbour imputation, we may

consider a simplified approach based on the IID assumption set out in section 2 and the expression for the variance of $\hat{\theta}_{\text{IMP}}$ in (13).

The simple estimator of the first term σ^2 / n:

$$n^{-1}\hat{\sigma}^2 = n^{-2}\sum_{s_1} w_i (u_i - \hat{\theta}_{\text{IMP}})^2 \qquad (20)$$

is approximately unbiased from Corollary 1 of Chen and Shao (2000). It follows that an approximately unbiased estimator of $\text{var}(\hat{\theta}_{\text{IMP}})$ is

$$\hat{V}(\hat{\theta}_{\text{IMP}}) = n^{-1}\hat{\sigma}^2 + n^{-2}\sum_{s_1}(w_i^2 - w_i)\hat{V}(u_i \mid \mathrm{z}_i) \qquad (21)$$

if we can construct an approximately unbiased estimator $\hat{V}(u_i \mid \mathrm{z}_i)$ of $V(u_i \mid \mathrm{z}_i)$. Various approaches to estimating $V(u_i \mid \mathrm{z}_i)$ seem possible. Following Fay (1999), we might consider the sample variance of u_j values for responding neighbours near to i with respect to z. An alternative approach would be to consider a model-based approach in which a model is fitted to $\psi(\mathrm{z}_i) = E(u_i \mid \mathrm{z}_i)$ for $i \in s$ giving $\hat{\psi}(\mathrm{z}_i)$ and we set $\hat{V}(u_i \mid \mathrm{z}_i) = \hat{\psi}(\mathrm{z}_i)[1 - \hat{\psi}(\mathrm{z}_i)]$. We have considered nonparametric methods of fitting $\psi(\mathrm{z}_i)$, but have found with the LFS data that these lead to very similar values of $\hat{V}(\hat{\theta}_{\text{IMP}})$ as a logistic regression model for $\psi(\mathrm{z}_i)$.

It may be possible to apply ideas in Chen and Shao (2001) or Kim and Fuller (2002) to extend the above approach to handle survey weights and a complex design. See Rancourt (1999) and Fay (1999) for other variance estimation approaches for nearest neighbour imputation and Little and Rubin (2002) for multiple imputation approaches.

7. Simulation Study

The aim of the study is to generate independent repeated samples $s^{(h)}$, $h = 1, \ldots, H$, with values y_i, y_i^*, x_i, r_i, $i \in s^{(h)}$ which are realistic in relation to the LFS application, considered further in section 8, to compute the corresponding estimates $\tilde{F}^{(h)}(y)$ for alternative approaches to missing data and values of y and to assess the performance of the estimators $\tilde{F}(y)$ empirically. In order to employ realistic values, the samples $s^{(h)}$ of size n were drawn with replacement (*i.e.*, using the bootstrap) from an actual sample of about 16,000 employees for the March–May 2000 quarter of the LFS (only main jobs of employees aged 18+ were considered and the very small number of cases with missing values on y_i^* or x_i were omitted). The values of x_i for each sample $s^{(h)}$ were taken directly from the values in the LFS sample. Variables were chosen for inclusion in x_i if they were either related to hourly pay, measurement error in y_i^* or response r_i (see Skinner *et al.* 2002) and included for example age, gender, household position, qualifications, occupation, duration of employment, full-time/part-time, industry and region (several of these variables were represented by dummy variables). We set $n = 15{,}000$, such that each $s^{(h)}$ was of a similar size as the original LFS sample, and $H = 1{,}000$. The values of y_i, y_i^* and r_i for each sample $s^{(h)}$ were generated from models, rather than directly from the LFS data, for the following reasons.

y_i: these values were generated from a model because they were frequently missing in the LFS. A linear regression model was used, relating $\ln(y_i)$ to $\ln(y_i^*)$ and x_i with a normal error and with 20 covariates including squared terms in $\ln(y_i^*)$ and age and interactions between $\ln(y_i^*)$ and 5 components of x_i. The model was fitted to the roughly 7,000 cases where y_i was observed.

y_i^*: these values were generated from a model to avoid duplicate values of (y_i^*, x_i) within each $s^{(h)}$, which it was considered might lead to an unrealistic distribution of distances between units for the nearest neighbour method. The model was a linear regression model relating $\ln(y_i^*)$ to x_i with a normal error and with 12 covariates, including a squared term in age and one interaction, fitted to the LFS data.

r_i: these values were generated from a model to ensure that the missing data mechanism was known. Several models were fitted. The only one reported here is a logistic regression relating r_i to $\ln(y_i^*)$ and x_i with 17 covariates including squared $\ln(y_i^*)$ and interactions between $\ln(y_i^*)$ and two covariates. The model was fitted to the LFS data. The missing data mechanism is MAR given the y_i^* and x_i for all the results presented except those in Table 5.

Estimates $\hat{\theta}_t^{(h)}$ of two parameters $(t = 1, 2)$ were obtained for each sample $s^{(h)}$,

$\theta_1 =$ proportion with pay below the national minimum wage (= £3.00 per hour aged $18-21$, £3.60 per hour aged 22+)

$\theta_2 =$ proportion with pay between minimum wage and £5/hour.

The true values are $\theta_1 = 0.056$ and $\theta_2 = 0.185$. The bias and standard error were estimated as $\hat{\text{bias}}(\hat{\theta}_t) = \bar{\theta}_t - \theta_t$ and $\hat{\text{s.e.}}(\hat{\theta}_t) = [H^{-1}\sum_{h=1}^{H}(\hat{\theta}_t^{(h)} - \bar{\theta}_t)^2]^{1/2}$, where $\bar{\theta}_t = H^{-1}\sum_h \theta_t^{(h)}$.

For the fractional imputation methods several different values for M were explored and $M = 10$ or 20 were chosen to achieve an increase in the efficiency whilst still being able to define a nearest neighbour imputation sensibly.

We first compare results for the alternative imputation approaches. Table 1 presents estimates of the biases of estimators of θ_1 and θ_2 for different imputation methods, for a MAR missing data mechanism. There is no evidence of significant biases for any of the nearest neighbour (NN) methods. The bias/standard error ratios are small and may be expected to be even smaller for estimates within domains *e.g.*, regions or age groups. We conclude that there is no evidence of important bias for these methods, provided the MAR mechanism holds and the model is correctly specified.

There is some evidence of statistically significant biases for each of the three methods based on imputation classes (HDIWR10, HDIWOR10, ABB10) perhaps because of the width of the classes, although the bias appears to be small relative to the standard error. Given the additional disadvantage of these methods, that the specification of the boundaries of the classes is arbitrary, these methods appear to be less attractive than the nearest neighbour methods. This finding contrasts with the preference sometimes expressed (*e.g.*, Brick and Kalton 1996, page 227) for stochastic methods of imputation, such as the HDI methods, compared to deterministic methods, such as nearest neighbour imputation, when estimating distributional parameters.

Corresponding estimates of standard errors are given in Table 2. We find as expected that the greatest standard error occurs for the single NN1 imputation method. The variance is reduced by around 10% using the penalty function method (NN1P). About $10-20\%$ reduction arises from using two imputations (NN2 or NN2 (4)) and around 20% reduction from using ten imputations (NN10, NN10 (20)), HDIWR10, HDIWOR10, ABB10). For a given number of imputations (2 or 10) there seem to be no obvious systematic effects of using a stochastic method (NN2 (4) or NN10 (20)) versus a deterministic method (NN2 or NN10). We would expect the standard errors for HDIWR10 to be no less than HDIWOR10, which is the case for $\hat{\theta}_1$ in table 2. The slight reduction for the standard error of estimator $\hat{\theta}_2$ is likely to be caused by a comparatively small number of simulation iterations $(H=1,000)$, which may not be fully sufficient for standard error estimation. We conclude that NN10 is the most promising approach, avoiding the bias of the imputation class methods and having appreciable efficiency gains over the methods generating one or two imputations.

Table 1
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Different Imputation Methods, Assuming MAR and Correct Covariates $(H=1,000)$

Imputation Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN1	$1.2*10^{-4}$	0.2 %	$0.9*10^{-4}$	0.0 %
	$(0.9*10^{-4})$		$(1.7*10^{-4})$	
NN1P[1]	$4.4*10^{-4}$	0.8 %	$0.3*10^{-4}$	0.0 %
	$(2.6*10^{-4})$		$(5.1*10^{-4})$	
NN2	$0.6*10^{-4}$	0.1 %	$1.6*10^{-4}$	0.0 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
NN2(4)	$1.4*10^{-4}$	0.2 %	$-2.5*10^{-4}$	−0.1 %
	$(0.9*10^{-4})$		$(1.5*10^{-4})$	
NN10	$0.2*10^{-4}$	0.0 %	$-1.2*10^{-4}$	−0.1 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
NN10(20)	$0.2*10^{-4}$	0.0 %	$0.7*10^{-4}$	0.0 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
HDIWR10	$2.8*10^{-4}$	0.5 %	$26.2*10^{-4}$	1.4 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
HDIWOR10	$2.5*10^{-4}$	0.4 %	$28.0*10^{-4}$	1.5 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
ABB10	$4.6*10^{-4}$	0.8 %	$29.8*10^{-4}$	1.6 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	

Standard errors of bias estimates are below the estimates in parentheses.
[1] Note: $H=100$ iterations were used due to computing time.

Table 2
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Different Imputation Methods, Assuming MAR and Correct Covariates ($H = 1,000$)

Imputation Method	s.e.$(\hat{\theta}_1)$	s.e.$(\hat{\theta}_2)$	$\frac{V(\hat{\theta}_1)}{V_{NN1}(\hat{\theta}_1)}$	$\frac{V(\hat{\theta}_2)}{V_{NN1}(\hat{\theta}_2)}$
NN1	$2.79*10^{-3}$	$5.43*10^{-3}$	1	1
NN1P[2]	$2.60*10^{-3}$	$5.15*10^{-3}$	0.87	0.91
NN2	$2.68*10^{-3}$	$5.05*10^{-3}$	0.91	0.86
NN2(4)	$2.73*10^{-3}$	$4.88*10^{-3}$	0.94	0.80
NN10	$2.56*10^{-3}$	$4.88*10^{-3}$	0.83	0.81
NN10(20)	$2.57*10^{-3}$	$4.79*10^{-3}$	0.84	0.77
HDIWR10	$2.52*10^{-3}$	$4.66*10^{-3}$	0.82	0.74
HDIWOR10	$2.48*10^{-3}$	$4.72*10^{-3}$	0.78	0.76
ABB10	$2.63*10^{-3}$	$4.87*10^{-3}$	0.88	0.80

[2] Note: $H = 100$ iterations were used due to computing time.

We next compare the NN10 imputation approach with propensity score weighting (PSW). We consider not only the case when the specification of the model used for imputation or weighting corresponds to the model used in the simulation, as in Table 1, but also some cases of misspecification. To ensure a fair comparison of weighting and imputation we use the same covariates when fitting both the models generating y_i and r_i. We first consider the estimated biases in Table 3. When the model for imputation (NN10) or the propensity scores is correctly specified neither method demonstrates any significant bias in the estimation of θ_1 or θ_2. Significant bias does arise, however, in both cases if the model is misspecified by failing to include covariates used in the simulation. The amount of bias is, however, noticeably greater for the weighting approach. For example, for the estimator $\hat{\theta}_1$ the bias is $3-7$ times higher under PSW than under NN10 depending on the misspecification. The impact of the misspecification seems higher for estimator $\hat{\theta}_2$, in particular for the PSW method. For this estimator, we found a $6-15$ times higher bias for PSW than for NN10.

Corresponding estimated standard errors of $\hat{\theta}_1$ and $\hat{\theta}_2$ are given in Table 4. These also tend to be greater for the weighting approach, showing an increase between $5-15\%$ in comparison to the imputation method. The increase in the standard error is higher for the second estimator $\hat{\theta}_2$, ranging from $12-15\%$, whereas for estimator $\hat{\theta}_1$ the increase is between $5-12\%$, depending on the misspecification. Consequently, the mean squared error is also higher for the weighting approach, with the increase ranging from 20% to 28% for the six values in Table 4. At least under the MAR assumption, the NN10 imputation approach appears to be preferable to propensity score weighting in terms of bias and variance.

Finally, we compare the properties of imputation (NN10) and propensity score weighting when the MAR assumption fails. We now simulate missingness according to the Common Measurement Error model assumption of section 3. The same logistic model with the same coefficients as in the previous simulation is used except that y_i^* is replaced as a covariate by y_i. Simulation estimates of biases and standard errors are presented in Table 5. We observe a non-negligible significant relative bias of around 5% for the imputation approach and a little higher for the propensity score weighting approach. The positive direction of the bias of $\hat{\theta}_1$ is as expected from arguments in Dickens and Manning (2004) and Skinner *et al.* (2002). MAR-based methods will tend to overestimate numbers of the low paid, if the CME assumption holds. This is because employees with observed y_i values tend to be lower paid than employees with missing y_i values and a MAR-based imputation method, even conditional on other variables, would tend to impute lower hourly pay values than would be the case under CME which allows for the dependency on true hourly pay. While the direction of the effect may be anticipated, the magnitude of the effect is of some importance for the robustness of MAR-based methods. The relative bias of 5% of the NN10 approach does not, however, appear to make the resulting estimates unusable.

Table 3
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1{,}000$)

Method	Assumed Covariates	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN10	M1 (correct)	$-0.18*10^{-4}$	$-0.03\,\%$	$-5.8*10^{-4}$	$-0.31\,\%$
		$(0.64*10^{-4})$		$(1.20*10^{-4})$	
	M2	$-1.31*10^{-4}$	$-0.24\,\%$	$-4.74*10^{-4}$	$-0.25\,\%$
		$(0.65*10^{-4})$		$(1.23*10^{-4})$	
	M3	$-1.66*10^{-4}$	$-0.30\,\%$	$-10.6*10^{-4}$	$-0.57\,\%$
		$(0.63*10^{-4})$		$(1.23*10^{-4})$	
Propensity Score Weighting	M1 (correct)	$0.15*10^{-4}$	$0.03\,\%$	$-2.62*10^{-4}$	$-0.14\,\%$
		$(0.72*10^{-4})$		$(1.35*10^{-4})$	
	M2	$-8.96*10^{-4}$	$-1.64\,\%$	$70.2*10^{-4}$	$3.80\,\%$
		$(0.68*10^{-4})$		$(1.40*10^{-4})$	
	M3	$-5.02*10^{-4}$	$-0.92\,\%$	$67.8*10^{-4}$	$3.66\,\%$
		$(0.68*10^{-4})$		$(1.41*10^{-4})$	

Note: M1 is the correct model
M2 excludes the interactions and the square terms from the correct model
M3 drops further covariates from model M2.

Table 4
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1{,}000$)

Method	Assumed Covariates	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)	MSE($\hat{\theta}_1$)	MSE($\hat{\theta}_2$)
NN10	M1 (correct)	$2.02*10^{-3}$	$3.80*10^{-3}$	$4.10*10^{-6}$	$1.49*10^{-5}$
	M2	$2.06*10^{-3}$	$3.88*10^{-3}$	$4.29*10^{-6}$	$1.54*10^{-5}$
	M3	$2.01*10^{-3}$	$3.89*10^{-3}$	$4.10*10^{-6}$	$1.63*10^{-5}$
Propensity Score Weighting	M1 (correct)	$2.27*10^{-3}$	$4.27*10^{-3}$	$5.16*10^{-6}$	$1.83*10^{-5}$
	M2	$2.17*10^{-3}$	$4.42*10^{-3}$	$5.51*10^{-6}$	$6.90*10^{-5}$
	M3	$2.16*10^{-3}$	$4.46*10^{-3}$	$4.94*10^{-6}$	$6.59*10^{-5}$

Table 5
Simulation Estimates of Biases and Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting. Under the (non-MAR) Common Measurement Error Model ($H = 1{,}000$)

Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)
NN10	$29.0*10^{-4}$	$5.1\,\%$	$92.0*10^{-4}$	$5.0\,\%$	$2.53*10^{-3}$	$4.70*10^{-3}$
	$(0.8*10^{-4})$		$(1.48*10^{-4})$			
Propensity Score Weighting	$32.3*10^{-4}$	$5.7\,\%$	$100*10^{-4}$	$5.7\,\%$	$2.31*10^{-3}$	$4.42*10^{-3}$
	$(0.73*10^{-4})$		$(1.40*10^{-4})$			

8. Application to the Labour Force Survey

In this section we consider the application of the methods developed in sections $2-4$ to LFS data. The LFS provides an important source of estimates of the distribution of hourly pay in the UK (Stuttard and Jenkins 2001). It is a quarterly survey of households selected from a national file of postal addresses with equal probabilities by stratified systematic sampling. All adults in selected households are included in the sample. The resulting sample is clustered by

household membership but not by geography. Each selected household is retained in the sample for interview on five successive quarters and then rotated out and replaced. Questions relating to hourly pay are asked in just the first and fifth interviews, generating data on this topic for about 16,000 employees per quarter.

Two measures of hourly pay are constructed, as outlined in Section 1. The derived hourly pay variable in the LFS is defined as follows: (a) employees are asked questions about their main job to determine earnings over a reference period, (b) questions are asked to determine hours worked over the reference period and (c) the result of (a) is divided by the result of (b). The direct variable is obtained by first asking whether the respondent is paid a fixed hourly rate and then, if the answer is positive, by asking respondents what this (basic) rate is. Skinner *et al.* (2002) discuss how the derived variable suffers from many sources of measurement error, as in similar surveys in other countries (Rodgers, Brown and Duncan 1993; Moore, Stinson and Welniak 2000). They conclude that the direct variable measures hourly pay much more accurately. A working assumption in this application is that the direct variable measures hourly pay without error. The problem with the direct variable, however, is that it is missing for respondents who state that they are not paid at a fixed hourly rate (and for item nonrespondents) and this missingness is positively associated with hourly pay. The proportion of LFS respondents with a (main) job who provide a response to the direct question is about 43%. This proportion tends to be higher for lower paid employees, for example the rate is 72% among those in the bottom decile of the derived variable. The direct variable is not collected for second (and further) jobs and we therefore restrict attention only to main jobs. The aim is to use the missing data methods developed in this paper to correct for the measurement error in hourly pay. Skinner *et al.* (2002) discuss the plausibility of the two missing data assumptions in section 3 for this application.

The methods in sections $2-4$ were developed under the assumption of an IID model and ignorable sampling. Employees are selected with equal probabilities in the LFS so the sampling may be viewed as ignorable with respect to the bias of point estimation but unit non-response is likely to be differential and survey weights are constructed to compensate for this non-response (ONS 1999). We propose to incorporate these survey weights into the estimator in (3) or equivalently to multiply the weights w_i in (9) by the survey weights. This is analogous to the way the pseudo-likelihood approach (Skinner 1989) weights estimators based upon an IID assumption. The aim is to use the methods of sections $2-4$ to compensate for bias due to measurement error and item non-response and the survey weights to compensate for bias due to sampling and unit nonresponse. We have not attempted to take account of the weights in the imputation methods and this could be explored in future research.

We now apply nearest neighbour imputation, hot deck imputation within classes and propensity score weighting to LFS data. All methods are weighted by the survey weights. Figure 1 compares an estimated distribution, which ignores measurement error (the bold line) with estimates based on three missing data methods (the three dotted lines). We suggest the latter estimates are more approximately unbiased than the former estimate. All three missing data adjustments show, as expected, a strong 'kink' in the distribution at the level of the national minimum wage unlike for the derived variable. Corresponding estimates of two low pay proportions of interest are presented in Table 6. The 'missing data adjustments' have a substantial impact in comparison to estimates based on the derived variable. The results suggest that the proportion of jobs paid at or below the national minimum wage rate may be overestimated by four or five times if measurement error is ignored. The differences between the missing data methods are much smaller. We can see that the estimates under propensity score weighting differ from estimates derived using imputation methods, at least for the June-August 1999 quarter. Note that this quarter of the LFS was subject to a lower response rate than subsequent quarters resulting from changes in the LFS questionnaire. It was found that for consecutive quarters, which are subject to about 43% response rate, weighting and imputation led to very similar estimates of low pay proportions, as illustrated in table 7 for the March–May 2000 quarter. The decrease in the proportion of low paid employees over time is a result of the impact of the National Minimum Wage legislation. In addition, different imputation and propensity score models are used to analyse the effects of various model specifications on estimates of low pay. From Table 6 we can see that there is an indication that different models can have an effect on the estimates. With increasing complexity of the model a reduction in the estimates for both point estimators is observed. This might reflect a departure from the MAR assumption for the simpler imputation models. At least for the 1999 quarter, the differences in the estimates between weighting and imputation methods seem to be greater than between models. Note that the estimates presented here might differ slightly from official UK estimates since, for example, the official estimates are based on different imputation models, treating outliers differently or imputing differently for certain professions.

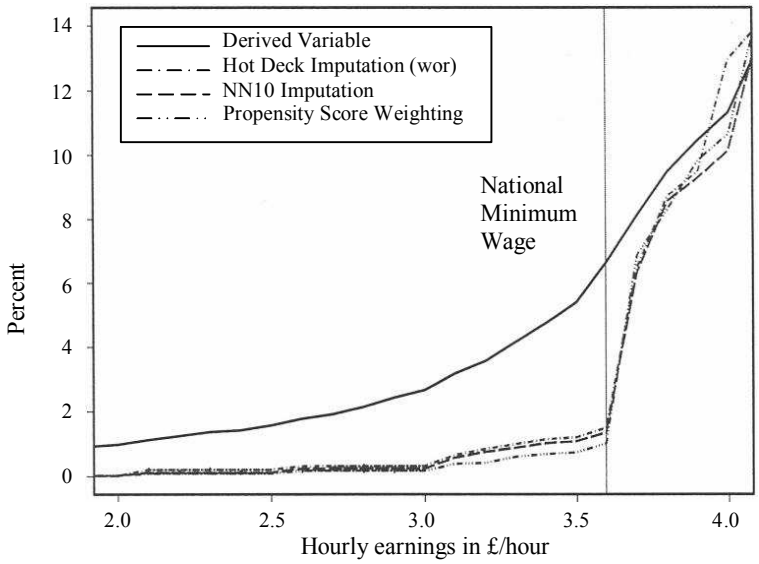


Figure 1. Alternative Estimates of the Distribution of Hourly Earnings From £2 to £4 for Age Group 22+, June-August 1999.

Table 6
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Different Propensity Score Models and Imputation Models Applied to LFS, June–August 1999

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Derived Variable	–	7.13	20.5
Propensity Score Weighting	M1	0.96	34.5
	M2	1.08	38.4
	M3	1.08	38.4
HDIWOR10	M1	1.44	32.1
	M2	1.41	32.9
	M3	1.50	33.2
NN10	M1	1.32	32.6
	M2	1.44	32.8
	M3	1.50	33.0

Note: M1 is the most complex model including square terms and interactions
M2 excludes the interactions and the square terms from model M1
M3 drops further covariates from model M2.

Table 7
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Propensity Score Weighting and Imputation Applied to LFS, March–May 2000

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Propensity Score Weighting	M1	0.54	27.10
HDIWOR10	M1	0.57	26.01
NN10	M1	0.55	26.61

9. Conclusions

In this paper we have considered the application of alternative missing data methods to correct for bias in the estimation of a distribution function arising from measurement error. Among imputation methods, nearest neighbour methods have performed most promisingly in terms of bias. These deterministic methods display no evidence of greater bias than stochastic imputation methods. Fractional imputation has shown appreciable efficiency gains compared to single imputation and appears more effective than penalizing the distance function or sampling without replacement with single imputation. In comparison to a propensity score weighting approach, the fractional nearest neighbour imputation has performed similarly, but has demonstrated slight advantages of robustness and efficiency. The simulation study suggested that the impact on the bias under a wrong model is greater for propensity score weighting and that the standard errors for the weighting approach were approximately $5-15\%$ times higher than for the imputation method.

Further research is being undertaken to develop and evaluate associated variance estimation methods, as well as alternative point estimation methods based upon the Common Measurement Error Model in section 2.

Acknowledgements

We are grateful to Danny Pfeffermann for comments on an earlier version of this paper.

References

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075-1082.

Chen, J., and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 113-131.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78, 451-462.

David, M.H., Little, R., Samuhel, M. and Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 168-173.

Dickens, R., and Manning, A. (2004). Has the national minimum wage reduced UK wage inequality? *Journal of the Royal Statistical Society*, Series A, 4, 613-626.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fay, R.E. (1999). Theory and application of nearest neighbour imputation in census 2000. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 112-121.

Fuller, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-141.

Kalton, G. (1983). *Compensating for missing survey data*. Michigan, Institute for Social Research.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, *Part A*, *Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kim, J.-K., and Fuller, W.A. (2002). Variance estimation for nearest neighbour imputation. Unpublished manuscript.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Luo, M., Stokes, L. and Sager, T. (1998). Estimation of the CDF of a finite population in the presence of a calibration sample. *Environmental and Ecological Statistics*, 5, 277-289.

Moore, J.C., Stinson, L.L. and Welniak, E.J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, 16, 331-361.

ONS (1999). *Labour Force Survey*. User Guide, Volume 1, Background and Methodology, London.

Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 131-138.

Rodgers, W.L., Brown, C. and Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked and hourly wages. *Journal of the American Statistical Association*, 88, 1208-1218.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., and Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith), Chichester, Wiley.

Skinner, C., Stuttard, N., Beissel-Durrant, G. and Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics*, 64, 653-676.

Stuttard, N., and Jenkins, J. (2001). Measuring low pay using the new earnings survey and the Labour Force Survey. *Labour Market Trends*, January 2001, 55-66.

Tenenbein, A. (1970). A double sampling scheme for estimating from binary data with misclassifications. *Journal of the American Statistical Association*, 65, 1350-1361.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 37-52
Statistics Canada, Catalogue No. 12-001-XIE

On Calibration Estimation for Quantiles

Torsten Harms and Pierre Duchesne [1]

Abstract

In this paper, we consider the estimation of quantiles using the calibration paradigm. The proposed methodology relies on an approach similar to the one leading to the original calibration estimators of Deville and Särndal (1992). An appealing property of the new methodology is that it is not necessary to know the values of the auxiliary variables for all units in the population. It suffices instead to know the corresponding quantiles for the auxiliary variables. When the quadratic metric is adopted, an analytic representation of the calibration weights is obtained. In this situation, the weights are similar to those leading to the generalized regression (GREG) estimator. Variance estimation and construction of confidence intervals are discussed. In a small simulation study, a calibration estimator is compared to other popular estimators for quantiles that also make use of auxiliary information.

Key Words: Calibration estimators; Quantiles; Ratio estimators; Difference estimators.

1. Introduction

In recent years, considerable attention has been given to the estimation of population distribution functions in the context of survey sampling. A particular target of this attention has been the median, which is often regarded as a more satisfactory location measure than the mean, especially when the variable of interest follows a skewed distribution. Traditional estimators of population means or totals can be usually substantially improved if relevant auxiliary information is made available. Consequently, the use of such auxiliary Information seems highly desirable in sample quantile estimators.

Using a model-based approach, Chambers and Dunstan (1986) considered quantile estimators based on an estimator of the distribution function which do incorporate auxiliary information. Rao, Kovar and Mantel (1990) have proposed design-based alternatives to the model-based approach. They used simulation experiments to compare two quantile estimators, based on ratio and difference estimators, to the simple design-based estimator which makes no use of the auxiliary information. It should be noted that neither of the two design-based proposals requires knowledge of the auxiliary information for each unit in the population; it rather suffices to know only the corresponding quantiles. While the model-based estimator proposed by Chambers and Dunstan (1986) can be more efficient than its design-based alternative if the model is correctly specified, Rao *et al.* (1990) have pointed out the advantage of the design-based estimators under model misspecification. Chambers, Dorfman and Hall (1992) have compared these two estimators theoretically with respect to their consistency, asymptotic bias and variance under a population model. Their main conclusion is that neither of the two methods is a sharp winner. Dorfman (1993) has reevaluated the simulation results obtained by Rao *et al.* (1990) and proposed a modified version of their methodology, using model-based arguments. Variance estimators in the model-based approach of Chambers and Dunstan (1986) and the design-based estimators of Rao *et al.* (1990) are discussed in Wu and Sitter (2001).

Other related works on quantile and median estimators include that of Kuk (1988) who proposes quantile estimators under pps (*proportional to size*) sampling and that of Kuk and Mak (1989) who use a method that is based on cross-classifying the individuals in the sample, according to the variable of interest and a single auxiliary variable. Meeden (1995) takes a different approach to construct a median estimator based on univariate auxiliary information, using the Bayesian concept of Polya sampling to impute all the target variable's unknown population values via a ratio-based approach. Rueda, Arcos and Martínez (2003) have recently built quantile estimators that extend ratio, difference and regression estimators in ways similar to those developed for the population mean.

In this paper, we follow the concept of calibration which was first introduced by Deville (1988) in order to derive a quantile estimator. The calibration approach has gained popularity in real applications, because the resulting estimators are easy to interpret and to motivate, relying, as they do, on sampling weights and natural calibration constraints. This approach was developed in the seminal work of Deville and Särndal (1992) as an alternative means of incorporating auxiliary information in the estimation of population totals. The so-called calibrated weights are found by minimizing a distance measure between the sampling weights and the new weights, which need to satisfy certain calibration constraints. For estimating totals the calibrated weights replace

1. Torsten Harms and Pierre Duchesne, Université de Montréal, Département de mathématiques et de statistique, CP 6128 Succursale Centre-Ville, Montréal, Québec, H3C 3J7, Canada. E-mail: duchesne@dms.umontreal.ca.

the original design weights used in Horvitz-Thompson type estimators. When the new weights are applied to the auxiliary variables available in the sample, they reproduce the known population totals of the auxiliary variables exactly; it is for this reason that the estimators in this class are called calibration estimators. See also Singh and Mohl (1996) who provide simple justifications of calibration estimators. They also present a very general and unifying treatment of calibration methods whose weights satisfy certain range restrictions and benchmark constraints.

Our fundamental aim is to propose calibration estimators for quantiles which are as easy to implement and interpret as the calibration estimators for totals developed by Deville and Särndal (1992). When compared to the quantile estimators available in the literature, the new calibration estimators should also be competitive with respect to their bias, variance, and coverage rates of the confidence intervals. Early calibration estimators for distribution functions and quantiles include those proposed by Kovačević (1997), who considered estimators of the distribution function calibrated on moments of the auxiliary variables. Harms (2003) has investigated a similar approach, with applications to the Finnish European Household Panel survey. Ren (2002) appears to have been the first to develop a unifying treatment of calibration estimators for distribution functions and quantiles. The calibration estimators for quantiles presented in this paper continue the work initiated by Ren (2002). We adhere to the original calibration paradigm for totals as closely as possible: when the parameter of interest is a total, it seems natural to calibrate on totals of the auxiliary variables. In the present context, since the parameter of interest corresponds to a quantile, the calibration constraints require that the weights are such that the sample quantile estimators of the auxiliary variables and their corresponding population quantiles are equal. In other words, the weighted quantile estimators for the auxiliary variables should yield exactly the population quantiles, which are assumed to be known. We present arguments which justify calibrating on quantiles, whenever the parameter of interest is itself a quantile. Interestingly, our methodology does not necessitate knowledge of the values of the auxiliary variables for all units in the population. Since the resulting estimators display a structural form very similar to the original calibration estimators for totals, it is expected that, under general conditions, the proposed estimators for quantiles will be asymptotically design-unbiased. Furthermore, these similarities allow us to derive variance estimators which admit a familiar form. Contrary to some of the other estimators, the proposed approach is also applicable to vectorial auxiliary variables (that is, when several auxiliary variables are available), while requiring only minimal auxiliary information. However, some restricttions may apply when the sample is highly unrepresentative of the sampled population or when the quantiles being estimated are very close to the population minimum or maximum. Note that highly unrepresentative samples can also cause problems for calibration estimators for totals commonly used; in such situations, the algorithm for computing calibration estimators may fail to converge for many distance measures of practical interest.

The organization of the paper is as follows: In section 2, some preliminaries are given, including a brief review of the calibration estimators for totals. The new calibration estimators for quantiles are developed in section 3.1. The standard distribution function can be interpreted as a Horvitz-Thompson estimator, providing a possible approach to the construction of a calibrated distribution function estimator. Quantile estimators are then naturally derived by inverting the distribution function estimator (see *e.g.*, Ren (2002)). As in calibration estimators for totals, design weights can be replaced by more general sampling weights, in order to take account the auxiliary information. However, for many situations of practical interest, it may happen that no solution exists for the calibration constraints when this kind of distribution function estimator is adopted, the reason being that this estimator corresponds to a step function. In order to avoid existence problems of solutions for the calibration constraints, a new distribution function estimator is introduced, based on the natural concept of interpolation. Under the common quadratic metric, an analytic representtation of the calibration weights is provided in section 3.2; variance estimators and confidence intervals are discussed in section 3.3. A practical aspect involves evaluating the methodology proposed with real populations and several sampling plans. Consequently, in section 4, we present a small simulation study where we compare our new approach, with respect to variance, bias and coverage rates of the confidence intervals, with that of Chambers and Dunstan (1986) as well as with some of the estimators proposed by Rao *et al.* (1990). Finally, concluding remarks are offered in section 5.

2. Some Preliminaries on Calibration Estimators

In this section, we present the fundamental concepts and notations useful for the sequel. We also give a brief review of calibration estimators for totals.

Let $U = \{1, ..., k, ..., N\}$ be a finite population of size N. Let $T_y = \sum_U y_k$ be the population total of the variable of interest y, (note that for a set A, $A \subseteq U$, $\sum_A$ will be used as shorthand for $\sum_{k \in A}$). A sample $s \subset U$ of size n is drawn according to a sampling plan. Let $\pi_k = \Pr(s \ni k)$ and $\pi_{kl} = \Pr(s \ni k, l)$ be the first and second order inclusion probabilities, respectively. We denote the design

weights $d_k = \pi_k^{-1}$ and $\hat{T}_{y,\mathrm{HT}} = \sum_s d_k y_k$ represents the Horvitz-Thompson (HT) estimator of T_y.

Let $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$ be a vector of auxiliary variables associated with unit k, $k \in U$. Calibration estimators naturally include auxiliary information in the estimation. Let $s = \{k_1, ..., k_n\}$, $s \subset U$. Starting with the vector of original weights $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$, new weights are found which, when applied to the auxiliary variables available in s, make it possible to retrieve the known population totals for the J auxiliary variables $\mathbf{T}_\mathbf{x} = \sum_U \mathbf{x}_k = (T_{x_1}, ..., T_{x_J})'$. The calibration estimator for totals are more precisely defined in Definition 1.

Definition 1 (Calibration estimator for totals). *Let* $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$ *be the design weights. The calibration estimator for totals takes the form* $\hat{T}_{y,\mathrm{cal}} = \sum_s w_{ks} y_k$, *where the weights* w_{ks}, $k \in s$ *are obtained as the following minimization problem with respect to the variable* $\mathbf{v} = (v_{k_1}, ..., v_{k_n})'$:

$$\mathbf{w} = \mathrm{argmin}_\mathbf{v}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \tag{1}$$

subject to the calibration constraints $\sum_s v_k \mathbf{x}_k = \mathbf{T}_\mathbf{x}$, *where* $\mathrm{D}(\cdot,\cdot)$ *denotes the distance measure and* $\mathbf{w} = (w_{k_1}, ..., w_{k_n})'$ *corresponds to the vector of the calibrated weights.*

For notational simplicity, we write $w_k \equiv w_{ks}$ in Definition 1 when no confusion is possible. It is common practice to let $x_{1k} \equiv 1$, $\forall k \in U$, and consequently $T_{x_1} = N$. This means that the calibrated weights satisfy the natural constraint $\sum_s w_k = N$. Many distance functions D are available in the literature (see, *e.g.*, Deville and Särndal (1992), Chen and Qin (1993), Thompson (1997)). Consider the quadratic distance function

$$\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s \frac{(v_k - d_k)^2}{d_k q_k}, \tag{2}$$

where q_k determines the importance of the unit $k \in s$ in the calibration problem. Heteroscedasticity problems can be handled using an appropriate choice of the q_k's. Solving the optimization problem (1) using the Lagrange multiplier technique (see Deville and Särndal (1992), among others), the weights $w_k = d_k(1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}_s)$ are obtained, where $\boldsymbol{\lambda}_s = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\mathbf{T}_\mathbf{x} - \hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}})$ and $\hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}}$ denotes the HT-estimator of $\mathbf{T}_\mathbf{x}$. This choice of distance function leads to the weights of the well-known generalized regression estimator (GREG) of Cassel, Särndal and Wretman (1976), which is studied in detail in Särndal, Swensson and Wretman (1992). Under minimal requirements for the disance measure D, Deville and Särndal (1992) have shown that all calibration estimators in this class are asymptotically equivalent to the GREG. For ease of interpretation and other cosmetic reasons, some users may want to have positive weights or restrict them to a specific interval (see also Singh and Mohl (1996)). In practical applications, these numerical features of the weights seem to be the main motivation for an alternative choice of D.

3. New Calibration Estimators

In this section we develop calibration estimators for quantiles, using ideas similar to those leading to the caliration estimators for population totals, as described in section 2. The new calibration estimators for quantiles are introduced in the next subsection, using interpolated distribution function estimators. Then, special attention is devoted to the quadratic distance function. The last subection presents variance estimation and the construction of confidence intervals.

3.1 Definition of the Calibration Estimators for Quantiles

Let $\mathbf{Q}_{\mathbf{x},\alpha} = (Q_{x_1,\alpha}, ..., Q_{x_J,\alpha})'$ denote the known vector of population quantiles for the vector of auxiliary variables $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$, $k \in U$. The Heavyside function $H(z)$ is given by:

$$H(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

The population distribution function of a scalar auxiliary variable x is defined in the usual way as $F_x(t) = N^{-1}\sum_U H(t - x_k)$, and the population quantile $Q_{x,\alpha}$ is obtained by letting $Q_{x,\alpha} = \inf\{t \mid F_x(t) \geq \alpha\}$.

The vector $\mathbf{Q}_{\mathbf{x},\alpha}$ contains quantiles of the auxiliary variables, obtained from information in past surveys or from available administrative sources. For example, for skewed distributions which are rather common in business and economic surveys, it seems more natural to keep in the record files the population medians rather than population means; in this case it seems natural to assume the knowledge of $\mathbf{Q}_{\mathbf{x},0.5}$. This suggests that, using the same approach as the one leading to calibration for totals described in section 2, the proposed estimator for the population quantile $Q_{y,\alpha}$ of the variable of interest y, noted $\hat{Q}_{y,\mathrm{cal},\alpha}$, could be obtained by inverting a certain estimator of the distribution function (that we discuss below), subject to calibration constraints such as $\hat{Q}_{x_j,\mathrm{cal},\alpha} = Q_{x_j,\alpha}$, $j = 1, ..., J$. Following the usual interpretation, if the calibrated weights allow us to retrieve the known population quantiles of the auxiliary variables then, under certain conditions, they should produce reasonable estimators for the quantile of the variable of interest y.

More precisely, the calibrated weights are obtained by solving the following optimization problem:

$$\mathbf{w} = \mathrm{argmin}_\mathbf{v}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \tag{3}$$

subject to the calibration constraints $\sum_s v_k = N$ and $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, \ldots, \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$.

The estimators $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha}$ and $\hat{Q}_{y,\text{cal},\alpha}$ rely on the vector of weights $\mathbf{w}$, stemming from the solution of the calibration problem (3). To calculate these estimators for quantiles, we need to construct $w-$weighted estimators of the distribution function for variables $\mathbf{x}$ and y. Based on the sampling weights $\mathbf{d}$, a natural estimator of the sampling distribution function is given by

$$\tilde{F}_y(t) = \sum_s d_k H(t - y_k) \Big/ \sum_s d_k, \tag{4}$$

which provides a consistent estimator of $F_y(t)$. Similarly, $F_{x_j}(t)$ can be consistently estimated by $\tilde{F}_{x_j}(t) = \sum_s d_k H(t - x_{jk}) / \sum_s d_k$, $j = 1, \ldots, J$. A $w-$weighted distribution function estimator of $F_{x_j}(t)$ is given by

$$\tilde{F}_{x_j,\text{cal}}(t) = \sum_s w_k H(t - x_{jk}) \Big/ \sum_s w_k. \tag{5}$$

A similar formula holds for $\tilde{F}_{y,\text{cal}}(t)$. These $w-$weighted estimators are considered in Ren (2002). However, if one estimates $Q_{x_j,\alpha}$ by $\hat{Q}_{x_j,\alpha} = \inf\{t \mid \tilde{F}_{x_j}(t) \geq \alpha\}$, or makes a similar estimation using a $w-$weighted version, then it is generally not possible to reach an exact solution of the calibration problem (3). Indeed, if the previous definition is used to estimate the quantiles by inverting the distribution function using the previous definitions, then the constraints in the optimization problem (3) will not, in general, be fulfilled unless the sample s contains precisely a unit k such that $x_{jk} = Q_{x_j,\alpha}$. When J is large, this problem can be more pronounced. Furthermore, even if the sample does contain such a value, it is sometimes not possible to obtain the weights needed to minimize the distance function, the reason being that under certain circumstances, the weights fulfilling the calibration constraints form an open set, whereas the optimal weights lie precisely on the border of this set. The following example illustrates this situation.

Example 1:
Consider a population U of size $N = 30$, such that the population median of x is $Q_{x,0.5} = 2$. A sample s of size $n = 3$ is drawn, and suppose that $x_k = k, \forall k \in s = \{1, 2, 3\}$. For simplicity, the distance measure $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s (v_k - d_k)^2$ is adopted; it is supposed that the sampling weights are $(d_1, d_2, d_3) = (15, 9, 6)$. Based on (5), the calibration constraint is $\tilde{Q}_{x,\text{cal},0.5} = \inf\{t \mid \tilde{F}_{x,\text{cal}}(t) \geq 0.5\} = 2$, which implies that $\sum_s w_k H(2 - x_k) \geq 15$ and $\sum_s w_k H(1 - x_k) < 15$. Equivalently, $w_1 + w_2 \geq 15$ and $w_1 < 15$. Thus we have to choose w_1 of the form $w_1 = 15 - \epsilon$, for $\epsilon > 0$. In this case, since $w_1 + w_2 + w_3 = 30$, we have that $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \epsilon^2 + (w_2 - 9)^2 + (w_2 - 9 - \epsilon)^2$, leading to the optimal solution $(w_1, w_2, w_3) = (15 - \epsilon, 9 + \epsilon/2, 6 + \epsilon/2)$. Consequently, for these weights $\mathrm{D}(\mathbf{v}, \mathbf{d}) = 3\epsilon^2/2$, which is obviously minimized when $\epsilon \to 0$. However, the limit reduces to $\mathbf{w} = (w_1, w_2, w_3) = (15, 9, 6)$ with $\mathrm{D}(\mathbf{w}, \mathbf{d}) = 0$, but based on these weights $\tilde{Q}_{x,\text{cal},0.5} = 1 \neq Q_{x,0.5} = 2$.

However, these difficulties can be naturally avoided by considering a smooth estimator of the distribution function. For simplicity, we consider here a distribution function estimator calculated using a linear interpolation (another possibility is discussed in section 5), which is precisely defined in Definition 2.

Definition 2 (Interpolated distribution function estimators). *Define*

$$\hat{F}_{y,\text{cal}}(t) = \frac{\sum_s w_k H_{y,s}(t, y_k)}{\sum_s w_k}, \tag{6}$$

$$\hat{F}_{x_j,\text{cal}}(t) = \frac{\sum_s w_k H_{x_j,s}(t, x_{jk})}{\sum_s w_k}, \tag{7}$$

where the Heavyside function H in (4) *and* (5) *is replaced by the slightly modified function*

$$H_{y,s}(t, y_k) = \begin{cases} 1, & y_k \leq L_{y,s}(t), \\ \beta_{y,s}(t) & y_k = U_{y,s}(t), \\ 0, & y_k > U_{y,s}(t), \end{cases} \tag{8}$$

where $L_{y,s}(t) = \max\{\{y_k, k \in s \mid y_k \leq t\} \cup \{-\infty\}\}$, $U_{y,s}(t) = \min\{\{y_k, k \in s \mid y_k > t\} \cup \{\infty\}\}$ *and* $\beta_{y,s}(t) = \{t - L_{y,s}(t)\}/\{U_{y,s}(t) - L_{y,s}(t)\}$. *The function* $H_{x_j,s}(t, x_k)$ *is defined similarly. The estimators* (6) *and* (7), *based on the functions* $H_{y,s}(t, y_k)$ *and* $H_{x_j,s}(t, x_k)$, *are called interpolated distribution function estimators of* $F_y(t)$ *and* $F_{x_j}(t)$, *respectively*.

The various quantities in (8) have easy interpretations: $L_{y,s}$ and $U_{y,s}$ represent the lower and upper neighbors of t in the sampled values $y_k, k \in s$, and $\beta_{y,s}(t)$ denotes the linear interpolation coefficient between these two quantities. In particular, for all $t \in \{y_k, k \in s\}$ we have $H_{y,s}(t, y_k) = H(t - y_k)$. Consequently, the relations $\hat{F}_{y,\text{cal}}(t) = \tilde{F}_{y,\text{cal}}(t)$ are satisfied for all $t \in \{y_k, k \in s\}$. For all the other values of t, $\hat{F}_{y,\text{cal}}(t)$ consists of a linear interpolation between these quantities. In the following example, Example 1 is revisited using the interpolated distribution function estimator (7).

Example 2:
In Example 1, using the interpolated version (7), the constraints are now $w_1 + w_2 + w_3 = 30$ and $(w_1 + w_2)/(w_1 + w_2 + w_3) = 0.5$. Consequently $w_3 = 15$, $w_1 + w_2 = 15$. Simple algebra shows that the optimal solution is $(w_1, w_2, w_3) = (10.5, 4.5, 15)$, which is now well-defined.

With the interpolated distribution function estimators, $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ and $\hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$ are now well defined $\alpha-$quantile estimators for all $\alpha \in (0, 1)$, as long as one can assure that the weights w_k are all strictly positive. Letting $\hat{Q}_{x_j,\text{cal},\alpha} = \hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$, we define the proposed calibration estimator

$\hat{Q}_{y,\text{cal},\alpha}$ for the quantile $Q_{y,\alpha}$, using the interpolated distribution function estimator given in Definition 2.

Definition 3 (Calibration estimator for quantiles). *Consider the optimization problem* (3), *subject to the calibration constraints* $\sum_s v_k = N$ *and* $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, \ldots, \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$. *Solving this optimization problem and denoting the resulting weights as* $\mathbf{w}$, *the proposed calibration estimator for quantiles of* $Q_{y,\alpha}$ *is defined by*

$$\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha), \tag{9}$$

where $\hat{F}_{y,\text{cal}}(t)$ *is given by* (6).

One of the appealing properties of the proposed estimator (9) is that it yields exact population quantiles when the relationship between y and a scalar auxiliary variable x is exactly linear. Assume that $y_k = a + bx_k$ holds perfectly for all units $k \in U$ and suppose that the units in the sample s are such that $x_k < Q_{x,\alpha} < x_l$ for some units x_k and $x_l, k, l \in s$. For the calibrated estimator (9), we have that $\hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = \alpha$. We need to distinguish the two cases, $b > 0$ and $b < 0$ (The case $b = 0$ is trivial since y_k is then identically equal to a constant). Firstly, consider the situation $b > 0$. Since the linear relation $y_k = a + bx_k$ is satisfied for all units k and since $b > 0$, the following relations hold: $L_{y,s}(a + bt) = a + bL_{x,s}(t); U_{y,s}(a + bt) = a + bU_{x,s}(t)$ and $\beta_{y,s}(a + bt) = \beta_{x,s}(t)$. These relations lead to $H_{y,s}(a + bt, y_k) = H_{x,s}(t, x_k)$. It follows that $\hat{F}_{y,\text{cal}}(a + bt) = \hat{F}_{x,\text{cal}}(t)$. Furthermore, $\hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = \alpha$ and using the relation $a + bQ_{x,\alpha} = Q_{y,\alpha}$, we deduce that $\hat{F}_{y,\text{cal}}(Q_{y,\alpha}) = \alpha$. Consequently, when an exact linear relationship holds and $b > 0, \hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha) = Q_{y,\alpha}$. Secondly, consider the case $b < 0$. We deduce in this case the following relations: $L_{y,s}(a + bt) = a + bU_{x,s}(t); U_{y,s}(a + bt) = a + bL_{x,s}(t); \beta_{y,s}(a + bt) = 1 - \beta_{x,s}(t)$ and $H_{y,s}(a + bt, y_k) = 1 - H_{x,s}(t, x_k)$. Since $b < 0$, the relationship between the quantiles of x and y is given by $a + bQ_{x,\alpha} = Q_{y,1-\alpha}$. Then, we deduce that $\hat{F}_{y,\text{cal}}(Q_{y,1-\alpha}) = \hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = 1 - \hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = 1 - \alpha$. Thus, in this situation, $Q_{y,1-\alpha}$ is estimated exactly by $\hat{Q}_{y,\text{cal},1-\alpha}$. This means that, when an exact relation holds, if $b > 0$ the proposed calibration estimator $\hat{Q}_{y,\text{cal},\alpha}$ yields perfect estimators with zero bias and variance of $Q_{y,\alpha}$. On the other hand, if $b < 0$ and calibrating on $Q_{x,\alpha}, Q_{y,1-\alpha}$ is estimated exactly by $\hat{Q}_{y,\text{cal},1-\alpha}$ (which makes sense because the perfect linear relationship between x and y is such that the slope parameter is negative).

Note that when $\hat{F}_{y,\text{cal}}$ and $\hat{F}_{x_j,\text{cal}}$ are invertible at points $Q_{y,\alpha}$ and $Q_{x_j,\alpha}$, the calibration constraints in (3) can be rewritten in terms of the distribution functions, that is the calibration constraints based on the quantiles are equivalent to $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha, j = 1, \ldots, J$. This means that the original calibration problem can be alternatively written in terms of distribution functions with the above constraints.

A natural question arises as to the existence of a solution to the optimization problem (3). Even when formulated with the interpolated distribution functions, it is not always possible to find a solution to (3). For example, if $Q_{x_j,\alpha}$ is smaller or larger than all values x_{jk} in the sample s, then $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha})$ will equal zero or one regardless of the choice of the weights $\mathbf{w}$. Thus in these cases it may happen that the calibration constraints cannot be fulfilled. However, when the sample's behavior differs widely from that of the target population, one should keep a very critical eye on any adjustment, and this situation can be considered somewhat extreme. In practice, this rarely occurs unless α is chosen very close to zero or one. Note that it may be impossible to obtain a solution when the sample size n is small. In these situations, the sample minimum or maximum could serve as a possible estimator or we could resort to the simple design-based estimator of the distribution function.

The second potential problem is that some weights w_k might be negative. In this case $\hat{F}_{y,\text{cal}}$ is no longer bijective. This is not a problem as long as $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ is still uniquely determined. This problem can be avoided by restricting all the weights to be strictly positive, using an appropriate metric $D(\cdot,\cdot)$. This approach has been adopted by Kovačević (1997) (for more details on distance functions yielding positive weights, see also Deville and Särndal (1992) and Singh and Mohl (1996)).

Remark 1:
The proposed distribution functions estimators (6) and (7) rely on a linear interpolation. In a unified way, the population distribution function, which is a step function as well, could also be defined using a linear interpolation. In practice, the two definitions differ only slightly in behavior, if the population N is sufficiently large. However, it should be noted that if the population size N is relatively small, it might be worth using an interpolation to define distribution functions.

Remark 2:
In the optimization problem (3), we calibrated on a particular quantile. This approach could be extended by allowing to calibrate on a finite set of quantiles, if such information is available. More precisely, suppose that for an auxiliary variable x, the α_m – quantiles $Q_{x,\alpha_m}, m = 1, \ldots, M$ are known, where $M < n - 1$. In this case, we could consider the calibration constraints $\hat{F}_{x,\text{cal}}(Q_{x,\alpha_m}) = \alpha_m, m = 1, \ldots, M$ and solve the optimization problem (3) with these additional calibration constraints. Naturally, this information yields a more complete description of the distribution of the auxiliary variables; so the efficiency of the calibration estimators is expected to be higher.

Remark 3:
The proposed calibration estimator (9) is obtained by calibrating on population quantiles. Another possibility has been considered by Ren (2002) who calibrated on population moments, up to order m, of the same distribution. More precisely, Ren (2002) has proposed calibration estimators for quantiles satisfying constraints of the form $\sum_s w_k x_k^m = \sum_U x_k^m$, $m = 0, 1, ..., M$. Calibration on different moments of the same distribution is closely related to calibrating on different quantiles of the same variable, and all these constraints provide a more complete description of the distribution of the auxiliary variable. For other generalizations of the calibration paradigm on moments, see also Ren and Deville (2000) and Harms (2003).

3.2 Analytical Solution of the Calibrated Weights when D is the Quadratic Metric

When the quadratic distance function (2) is adopted, an explicit solution of the optimization problem (3) can be derived. This situation is similar to the calibration estimators for totals, where the weights of the GREG estimator are explicitly obtained under the metric (2). A careful analysis of the estimation problem for quantiles reveals important similarities, the reason being that the estimators given by (7) are weighted sums of the variables $\{H_{x_j,s}(t, x_{jk}),\ k \in s\}$, $j = 1, ..., J$. This is stated in Proposition 1.

Proposition 1 (Calibrated weights for the quadratic metric). *Consider the quadratic distance function* (2)*. The vector of weights* $\mathbf{w}$ *which solves the optimization problem* (3) *satisfies the relation*:

$$w_k = d_k(1 + q_k \mathbf{a}_k' \boldsymbol{\lambda}_s), k \in s, \tag{10}$$

where the vector $\boldsymbol{\lambda}_s = (\lambda_0, ..., \lambda_J)'$ *is determined via the* $J+1$ *constraints as*:

$$\boldsymbol{\lambda}_s = \left(\sum_s d_k\ q_k\ \mathbf{a}_k\ \mathbf{a}_k'\right)^{-1}\left(\mathbf{T}_\mathbf{a} - \sum_s d_k\ \mathbf{a}_k\right), \tag{11}$$

with $\mathbf{T}_\mathbf{a} = (N, \alpha, ..., \alpha)'$ *and the components of* $\mathbf{a}_k = (1, a_{1k}, ..., a_{Jk})'$ *are given by*

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

with $j = 1, ..., J$.

Proof. To prove Proposition 1, first note that, since the first constraint $\sum_s w_k = N$ must be satisfied, it follows that $\hat{F}_{x_j,\text{cal}}(t) = N^{-1}\sum_s w_k H_{x_j,s}(t, x_{jk})$. Proceeding as in Deville and Särndal (1992), we can show that the vector $\mathbf{a}_k = (1, a_{1k}, ..., a_{Jk})'$ satisfies

$$\mathbf{a}_k = \left(1, \frac{\partial \hat{F}_{x_1,\text{cal}}}{\partial w_k}, ..., \frac{\partial \hat{F}_{x_J,\text{cal}}}{\partial w_k}\right)'\Bigg|_{\sum_s w_k = N;\ \hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha,\ j=1,...,J}, \tag{12}$$

that we now evaluate explicitly. Evaluating the derivatives, we have that $a_{jk} = N^{-1}H_{x_j,s}(t, x_{jk})$, $j = 1, ..., J$, evaluated at $t = Q_{x_j,\alpha}$. This leads to

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

$j = 1, ..., J$, as announced.

In (11), $\mathbf{T}_\mathbf{a}$ can be interpreted as the expected value of $\sum_s d_k\ \mathbf{a}_k$. The derived weights (10) in the distribution function estimator (6) rely on the variables $\mathbf{a}_k$, $k \in s$ defined by (12). Note that they correspond to a certain transformation of the auxiliary variable $\mathbf{x}_k$. The difference between the weights for totals and quantiles relies on this variable $\mathbf{a}_k$; when $\mathbf{a}_k$ is replaced by $\mathbf{x}_k$, we retrieve the original weights for totals. Consequently, it is useful to interpret this new variable. When estimating a total, the impact on the j^{th} calibration constraint is measured by x_{jk}, for each unit $k \in s$. In our framework, the impact of the unit k is now given by N^{-1} if $x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha})$; it corresponds to the factor $N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha})$ when $x_{jk} = U_{x_j,s}(Q_{x_j,\alpha})$ and it is null elsewhere. In section 5, we shall discuss other estimation problems, leading to different variables $\mathbf{a}_k$.

Noting the similarities between the estimation of totals and quantiles, variance estimation can also be considered. This issue is addressed in the next subsection.

3.3 Variance Estimation and Confidence Intervals

As described in the previous section, the estimator $\hat{Q}_{y,\text{cal},\alpha}$ displays several similarities to the usual GREG estimator for population totals. The transformed variables given by (12) provide the main difference between the calibration estimators for quantiles and totals. Interestingly, because of the structural similarity with the original calibration estimators, it is straightforward to derive a confidence interval for the proposed estimator $\hat{Q}_{y,\text{cal},\alpha}$. We consider the construction of confidence intervals following Woodruff's (1952) approach. The confidence interval is given in Result 1.

Result 1 (Woodruff confidence interval for the calibration estimator for quantiles). *The confidence interval based on Woodruff's* (1952) *approach, using the calibration estimator* (9) *for the quantile* $Q_{y,\alpha}$ *is given by*

$$[\hat{F}^{-1}_{y,\mathrm{cal}}(\hat{c}_{1y}),\ \hat{F}^{-1}_{y,\mathrm{cal}}(\hat{c}_{2y})], \qquad (13)$$

where $\hat{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\}]^{1/2}$ *and* $\hat{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\}]^{1/2}$. *The resulting procedure yields an approximate confidence interval for* $Q_{y,\alpha}$ *at a specified* $1-\gamma$ *confidence level.*

Proof. Assuming that $\hat{F}_{y,\mathrm{cal},\alpha}(Q_{y,\alpha})$ is approximately normally distributed, it follows that $\Pr(c_{1y} \le \hat{F}_{y,\mathrm{cal},\alpha}(Q_{y,\alpha}) \le c_{2y})$ should approximately be equal to $1-\gamma$, if one chooses

$$c_{1y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\}]^{1/2}, \qquad (14)$$

$$c_{2y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\}]^{1/2}, \qquad (15)$$

where z_γ denotes the γ^{th} – quantile of the $N(0,1)$ standard normal distribution. Since $\hat{F}_{y,\mathrm{cal},\alpha}(Q_{y,\alpha})$ represents essentially a sample mean, a possible variance estimator justified by the classical Taylor linearization is given by

$$\hat{V}\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\} = N^{-2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}(w_k\, e_k)\,(w_l\, e_l), \qquad (16)$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$; the weights $w_k, k \in s$, correspond to the calibrated weights (3) which reduce to (10) when D is the quadratic distance function (2); the residuals are given by $e_k = H_{y,s}(\hat{Q}_{y,\mathrm{cal},\alpha},\ y_k) - \mathbf{a}_k'\,\hat{\mathbf{B}}_s$ where

$$\hat{\mathbf{B}}_s = \left(\sum_s w_k\ q_k\ \mathbf{a}_k\ \mathbf{a}_k'\right)^{-1}\sum_s w_k\ q_k\ \mathbf{a}_k\ H_{y,s}(Q_{y,\mathrm{cal},\alpha},\ y_k)$$

represents the regression coefficient estimator. Since the constants c_{1y} and c_{2y} given by (14) and (15) rely on $V\{\hat{F}_{y,\mathrm{cal}}(Q_{y,\alpha})\}$, we can estimate these quantities using the variance estimator (16).

In Result 1, note that Deville and Särndal (1992) advocated a w – weighted variance estimator similar to (16) for estimating the variance of the calibration estimators of the population totals. The performance of the proposed calibration estimator (9) and the confidence interval given by (13) are studied empirically in section 4.

4. Simulation Results

From a practical point of view, it is natural to inquire about the finite sample properties of the new calibration estimators and to compare them to popular estimators for quantiles available in the literature. In this section, simulation experiments are undertaken, to illustrate empirically the new estimators. In particular, their empirical bias and variance in real populations are investigated. The coverage properties of the confidence intervals represent another question of practical interest, which is also studied.

In partial answer to these questions, we carried out three small simulation studies. For several sampling plans and for real populations, the proposed calibration estimator for quantiles is compared to its popular competitors. In the next subsection 4.1, we describe in detail the populations investigated and we discuss the sampling plans chosen. In subsection 4.2, the estimators included in the empirical study are presented and, in subsection 4.3, the frequentist measures (empirical bias, variance and mean squared error, coverage rates of the confidence intervals) are described. Our empirical results are analyzed in subsection 4.4.

4.1 Description of the Real Populations and the Sampling Plans

The real populations are displayed in Figures 1 to 6. The first population, noted MU284, is taken from Särndal *et al.* (1992, Appendix B). This population consists of $N = 284$ municipalities in Sweden. We retain as variable of interest the population in 1985 (variable P85), and we assume that the auxiliary information available is the population in 1975 (variable P75). Both variables are measured in thousands. In Figure 1, the variable P85 is expressed as a function of P75; as expected, the relationship between P85 and P75 is strongly linear. The variable P85 follows a highly skewed distribution, as shown in Figure 2. In this population, 500 samples were drawn according to simple random sampling without replacement (SRS). In addition, the same study was carried out under a sampling plan with unequal probabilities, the Poisson (PO) sampling scheme. The properties of the PO sampling plan are described in Särndal *et al.* (1992). Due to the wide range of values for y, it was not possible to construct sample selection probabilities π_k of the form $\pi_k \propto y_k$, since this would mean that some π_k had to be greater than one. For the purpose of our illustration, we determined selection probabilities using the relation $\pi_k \propto 0.2y_k + 0.05$ (we recognize that these π_k's are idealized, since y_k is not available in practice). Under the SRS sampling plan (PO sampling plan), we considered the sample sizes (expected sample sizes) $n = 25$ and $n = 50$.

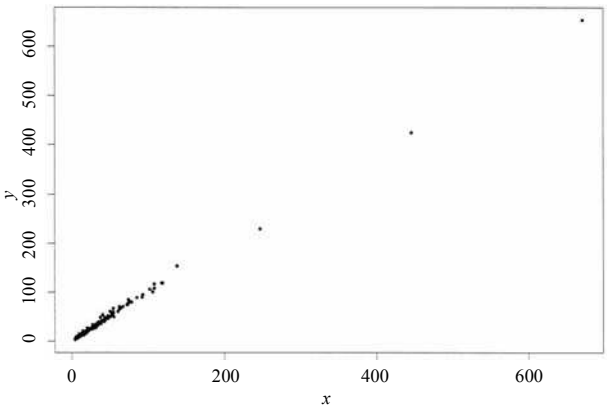


Figure 1. The Population MU284, where y = P85 and x = P75.

For the second study, we chose the MU284 population, but now made the variable of interest y = RMT85, which

represents the revenues from 1985 municipal taxation (in millions of kronor). Here the auxiliary variable chosen is $x = \text{REV84}$, which denotes real estate values according to 1984 assessments for each municipality (in millions of kronor). As can be seen in Figure 3, the relationship between x and y is somewhat spread out for larger values of x. The histogram of the variable RMT85 reveals that it follows a skewed distribution (Figure 4). For this study, 500 samples were drawn according to the SRS scheme of size $n = 25$ and $n = 50$.

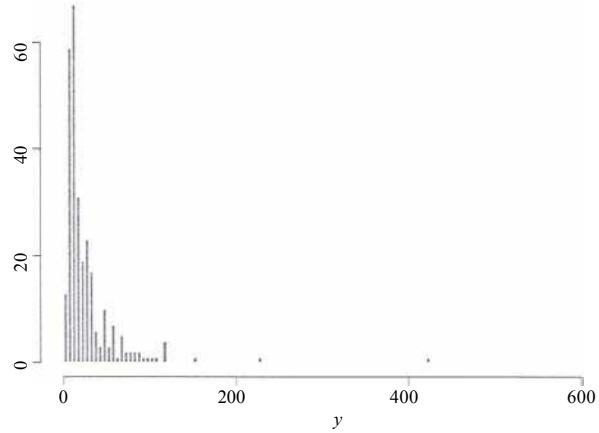


Figure 2. Histogram of the Variable P85 in the MU284 Population.

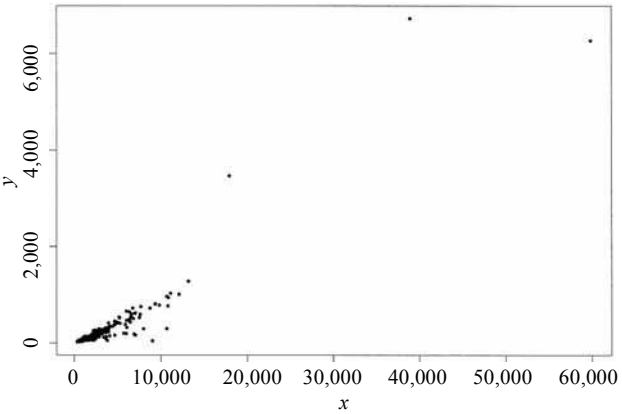


Figure 3. The Population MU284, where $y =$ RMT85 and $x =$ REV84.

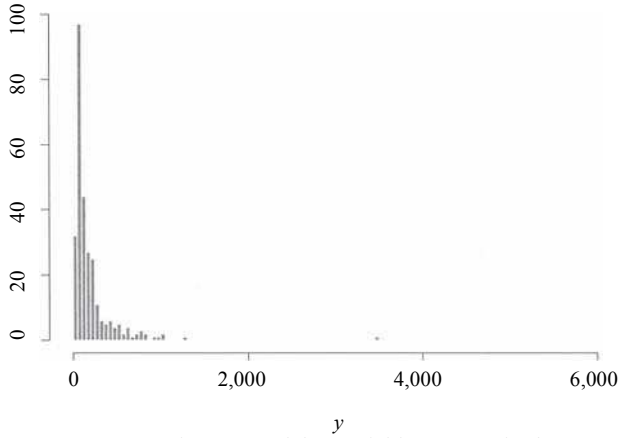


Figure 4. Histogram of the Variable RMT85 in the MU284 Population.

The third population is based on a random subsample of the *Survey of Labor and Income Dynamics*, noted SLID982. The survey was conducted at Statistics Canada in 1998. For simplicity's sake, only entries with no missing values were selected. The size of the subsample is $N = 2,000$ and for our purpose this is assumed to be a population (the original sample size of this survey is approximately 60,000). Taxable income (in thousands of dollars) is the target variable and the auxiliary variable is the duration in months of the current employment. From Figure 5, the linear relationship between taxable income and length of employment is less pronounced. However, the two variables do not appear to be independent. In Figure 6, the variable of interest exhibits a strong coefficient of skewness. We have drawn 500 samples from the SLID982 population, according to SRS and PO sampling plans. The sample sizes (expected sample size) $n = 100$ and $n = 200$ were considered. For PO sampling, the first order probabilities, $\pi_k, k \in U$, were defined according to two rules. Under the first rule, the π_k's were created such that π_k is approximately proportional to the variable of interest, that is taxable income (for the purpose of our study we assume that it is possible to create such π_k's). Since some y_k are negative in this population, we chose $p_{1k} = y_k - \min\{y_k,\ k \in U\} + 1$ and we defined $\pi_k = E(n_s)p_{1k} / \sum_U p_{1k}$, where $E(n_s)$ stands for the expected sample size, in our case $E(n_s) = 100$ and 200. Under the second rule, the π_k's were proportional to the entries in Table 1. This means that for each $k \in U$, there exists a factor p_{2k}, which is determined by the age-sex group of individual k. Then $\pi_k = E(n_s)p_{2k} / \sum_U p_{2k}$, where the factors p_{2k} are given in Table 1. The factors p_{2k} in Table 1 are based on a hypothetical sampling plan, in which we assume that these factors provide suitable size measures for the units in the various age-sex classes (see *e.g.*, Särndal *et al.* (1992, page 87)); for these units, more males than females are likely to be selected and, for both sexes, adults in the 27 to 37 and 38 to 46 age range are more likely to be included in the sample.

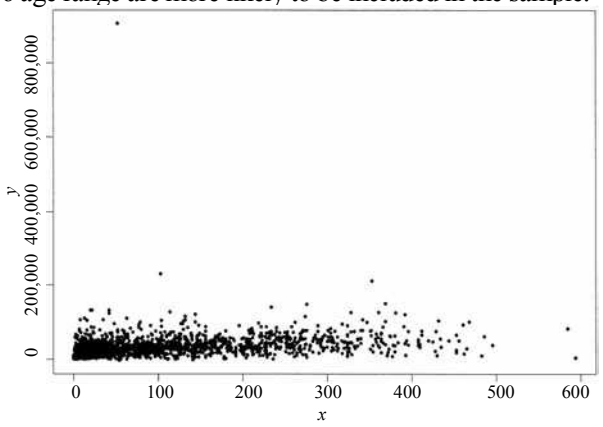


Figure 5. The SLID982 Population, where the Dependent Variable is the Taxable Income and Independent Variable is the Duration of Current Employment (in Months).

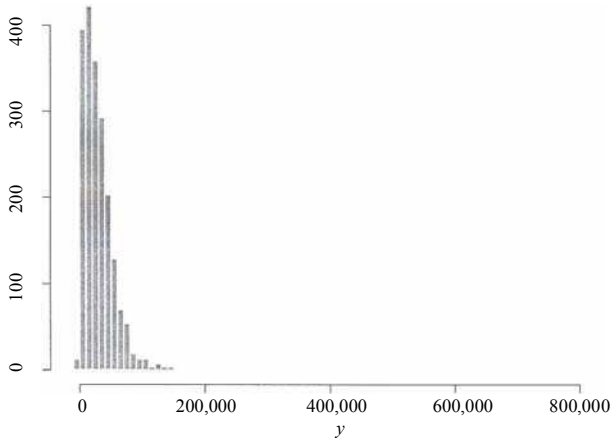


Figure 6. Histogram of the Taxable Income in the SLID982 Population.

Table 1
Factor p_{2k} by Age and Sex of Individual k, in the SLID982 Population

		Age			
		16–25	27–37	38–46	47–69
Sex	Male	3	6	5	4
	Female	1	2	3	2

In these three studies, we estimate the quartiles, that is the population parameters $Q_{y,\alpha}$ with $\alpha = 0.25$, 0.5 and 0.75. Since the variables of interest display highly skewed distributions, it might be particulary interesting to study the quantile corresponding to $\alpha = 0.75$, in addition to the median and the first quartile. The next section describes the estimators included in the study.

4.2 Estimators Included in the Empirical Study

Since one of our goals is to propose estimators with reasonable properties with respect to bias, variance and coverage rates of the confidence intervals, we compare the new estimator defined by (9) based on the metric (2) to some of the popular quantile estimators proposed in the literature.

First, we include the simple design-based estimator based on the inversion of the estimator $\hat{F}_y(t) = \sum_s d_k\, H_{y,s}(t, y_k) / \sum_s d_k$:

$$\hat{Q}_{y,\mathrm{HT},\alpha} = \hat{F}_y^{-1}(\alpha). \tag{17}$$

The estimator (17) does not make use of auxiliary information. A possible variance estimator is

$$\hat{V}\{\hat{F}_y(Q_{y,\alpha})\} = \hat{N}^{-2} \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_k) - \alpha)}{\pi_k} \right\} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_l) - \alpha}{\pi_l} \right\},$$

where $\hat{N} = \sum_s d_k$, and confidence intervals can be calculated using

$$[\hat{F}_y^{-1}(\tilde{c}_{1y}),\ \hat{F}_y^{-1}(\tilde{c}_{2y})],$$

where

$$\tilde{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}, \tag{18}$$

$$\tilde{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}. \tag{19}$$

For more details, see Särndal *et al.* (1992, page 202).

We also include in our empirical study the model-based estimator of Chambers and Dunstan (1986), which is motivated by a linear superpopulation model $y_k = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_k + \epsilon_k$, $k \in U$, where ϵ_k forms an identically and independently distributed sequence of random variables with mean zero and finite variance. Their estimator is defined as

$$\hat{Q}_{y,\mathrm{CD},\alpha} = \inf\{t \mid \hat{F}_{y,\mathrm{CD}}(t) \geq \alpha\}, \tag{20}$$

where $\hat{F}_{y,\mathrm{CD}}(t) = N^{-1}\{\sum_s H(t - y_k) + \sum_{U/s} \hat{G}(t - \hat{y}_k)\}$ represents a model-based estimator of the distribution function,

$$\hat{G}(u) = n^{-1} \sum_s H(u - \hat{\epsilon}_k) \tag{21}$$

denotes the empirical distribution function of the residuals $\hat{\epsilon}_k = y_k - \hat{y}_k$, $k \in s$, and $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_k$, $k \in U/s$ correspond to the least-squares predictions. Since the estimator (20) basically imputes the unknown y_k for $k \in U/s$, note that it necessitates a complete knowledge of $\mathbf{x}_k$ for $k \in U$.

The construction of a confidence interval for $\hat{Q}_{y,\mathrm{CD},\alpha}$ relies on estimating the variance $V\{\hat{F}_{y,\mathrm{CD}}(t)\}$. However, this variance estimation problem creates difficulties, since any analytical variance formula depends on the assumed model. Furthermore, such analytical expressions involve kernel density estimators, which are numerically intensive and depend on a kernel function and a bandwidth. For all these reasons, we decide to implement the delete-one jackknife variance estimators studied in Wu and Sitter (2001), who have shown the consistency of the proposed variance estimators. In the context of survey sampling, various resampling methods, including the jackknife, are introduced in Kovar, Rao and Wu (1988). The jackknife technique involves deleting a unit and re-calculating the estimator. Let $s_i = s/\{i\}$ be the sample without unit i. Consider $\hat{\beta}_{0i}$ and $\hat{\boldsymbol{\beta}}_i$, the regression estimators of β_0 and $\boldsymbol{\beta}$ calculated on s_i. Under a simple regression model, define

$$F_i^* = (n-1)^{-1} \sum_{k \in s_i} \left[N^{-1} \sum_{l \in U/s} H\{\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{\beta}_i(x_l - x_k) - y_k\} \right].$$

A consistent variance estimator of $V\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\mathrm{CD},\alpha})\}$ is given by

$$\hat{V}_{y,\mathrm{CD}}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\} = \frac{n-1}{n}\sum_{i\in s}\ (F_i^* - \bar{F}^*)^2 + \frac{f(1-f)}{N-n}\sum_{k\in U/s} \hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\,\{1-\hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\},$$

where $f = n/N$ is the sampling fraction, $\bar{F}^* = n^{-1}\sum_s F_i^*$, and $\hat{G}$ is given by (21). Based on $\hat{V}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\}$, it is now possible to calculate the confidence intervals for $Q_{y,\alpha}$ using the inversion approach.

Since our method necessitates only the knowledge of the vector of quantiles $\mathbf{Q}_{\mathbf{x},\alpha}$, we include in our study the ratio and difference estimators for the quantiles studied in Rao *et al.* (1990):

$$\hat{Q}_{y,\mathrm{ra},\alpha} = Q_{x,\alpha}(\hat{Q}_{y,\mathrm{HT},\alpha}/\hat{Q}_{x,\mathrm{HT},\alpha}), \tag{22}$$

$$\hat{Q}_{y,\mathrm{diff},\alpha} = \hat{Q}_{y,\mathrm{HT},\alpha} + \hat{R}(Q_{x,\alpha} - \hat{Q}_{x,\mathrm{HT},\alpha}), \tag{23}$$

where $\hat{Q}_{y,\mathrm{HT},\alpha}$ is given by (17) and $\hat{Q}_{x,\mathrm{HT},\alpha}$ is calculated similarly; the ratio estimator given by $\hat{R} = \sum_s d_k\, y_k / \sum_s d_k\, x_k$ provides a consistent estimator of $R = \sum_U y_k / \sum_U x_k$. Note that the estimators (22) and (23) are elaborated based on a scalar auxiliary variable, that is $J = 1$. Valid variance estimators of (22) and (23) are given by:

$$\hat{V}(\hat{Q}_{y,\mathrm{ra},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \left(\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\right)^2 \hat{V}(\hat{Q}_{x,\mathrm{HT},\alpha}) - 2\,\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\ \hat{Q}_{x,\mathrm{HT},\alpha}),$$

$$\hat{V}(\hat{Q}_{y,\mathrm{diff},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \hat{R}^2\,\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) - 2\hat{R}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\ \hat{Q}_{x,\mathrm{HT},\alpha}).$$

These variance estimators rely on the variance of $\hat{Q}_{y,\mathrm{HT},\alpha}$, and the covariance between $\hat{Q}_{y,\mathrm{HT},\alpha}$ and $\hat{Q}_{x,\mathrm{HT},\alpha}$ which are estimated using Woodruff's (1952) approach:

$$\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) = \frac{W_y^2}{4z_{1-\gamma/2}^2},$$

$$\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\ \hat{Q}_{x,\mathrm{HT},\alpha}) = \frac{W_y W_x \hat{C}\{\hat{F}_x(Q_{x,\alpha}),\ \hat{F}_y(Q_{y,\alpha})\}}{4z_{1-\gamma/2}^2[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}},$$

where $W_y = \hat{F}_y^{-1}(\tilde{c}_{2y}) - \hat{F}_y^{-1}(\tilde{c}_{1y})$ and $W_x = \hat{F}_x^{-1}(\tilde{c}_{2x}) - \hat{F}_x^{-1}(\tilde{c}_{1x})$ denote the Woodruff intervals associated with y and x, with $\tilde{c}_{1y}$ and $\tilde{c}_{2y}$ defined by (18) and (19), $\tilde{c}_{1x} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$, $\tilde{c}_{2x} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$ and

$$\hat{C}\{\hat{F}_y(Q_{y,\alpha}),\ \hat{F}_x(Q_{x,\alpha})\} = \hat{N}^{-2}\sum\sum_s\ \frac{\Delta_{kl}}{\pi_{kl}}\left\{\frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha},\ y_k) - \alpha}{\pi_k}\right\}\left\{\frac{H_{x,s}(\hat{Q}_{x,\mathrm{HT},\alpha},\ x_l) - \alpha}{\pi_l}\right\}.$$

Summarizing, we expect $\hat{Q}_{y,\mathrm{CD},\alpha}$ to perform well when the linear model describes the population adequately. This motivates the comparison of the new methodology with a model-based estimator. Furthermore, it seems of interest to evaluate $\hat{Q}_{y,\mathrm{cal},\alpha}$ and the leading design-based proposals, such as $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{ra},\alpha}$. The estimators $\hat{Q}_{y,\mathrm{cal},\alpha}$, $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{ra},\alpha}$ use $Q_{x,\alpha}$ only to improve the estimations and they take into account the sampling plan; these estimators are natural competitors. Note that the different estimators included in our study are elaborated under different assumptions on the dimension of the vector of the auxiliary variable $\mathbf{x}$, and on the availability of $\mathbf{x}_k$. Table 2 provides a comparison of the different estimators described in this section.

Table 2
Comparison of the Proposed Calibration Estimators and of Some Leading Estimators for Quantiles Proposed in the Literature, with Respect to the Dimension J of $\mathbf{x}$ and the information requirement on $\mathbf{x}$

Estimator	Dimension of $\mathbf{x}$	Information requirements on $\mathbf{x}$
$\hat{Q}_{y,\mathrm{HT},\alpha}$	n.a.	none
$\hat{Q}_{y,\mathrm{CD},\alpha}$	$J \geq 1$	$\mathbf{x}_k, k \in U/s$
$\hat{Q}_{y,\mathrm{ra},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{diff},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{cal},\alpha}$	$J \geq 1$	$\mathbf{Q}_{\mathbf{x},\alpha}$

4.3 Frequentist Measures

Our goal it to evaluate the estimators with respect to bias and variance. Other important considerations are the mean squared error (MSE) and the coverage rates of the confidence intervals.

Let $\hat{Q}_{y,\alpha}$ be an estimator of the population quantile $Q_{y,\alpha}$. Assume $\hat{Q}_{y,\alpha}^{(v)}$ is the estimator of the quantile calculated using the sample v, $v = 1, \ldots, K$. The Monte Carlo mean E_{MC}, the Monte Carlo bias B_{MC}, and the Monte Carlo variance V_{MC} are given by the usual formulas, that is

$$E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\hat{Q}_{y,\alpha}^{(v)},$$

$$B_{\mathrm{MC}} = E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) - Q_{y,\alpha},$$

$$V_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\{\hat{Q}_{y,\alpha}^{(v)} - E_{\mathrm{MC}}(\hat{Q}_{y,\alpha})\}^2.$$

Our main criterion for determining efficiency is the Monte Carlo MSE, defined by $\mathrm{MSE}_{\mathrm{MC}} = K^{-1}\sum_{v=1}^{K}(\hat{Q}_{y,\alpha}^{(v)} - Q_{y,\alpha})^2$. The confidence intervals are calculated at the 95% confidence level, according to the procedures described in the previous sections. For an estimator $\hat{Q}_{y,\alpha}^{(v)}$ and its variance estimator $\hat{V}^{(v)}$, $v = 1, \ldots, K$, the coverage rates at the 95% confidence level are calculated as

$$\mathrm{CR}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K} I\left(\left\{Q_{y,\alpha} \in \left[\hat{Q}_{y,\alpha}^{(v)} - 1.96\sqrt{\hat{V}^{(v)}},\ \hat{Q}_{y,\alpha}^{(v)} + 1.96\sqrt{\hat{V}^{(v)}}\right]\right\}\right),$$

where $I(A)$ is the indicator function of the set A. The coverage rates are given below the column CR. We recall that we adopt $K = 500$ for all studies.

4.4 Discussion of the Empirical Results

The results are presented in Tables 3 to 8. We first discuss the results from Tables 3 to 4, when sampling the MU284 population with SRS and PO sampling plans. As can be seen, all the estimators display a similar behavior in both studies. The model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ appears to be the most efficient among those analyzed when examining $\alpha = 0.75$ and is in general very efficient. This was expected, since the relationship between $x = \mathrm{P75}$ and $y = \mathrm{P85}$ is strongly linear and the model-based estimator assumes a simple regression model. However, for $\alpha = 0.25$ the differences in efficiency are less pronounced with respect to the other estimators based on auxiliary information. Among the estimators using only $Q_{x,\alpha}$ as information on the auxiliary variable, a rather similar performance is obtained. When the sample size is small, coverage rates usually deviate from the 95% nominal level. This is particulary true for the coverage rates of $\hat{Q}_{y,\mathrm{cal},\alpha}$, which are somewhat underestimated. However, some improvement is observed at $n = 50$, illustrating the consistency of the procedures studied. On the other hand, those of $\hat{Q}_{y,\mathrm{ra},\alpha}$ and $\hat{Q}_{y,\mathrm{diff},\alpha}$ are always one. This suggests that the variances are overestimated for these estimators. Due to an important component of bias in the MSE, the coverage rates of the model-based estimator sometimes deteriorate as the sample size increases. The best coverage rates are obtained by using the simple HT estimator, $\hat{Q}_{y,\mathrm{HT},\alpha}$, which is however less efficient than the other estimators.

Table 5 shows the result for the second population, which is the MU284 population but with $y = \mathrm{RMT85}$ and $x = \mathrm{REV84}$. Figure 3 seems to show a heteroscedasticity phenomenon in this population. In view of this, since the ratio estimator is justified when the underlying population displays such behavior, it is not surprising that the ratio estimator $\hat{Q}_{y,\mathrm{ra},\alpha}$ performs well in this particular situation; if outperforms $\hat{Q}_{y,\mathrm{diff},\alpha}$ in several cases. For a small sample size, the ratio estimator generally behaves better than $\hat{Q}_{y,\mathrm{cal},\alpha}$. However, for $n = 50$, the calibration estimator appears to perform as well or slightly better than the ratio estimator. In this experiment, the bias and variance of the model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ increase the MSE substantially. Furthermore, in some cases, confidence intervals for this estimator could not be obtained, since the Woodruff method is not appropriate in cases with extremely large variance (the Woodruff interval becomes too large and the linearity of the distribution function within this interval can thus no longer be assumed). We suspect that a model taking into account heteroscedasticity might improve the performance of the model-based estimator. This highlights the fact that to obtain high efficiency with model-based estimators, the model must be correctly specified.

The results in Table 6 to 8 concern the SLID982 population, under SRS and PO sampling plans with two rules for the π_k's. All the estimators in Table 6 perform reasonably well in estimating the first quartile and the median, except for the ratio estimator $\hat{Q}_{y,\mathrm{ra},\alpha}$ which is the least efficient. Since the relationship between the dependent and independent variables is not precisely a linear model, this may partially explain the poor performance of the ratio estimator in this case. The relationship between x and y is not proportional and so the difference estimator $\hat{Q}_{y,\mathrm{diff},\alpha}$ appears preferable to $\hat{Q}_{y,\mathrm{ra},\alpha}$. However, for $\alpha = 0.75$, these estimators show the highest MSE, being both the least efficient. Interestingly, in this part of the experiment $\hat{Q}_{y,\mathrm{cal},\alpha}$ dominates the design-based estimators in terms of MSE. However, for small α, $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{cal},\alpha}$ perform similarly. It should be noted that for a larger sample size, $\hat{Q}_{y,\mathrm{cal},\alpha}$ and $\hat{Q}_{y,\mathrm{CD},\alpha}$ give the best efficiencies for the median and the third quartile. In fact, the model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ slightly outperforms $\hat{Q}_{y,\mathrm{cal},\alpha}$, but it should be noted that it uses more auxiliary information than $\hat{Q}_{y,\mathrm{cal},\alpha}$.

Tables 7 and 8 present results under PO sampling plans. In general, design-based estimators perform much like those under SRS sampling plan. This is not the case for the model-based estimator; it is less efficient, likely because it does not incorporate the information about the sampling plan. More precisely, Table 7 presents simulation results under PO sampling, using the first rule for the π_k's, $k \in U$. Coverage rates of the model-based estimator are particularly disappointing in this experiment; the components of bias were too important in the MSE. The design-based estimators provide much closer empirical coverage rates, to the nominal 95% confidence level. For moderate and large α, $\hat{Q}_{y,\mathrm{cal},\alpha}$ is the most efficient estimator. In fact, the calibration estimator $\hat{Q}_{y,\mathrm{cal},\alpha}$ performs well in this

experiment. Finally, Table 8 presents results obtained under PO sampling with the second rule for the π_k's. In this case, $\hat{Q}_{y,\text{ra},\alpha}$. is the least efficient estimator for the first quartile and the median, and $\hat{Q}_{y,\text{diff},\alpha}$ is the least efficient for $\alpha = 0.75$. In general, $\hat{Q}_{y,\text{cal},\alpha}$ dominates the other estimators in this situation, offering the highest efficiency.

Table 3
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = P85, x = P75, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	−0.0343	0.5075	0.5077	0.886	−0.0499	0.2437	0.2457	0.828
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0266	2.3196	2.3157	0.952	0.0035	1.1087	1.1065	0.936
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.1444	0.3869	0.4070	1.000	−0.0774	0.1684	0.1741	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1486	0.3901	0.4114	1.000	−0.0734	0.1723	0.1774	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	0.4855	0.2791	0.5143	0.906	0.5485	0.1981	0.4985	0.824
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.2762	1.6499	1.7229	0.918	−0.2835	0.9585	1.0370	0.944
	$\hat{Q}_{y,\text{HT},\alpha}$	0.2605	12.5161	12.5589	0.922	−0.0064	5.8466	5.8349	0.916
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.2586	0.8828	0.9479	1.000	−0.4296	0.6701	0.8533	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.2775	0.9898	1.0648	1.000	−0.4331	0.7492	0.9352	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	0.9431	0.4054	1.2940	0.866	0.9884	0.2410	1.2175	0.714
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.6229	3.3241	3.7055	0.614	−0.3661	1.8107	1.9411	0.710
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.1414	53.1951	53.1088	0.948	−0.3692	18.8586	18.9572	0.964
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.7925	3.0021	3.6242	1.000	−1.0004	1.4594	2.4573	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.8230	3.4379	4.1083	1.000	−1.0396	1.5267	2.6044	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	0.4343	0.5108	0.6984	0.954	0.4485	0.2618	0.4624	0.974

Table 4
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = P85, x = P75, Under PO Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	−0.0441	0.4886	0.4896	0.888	−0.0169	0.2601	0.2599	0.828
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.1698	2.2825	2.3068	0.936	−0.0384	1.1828	1.1819	0.928
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.1509	0.3857	0.4076	1.000	−0.0913	0.2100	0.2179	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1634	0.3821	0.4080	1.000	−0.0877	0.2149	0.2221	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	0.6709	0.3310	0.7805	0.896	0.8792	0.1339	0.9066	0.554
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.3610	1.4881	1.6155	0.920	−0.3236	0.8833	0.9863	0.936
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0612	11.3969	11.3778	0.926	−0.2712	5.2672	5.3302	0.906
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.3735	1.0009	1.1385	1.000	−0.4130	0.5486	0.7181	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.3962	1.1271	1.2818	1.000	−0.4217	0.5962	0.7729	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	1.1740	0.4947	1.8719	0.820	1.3297	0.2146	1.9822	0.474
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.6420	2.6605	3.0674	0.608	−0.4476	1.6212	1.8183	0.708
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.6200	51.2934	51.5752	0.956	−0.6632	17.3625	17.7677	0.966
	$\hat{Q}_{y,\text{ra},\alpha}$	−0.8686	2.8841	3.6329	1.000	−0.9683	1.6494	2.5837	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.9025	2.9826	3.7911	1.000	−1.0177	1.6340	2.6665	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	0.4620	0.4501	0.6627	0.982	0.5388	0.2329	0.5228	0.980

Table 5
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = RMT85, x = REV84, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	1.0161	51.5421	52.4714	0.892	0.6499	24.0662	24.4404	0.954
	$\hat{Q}_{y,\text{HT},\alpha}$	0.3733	110.2572	110.1760	0.960	0.3383	47.2921	47.3120	0.962
	$\hat{Q}_{y,\text{ra},\alpha}$	3.0025	65.4135	74.2979	0.998	2.3856	30.7284	36.3580	0.992
	$\hat{Q}_{y,\text{diff},\alpha}$	2.5952	107.7891	114.3084	0.994	2.4083	55.6977	61.3862	0.986
	$\hat{Q}_{y,\text{CD},\alpha}$	−16.5165	1661.0257	1930.4983	0.990	−17.3217	820.7447	1119.1443	0.960
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−1.6219	215.0326	217.2330	0.870	−0.3419	118.2125	118.0930	0.922
	$\hat{Q}_{y,\text{HT},\alpha}$	0.0075	763.6236	762.0964	0.910	−0.3977	331.2357	330.7314	0.914
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7712	212.8298	212.9988	0.996	−0.2810	136.4382	136.2443	0.996
	$\hat{Q}_{y,\text{diff},\alpha}$	0.3415	283.6718	283.2210	0.998	−1.0104	201.3707	201.9889	0.998
	$\hat{Q}_{y,\text{CD},\alpha}$	17.6124	190.0045	499.8199	n.a.	13.5037	100.2106	282.3611	0.566
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−5.3477	1023.6924	1050.2431	0.826	−4.7339	443.0660	464.5896	0.926
	$\hat{Q}_{y,\text{HT},\alpha}$	−4.6352	3526.8202	3541.2514	0.938	−5.8890	1242.4858	1274.6812	0.940
	$\hat{Q}_{y,\text{ra},\alpha}$	−1.4390	980.5573	980.6669	0.994	−2.0070	555.5135	558.4305	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−5.3988	1464.7867	1491.0041	0.996	−3.9008	744.1604	757.8881	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	49.3038	2753.8212	5179.1826	n.a.	49.4089	1488.9734	3927.2324	0.596

Table 6
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.1360	3.0390	3.0514	0.956	0.2331	1.6787	1.7297	0.934
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0596	3.6099	3.6062	0.946	0.0499	1.9277	1.9263	0.918
	$\hat{Q}_{y,\text{ra},\alpha}$	0.3067	6.8815	6.9618	0.970	0.0910	3.0743	3.0764	0.958
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.0504	2.9691	2.9657	0.980	0.0198	1.6139	1.6111	0.952
	$\hat{Q}_{y,\text{CD},\alpha}$	1.1042	2.1180	3.3329	0.922	1.1392	1.2937	2.5888	0.826
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.4034	6.3364	6.4865	0.966	−0.1402	2.9940	3.0076	0.940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4157	7.4589	7.6168	0.918	−0.1894	3.5865	3.6151	0.928
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7015	41.8314	42.2399	0.958	0.2238	18.7005	18.7131	0.952
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.4859	14.2083	14.4160	0.970	−0.2740	6.6184	6.6803	0.974
	$\hat{Q}_{y,\text{CD},\alpha}$	0.5702	3.5420	3.8601	0.952	0.6697	1.7559	2.2009	0.932
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.4164	12.4657	12.6142	0.952	−0.2384	5.9118	5.9568	0.950
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.5913	12.5456	12.8701	0.930	−0.3519	6.5496	6.6603	0.926
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7404	48.6836	49.1345	0.954	0.2967	18.5786	18.6294	0.966
	$\hat{Q}_{y,\text{diff},\alpha}$	0.3288	53.6456	53.6464	0.954	0.1841	21.7552	21.7456	0.966
	$\hat{Q}_{y,\text{CD},\alpha}$	0.5966	8.3416	8.6809	0.954	0.5413	4.3692	4.6535	0.936

Table 7
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the First Rule for the Construction of the π_k, $k \in U$. The number of replications is set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.1393	4.8403	4.8500	0.956	0.1603	2.8293	2.8493	0.922
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0477	5.8276	5.8182	0.934	−0.0227	3.5939	3.5872	0.924
	$\hat{Q}_{y,\text{ra},\alpha}$	0.1648	9.5171	9.5252	0.980	0.1263	4.8687	4.8749	0.972
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1418	4.7045	4.7152	0.960	−0.0464	2.9213	2.9176	0.936
	$\hat{Q}_{y,\text{CD},\alpha}$	3.9150	3.5279	18.8477	0.584	3.9114	1.9163	17.2112	0.194
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1746	8.2437	8.2577	0.944	−0.2413	3.6477	3.6986	0.940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.2824	10.1117	10.1712	0.916	−0.3343	4.5023	4.6050	0.936
	$\hat{Q}_{y,\text{ra},\alpha}$	0.6558	50.4938	50.8228	0.944	0.4263	26.5883	26.7169	0.948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.5975	17.0315	17.3544	0.972	−0.3496	8.9060	9.0104	0.970
	$\hat{Q}_{y,\text{CD},\alpha}$	4.3173	4.4061	23.0363	0.484	4.0937	2.0711	18.8252	0.184
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.2229	12.1861	12.2114	0.942	−0.2113	6.5823	6.6138	0.952
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4150	14.2935	14.4371	0.934	−0.2786	7.6597	7.7220	0.934
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7861	47.3844	47.9077	0.980	−0.1344	19.5992	19.5781	0.958
	$\hat{Q}_{y,\text{diff},\alpha}$	0.4347	52.3845	52.4687	0.972	−0.3409	23.8277	23.8962	0.958
	$\hat{Q}_{y,\text{CD},\alpha}$	4.4114	7.7023	27.1478	0.654	4.3549	4.1566	23.1136	0.392

Table 8
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the Second Rule for the Construction of the π_k, $k \in U$. The Number of Replications is Set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.2392	3.4402	3.4906	0.962	0.1674	1.5214	1.5464	0.952
	$\hat{Q}_{y,\text{HT},\alpha}$	0.0267	4.0027	3.9954	0.940	−0.0370	1.6995	1.6975	0.958
	$\hat{Q}_{y,\text{ra},\alpha}$	0.4402	7.4350	7.6139	0.970	0.1850	3.0687	3.0968	0.978
	$\hat{Q}_{y,\text{diff},\alpha}$	0.0528	3.2842	3.2804	0.972	−0.0127	1.4718	1.4690	0.964
	$\hat{Q}_{y,\text{CD},\alpha}$	2.1458	3.0460	7.6444	0.876	1.9785	1.3010	5.2130	0.690
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1410	6.5627	6.5695	0.942	−0.2850	2.9662	3.0415	0.954
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.2133	7.6604	7.6906	0.928	−0.2876	3.6017	3.6772	0.926
	$\hat{Q}_{y,\text{ra},\alpha}$	1.0245	43.2773	44.2402	0.930	−0.3075	17.7242	17.7833	0.948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1973	14.5261	14.5360	0.958	−0.6111	6.2988	6.6596	0.978
	$\hat{Q}_{y,\text{CD},\alpha}$	2.2140	4.5617	9.4543	0.834	1.8882	2.0393	5.6005	0.738
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1985	12.6334	12.6476	0.952	−0.0022	5.6442	5.6329	0.966
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4012	13.5045	13.6384	0.922	−0.1078	6.2239	6.2231	0.934
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7968	44.0650	44.6118	0.958	0.3727	19.1830	19.2836	0.960
	$\hat{Q}_{y,\text{diff},\alpha}$	0.4613	49.6620	49.7755	0.960	0.2340	22.1292	22.1397	0.966
	$\hat{Q}_{y,\text{CD},\alpha}$	2.6329	9.6723	16.5850	0.854	2.6729	4.1179	11.2541	0.738

5. Concluding Remarks

In this paper, we have developed quantile estimators based on the calibration paradigm. The estimators are particularly easy to implement and to interpret, since they focus on weights and calibration constraints. Furthermore, they require only the population quantiles of the auxiliary variables, which can be vectorial. When the quadratic metric is adopted, analytic expressions can be obtained for calibrated weights as well as variance estimators, which are similar to those for the calibration estimator for totals. From a practical point of view, an appealing consequence of the new methodology is that the proposed estimators are easy to calculate; it suffices to transform the auxiliary variables and then use existing software to compute the calibration estimators.

In a small simulation study, we compared the calibration estimator for quantiles, under the quadratic metric, to other leading quantile estimators available in the literature. The proposed estimator performed reasonably well in our empirical experiments; its performance was often preferable or at least similar to that of other estimators using the same amount of information. The model-based estimator incorporating much more information about the auxiliary variables appeared preferable under SRS sampling and a correctly specified model, but was outperformed by the new estimator when the first order inclusion probabilities were unequal. In general, the proposed estimator compared very well with the design-based alternatives of Rao *et al.* (1990).

While, in this paper, we have concentrated on the estimation of quantiles by calibrating on known population quantiles for the auxiliary variables, calibration estimators can be extended to other important estimation problems of interest in survey sampling. The formulation of these problems all lead to different transformed variables, that we have noted $\mathbf{a}_k$ in this paper. For example, it is possible to formulate a calibration problem for the well-known Gini coefficient and then show that the solution to this calibration problem will give weights analogous to those derived in this paper; however these weights can only be determined numerically. More work is needed in this direction, in order to extend calibration estimators to a more general framework, which would include totals, quantiles, and Gini coefficients as special cases. Another challenging research avenue concerns the choice of the distribution function estimator. In this paper, we have advocated a distribution function estimator calculated using a linear interpolation. Alternatively, we could consider kernel distribution function estimator (see *e.g.*, Altman and Léger (1995)). Kernel density estimation from complex surveys is elaborated in Bellhouse and Stafford (1999). This means that, in $\hat{F}_{y,\,\mathrm{cal}}(t)$, the function $H_{y,\,s}(t, y_k)$ could be replaced by a general kernel, which would, however, depend on an additional parameter, the bandwidth. Note that the linear interpolation in the present paper avoids the choice of a bandwidth, which is often a delicate matter. Developing a general framework for calibration problems of a certain functional, and kernel distribution function estimators, are left for future studies.

Acknowledgements

The authors thank two anonymous referees for their thoughtful comments and suggestions, which greatly enhanced the paper. Discussions and comments from Raymond Chambers, Christian Léger, Éric Rancourt, Ulrich Rendtel and participants in the 32nd meeting of the Statistical Society of Canada and of the 2004 Joint Statistical Meeting are gratefully acknowledged. The first author was supported by a scholarship from the German Academic Exchange Service (DAAD) and the second author by grants from the National Science and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les technologies du Québec (Canada).

References

Altman, N., and Léger, C. (1995), Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195-214.

Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

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

Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of finite population distribution functions. *Biometrika*, 79, 577-582.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80, 107-116.

Deville, J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquêtes par sondage. In *Essais en l'Honneur d'Edmont Malinvaud*, (Eds, A. Monfort, and J.J. Laffond), *Economica*, Paris, 915-929.

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

Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.

Harms, T. (2003), Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators, Chintex working paper #13, Federal Statistical Office, Germany.

Kovačević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, 139-144.

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 (Supp.), 25-45.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75, 97-103.

Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society*, Series B (Methodological), 51, 261-269.

Meeden, G. (1995). Median estimation using auxiliary information. *Survey Methodology*, 21, 71-77.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, *INSEE Méthodes*, Tome 1, 100, 263-289.

Ren, R., and Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments, II^{ème} Colloque Francophone sur les Sondages. Bruxelles.

Rueda, M.M., Arcos A. and Martínez, M.D. (2003). Difference estimators of quantiles in finite populations. *Test*, 12, 481-496.

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

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

Thompson, M. (1997). *Theory of Sample Surveys*. Chapman & Hall, New York.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 625-646.

Wu, C., and Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, 29, 289-308.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 53-64
Statistics Canada, Catalogue No. 12-001-XIE

A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data

David Haziza and Jon N.K. Rao [1]

Abstract

In the presence of item nonreponse, two approaches have been traditionally used to make inference on parameters of interest. The first approach assumes uniform response within imputation cells whereas the second approach assumes ignorable response but make use of a model on the variable of interest as the basis for inference. In this paper, we propose a third appoach that assumes a specified ignorable response mechanism without having to specify a model on the variable of interest. In this case, we show how to obtain imputed values which lead to estimators of a total that are approximately unbiased under the proposed approach as well as the second approach. Variance estimators of the imputed estimators that are approximately unbiased are also obtained using an approach of Fay (1991) in which the order of sampling and response is reversed. Finally, simulation studies are conducted to investigate the finite sample performance of the methods in terms of bias and mean square error.

Key Words: Bias-adjusted estimator; Deterministic regression imputation; Imputation model approach; Item nonresponse; Nonresponse model approach; Random regression imputation; Variance estimation.

1. Introduction

Item nonresponse occurs in a survey when a sampled element participates in the survey but fails to provide responses on one or more of the survey items (Brick and Kalton 1996). It is usually handled by some form of imputation which involves "filling in" missing values for each item. Imputation may achieve an effective bias reduction, provided suitable auxiliary information is available for all the sampled elements and appropriately incorporated in the imputation model and/or the non-response model.

Imputation offers the following desirable features, among others: (i) it leads to the creation of a complete data file, and (ii) it permits the use of the same survey weights for all items which ensures that the results obtained from different analyses of the completed data set are consistent with one another, unlike the results of analyses from an incomplete data set. However, imputation also presents the following difficulties, among others: (a) marginal imputation for each item distorts the relationship between items, and (b) treating the imputed values as if they were true values may lead to serious underestimation of the variance of imputed estimators, especially when the nonresponse rate is appreciable. Methods that address (a) and (b) have been proposed in the literature.

In this paper, we focus on marginal imputation that is commonly used in many surveys. We first consider deterministic linear regression imputation that includes mean and ratio imputation as special cases. In this method a missing value is replaced by the predicted value obtained by fitting a linear regression model using respondent values and auxiliary variables collected on all the sampled elements. We also consider the case of random linear regression imputation that may be viewed as a deterministic regression imputation plus an added random residual. It includes random hot-deck imputation as a special case.

Let U be a finite population of possibly unknown size N. The objective is to estimate the population total $Y = \sum_U y_i$ of an item y when imputation has been used to compensate for nonresponse on the item values y_i. For brevity, $\sum_A$ will be used for $\sum_{i \in A}$, where $A \subseteq U$. Suppose a probability sample, s, of size n is selected according to a specified design $p(s)$ from U. Under complete response to item y, a design-unbiased estimator of Y is given by the well-known Horvitz-Thompson estimator

$$\hat{Y} = \sum_s w_i y_i, \qquad (1)$$

with sampling (or design) weights $w_i = 1/\pi_i$, where π_i denotes the inclusion probability of population unit i in the sample s, $i = 1, \ldots, N$. Rao (2005) suggested that (1) should be called the Narain-Horvitz-Thompson (NHT) estimator in recognition of the fact that Narain (1951) also discovered (1) independently of Horvitz and Thompson (1952).

In the presence of nonresponse to item y, we use imputation and define an imputed estimator $\hat{Y}_I$ as

$$\hat{Y}_I = \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* = \sum_s w_i \tilde{y}_i, \qquad (2)$$

where y_i^* denotes the value imputed for missing y_i, a_i denotes the response indicator equal to 1 if unit i responds to item y and 0 otherwise and $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$. The

1. David Haziza, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada. K1S 5B6.

imputed estimator (2) can be implemented from the imputed data file containing the survey weights w_i and the $\tilde{y}_i$ only, without the knowledge of response indicators a_i. However, the response indicators will be required for variance estimation. Let $p_i = P(a_i = 1)$ be the item y response probability for unit i. In this paper, we assume that the units respond independently of one another, *i.e.*, $p_{ij} = P(a_i = 1, a_j = 1) = p_i p_j$ if $i \neq j$.

As for any method of compensating for missing data, imputation requires some assumptions about the response mechanism and/or the imputation model. In the presence of imputed data, two different approaches are generally used for making inference on totals, means and other parameters of interest: (i) Imputation model (IM) approach; (ii) Nonresponse Model (NM) approach. Approach (i) is also called model-assisted approach (Särndal 1992) and approach (ii) design-based approach (Shao and Steel 1999). NM approach is based on partitioning the population U into J imputation cells and then imputing nonrespondents y–values within each cell using respondent y–values within the same cell as donor values, independently across the J cells. The following assumption is made:

Assumption NM: Response probability for a given item of interest is constant within imputation cells. That is, $p_i = p_v$, say, where the subscript v denotes the imputation cell.

In the NM approach, explicit assumptions about the response mechanism are made. It follows that inference under assumption NM is with respect to repeated sampling and uniform response mechanism within cells. Approach NM has been studied by Rao (1990, 1996), Rao and Shao (1992), Rao and Sitter (1995) and Shao and Steel (1999), among others. For simplicity, we assume a single imputation cell so that $p_i = p$ under assumption NM.

IM approach is based on the following assumption:

Assumption IM: Item values are missing at random (MAR) in the sense that the response probability does not depend on the item value being imputed but may depend on auxiliary variables used for imputation. Further, a model that generates the item values y_i is assumed.

In the IM approach, explicit assumptions about the distribution of item values y_i is made through a model called the "imputation model". It follows that inference under assumption IM is with respect to repeated sampling and the assumed model that generates the finite population of y–values and nonrespondents to item y. Underlying response mechanism is not specified, except for the MAR assumption, unlike in the NM approach. The assumed response mechanism under assumption IM is much weaker than the uniform response within cells under assumption NM, but inferences under assumption IM depends on the assumed population model. IM approach has been studied by Särndal (1992), Deville and Särndal (1994) and Shao and Steel (1999), among others.

Under linear regression imputation, IM approach assumes the following linear regression imputation model:

$$E_m(y_i) = \mathbf{z}_i'\boldsymbol{\gamma},\ V_m(y_i) = \sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z}_i),$$
$$\mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j, \qquad (3)$$

where $\boldsymbol{\gamma}$ is k–vector of unknown parameters, $\mathbf{z}_i$ is a k–vector of auxiliary variables available for all $i \in s$, $\boldsymbol{\lambda}$ is a k–vector of specified constants, σ^2 is an unknown parameter and E_m, V_m, and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the imputation model. The restriction $\sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z}_i)$ does not severely restrict the range of imputation models.

In this paper, we propose a third approach, called the Generalized Nonresponse Model (GNM) approach. GNM approach is based on the following assumption:

Assumption GNM: Item values are missing at random (MAR) and response probability is specified as a function of auxiliary variables, $\mathbf{u}_i$, observed on all the sample elements, and unknown parameters $\boldsymbol{\eta}$.

In this paper, we assume that the probability of response, p_i, for unit i, is linked to an l–vector of auxiliary variables $\mathbf{u}_i$ according to a logistic model so that

$$p_i = f(\mathbf{u}_i'\boldsymbol{\eta}) = \exp(\mathbf{u}_i'\boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i'\boldsymbol{\eta}), \qquad (4)$$

where $\boldsymbol{\eta}$ is the l–vector of model parameters. Model (4) is the assumed nonresponse model. It can be validated from the values a_i and $\mathbf{u}_i$ for $i \in s$. Note that a_i and $\mathbf{u}_i$ are item specific. Also, note that assumption NM is a special case of assumption GNM. As in NM approach, explicit assumptions about the response mechanism are made and inference under assumption GNM is with respect to repeated sampling and the assumed response mechanism.

Recall that imputation is designed to reduce the nonresponse bias, assuming that the available auxiliary variables can explain the item to be imputed and/or the item response probability. Hence, in practice, the choice of the approach (IM or GNM) should be dictated by the quality of the imputation model and the nonresponse model. The choice between modeling the item response probability and modeling the item of interest will depend on how much reliance one is ready to place on the two models. Although it may seem intuitively more appealing to model the item of interest, there are some cases encountered in practice for which it may be easier to model the response probability (GNM approach). For example, the Capital Expenditures Survey at Statistics Canada produces data on investment made in Canada, in all types of Canadian industries. For this survey, two important variables of interest are capital

expenditures on new construction (CC) and capital expenditures on new machinery and new equipment (CM). In a given year, a large number of businesses have not invested any amount of money on new construction or new machinery. As a result, the sample data file contains a large number of zeros for the two variables CC and CM. In this case, modeling the variables of intrest (CC or CM) may prove to be difficult.

Survey design weights are generally used in linear regression imputation. The resulting imputed estimator of a population total is "robust" in the sense that it is approximately unbiased under either assumption NM or assumption IM. However, the imputed estimator is generally biased under assumption GNM. In this paper, we propose a new method of linear regression imputation that is robust in the sense of leading to approximately unbiased estimators under either assumption GNM or assumption IM.

Section 2 develops a new method of deterministic linear regression imputation as well as random linear regression imputation, and demonstrates the robustness property in estimating a population total *Y*. Results of a simulation study on the finite-sample performance of the imputed estimator under the new method of imputation are reported in section 3. Variance estimators are derived in section 4, using the 'reverse' approach of Fay (1991) in which the order of sampling and response is reversed:

Population → census with nonrespondents → sample with nonrespondents.

Simulation results on variance estimators are also given. Finally, the case of domain means is investigated in section 5.

2. Estimation of a Total

In this section, we study the bias of the imputed estimator $\hat{Y}_I$. The total error, $\hat{Y}_I - Y$, may be decomposed as

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \tag{5}$$

The term $\hat{Y} - Y$ in (5) is called the sampling error, whereas the term $\hat{Y}_I - \hat{Y}$ is called the nonresponse/imputation error. Note that there is no imputation error under deterministic imputation. Since the sampling error does not depend on nonresponse and imputation method, we focus on the nonresponse/imputation error $\hat{Y}_I - \hat{Y}$ and evaluate its properties conditionally on the sample *s*. Under the NM or GNM approach, the conditional nonresponse bias is defined as $E_r(\hat{Y}_I - \hat{Y} \mid s)$, where $E_r(.)$ denotes the expectation with respect to the response mechanism. Under the IM approach, the conditional nonresponse bias is defined as $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$ under MAR assumption.

2.1 Deterministic Regression Imputation

Deterministic regression imputation uses the imputed values

$$y_i^* = \mathbf{z}_\mathbf{i}' \hat{\boldsymbol{\gamma}}_r \tag{6}$$

for missing y_i, where

$$\hat{\boldsymbol{\gamma}}_r = \left(\sum_s w_i a_i \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}) \right)^{-1} \sum_s w_i a_i \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}) \tag{7}$$

is the weighted least squares estimator of $\boldsymbol{\gamma}$ in the model (3), based on the sample elements responding to item *y*. Using (6), the imputed estimator (2) can be written as

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \hat{\boldsymbol{\gamma}}_r, \tag{8}$$

where $\hat{Y}_r = \sum_s w_i a_i y_i$, $\hat{\mathbf{Z}} = \sum_s w_i \mathbf{z}_\mathbf{i}$ and $\hat{\mathbf{Z}}_\mathbf{r} = \sum_s w_i a_i \mathbf{z}_\mathbf{i}$. Note that the imputed estimator (8) is similar to a regression estimator in the case of two-phase sampling.

Under assumption NM, $E_r(a_i \mid s) = p$ and the conditional nonresponse bias, $E_r(\hat{Y}_I - \hat{Y} \mid s)$, is approximately equal to 0. Furthermore, under assumption IM and regression model (3), the conditional nonresponse bias $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$, is equal to 0. However, under assumption GNM, the conditional nonresponse bias is given by

$$E_r(\hat{Y}_I - \hat{Y} \mid s) \approx -\sum_s w_i (1 - p_i)(y_i - \mathbf{z}_\mathbf{i}' \hat{\boldsymbol{\gamma}}_p) \equiv \mathrm{B}(\hat{Y}_I \mid s), \tag{9}$$

where

$$\hat{\boldsymbol{\gamma}}_p = \left(\sum_s w_i p_i \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}) \right)^{-1} \sum_s w_i p_i \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}). \tag{10}$$

This result follows from the fact that under assumption GNM, $E_r(a_i \mid s) = p_i$. Hence, the choice of imputed values (6) is, in general, not suitable under assumption GNM. For the special case of assumption NM with $p_i = p$, the last term in (9) vanishes, noting that $(\sum_s w_i \mathbf{z}_\mathbf{i}')\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}))\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})) = \sum_s w_i y_i$.

2.2 A Bias-Adjusted Estimator

We assume for now that the response probabilities p_i are known. A natural approach for eliminating the bias of $\hat{Y}_I$ under assumption GNM is to consider a bias-adjusted estimator of the form

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I \mid s), \tag{11}$$

where $\hat{B}(\hat{Y}_I \mid s)$ is an estimator of $\mathrm{B}(\hat{Y}_I \mid s)$:

$$\hat{B}(\hat{Y}_I \mid s) = -\sum_s w_i a_i \frac{(1 - p_i)}{p_i}(y_i - \mathbf{z}_\mathbf{i}' \hat{\boldsymbol{\gamma}}_r). \tag{12}$$

Note that $E_r[\hat{B}(\hat{Y}_I \mid s) \mid s] \approx \mathrm{B}(\hat{Y}_I \mid s)$ under assumption GNM. Substituting (12) in (11), we get a bias-adjusted estimator as

$$\hat{Y}_I^a = \sum_s \frac{w_i}{p_i} a_i y_i + \left(\sum_s w_i \mathbf{z}_\mathbf{i}' - \sum_s \frac{w_i}{p_i} a_i \mathbf{z}_\mathbf{i}' \right) \hat{\boldsymbol{\gamma}}_r. \tag{13}$$

Note that (13) is also in the form of a two phase regression estimator.

In practice, response probabilities p_i are unknown. Suppose we can obtain estimators $\hat{p}_i$ of p_i by modelling p_i according to the nonresponse model (4). Then, a bias-adjusted estimator is obtained by replacing p_i in (13) with $\hat{p}_i$. This estimator is also approximately conditionally unbiased under assumption IM. Hence, the bias-adjusted estimator (13) is robust in the sense of validity under either assumption IM or assumption GNM. However, unlike the imputed estimator $\hat{Y}_I$ given by (2), the bias-adjusted estimator $\hat{Y}_I^a$ cannot be computed without the knowledge of the response identifiers, a_i, and the estimated response probabilities, $\hat{p}_i$. Hence, both the response indicators and the estimated response probabilities must be provided with the imputed data file to implement $\hat{Y}_I^a$, which may not be the case in practice. This drawback of $\hat{Y}_I^a$ can be eliminated by using the new imputation method, given in section 2.3, that leads to an approximately unbiased estimator under either assumption GNM or assumption IM without the knowledge of a_i and $\hat{p}_i$ on the imputed data file. However, for variance estimation, access to a_i and $\hat{p}_i$ is needed.

2.3 Modified Deterministic Regression Imputation

We assume for now that the response probabilites p_i are known. We then use the imputed values

$$y_i^* = \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}} \tag{14}$$

for missing y_i and obtain the form of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ that leads to an approximately unbiased estimator under assumption GNM.

2.3.1 Approximately Unbiased Estimator

The following lemma gives the form of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ that leads to an approximately unbiased estimator under assumption GNM.

Lemma 1: *Under assumption GNM, the choice of* $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ *that leads to* $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ *is given by*

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N} = \left[\sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right]^{-1} \sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i}y_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}). \tag{15}$$

Proof: The conditional nonresponse bias of $\hat{Y}_I$ with $y_i^* = \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ under assumption GNM is given by

$$E_r(\hat{Y}_I - \hat{Y} \mid s) = -\sum_s w_i(1-p_i)(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}).$$

Noting that $(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}) = 1$, it follows that $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ if $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ satisfies

$$\boldsymbol{\lambda}'\left[\sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i}(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}})/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right] = 0. \tag{16}$$

The choice $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$ satisfies (16).

Note that $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$ is unknown since the y–values are only observed for $i \in s_r$ and the response probabilities p_i are unknown. An estimator of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$, based on the responding units and estimated response probabilites $\hat{p}_i$, is given by

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}} = \left[\sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right]^{-1} \sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i}y_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}). \tag{17}$$

We have $E_r(\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}} \mid s) \approx \tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$ so that $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ is conditionally approximately unbiased for $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$ under assumption GNM. Hence, using the imputed values

$$y_i^* = \mathbf{z}_\mathbf{i}'\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}} \tag{18}$$

in (2) with $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ given by (17), leads to an approximately unbiased estimator of the total Y under assumption GNM. Note that $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ is a weighted least square estimator of $\boldsymbol{\gamma}$ with respect to a new set of weights, $\tilde{w}_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})$, where $\tilde{w}_i = w_i((1-\hat{p}_i)/\hat{p}_i)$. Hence, the procedure increases the weights w_i for those units with $\hat{p}_i < 1/2$ and decreases the weights for those units with $\hat{p}_i > 1/2$. The imputed estimator can be implemented from the imputed data file containing the sampling weights w_i and the $\tilde{y}_i$ only; response identifiers a_i and estimated response probabilities, $\hat{p}_i$, are not required. However, a_i and $\hat{p}_i$ are needed for variance estimation. Note that the producer of the imputed data file uses the information on a_i and $\mathbf{u}_\mathbf{i}$ to fit the response model (4) and generate the imputed values y_i^* given by (18).

The use of imputed values (18) also leads to an approximately unbiased estimator of Y under assumption IM. First, under the regression model (3), noting that $E_m(y_i \mid s) = \mathbf{z}_\mathbf{i}'\boldsymbol{\gamma}$ and $E_m(\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}} \mid s) = \boldsymbol{\gamma}$, we have $E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ and $E_r E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ without specifying the underlying MAR response mechanism. Hence, the use of imputed values (18) leads to a robust imputed estimator in the sense of validity under both approaches. Finally, it is interesting to note that the imputed values (18) can also be obtained using the method of calibration imputation (Beaumont 2005). Calibration imputation consists of finding final imputed values as close as possible to original imputed values according to some distance function, subject to the calibration constraint.

Two particular cases of modified regression imputation (18) are of interest: (i) modified ratio imputation with $\mathbf{z}_\mathbf{i} = z_i$ and $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = z_i$; (ii) modified mean imputation with $\mathbf{z}_\mathbf{i} = 1$ and $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = 1$. In case (i), the imputed values (18) reduce to

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i z_i} z_i. \tag{19}$$

In case (ii), the imputed values (18) reduce to

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i}. \qquad (20)$$

Under uniform response $p_i = p$, the imputed values (19) and (20) reduce to $(\sum_s w_i a_i y_i / \sum_s w_i a_i z_i) z_i$ and $\bar{y}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i$ respectively, which are the usual values that survey practioners use for ratio and mean imputation (Rao and Sitter 1995).

2.3.2 Optimal Choice of $\tilde{\gamma}_s$

We now turn to the "optimal" choice of $\tilde{\gamma}_s$ by minimizing the conditional mean square error of the imputed estimator $\hat{Y}_I$ with $y_i^* = \mathbf{z}_i' \tilde{\gamma}_s$. The conditional mean square error of the imputed estimator $\hat{Y}_I$ is given by

$$\begin{aligned} \mathrm{MSE}_r(\hat{Y}_I \mid s) &= V_r(\hat{Y}_I \mid s) + [\mathrm{Bias}(\hat{Y}_I \mid s)]^2 \\ &= \sum_s w_i^2 p_i (1-p_i)(y_i - \mathrm{z}_i' \tilde{\gamma}_s)^2 \\ &\quad + \left[\sum_s w_i (1-p_i)(y_i - \mathrm{z}_i' \tilde{\gamma}_s) \right]^2, \qquad (21) \end{aligned}$$

where $V_r(.\mid s)$ denotes the conditional nonresponse variance with respect to the response mechanism, given the sample s. We search for $\tilde{\gamma}_s$ that minimizes $\mathrm{MSE}_r(\hat{Y}_I \mid s)$.

The optimal choice, $\tilde{\gamma}_{\mathbf{opt}}$, of $\tilde{\gamma}_s$ is complex, but in the special case of ratio imputation, $\tilde{\gamma}_{\mathbf{opt}}$ reduces to

$$\tilde{\gamma}_{\mathbf{opt}} = \frac{\sum_s w_i (1-p_i) y_i \sum_s w_i (1-p_i) z_i + \sum_s w_i^2 p_i (1-p_i) y_i z_i}{\left[\sum_s w_i (1-p_i) z_i \right]^2 + \sum_s w_i^2 p_i (1-p_i) z_i^2}. \qquad (22)$$

Assume that the sampling weights w_i satisfy $\max(n / Nw_i) = O(1)$ and that a positive constant C exists such that $C < p_i$. Then,

$$\begin{aligned} \tilde{\gamma}_{\mathbf{opt}} &= \frac{\sum_s w_i (1-p_i) y_i}{\sum_s w_i (1-p_i) z_i} + O\left(\frac{1}{n}\right) \\ &= \tilde{\gamma}_{s,N} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for large sample sizes, the choice $\tilde{\gamma}_{s,N}$ is nearly optimal for ratio imputation. Similarly, $\tilde{\gamma}_{s,N}$ is nearly optimal for mean imputation which is a special case of ratio imputation.

2.4 Random Regression Imputation

Random imputation can be viewed as deterministic imputation plus a random noise. Let s_r and s_m denote the sets of sample respondents and nonrespondents respectively, and let $e_j = (y_j - \mathbf{z}_j' \hat{\gamma}_r) / (\boldsymbol{\lambda}' \mathbf{z}_j)^{1/2}$ be the standardized residuals for the respondents $j \in s_r$ under deterministic regression imputation. Further, $e_i^* = e_j$ with $P(e_i^* = e_j) = w_j / \sum_s w_i a_i$ independently for each $i \in s_m$. Then, random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \hat{\gamma}_r + \epsilon_i^*, i \in s_m$, where $\epsilon_i^* = (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} (e_i^* - \bar{e}_r)$ with $\bar{e}_r = \sum_s w_j a_j e_j / \sum_s w_j a_j$. Let $E_*(.)$ denote the expectation with respect to the random imputation process. We have $E_*(\epsilon_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals (8). Hence, the imputed estimator $\hat{Y}_I$ is approximately unbiased under either assumption NM or assumption IM. It may be noted that random regression imputation covers random (weighted) hot-deck imputation as a special case. To see this, consider the mean imputation model $E_m(y_i) = \gamma, V_m(y_i) = \sigma^2$ and $\mathrm{Cov}_m(y_i, y_j) = 0, i \neq j$. We have $\hat{\gamma}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i = \bar{y}_r$, the weighted mean of the respondent y–values, and $e_j = y_j - \bar{y}_r$. Therefore, $y_i^* = \bar{y}_r + \epsilon_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $w_j / \sum_s w_i a_i$.

The imputed estimator based on random regression imputation is asymptotically biased under assumption GNM. To obtain an approximately unbiased estimator for Y, we propose modified random regression imputation. Let $\tilde{e}_j = (y_j - \mathbf{z}_j' \tilde{\gamma}_r) / (\boldsymbol{\lambda}' \mathbf{z}_j)^{1/2}$ and $\tilde{e}_i^* = \tilde{e}_j$ with $P(\tilde{e}_i^* = \tilde{e}_j) = \tilde{w}_j / \sum_s \tilde{w}_i a_i$ independently for each $i \in s_m$, where $\tilde{\gamma}_r$ is given by (17) and $\tilde{w}_i = w_i (1 - \hat{p}_i) / \hat{p}_i$. Then, modified random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \tilde{\gamma}_r + \tilde{\epsilon}_i^*$, where $\tilde{\epsilon}_i^* = (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} (\tilde{e}_i^* - \tilde{e}_r)$ with $\tilde{e}_r = \sum_s \tilde{w}_j a_j \tilde{e}_j / \sum_s \tilde{w}_j a_j$. We have $E_*(\epsilon_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals the imputed estimator under modified deterministic regression imputation. Hence, the imputed estimator $\hat{Y}_I$ is approximately unbised under either assumption GNM or assumption IM. For the special case of mean imputation model, we have $\tilde{\gamma}_r = \sum_s \tilde{w}_i a_i y_i / \sum_s \tilde{w}_i a_i$ and $y_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $\tilde{w}_j / \sum_s \tilde{w}_i a_i$.

3. Simulation Studies

We performed two simulation studies to investigate the finite sample performance of the proposed deterministic modified regression and modified random regression imputation methods in terms of relative bias and relative root mean square error. The first simulation study compares the performance of the traditional deterministic regression imputation and the proposed modified deterministic regression imputation when the imputation model and/or the nonresponse model are not correctly specified. The second simulation study compares the performance of the imputed estimator obtained by using imputation classes based on the estimated response probabilities and weighted mean imputation (traditional) with the imputed estimator obtained by using the proposed modified deterministic regression imputation method.

3.1 Simulation Study 1

We generated a finite population of size $N = 1{,}000$ containing 3 variables: a variable of interest y and two auxiliary variables z_1 and z_2. To do so, we first generated z_1 and z_2 independently from an exponential distribution with mean 4 and 30 respectively. Then the y–values were generated according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The values of the parameters γ_0, γ_1 and γ_2 were respectively set to 20, 2 and 0.1 and the variance σ^2 was chosen to lead to a model R^2–value approximately equal to 0.75. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 5{,}000$ simple random samples without replacement of size $n = 100$ from the finite population. In each sample, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic regression model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic regression model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_{1i} and a_{2i} were generated independently from a Bernoulli distribution with parameters p_{1i} and p_{2i}, respectively. Note that in the case of the nonresponse mechanism 2, the response mechanism is nonignorable in the sense that the probability of response depends on the variable of interest y.

To compensate for the nonresponse to item y, we used the traditional deterministic regression imputation for which the imputed values are given by (6) and the modified deterministic regression imputation for which the imputed values are given by (18). Imputations were based on the models for y and for p listed in Table 1 as $y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}$ and $p_{(1)}, p_{(2)}, p_{(3)}$. Note that $p_{(1)}$ corresponds to response mechanism 1 and $y_{(1)}$ to the model generating the population.

From each simulated sample, we calculated the imputed estimator $\hat{Y}_I$ given by (2) with the imputed values (6) and (18), based on selected combinations of the models $y_{(a)}$ and $p_{(b)}; a = 1, \ldots, 4; b = 1, 2, 3$. As a measure of the bias of an imputed estimator $\hat{Y}_I$, we used the percent simulated relative bias (RB) given by

$$\mathrm{RB}(\hat{Y}_I) = \frac{\mathrm{Bias}(\hat{Y}_I)}{Y} \times 100, \tag{23}$$

where

$$\mathrm{Bias}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_I^{(r)} - Y \tag{24}$$

and $\hat{Y}_I^{(r)}$ denotes the value of $\hat{Y}_I$ for the r–th simulated sample. As a measure of variability of an imputed estimator $\hat{Y}_I$, we used the percent simulated relative root mean square error (RRMSE) given by

$$\mathrm{RRMSE}(\hat{Y}_I) = \frac{\sqrt{\mathrm{MSE}(\hat{Y}_I)}}{Y} \times 100, \tag{25}$$

where

$$\mathrm{MSE}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_I^{(r)} - Y)^2. \tag{26}$$

Table 1
Models Used for Imputation

Models for y	Intercept	z_1	z_2
$y_{(1)}$	Yes	Yes	Yes
$y_{(2)}$	Yes	No	Yes
$y_{(3)}$	Yes	Yes	No
$y_{(4)}$	No	Yes	Yes
Models for p_i	Intercept	z_1	z_2
$p_{(1)}$	Yes	Yes	No
$p_{(2)}$	Yes	No	Yes
$p_{(3)}$	No	Yes	No

Results on relative bias and RRMSE are shown in Table 2 for the the samples generated by reponse mechanism 1 and in Table 3 for the samples generated by the response mechanism 2. From Table 2, it is clear that, when the imputation is performed according to the correct model (*i.e.*, $y_{(1)}$), traditional deterministic regression imputation leads to an approximately unbiased estimator and it is more efficient than the modified deterministic regression imputation in terms of RRMSE. As noted by a referee, modified deterministic regression imputation can lead to more efficient estimators than traditional deterministic regression. That is, there are scenarios (not considered here) for which the proposed modified deterministic regression imputation method may be more efficient than the traditional deterministic regression imputation method.

When the imputation model is incorrectly specified (*e.g.*, $y_{(2)}$ and $y_{(4)}$), deterministic imputation leads to biased estimators whereas the bias of the modified determinisic imputation is small to negligible, provided the nonresponse model is correctly specified (*i.e.*, $p_{(1)}$). As a result, RRMSE for the deterministic imputation is larger than that for the

modified deterministic regression imputation. When both imputation and nonresponse models are not correctly specified (*e.g.*, $y_{(4)} - p_{(2)}$), all the estimators are biased.

From Table 3, it is clear that, for the case of mechanism 2, the imputed estimator obtained under modified regression imputation performs equally or better than the imputed estimator obtained under traditional regression imputation in all the scenarios. This result is not suprising since achieving an effective bias reduction in the case of nonignorable nonresponse requires the use of all the appropriate auxiliary information available. The auxiliary information used in the case of the proposed modified regression imputation is richer than the one used in the case of regression imputation since it uses the auxiliary variables that are related to both the variable of interest y and the response probability whereas regression imputation uses only the auxiliary variables related to the variable of interest y.

Table 2
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 1

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	0.19	−0.01	1.85	2.33
$y_{(2)} - p_{(1)}$	5.20	0.16	5.60	2.66
$y_{(3)} - p_{(1)}$	0.17	−0.04	1.87	2.37
$y_{(4)} - p_{(1)}$	−14.80	−3.50	15.00	6.70
$y_{(1)} - p_{(2)}$	0.19	0.12	1.85	1.86
$y_{(4)} - p_{(2)}$	−14.80	−14.80	15.00	14.60
$y_{(1)} - p_{(3)}$	0.19	0.05	1.85	1.88

Table 3
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 2

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	1.84	1.83	2.55	2.54
$y_{(2)} - p_{(1)}$	4.46	1.84	4.89	2.65
$y_{(3)} - p_{(1)}$	2.03	2.02	2.70	2.70
$y_{(4)} - p_{(1)}$	−4.58	−3.04	5.07	3.81
$y_{(1)} - p_{(2)}$	1.84	1.84	2.55	2.55
$y_{(4)} - p_{(2)}$	−4.58	−1.70	5.07	2.88
$y_{(1)} - p_{(3)}$	1.84	1.84	2.55	2.55

3.2 Simulation Study 2

We generated a finite population of size $N = 1{,}000$ containing 3 variables: a variable of interest y and three auxiliary variables z_1, z_2 and z_3, by first generating z_1, z_2 and z_3 independently from an exponential distribution with mean 100 and then generating the y–values according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{1i}^2 + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The values of the parameters $\gamma_0, \gamma_1, \gamma_2$ and γ_3 were respectively fixed to 20, 10, 0.5 and 10. The variance σ^2 was chosen to lead to a model R^2 approximately equal to 0.66. The objective is to estimate the population mean $\bar{Y} = \sum_U y_i / N$. In order to focus on the nonresponse/imputation error, we considered the case of a census, *i.e.*, $n = N = 1{,}000$. From the simulated population, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

The values of λ_0, λ_1 and λ_2 were chosen to give an overall response rate approximately equal to 70%. Response indicators a_{1i} and a_{2i} were then genrated independently $R = 1{,}000$ times from a Bernoulli distribution with parameters p_{1i} and p_{2i}, respectively.

To compensate for nonresponse, two strategies were used: The first strategy consisted in dividing the sample, s, into imputation classes $s_1, s_2, \ldots, s_C$ based on the auxiliary variables z_1, z_2 and z_3. To form the classes, we used the score method which may be described as follows: Using the auxiliary information, we first estimated the response probabilites, p_i, to obtain $\hat{p}_i$ for both the respondents and the nonrespondents using logistic regression on z_1, z_2 and z_3. Using the $\hat{p}_i$'s, we then partitioned the population into C classes using the procedure FASTCLUS of SAS (that uses the k–means classification algorithm). The score method leads to a partition of the population in such a way that, within classes, units (respondents and nonrespondents) are homogeneous with respect to $\hat{p}_i$–values. The second strategy used the proposed modified regression imputation method based on the auxiliary variables z_1, z_2 and z_3. The goal of the simulation study is to compare the performances of two imputed estimators of the population mean $\bar{Y}$: (a) Imputed estimator based on the C imputation classes:

$$\bar{y}_I^C = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic}, \qquad (27)$$

where

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c}\left[\sum_{s_c} w_i a_i y_i + \sum_{s_c} w_i (1 - a_i) y_i^*\right],$$

and $\hat{N}_c = \sum_{s_c} w_i$. We used weighted mean imputation within classes; *i.e.*, $y_i^* = \sum_{s_c} w_i a_i y_i / \sum_{s_c} w_i a_i$.

(b) Imputed estimator based on the proposed modified regression imputation, denoted $\bar{y}_I$:

$$\bar{y}_I = \frac{1}{\hat{N}}\left[\sum_s w_i a_i y_i + \sum_s w_i (1-a_i) y_i^*\right], \qquad (28)$$

where the imputed values y_i^* are given by (18) using $\mathbf{z}_\mathbf{i}' = (z_{1i}, z_{2i})'$ and $\hat{N} = \sum_s w_i$. For mechanism 1, the response probabilities p_i were correctly estimated using the variable z_1 and z_3 whereas the variables z_1, z_2 and z_3 were used to estimate p_i for mechanism 2.

Note that $w_i = 1$ in this simulation study for all $i \in U$ because no sampling is involved. Finally, Table 4 compares these estimators in terms of relative bias, given by (23) and RRMSE, given by (25). From Table 4, it is clear that the proposed imputed estimator (28) performs considerably better than the estimator (27) based on imputation classes in terms of RRMSE for both mechanism 1 and mechanism 2.

Table 4
Relative Bias (%) and RRMSE (%) of Imputed Estimators

Imputed estimator*	Number of classes	RB	RRMSE
$\bar{y}_I^C$ (mechanism 1)	1	14.4	14.5
	5	−0.02	4.26
	10	−0.85	7.33
	20	−0.20	8.61
	30	−0.03	8.61
	40	0.03	9.09
	50	0.06	9.44
$\bar{y}_I$ (mechanism 1)	–	1.11	1.90
$\bar{y}_I^C$ (mechanism 2)	1	29.0	29.1
	5	21.4	21.4
	10	21.0	21.1
	20	20.9	21.0
	30	20.9	21.0
	40	21.0	21.0
	50	21.0	21.0
$\bar{y}_I$ (mechanism 2)	–	10.9	10.9

* $\bar{y}_I^C$ given by (27) and $\bar{y}_I$ given by (28).

4. Variance Estimation

In this section, we derive a variance estimator of the imputed estimator $\hat{Y}_I$, using the reverse approach of Fay (1991). The total variance of $\hat{Y}_I$ under a particular deterministic imputation method, is given by

$$V(\hat{Y}_I - Y) = E_r V_p(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p(\hat{Y}_I - Y \mid \mathbf{a}), \qquad (29)$$

where $\mathbf{a} = (a_1, \ldots, a_N)'$ is the vector of response indicators, (Shao and Steel 1999). An estimator of the overall variance $V(\hat{Y}_I - Y)$ in (29) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ conditional on the response indicators a_i, and v_2 is an estimator of $V_r[E_p(\hat{Y}_I - Y \mid \mathbf{a})]$. The estimator v_1 does not depend on the response mechanism or the imputation model, and hence v_1 is valid under either assumption GNM or assumption IM.

Under the corresponding random imputation, the variance of the imputed estimator $\hat{Y}_I$ is given by

$$V(\hat{Y}_I - Y) = E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a}) + E_r E_p V_*(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a}), \qquad (30)$$

where $V_*(.)$ denotes the variance operator with respect to random imputation. We assume that $E_*(\hat{Y}_I \mid \mathbf{a})$ agrees with the imputed estimator for the deterministic case. Hence, $E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ is estimated by v_1 for the deterministic case. Similarly, $V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ is estimated by v_2 for the deterministic case. The additional contribution to variance due to random imputation comes from the component $E_r E_p V_*(\hat{Y}_I - Y \mid \mathbf{a})$, which is estimated by $v_* = V_*(\hat{Y}_I - Y \mid \mathbf{a})$. Hence, it follows from (30) that the overall variance $V(\hat{Y}_I - Y)$ is estimated by $v_t = v_1 + v_* + v_2$. The term v_* is absent for deterministic imputation.

4.1 Known p_i

In this section, we assume that the reponse probabilites p_i are known. We first consider the case of modified deterministic regression imputation in section 4.1.1. The case of modified random regression imputation is studied in section 4.1.2.

4.1.1 Modified Deterministic Regression Imputation

Under modified deterministic regression imputation, the imputed estimator with known p_i may be written as

$$\hat{Y}_{Ip} = \sum_s w_i a_i y_i + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}}, \qquad (31)$$

where

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} = \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]^{-1} \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]. \qquad (32)$$

To obtain v_1, we use standard Taylor linearization which leads to

$$\hat{Y}_{Ip} - Y \approx \sum_s w_i \tilde{\xi}_{ip}, \qquad (33)$$

where

$$\tilde{\xi}_{ip} = a_i y_i + (1-a_i) \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\mathbf{T}}_\mathbf{p}^{-1} a_i \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})} \mathbf{z}_\mathbf{i} (y_i - \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}})$$

with $\tilde{\mathbf{T}}_\mathbf{p} = \sum_s w_i a_i ((1-p_i)/p_i) \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})$. Denoting the variance estimator of the full sample estimator as

$\hat{Y} = \sum_s w_i y_i$ as $v(y)$, it follows from (33) that an estimator of $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ is given by

$$v_1 = v(\tilde{\xi}_p), \tag{34}$$

which is obtained by replacing y_i by $\tilde{\xi}_{ip}$ in the formula for $v(y)$.

To obtain the second component v_2, first note that

$$E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a}) \approx \sum_s a_i y_i + \sum_U (1 - a_i)\boldsymbol{\gamma}_p - Y,$$

where

$$\boldsymbol{\gamma}_p = \left[\sum_U a_i \frac{(1-p_i)}{p_i}\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1} \sum_U a_i \frac{(1-p_i)}{p_i}\mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}).$$

Using Taylor linearization, it can be shown that

$$V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})] \approx \sum_U p_i(1-p_i)\zeta_i^2, \tag{35}$$

where

$$\zeta_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\mathbf{Z} - \mathbf{Z_r})'\mathbf{T_p^{-1}}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\boldsymbol{\gamma}_p)$$

with $\mathbf{Z} = \sum_U \mathbf{z_i}$, $\mathbf{Z_r} = \sum_U a_i\mathbf{z_i}$ and $\mathbf{T_p} = \sum_U a_i((1-p_i)/p_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})$. The component v_2 is then obtained by estimating the unknown quantities in (35), which leads to

$$v_2 = \sum_s w_i a_i (1-p_i)\hat{\zeta}_i^2, \tag{36}$$

where

$$\hat{\zeta}_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})'\tilde{\mathbf{T}}_\mathbf{p}^{-1}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{rp}).$$

An estimator of the total variance v_t is obtained as the sum of (34) and (36): $v_t = v_1 + v_2$. In practice, the response probabilities are unknown. As a result, it is not possible to calculate the variance estimator v_t. A simple solution consists in replacing p_i by the estimated response probabilities $\hat{p}_i$ in (34) and (36) and use the resulting v_t as the variance estimator of $\hat{Y}_I$. As we show in a simulation sudy in section 4.3, this simple method gives acceptable results.

4.1.2 Modified Random Regression Imputation

We first note that

$$V_*(y_i^*) = (\boldsymbol{\lambda}'\mathbf{z_i})\sum_s w_j \frac{(1-p_j)}{p_j} a_i(\tilde{e}_j - \tilde{e}_r)^2 \Big/ \sum_s w_j \frac{(1-p_j)}{p_j} a_j \equiv \tilde{s}_e^2$$

and $\mathrm{Cov}_*(y_i^*, y_j^*) = 0, i \neq j$. Hence, from (2) the component v_*, due to random imputation, is given by

$$v_* = \sum_s w_i^2(1-a_i)V_*(y_i^*) = \sum_s w_i^2(1-a_i)\tilde{s}_e^2. \tag{37}$$

An estimator of the total variance is obtained as the sum of (34), (36) and (37): $v_t = v_1 + v_2 + v_*$. Once again, since the response probabilities p_i are unknown, it is not possible to compute v_* in (37). We propose to replace p_i in (37) by the estimated response probabilities $\hat{p}_i$.

4.2 Unknown p_i

We use Binder's method (Binder 1983) to derive the component v_1 when the response probabilities p_i are estimated. We assume that $p_i = f(\mathbf{u_i}'\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is l–vector of unknown parameters, $\mathbf{u_i}$ is a l–vector of auxiliary variables available for all $i \in s$. For example, in the case of logistic regression, $f(\mathbf{u_i}'\boldsymbol{\eta}) = \exp(\mathbf{u_i}'\boldsymbol{\eta})/\exp(1 + \mathbf{u_i}'\boldsymbol{\eta})$. The estimated response probabilities are given by $\hat{p}_i = f(\mathbf{u_i}'\hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{\theta} = (\boldsymbol{\eta}_N', \boldsymbol{\gamma}_N', Y)'$, where $\boldsymbol{\eta}_N$ and $\boldsymbol{\gamma}_N$ are census parameter corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, respectively. An estimator of $\boldsymbol{\theta}$ given by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}', \tilde{\boldsymbol{\gamma}}_r', \hat{Y}_I)'$ can be expressed as a solution of the sample estimating equations

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = 0,$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta}) = \left(\hat{\mathbf{S}}_1(\boldsymbol{\theta}), \hat{\mathbf{S}}_2(\boldsymbol{\theta}), \hat{S}_3(\boldsymbol{\theta})\right)'$ with

$$\hat{\mathbf{S}}_\mathbf{1}(\boldsymbol{\theta}) = \sum_s w_i \mathbf{u_i}[a_i - f(\mathbf{u_i}'\boldsymbol{\eta}_N)] = \mathbf{0},$$

$$\hat{\mathbf{S}}_\mathbf{2}(\boldsymbol{\theta}) = \sum_s w_i a_i \mathbf{z_i}\frac{(1 - f(\mathbf{u_i}'\boldsymbol{\eta}_N))}{f(\mathbf{u_i}'\boldsymbol{\eta}_N)}(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N)/(\boldsymbol{\lambda}'\mathbf{z_i}) = \mathbf{0}$$

and

$$\hat{S}_3(\boldsymbol{\theta}) = Y - \sum_s w_i \mathbf{z_i}'\boldsymbol{\gamma}_N - \sum_s w_i a_i(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N) = 0.$$

Let $\hat{\mathbf{J}}(\boldsymbol{\theta}) = (\partial\hat{\mathbf{S}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ be the $(k+l+1)\times(k+l+1)$ matrix of partial derivative. We have

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]\boldsymbol{\Sigma}(\boldsymbol{\theta})[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]',$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the $(k+l+1)\times(k+l+1)$ symetric matrix whose ij element is the covariance between $\hat{S}_i(\boldsymbol{\theta})$ and $\hat{S}_j(\boldsymbol{\theta})$ with respect to sampling given the vector of response indicator $\mathbf{a}$. If $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is replaced by a consistent estimator $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, say, we obtain a consistent variance estimator $\mathbf{v}(\hat{\boldsymbol{\theta}})$ given by

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})[\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]'.$$

Since we are interested in the variance estimator, v_1, of $\hat{Y}_I$, we need the final row, $\mathbf{b}$, say, of $\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. It follows that

$$v_1 = \mathbf{b}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{b}'. \tag{38}$$

To obtain the component v_2, we assume that the sampling weights w_i satisfy $\max(n/Nw_i) = O(1)$ and that there exists a positive constant C such that $C < p_i$. Furthermore, we assume that $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p(n^{-1/2})$. By Taylor linearization, we have

$$\hat{Y}_I = \hat{Y}_{Ip} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\sum_s p_i^{-1}(y_i - \tilde{\boldsymbol{\gamma}}_a)\partial\frac{f(\mathbf{u_i}'\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} + O_p(N/n),$$

where

$$\tilde{\gamma}_a = \left[\sum_U (1-a_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1}\left[\sum_U (1-a_i)\mathbf{z_i}y_i/(\boldsymbol{\lambda}'\mathbf{z_i})\right].$$

Assuming that $f(\mathbf{u_i}'\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ is uniformly bounded, we have

$$E_p(\hat{Y}_I) = E_p(\hat{Y}_{Ip}) + O_p(N/n^{1/2}).$$

Hence, the component $V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})]$ is approximately given by (35) and v_2 is given by (36) with p_i replaced by $\hat{p}_i$. In the case of modified random regression imputation, the component due to random imputation will be estimated by (37) with p_i replaced by $\hat{p}_i$.

4.3 Simulation Study

We performed a limited simulation study to assess the performance of the variance estimators considered in sections 4.1 and 4.2. We generated a population of size $N = 2{,}500$ containing two variables y and z. First, the variable z was generated from a Gamma distribution with scale parameter equal to 4 and shape parameter equal to 10. The y–values were then generated according to the ratio model

$$y_i = \gamma z_i + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The value of the parameter γ was set to 2 and the variance σ^2 was chosen to lead to a model R^2–value approximately equal to 0.81. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 10{,}000$ simple random samples without replacement from the finite population using the following sampling fractions n/N: 0.05; 0.1 and 0.25. In each sample, nonresponse to item y was generated according to the following response mechanism: Response probability p_i for unit i is given by the logistic model

$$\log\frac{p_i}{1-p_i} = \lambda_0 + \lambda_1 z_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_i were then generated independently from a Bernoulli distribution with parameters p_i.

To compensate for the nonresponse to item y, we used the modified deterministic ratio imputation for which the imputed values are given by (19). From each simulated sample, we calculated the imputed estimator $\hat{Y}_I$ given by (2) with the imputed values (19). As a measure of the bias of a variance estimator v, we used the relative bias $[E(v) - \mathrm{MSE}(\hat{Y}_I)]/\mathrm{MSE}(\hat{Y}_I)$. Let v_{naive} denotes the total variance estimator obtained by summing (34) and (36) when the response probabilties p_i are replaced by the estimated response probabilities $\hat{p}_i$ and v_{correct} denotes the total variance estimator obtained by summing (38) and (36) with p_i replaced by $\hat{p}_i$. Table 5 gives the relative bias (in %) of the two variance estimators. It is clear from Table 5 that both variance estimators lead to underestimation, but v_{correct} is slightly better in terms of underestimation. Also, both variance estimators performed well with a relative bias less than –10%. Hence, the simpler variance estimator v_{naive} might be suitable in practice.

Table 5
Relative Bias (%) of the Variance Estimators

f	RB(v_{naive})	RB(v_{correct})
0.05	–6.3	–5.1
0.10	–5.8	–4.1
0.25	–4.3	–3.2

5. Estimation of Domain Means

In practice, estimates for various domains (subpopulations) are often needed. For example, in the Canadian Labour Force Survey, estimates of unemployment are required by age-sex group and by industry at the provincial level. To compensate for item nonresponse, the proposed modified regression imputation may be used. However, the domains must be specified in advance at the imputation stage. In other words, the domain indicators must be part of the imputation model. In practice, domains are generally not specified at the edit and imputation stage and domain estimates are obtained from imputed data based on imputation models without the domain indicators. As a result, the imputed estimators for domains are generally biased. We propose a bias-adjusted estimator, along the lines of section 2.2, to remedy this problem. The bias-adjusted estimator can be obtained at the estimation stage and does not require the specification of the domains at the imputation stage.

A vector of domain means may be expressed as

$$\overline{\mathbf{Y}}_{(\mathbf{d})} = \left(\sum_U \mathbf{x_i}\mathbf{x_i}'\right)^{-1}\sum_U \mathbf{x_i}y_i, \tag{39}$$

where $\mathbf{x} = (x_{1i}, \ldots, x_{di}, \ldots, x_{Di})'$ is a vector of domain indicators, x_{di}, such that $x_{di} = 1$ if $i \in$ domain d and $x_{di} = 0$, otherwise. We assume that $\mathbf{x}$ is known for all the units $i \in s$. In other words, only item y may be missing. In the absence of nonresponse, an approximately unbiased estimator of $\overline{\mathbf{Y}}_{(d)}$ is given by

$$\hat{\overline{\mathbf{Y}}}_{(\mathbf{d})} = \left(\sum_s w_i\mathbf{x_i}\mathbf{x_i}'\right)^{-1}\sum_s w_i\mathbf{x_i}y_i. \tag{40}$$

In the presence of nonresponse to item y, an imputed estimator of $\overline{\mathbf{Y}}_{(d)}$ is given by

$$\begin{aligned}\hat{\overline{\mathbf{Y}}}_{\mathbf{I(d)}} &= \hat{\mathbf{T}}^{-1}\left[\sum_s w_i a_i\mathbf{x_i}y_i + \sum_s w_i(1-a_i)\mathbf{x_i}y_i^*\right] \\ &= \hat{\mathbf{T}}^{-1}\sum_s w_i a_i\mathbf{x_i}\tilde{y}_i,\end{aligned} \tag{41}$$

where $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x_i}\mathbf{x_i}'$. Note that the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ in (41) does not require the response identifiers, a_i. Haziza and Rao (2005) showed that the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ is biased under assumption NM. They proposed a bias-adjusted estimator which is approximately unbiased under either assumption NM or assumption IM. In this section, we propose an extension of the Haziza-Rao bias-adjusted estimator which is approximately unbiased under either assumption GNM or assumption IM.

It is easily seen that, under assumption GNM, the conditional nonresponse bias of the imputed estimator (41) that uses the modified deterministic regression imputation (18) is given by

$$\text{Bias}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s w_i(1-p_i)\mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{s,N})\right], \quad (42)$$

where $\tilde{\boldsymbol{\gamma}}_{s,N}$ is given by (15). An approximately conditionally unbiased estimator of the bias in (42) is given by

$$\hat{B}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s \tilde{w}_i a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}})\right], \quad (43)$$

where $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ is given by (17). A bias-adjusted estimator, $\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}}$, is then obtained as $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} - \hat{B}(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s)$, which leads to

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{\mathbf{T}}^{-1}\left[\sum_s \frac{w_i}{\hat{p}_i} a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}) + \sum_s w_i \mathbf{x_i}\mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}\right]. \quad (44)$$

The bias-adjusted estimator (44) is approximately unbiased under either IM or GNM. Hence, it is robust in the sense of validity under both assumption IM or assumption GNM. However, it requires both the response identifiers a_i and the estimated response probabilities $\hat{p}_i$, unlike the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ in (41).

It is possible to obtain a bias-adjusted estimator of the form (44) if we use the traditional deterministic regression imputation instead. It is interesting to note that the bias-adjusted estimator is identical to the estimator obtained using calibrated imputation (Beaumont 2005). The latter estimator does not require the knowledge of a_i and $\hat{p}_i$ in the imputed data file but the domains must be specified at the imputation stage, which may not be feasible in practice.

If the nonresponse model (4) contains only the intercept, we have $\hat{p}_i = \hat{p}$, where $\hat{p}$ denotes the overall response rate. In this case, the bias-adjusted estimator (44) reduces to

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{p}^{-1}\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} + (1-\hat{p}^{-1})\hat{\mathbf{T}}^{-1}\sum_s w_i \mathbf{x_i}\mathbf{z_i}'\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}, \quad (45)$$

noting that $\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}} = \hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}$, where, under deterministic regression imputation,

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}} = \left(\sum_{i\in s} w_i \mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right)^{-1} \times \left[\sum_{i\in s} w_i a_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}) + \sum_{i\in s} w_i(1-a_i)\mathbf{z_i} y_i^*/(\boldsymbol{\lambda}'\mathbf{z_i})\right] = \hat{\boldsymbol{\gamma}}_{\boldsymbol{r}}.$$

Haziza and Rao (2005) obtained the bias-adjusted estimator (45).

Concluding Remarks

For simplicity, we focussed on a single imputation class but our GNM method readily extends to multiple imputation classes by using separate imputations across classes. For example, we could use weighted mean imputation within classes using our modified weights $\tilde{w}_i$. Also, our method can be extended to the case of composite imputation (Sitter and Rao 1997; Shao and Steel 1999) which uses different imputations for missing item values depending on the auxiliary information available. For example, ratio imputation is used when an auxiliary variable *x* is observed and some other imputation when x is not observed. In this case, the IM approach based on the ratio model relating y to x will not be applicable unlike in the case where *x* is observed on all the sampled units.

Acknowledgments

J.N.K. Rao's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank the reviewers for useful comments and suggestions.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 15, 279-292.

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Deville, J.C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

Haziza, D., and Rao, J.N.K. (2005). Inference for domains under imputation for missing survey data. *Canadian Journal of Statistics*, 33, 149-161.

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

Narain, R.D. (1951). On sampling without replacement with variying probabiltities. *Journal of the Indian Society of Agricultural Statistics*, 2, 169-174.

Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association*, 91, 499-506.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 117-138.

Rao, J.N.K., and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Särndal, C.-E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., and Steel, P. (1999). Variance Estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 65-76
Statistics Canada, Catalogue No. 12-001-XIE

A Model for Estimating and Imputing Nonrespondent Census Households under Sampling for Nonresponse Follow-up

Elaine L. Zanutto and Alan M. Zaslavsky [1]

Abstract

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000. Sampling for NRFU involves sending field enumerators to only a sample of the housing units that did not respond to the initial mailed questionnaire, thereby reducing costs but creating a major small-area estimation problem. We propose a model to impute the characteristics of the housing units that did not respond to the mailed questionnaire, to benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution. We evaluate the performance of our loglinear model through simulation. Results show that, when compared to estimates from alternative models, our loglinear model produces estimates with much smaller MSE in many cases and estimates with approximately the same size MSE in most other cases. Although sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas.

Key Words: Missing data; Small area estimation; Iterative proportional fitting; Log-linear models; ECM.

1. Introduction

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000 (U.S. Bureau of the Census 1997a, b). Under current procedures used in 99% of households, the Census Bureau first mails or personally delivers a questionnaire, to be returned by mail. Then field enumerators attempt to contact all mail nonrespondents (about 35% of those mailed). The workload of about 42 million households makes this one of the most expensive census operations.

Sampling for NRFU involves sending field enumerators to only a sample of the nonresponding housing units. This sample is either an unclustered element sample of nonresponding housing units (the "unit sample") or a cluster sample consisting of all nonresponding units in a sample of the census blocks (small areas approximating a city block or some compact rural area, averaging about 15 housing units). This second stage of followup leads to the completion of a questionnaire (through proxy response or imputation, if necessary) for all sample housing units, except those that are resolved to be vacant.

The potential cost savings of sampling are large, but it would require estimating the characteristics of a huge number of nonsampled nonresponding households, posing a major small-area estimation problem (Ghosh and Rao 1994; Rao 2003). We show that using appropriate models to impute the characteristics of the nonsample nonrespondent households, we may benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution.

Although, for complex legal reasons, sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of

1. Elaine L. Zanutto, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, U.S.A. E-mail: zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A. E-mail: zaslavsky@hcp.med.harvard.edu.

respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas. The related methodologies of Purcell and Kish (1980) and Zhang and Chambers (2004) also use loglinear models to estimate small-area cross-classified counts assuming that the total populations are known and that auxiliary cross-classified data is available at the small area level. We have an additional source of information, specifically the characteristics of the nonrespondents in the NRFU sample. This allows us to model the relationship between respondents and nonrespondents directly in some blocks.

Section 2 summarizes proposed strategies for imputing missing data in this situation. Section 3.1 describes our general sampling and estimation procedure. We present our estimation and imputation model in Section 3.2, our smoothing and estimation procedures in Section 3.3, and evaluate our model by simulation in Section 4. Methods for MSE estimation are summarized in Section 5, and Section 6 presents conclusions.

2. Previous Proposals for Imputing Census Nonrespondents

Several methods have been proposed for imputing the characteristics of nonresponding housing units. "Top-down" strategies first estimate counts for aggregates of households and then allocate them to small areas in a manner that maintains consistency with the aggregates. Simple ratio models (Fuller, Isaki and Tsay 1994, henceforth "FIT"), Poisson regression models (Bell and Otto 1994), or more complex loglinear models (as we propose here and in Zanutto and Zaslavsky 1995b, a) are used to estimate counts for small areas and detailed demographic groups for which direct estimates are not possible. Like us, FIT classify households into a modest number of types defined by important characteristics (*e.g.*, number of people, race, tenure) and then estimate the number of households of each type among nonsample nonrespondents. A complete census roster is then generated by imputing the estimated number of households of each type. The main difference between our approach and that of FIT is that by using a loglinear model rather than a stratified ratio model, we obtain more flexibility in the detail of constraints imposed at various levels of geography. Bell and Otto (1994) estimate the number of people over 18 years old of each race (Hispanic, non-Hispanic Black, Other) in each nonsample nonrespondent housing unit but do not consider how to group imputed persons into households or how to impute household-level characteristics such as tenure. These *ad hoc* "top-down" models incorporate at most a few household characteristics and hence do not explicitly model household structure, but they are designed to maintain the consistency of the aggregates that are considered most important.

Schafer (1995) develops a "bottom-up" strategy in which households are built up from individual persons and their characteristics and relationships, each of which must be described by its own model. These models describe the population in more detail and can support full probability (*e.g.*, Bayesian) inferences about unobserved characteristics. However, this approach, unlike the other, requires that a fairly complex set of models be built before any imputations can be made. Furthermore, in this framework it is more difficult to maintain consistency between microdata and aggregate controls. A combined strategy, however, could use our models to produce nearly unbiased estimates by household types and Schafer's models to complete the imputations.

3. Estimation Procedures and Models

3.1 Overview

In the first step of the imputation procedure, counts of the number of nonsample nonrespondent households of each type are predicted using a combination of logistic and loglinear models for each block. This step is the topic of this paper (and of FIT).

For modeling we classified households into types based on a few important characteristics. Here we use 19 types, one of which is "vacant." The remaining 18 are defined by the cross-classification of households by three size categories (1–2 people, 3–4 people, 5 or more people), three race categories (Hispanic, non-Hispanic Black, Other), and two tenure categories (owner, renter).

To predict the number of vacant housing units among nonsample nonresponding units in each block we (and FIT) fitted a logistic regression model, recognizing that the relationship between respondent and nonrespondent households is different for vacant than for nonvacant housing units. Respondent vacants are simply those that were identified as vacant by a postal service letter carrier, leading to mail return of the original questionnaire. Their distribution is likely to depend largely on housing characteristics related to postal delivery, telling us little about the distribution of nonrespondent vacants.

After modeling vacancies, we fitted a loglinear model to predict the distribution of the nonvacant household types in the remaining nonsample nonrespondent households at three geographical levels. The block is the smallest unit and the one for which estimated counts are calculated. The "estimation domain" is the largest unit and is the area in which estimation is conducted independent of other such domains; in our application to the 1990 census, this is the area for which the census was administered from one of 449 district field offices (DO) representing about 200,000

households on average. Finally, we call an intermediate level of geography an "area", comprising a relatively homogeneous collection of contiguous blocks within an estimation domain. In standard Census Bureau geography these might be census tracts, block groups, or Address Register Areas.

We lay out briefly the remaining steps that would be followed to obtain census products using the estimates. In the second step of the imputation procedure the predicted counts would be rounded to integers. Unbiased schemes (*i.e.*, stochastic procedures that in expectation impute the predicted number of units in each cell) for "controlled rounding" (*i.e.*, rounding in a two-way table while preserving marginal totals) were developed by Cox (1987) and George and Penny (1987). However, more research is needed to determine if these methods can be modified to round households counts while preserving all the margins corresponding to effects in the loglinear model. This is an active research topic due to its importance to statistical nondisclosure.

Finally, detailed person and household information would be imputed for nonrespondent households by substituting donor households with similar characteristics. Donors can be chosen from the sampled nonrespondents, the respondents, or a combination of both sources. Finally, tabulations and microdata samples would be prepared from the completed rosters.

3.2 Loglinear Model

We fitted a loglinear model to estimate the prevalence of the various types of households among nonsample nonrespondent households in a DO, using data from the respondents and from the nonrespondents in the NRFU sample for that DO. The model predicts household types for nonsample nonrespondent households in each block by using information about the characteristics of respondent households in the same block and the characteristics of nonrespondent households, measured by the NRFU sample, in surrounding blocks. To accomplish this, the loglinear model contains interactions among the household characteristics that define household type and response status at various levels of geography.

This modeling strategy is motivated by the fact that when a hierarchical loglinear model (*i.e.*, one in which for every included interaction effect, all main effects or interactions marginal to it are also included) is fitted by maximum likelihood, the fitted values for every margin or mean corresponding to an effect in the model are equal to the corresponding observed margins or means (Birch 1963). Therefore, predictions for household types agree with observed rates for the characteristics included in the model, at the levels of geography and response status corresponding to the interactions included in the model. Also, because model predictions for the included effects are constrained to agree with observed rates based on a probability sample (the NRFU sample), the corresponding estimates are consistent and approximately unbiased. (Exact unbiasedness is not obtained because of the nonlinearity of the prediction model and because the number of nonsample nonrespondent households in a block might be associated with some characteristics of the nonresponding households in the block.)

The loglinear model includes nested geographical factors for blocks and areas. It also includes crossed factors representing the demographic characteristics of households: first-stage response indicator (respondent or nonrespondent household), household type index, and model expressions in the variables that define household types. These model expressions are submodels of the fully interacted model which defines household type (*i.e.*, race × size × tenure).

We use the following notation:

i = block index ($i = 1, \ldots,$ number of blocks in the DO),

j = index of household type ($j = 1, \ldots,$ number of types),

r = first-stage (mail) response indicator, $r = 0$ for nonresponding households and $r = 1$ for respondents,

$a = a(i)$ = index for the area containing block i ($a = 1, \ldots,$ number of areas),

$x_k = x_k(j)$, $k = 1, 2, 3, 4$ = model expressions in the variables that define household types where x_1 represents the full cross-classification defining household types, x_2 and x_3 are model expressions which are marginal to x_1, and x_4 is a model expression which is marginal to x_3. (This terminology is explained below.)

We assume a loglinear model of the following form:

$$n_{ijr} \sim \text{Poisson}(m_{ijr}),\ \log(m_{ijr}) = z_{ijr}^T \beta \qquad (1)$$

where n_{ijr} and m_{ijr} are respectively the observed and expected counts for block i, household type j and response status r, and Z is the design matrix corresponding to the following model formula:

$$x_1 + i * x_2 + i * r + r * x_3 + r * a * x_4. \qquad (2)$$

In the standard generalized linear models notation of Wilkinson and Rogers (1973), the "*" operator indicates that the main effects and all interactions that are marginal to the given interaction are included in the model, so that this model contains main effects for model expression x_1, response indicator r, and block indices i and the interactions $i * x_2$, $i * r$, $r * x_3$, and $r * a * x_4$.

Because, in (1), x_4 interacts with area, the smallest level of aggregation for the non-respondent data, it should represent a fairly coarse classification of households including only those household characteristics that are most important to impute accurately at the area level. The x_3 expression may include terms not included in x_4, since it is fitted at a higher level of geography where there is more data available. Similarly, the x_1 expression might include the most interactions, including the interaction of all variables that define household type, since it is fitted at the largest level of geography, using all available data. Finally, x_2, which can be different than x_3 since it interacts with i instead of r, should be less detailed than x_1 since it interacts with block, a much smaller level of geography. These guidelines are motivated by the fact that estimates of interactions with i, r, or a are determined by relatively few observations and should be kept simple. Choosing x_2, x_3, and x_4 as described above should improve the precision of model estimates while preserving the most important margins.

As an example of possible $x_1, \ldots, x_4$ terms, suppose that we define household type by a race × size × tenure cross-classification. Then one possible specification of x_1, x_2 and x_3 is $x_1 =$ race $*$ size $*$ tenure, $x_2 =$ race $*$ size + tenure, $x_3 =$ size $*$ tenure, and $x_4 =$ race + size + tenure. Allowing the $x_1, \ldots, x_4$ terms to be model expressions, rather than just simple interactions, gives us a concise way to represent a model containing all the desired interactions. For example, a model containing an $i * x_2$ term, where x_2 is specified above, includes both a block × race × size interaction and a block × tenure interaction.

A heuristic interpretation of our loglinear model is that we estimate the detailed distribution of household types across the whole area (x_1) and then shift that distribution to allow for the general characteristics of the block (x_2), the general differences between responding and nonresponding households (x_3), and the most important differences between responding and nonresponding households in the particular area (x_4). All interactions could be included except those of the form $r * i * x$, where x represents a model expression in the variables that define household type (*i.e.*, such as x_1, x_2, x_3, or x_4). Interactions of this form depend on the margins determined only by non-respondent households in a single block and these are unavailable in nonsample blocks under the block sample design, and based on a very small sample under the unit sampling design. Therefore our model specification excludes all $r * i * x$ effects, which are always inestimable (or poorly estimated, in the household sample design). This model generalizes two simple theories which are contained as submodels. First, if there are no differences between blocks (*i.e.*, the loglinear $i * x_2$ and $a * x_4$ interactions are zero) then nonrespondent households in each block are imputed according to the overall proportion of nonrespondent households in each of the x_3 categories in the NRFU sample, through the $r * x_3$ effect. In other words, the imputations are made using the same proportions in each block. Second, if there are no differences between respondents and nonrespondents (*i.e.*, no $r * x_3$ or $r * x_4$ interactions) then nonrespondents are imputed in the same proportions as observed in the respondents in each block.

Our general model formulation can accommodate many definitions of area and household type and choices of model expressions. Areas should be defined to be large enough to contain adequate data to estimate the corresponding interactions, but also relatively homogeneous. For example, areas could be defined by a combination of geographical contiguity and stratification by block-level covariates (such as percent minority), in order to obtain more homogeneous areas whose differences could be described by modeling. Generalization to more than two levels of geography within the estimation domain is also straightforward. Thus, for example, we could interact another model expression x_5 with a geographical unit intermediate between the area and the block.

Fitting the model by maximum likelihood, the following quantities are made equal to the corresponding observed values: (1) fitted block counts (through the main effect for block, i), (2) response rates by block (through the $r * i$ term), (3) household characteristic means overall (for x_1 characteristics through the main effect term for x_1) and (4) by block (for x_2 characteristics through the $i * x_2$ term), and (5) household characteristic means for nonrespondents overall (for x_3 characteristics, through the $r * x_3$ term) and (6) for nonrespondents by area (for x_4 characteristics, through the $r * a * x_4$ term). Thus, this model generalizes the model used by FIT of block × type independence, yielding unbiasedness at smaller levels of aggregation, assuming that the margins and averages are estimated unbiasedly from the data. The estimate for area is not exactly the same as the usual unbiased estimate obtained by direct estimation from the NRFU sample because the model makes observed and fitted margins agree for the households in sample. In effect, there is covariance (regression) adjustment that shifts the aggregate to account for observed differences between respondent households in sample blocks and respondent households in nonsample blocks, or in the unit sampling design, between respondent households in blocks with households in the NRFU sample and blocks without households in the NRFU sample.

The idea of modeling household characteristics using low-dimensional covariates at the block level and in more detail at more aggregated levels is similar in concept, although not in details, to the model described in Zaslavsky

(2004). For use of loglinear weights to match sample estimates of aggregates, see Brackstone and Rao (1976), Oh and Scheuren (1983), and Zaslavsky (1988).

3.3 Estimation and Smoothing

We fit the model by maximum likelihood estimation under the Poisson sampling model, which is equivalent to fitting a multinomial logistic regression model. The fitting is complicated by the fact that the data do not form a complete block × response × type table because we have counts by block, but not characteristics for nonsample nonresponding households. In the block sampling design we lack characteristics of all nonrespondents in some blocks and in the unit sampling design we lack characteristics of some nonrespondents in almost all blocks. To fit the model we use a modified iterative proportional fitting (IPF) algorithm adapted to data that are partially classified in a part of the dataset (Appendix).

With some data sets, some parameters may be inestimable because the maximum likelihood estimates lie on the boundary of the parameter space (infinite on the loglinear scale, indicating a zero on the count scale) or because there is no information for the parameter. Tailoring the model specification in each estimation domain to remove inestimable parameters is impractical in a census production setting.

By introducing a small amount of prior information, estimability of all parameters can be guaranteed. To do this, we append a small amount of "pseudo-data" to the data for each area, whose proportions by type are equal to those for some surrounding area (the DO, in our simulations), by adding these counts to the data table before fitting the model. This implements an empirical Bayes analysis for multinomial data with distribution $f(n_1, ..., n_H \mid p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{n_i}$, where $n_1, ..., n_H$ are the observed number of households of each type in a block or area. If $\{p_i\}$ have a joint Dirichlet prior distribution, $f(p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{\alpha_i - 1}$, $\alpha_i \geq 0$, the resulting posterior distribution for the p_i's is Dirichlet with parameters $\alpha_i + x_i$ (Gelman, Carlin, Stern and Rubin 1995, page 76) and posterior mode proportional to the parameters. Thus, this empirical Bayes procedure is equivalent to adding $\sum \alpha_i$ households to the area, where α_i of these households are of the i^{th} type. We fix the α_i's to be proportional to the observed proportions of each household type in some surrounding area, so the observed distribution of household types is smoothed by mixing it with the distribution for a surrounding area, thus avoiding introducing bias at the level of the larger area. This prior specification induces a prior on the parameters of the loglinear model. See Rubin and Schenker (1987), Zaslavsky (1988), and example and historical references in Clogg, Rubin, Schenker, Schultz and Weidman (1991) for similar use of smoothing.

After estimating the model parameters, the next step is to calculate predicted counts for each household type for the nonrespondent households that are not in the NRFU sample. Using the IPF algorithm, the predictions for the nonsample nonrespondent households are obtained automatically by applying the same fitting proportions to the partially observed part of the table as to the fully observed part of the table, so no further calculation is required (Appendix).

4. Simulations

4.1 Overview

Our simulation study evaluated the bias, variance and MSE of the estimates of estimated demographic aggregates (such as the number of households by race, size and tenure) at various levels of geography, using estimated household compositions for non-respondent households that are not in the NRFU sample. Analytic evaluations are infeasible, given the complexity of the models and sampling scheme, the dependence of the performance of the model on the actual geographical distribution of household types, and the number of variations of the model that could be examined.

We used block-level data from three DOs from the 1990 U.S. Decennial Census; these constituted our estimation areas. The simulations are similar in structure to those described by Schindler (1993) or FIT.

The steps of the simulation are as follows:

1. Blocks or nonrespondent housing units are sampled according to the NRFU sampling scheme.
2. A logistic regression model for vacant households is fitted to the respondent households and the sampled nonrespondent households.
3. The predicted number of nonrespondent households that are vacant is calculated for each block.
4. A model for nonvacant types is fitted using the respondent households and the sampled nonrespondent households.
5. The predicted number of nonsample nonrespondent households of each nonvacant type are calculated for each block.
6. Aggregates of interest are calculated based on the predicted counts, and compared to the truth using loss functions.

In our simulations, repeating these steps 30 times yielded estimates of RMSE (defined in section 4.3) with adequate accuracy to evaluate the performance of our model relative to the alternative models. Specifically, the estimated coefficients of variation of the estimated differences in RMSE for the stratified ratio method (described below) and

loglinear method are less than 0.05, except when the difference between estimated RMSEs is very small, resulting in a large coefficient of variation.

The performance of our proposed model is compared with two alternative estimation methods, under both the unit and block sampling designs. Each method first fits a logistic regression model to estimate the number of nonrespondent households that are vacant in each block. The first alternative, the "unstratified ratio method", imputes households for nonsample nonrespondent households in each block in proportion to the distribution of household types among nonrespondent households in the follow-up sample for the entire DO. The second alternative, the "stratified ratio method", is a version of that in FIT. We first form strata of approximately 82 blocks based on the racial composition of the blocks, as described by FIT. (We use both respondent and nonrespondent data to form strata, assuming, as in FIT, that similar information would be available from administrative records. Stratification based only on respondent information yielded similar results.) Then, in each stratum, nonsample nonrespondent households are imputed to nonvacant types in proportion to the frequency of the type in the follow-up sample for that stratum.

We simulate each estimation method using a NRFU sampling rate of 30%. In each stratum, we simulate NRFU sampling by selecting a 30% simple random sample of blocks for the block sampling design, and a 30% simple random sample of nonrespondent households in each stratum for the unit sampling design. The characteristics of the nonrespondent households in these samples is assumed to be known (*i.e.*, as a result of follow-up operations). For both our loglinear model method and the stratified ratio method, we select a 30% sample of blocks or nonrespondent households using simple random sampling without replacement from each area.

We considered several loglinear model formulations. The best model for both the block and unit sampling designs, by the criteria described in Section 4.3, uses $x_1 = \text{size} * \text{race} * \text{tenure}$, $x_2 = \text{race} * \text{tenure} + \text{size}$, $x_3 = \text{race} * \text{size}$, $x_4 = \text{tenure}$. This model is used in the simulations.

To ensure the model can be fitted in every case and to speed the convergence of the IPF, we smooth the data by adding one hypothetical respondent household ("pseudo-data") to each block. This household is divided among the 18 nonvacant household types according to the overall DO proportions of respondent households. Estimates using 5 households for smoothing were about as accurate as with one, and more aggressive smoothing (adding 10, 15, 20 or 25 households per block) slightly increases errors in the estimates. Also, although adding only a small fraction of a household to each block is sufficient to ensure that the model can be fitted in every case, using less than 1 household per block drastically slowed convergence and slightly increased the error in the estimates.

The three estimation procedures used the same logistic regression model for vacancies. The covariates for each block are the mail nonresponse rate, the percentages of respondent households that are (separately) renters, apartment dwellers, and of a minority race (either Black or Hispanic), the average value of owner-occupied homes, the average monthly rent for rental units, indicator variables for each of the areas, and interactions between percentage of respondent renters and average monthly rent, percentage of respondent renters and average monthly rent squared (mean-centered), percentage of respondent owners and average home values, and percentage of respondent owners and average home values squared (mean-centered). To avoid computational problems arising from blocks with no nonrespondent vacant households, one hypothetical non-respondent household is added to each block divided between vacant and nonvacant according to their proportions in the sampled nonrespondent households in the DO.

4.2 Data

We use short-form data from the 1990 census for three DOs, whose characteristics are described in Table 1. The race of a household is determined by the most prevalent race in the household, usually (98% of households) the only race. In DO 1 we grouped consecutive (and therefore contiguous) block groups (clusters of contiguous blocks) into 94 areas containing an average of 52 blocks and 1100 households. For DOs 2 and 3, block group information was unavailable so we formed areas by grouping consecutive blocks into clusters containing an average of 50 blocks (on average, 548 households per area in DO 2 and 918 households per area in DO 3).

Table 1
Characteristics of the Census District Office Areas Used in the Simulations

	DO1	DO2	DO3
Household	112,966	169,321	149,567
Blocks	4,907	15,470	8,167
Pseudo-areas	94	309	163
Non-Hispanic Black	14.4%	28.5%	1.3%
Hispanic	6.1%	1.0%	6.6%
Other	73.5%	59.4%	81.5%
Owner	63.8%	59.5%	52.6%
Renter	30.2%	29.4%	36.7%
Vacant	6.0%	11.1%	10.7%
Size 1 (1–2 people)	50.4%	46.9%	55.2%
Size 2 (3–4 people)	31.6%	31.6%	26.2%
Size 3 (5+ people)	12.0%	10.4%	7.9%
Response Rate	72.6%	65.3%	56.7%

4.3 Measures of Bias, Variance, and Mean Squared Error

Loss functions for our evaluations are based on the relative error for household category j (a type or combination of types) in geographic area i (a block or collection of blocks):

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}} \tag{3}$$

where Y_{ij} is the true number of households of category j in geographical unit i, $\hat{Y}_{ijs}$ is the corresponding number of households estimated from sample s (including those observed in the sample and estimated by the model), and Y_{i+} is the total number of households in geographical unit i.

We summarize bias in estimated counts for category j and a level of geography (block, area, DO) with Root Mean Weighted Squared Bias (RMWSB):

$$\hat{\text{RMWSB}}_j^2 = \frac{\sum_i Y_{i+} \left\{ \left(\frac{1}{S}\sum_s d_{ijs} \right)^2 - \frac{1}{S(S-1)} \left(\sum_s d_{ijs}^2 - \frac{1}{S}\left(\sum_s d_{ijs}\right)^2 \right) \right\}}{\sum_i Y_{i+}} \tag{4}$$

where S is the number of samples drawn and $i = 1, \ldots, I$ where I is the number of geographical units. The second term in the numerator removes a bias due to the finiteness of the simulation. From a design-based perspective, we regard the composition of each area as a fixed quantity, and only sampling is random. Then bias is defined as the average difference, over all possible samples, between the truth for an area and the corresponding estimates, essentially the model error for that area. Such error is inevitable since the composition of the nonrespondents in any block is not entirely predictable. A more serious type of bias would involve systematic error in estimates for a collection of blocks with similar composition. Although we have not checked for all possible types of bias in this sense, the model specification protects us against bias at higher levels of aggregation because model estimates are constrained to agree (approximately) with unbiased estimates for areas and DOs.

As a measure of overall error, we calculate the Root Mean Weighted Mean Squared Error (RMWMSE) for each household category j, which is given by

$$\hat{\text{RMWMSE}}_j^2 = \frac{\sum_i Y_{i+} \left(\frac{1}{S}\sum_s d_{ijs}^2 \right)}{\sum_i Y_{i+}} \tag{5}$$

where Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, i, and S are defined as above. (The two "means" refer to mean over geographical units (i) and over samples (s).) We obtain a measure of the standard deviation of the estimates for household category j by calculating the Root Mean Weighted Variance (RMWV):

$$\hat{\text{RMWV}}_j^2 = \frac{\sum_i Y_{i+} \left\{ \frac{1}{S-1} \left(\sum_s d_{ijs}^2 - \frac{1}{S}\left(\sum_s d_{ijs}\right)^2 \right) \right\}}{\sum_i Y_{i+}} = \hat{\text{RMWMSE}}_j^2 - \hat{\text{RMWSB}}_j^2 . \tag{6}$$

Note that these MSE, bias, and standard deviation measures are all estimates of expectations with respect to repeated NRFU sampling from the given finite population of blocks. These loss functions can be applied at various levels of geography, reflecting the fact that the main use of block level estimates is aggregation to form estimates at higher levels of geography. With this in mind, these measures were also chosen because they weight errors by the size of the geographical unit. This leads to consistent estimates of error when aggregating over geographical units, which is appropriate due to the arbitrariness of unit boundaries (Zaslavsky 1993). We base our measures on errors relative to the total area i population rather than the population in the target category only, because the latter denominator inflates the importance of small errors in blocks where the category rarely or never appears.

4.4 Results

For simulated NRFU sampling using both the block and unit sampling designs, estimates of the number of households with each characteristic are calculated at block, area, and DO levels of geography using each of the three estimation methods. The results for each method are represented by the shaded bars in Figure 1 for the unit sampling design. (Results for the block sampling design are not shown here, but the pattern of results are similar with the RMWMSE being about 10% greater for all estimates.) In this figure, each row of bar charts displays the RMWMSE for block, area, and DO level estimates for one of the three DOs. Each group of three bars represents the RMWMSE for estimates of the total number of households for each of the tenure categories, the household size categories and the race categories using each of the three methods. Because all three methods use the same logistic regression model to predict the number of vacant nonsample nonrespondents in each block, the vacant category is omitted from the plots.

RMWMSE with both the stratified ratio method and the loglinear model was much smaller than with the unstratified ratio method for most household characteristics at the block and area level. Therefore, we confine further discussion to comparison of the two former methods.

The most dramatic differences appear for the tenure categories at the block and area levels. In each DO, block and area level estimates of the tenure categories from the loglinear model have much smaller RMWMSE than the estimates from the stratified ratio method, primarily because the former had much smaller bias (RMWSB). Standard deviations (RMWV) were slightly larger for the loglinear model under the unit sampling design, but about equal for the two methods under the block sampling design. The loglinear model had smaller bias for the tenure categories at the area level because tenure is included in the model as an area-level effect, x_4. Stratification on race in the ratio method reduces RMWMSE for the race categories at the block level, but the two methods have comparable RMWMSE for the race categories at the area and DO levels. The stratified ratio method loses its advantage over the loglinear model at the area level because the former does not use any area-level information. Both methods generally produce estimates with comparable RMWMSE at all levels of geography for the size categories.

The statistical significance (under the simulations) of differences in RMWMSE between the methods was evaluated using t–tests. Almost all differences at the block and area levels, excluding the vacant category, have two-tailed $p-$values ≤ 0.001 and therefore cannot be attributed to simulation error.

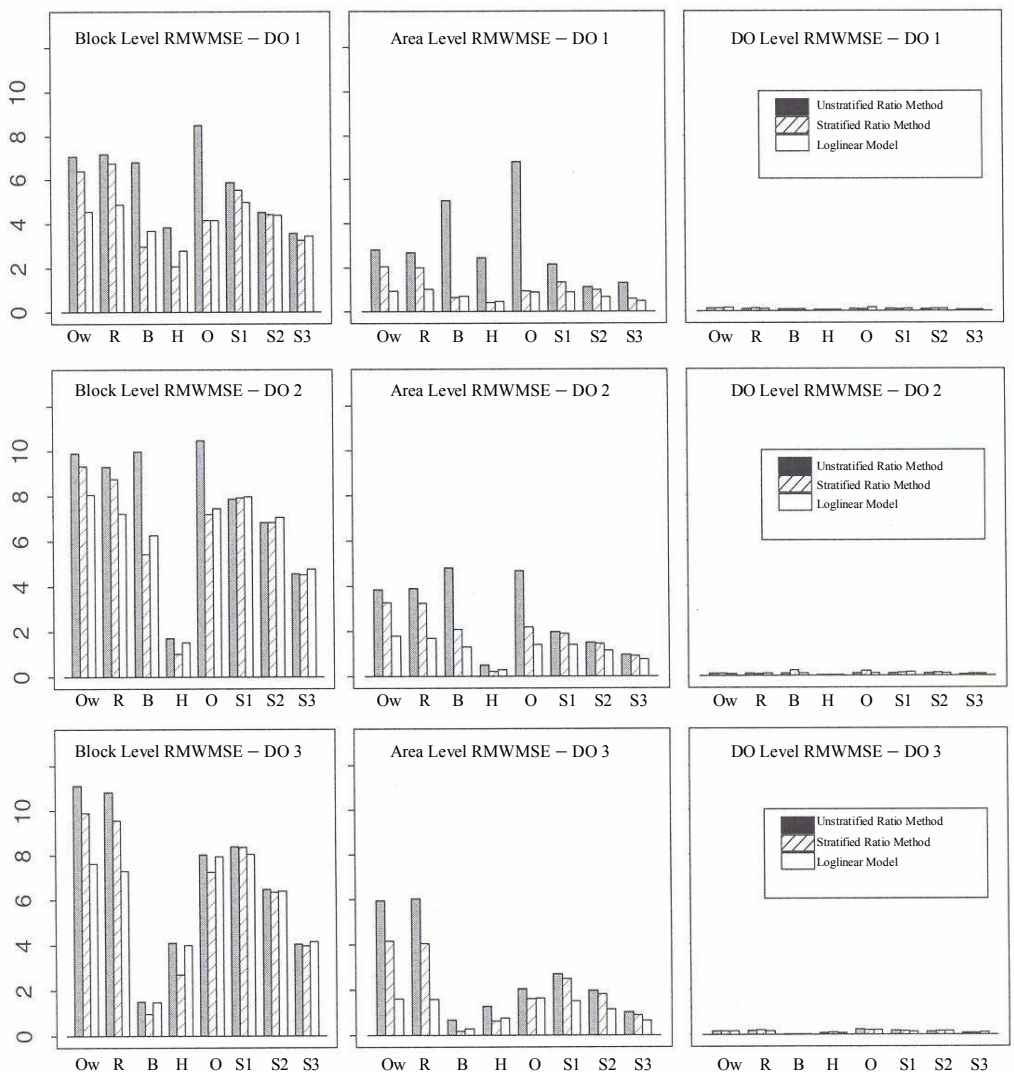


Figure 1. RMWMSE for block, area, and DO level estimates for each household characteristic, using the unit sampling design for DO 1, 2, and 3, with 30 simulated samples ("Ow" = Owner, "R" = Renter, "B" = Black, "H" = Hispanic, "O" = Other race, "S1" = Size group 1 (1 – 2 people), "S2" = Size group 2 (3 – 4 people), "S3" = Size group 3 (5 or more people)).

5. Assessment and Prediction of Model Error

Methods for estimation of MSE of fitted estimates using sample data are briefly summarized here due to space limitation; methods and findings are available from the first author.

First, we developed analytic approximations that predict the effect of changing the sampling rate on the accuracy of our estimates without requiring additional simulations at each rate. These can be useful for sample design. We approximate the RMWMSE of block, area, and DO level estimates at a new sampling rate under both the block and unit sampling designs, assuming simulation results using one sampling rate are already available, by combining estimates of bias and variance at the current sampling rate using two rescaling factors. The first factor reflects the changed proportion of housing units that require estimation under the new sampling rate, which affects the bias and variance of the combined estimates. The other reflects the effect of the sampling rate on the variance of the estimates for the nonresponding units. Simulations demonstrated the accuracy of predictions for RMWMSE using these approximations, except for some extreme extrapolations.

Using these results, we developed a cross-validation procedure to facilitate within-sample estimates of RMWMSE for use in a production setting where the true characteristics of the nonsample nonrespondent households are not known. The follow-up sample in each areas is divided randomly into C cross-validation groups (of blocks for block sampling, and of households for unit sampling). Each cross-validation group is dropped out in turn and the model is fitted to the nonrespondents in the remaining $C-1$ cross-validation groups and the respondents in all C groups. We can then estimate RMWMSE under the design simulated by the cross-validation and project this estimate to the actual sampling rate, or some other rate of interest, using the approximations described in the preceding paragraph. Simulations show that this produces accurate estimates of RMWMSE at block and DO levels of geography, with some overestimation at the area level. This method also provides separate estimates of bias and variance that are shown by simulation to be very accurate. These are useful for assessing model adequacy since a poorly-fitting model would be betrayed by a large component of MSE due to bias.

6. Conclusions

In the preceding sections, we have presented a model-based approach to imputation of the characteristics of nonresponding households in a census that were not sampled for nonresponse followup. In simulations, our loglinear model produces estimates with much smaller error than two alternatives for some estimands, and is about equivalent for others. These conclusions hold for both the block and unit sampling designs. An advantage of our approach is that models can be specified to constrain only a few marginal tables or interactions of characteristics at the finest levels of geography, where the data are sparse, while fitting more detailed distributions of characteristics at higher levels of geographic aggregation at which more data are available. This is consistent with typical practice in release of census data, which include minimal characteristics at the block level but increasingly more detailed characteristics for larger units.

Many important uses of the census involve estimation of the population and its characteristics for small domains such as legislative districts and planning areas for social services (such as schools and clinics) and commercial development. Even though these domains will not always align with the areas used in census estimation, controlling the census estimates to match unbiased estimates at several levels of geography makes it more likely that estimates for policy-relevant domains assembled from wholes or parts of these areas will also be nearly unbiased. Our method has more predictable aggregate properties than complex alternatives such as hierarchical spatial modeling. Although the latter might produce estimates with smaller MSE at the lowest levels of geography, fitting such models and checking their biases at various levels of geographic aggregation would require extensive local tuning which is likely to be impractical in a census production setting.

Our methodology is illustrated here in the context of a NRFU sampling for the U.S. Decennial Census, but our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for nonresponse followup with hierarchically structured populations. We can also incorporate administrative records as covariates for predicting the characteristics of the corresponding nonrespondent households (Zanutto and Zaslavsky 2002). In that scenario, data from households in the NRFU sample for which we have both census and administrative records information are used to estimate the systematic differences between the two information sources. Under the same models, we impute the characteristics of nonsample nonrespondent households. Using administrative records through this modeling approach can improve the accuracy of small area (block-level) estimates.

Although the discussion of sampling in the United States census has been politically contentious, nonetheless in the long run it seems likely that some form of estimation will be used for nonrespondents. The potential might be even greater in countries where population estimation already makes substantial use of administrative records (Redfern 1989). Methods such as those described here that can

combine information across data sources while reflecting local diversity will be essential to such efforts.

Appendix

Iterative Proportional Fitting with Partially Cross-Classified Data

A standard approach to fitting loglinear models to partially cross-classified data uses an EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 2002, chapter 8), in which in alternate steps (1) the expected counts are imputed under the model and (2) the model is refitted to the observed and imputed data, using iterative proportional fitting (IPF) (Darroch and Ratcliff 1972) for models without closed-form solutions. In the more efficient ECM modification of this algorithm, only a single cycle of the IPF algorithm is taken at each step (Meng and Rubin 1993).

For our application we developed a modified IPF algorithm that is faster than the EM and ECM algorithms for our models, which always include a block × response interaction and never include any block × type × response interactions. We found that our modified IPF algorithm converges in approximately one half to two thirds the number of cycles that ECM requires with less computation per step (Zanutto 1998, Part 1, Appendix A). (Convergence is declared when the predicted and observed values of the minimal sufficient statistics of the model are sufficiently close.)

Our algorithm takes advantage of the fact that partially classified observations contribute to the likelihood only through the total number of nonrespondent households in each block. Therefore, to maximize this part of the likelihood we need only ensure that the fitted number of nonrespondents in each block equals the observed number, which is automatic because the block × response interaction is always included in our model.

The modified IPF algorithm fits the model to the fully classified observations using an ordinary IPF algorithm, ignoring the partially cross-classified observations. For the block sampling design, this means that the model is fitted using the fully observed part of the block × type × response table using an ordinary IPF algorithm, ignoring the partially classified part of the table. Predictions for the partially cross-classified cells are obtained by applying the same fitting proportions to those cells as to the fully observed part of the table. Finally, predictions for the partially cross-classified cells are scaled so that the fitted number of nonrespondents in each block equals the observed number. For the unit sampling design, the same algorithm is used, viewing the collection of respondent households and nonrespondent households in the follow-up sample as analogous to the fully-observed part of the table in the block sampling design and viewing the blocks with no nonrespondents in the follow-up sample as analogous to the out-of-sample blocks in the block sampling design. This gives predictions for nonrespondent households in blocks with no nonrespondents in the follow-up sample. Predictions for nonrespondent households in blocks with one or more nonrespondent households in the follow-up sample are obtained by applying the predicted distribution of household types among sampled nonrespondent households in each of these blocks to the corresponding nonsample nonrespondent households in these blocks. For more details about in the unit sampling case, see Zanutto and Zaslavsky (2002).

We now illustrate the IPF algorithm for the block sampling design under a Poisson model like (1) with $\log(m_{ijr}) = z_{ijr}^T \beta$ where m_{ijr} represents the expected number of households in block i of household type j of response status r, and Z is the design matrix corresponding to the model expression $i * x + i * r + r * x$. This is a simplified version of the model in (2) with only one level of geography and only one "x" representing the full cross-classification defining household types. We observe n_{ijr} if $r = 1$ or if $r = 0$ and $i \in S$, but only n_{i+0} if $i \notin S$, where S represents the set of blocks selected for the NRFU sample.

The IPF algorithm to fit this model starts with initial estimates $\hat{m}_{ijr}^0 = 1$ for all i, j, r and contains the following three steps in cycle t:

$$\text{Step 1: } \hat{m}_{ijr}^{t+\frac{1}{3}} = \begin{cases} \hat{m}_{ijr}^t \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^t} \right) & \text{if } i \in S \text{ or if } i \notin S,\ r = 1 \\ \hat{m}_{ijr}^t & \text{if } i \notin S,\ r = 0 \end{cases}$$

$$\text{Step 2: } \hat{m}_{ijr}^{t+\frac{2}{3}} = \begin{cases} \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij+}}{\hat{m}_{ij+}^{t+\frac{1}{3}}} \right) & \text{if } i \in S \\ \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij1}}{\hat{m}_{ij1}^{t+\frac{1}{3}}} \right) & \text{if } i \notin S \end{cases}$$

$$\text{Step 3: } \hat{m}_{ij1}^{t+1} = \hat{m}_{ij1}^{t+\frac{2}{3}} \left(\frac{n_{+j1}}{\hat{m}_{+j1}^{t+\frac{2}{3}}} \right)$$

$$\hat{m}_{ij0}^{t+1} = \hat{m}_{ij0}^{t+\frac{2}{3}} \left(\frac{\sum_{i \in S} n_{ij0}}{\sum_{i \in S} \hat{m}_{ij0}^{t+\frac{2}{3}}} \right).$$

The scaling factors in each step are based only on observed counts.

These steps are repeated until the estimates of the minimal sufficient statistics for the model, excluding $\hat{m}_{i+r}$ for $i \notin S$, $r = 0$ (*i.e.*, $\hat{m}_{i+r}$ for $i \in S$ and $i \notin S, r = 1$, $\hat{m}_{ij+}$ for $i \in S$, $\hat{m}_{ij1}$ for $i \notin S$, $\hat{m}_{+j1}$, and $\sum_{i \in S} \hat{m}_{ij0}$) are sufficiently close to their observed values. Denoting the step at which this occurs as t^*, the final step in this algorithm is to set

$$\hat{m}_{ijr}^{t^*+1} = \begin{cases} \hat{m}_{ijr}^{t^*} \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^{t^*}} \right) & \text{if } i \notin S,\ r = 0 \\ \hat{m}_{ijr}^{t^*} & \text{otherwise,} \end{cases}$$

to ensure that estimated block×response margin $(i * r)$ for $i \notin S,\ r = 0$ equals the observed margin.

This IPF algorithm produces estimates that converge to the maximum likelihood estimates of the model parameters (Zanutto 1998, Part 1, Appendix A). The second case in Step 2 is not needed to maximize the likelihood but is included to obtain predictions for the nonsample nonrespondent cells (*i.e.*, $i \notin S,\ r = 0$).

References

Bell, W.R., and Otto, M.C. (1994). Investigation of a model-based approach to estimation under sampling for nonresponse in the decennial census. Unpublished paper presented at the Joint Statistical Meetings, Toronto.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society*, Series B, Methodological, 25, 220-233.

Brackstone, G.J., and Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

Darroch, J.N., and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-22.

Fuller, W.A., Isaki, C.T. and Tsay, J.H. (1994). Design and estimation for samples of census nonresponse. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 289-305.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall Ltd.

George, J.A., and Penny, R.N. (1987). Initial experience in implementing controlled rounding for confidentiality control. In *Proceedings of the Bureau of the Census Annual Research Conference*, Volume 3. Washington, DC: U.S. Bureau of the Census, 253-262.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons, Inc.

Meng, X.-L., and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys* (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 143-184.

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.

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

Redfern, P. (1989). European experience of using administrative data for censuses of population: The policy issues that must be addressed. *Survey Methodology*, 15, 83-99.

Rubin, D.B., and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*, 3, 375-387.

Schafer, J.L. (1995). Model-based imputation of census short-form items. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: Bureau of the Census, 267-299.

Schindler, E. (1993). Sampling for the count; sampling for non-mail returns. Unpublished report, U.S. Bureau of the Census.

U.S. Bureau of the Census (1997a). Census 2000 operational plan. Washington, DC.

U.S. Bureau of the Census (1997b). Report to Congress–the plan for Census 2000. Washington, DC.

Wilkinson, G.N. and Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399.

Zanutto, E. (1998). *Imputation for Unit Nonresponse*: *Modeling Sampled Nonresponse Follow-up*, *Administrative Records*, *and Matched Substitutes*. Ph.D. thesis, Harvard University, Cambridge, Massachusetts.

Zanutto, E., and Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 608-613.

Zanutto, E., and Zaslavsky, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse followup. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 673-686.

Zanutto, E., and Zaslavsky, A.M. (2002). Using administrative records to improve small area estimation: An example from the U.S. Decennial Census. *Journal of Official Statistics*, 18, 559-576.

Zaslavsky, A.M. (1988). Representing local area adjustments by reweighting of households. *Survey Methodology*, 14, 265-288.

Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.

Zaslavsky, A.M. (2004). Representing the Census undercount by multiple imputation of households. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Eds. A. Gelman and X.-L. Meng). West Sussex, England: John Wiley & Sons, Inc. 129-140.

Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society*, Series B, 66, 479-496.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 77-85
Statistics Canada, Catalogue No. 12-001-XIE

The 2006 Reverse Record Check Sample Allocation

Alain Théberge [1]

Abstract

Sample allocation can be optimized with respect to various goals. When there is more than one goal, a compromise allocation must be chosen. In the past, the Reverse Record Check achieved that compromise by having a certain fraction of the sample optimally allocated for each goal (for example, two thirds of the sample is allocated to produce good-quality provincial estimates, and one third to produce a good-quality national estimate). This paper suggests a method that involves selecting the maximum of two or more optimal allocations. By analyzing the impact that the precision of population estimates has on the federal government's equalization payments to the provinces, we can set four goals for the Reverse Record Check's provincial sample allocation. The Reverse Record Check's subprovincial sample allocation requires the smoothing of stratum-level parameters. This paper shows how calibration can be used to achieve this smoothing. The calibration problem and its solution do not assume that the calibration constraints have a solution. This avoids convergence problems inherent in related methods such as the raking ratio.

Key Words: Calibration; Raking ratio; Reverse record check; Sample allocation; Smoothing.

1. Introduction

The Canadian Census of Population is conducted every five years, most recently in 2001. The Reverse Record Check (RRC) measures the undercoverage and part of the overcoverage in the Census. For the next RRC in 2006, it is hoped that most of the census overcoverage will be measured by another survey, the Automated Match Study, which is more efficient for this task. This should make it possible to optimize the RRC sample allocation for undercoverage measurement. RRC coverage estimates are used in conjunction with census counts to produce population estimates. The population estimates are used for various purposes; for example, the federal Department of Finance uses them to calculate the equalization payments that the federal government makes to the provincial governments.

Traditionally, one consideration in allocating the RRC sample among the provinces has been to balance the need for a good-quality estimate of the national rate of persons missed by the Census and the need for good-quality estimates of provincial rates for use in producing Statistics Canada's population estimates.

It was hoped that this approach would also meet the need for good-quality equalization payment estimates (they are estimates because they depend on population estimates), but this has never been verified. The federal government makes equalization payments to the have-not provinces. In this paper, we examine the impact that the provincial sample allocation has on the quality of equalization payment estimates.

If the variance of a variable of interest is the same in every province, we can obtain an optimal allocation for a minimum-variance national estimate if the sample size is proportional to the frame size for province p, N_p. An allocation that produces provincial estimates of equal variance is one where the sample size is constant (proportional to N_p^0). One way often used to balance the two needs is to make the sample size proportional to $N_p^{1/2}$. A different method of achieving this balance has been used in the past by the RRC: part of the sample is allocated so as to yield provincial estimates of equal variance, and the other part is allocated so as to produce a minimum-variance national estimate. Traditionally, about two thirds of the sample is allocated in such a way as to produce provincial estimates of equal variance.

In this paper, we propose a new method of obtaining a provincial allocation which balances two or more goals. That method involves computing a distinct allocation for each goal, possibly with a different total sample size for each allocation; we obtain the final allocation, which should satisfy every goal, by taking the maximum sample size over each of the distinct allocations for each province.

The optimal subprovincial allocation is simply given by the Neyman allocation. The difficulty lies in predicting the variance in relatively small strata, or more precisely, in predicting the totals (number of persons missed by the Census, number of RRC non-respondents) on which the variance depends. For each province, the approach taken in this paper is to start with more stable national values at the cell level (age × sex × marital status) and scale them so that the totals agree with the provincial values for each age group, for each sex and for each marital status. This goal is reminiscent of an iterative raking procedure introduced by Deming and Stephan (1940), also used by Brackstone and Rao (1976). Deville and Särndal (1992) showed how

1. Alain Théberge, Social Survey Methods Division, Statistics Canada, R.H. Coats Bldg, 15th Floor, Ottawa, Ontario, Canada, K1A 0T6.

calibration can be used to achieve the same result. In the case of the RRC, calibration will be used even though the cells cannot be put in a convenient three-dimensional matrix because the age groups differ for each marital status. The raking ratio method sometimes fails to converge because the constraints cannot be satisfied. By stating the calibration problem as in Théberge (1999), we allow for the possibility that the constraints are inconsistent, and this does not cause convergence problems. In addition, if we use the Moore-Penrose inverse as part of the solution, the constraints can be linearly dependent.

In the next section, we will explore the relationship between population estimates and equalization payments. As we will see, the sample allocation problem entails balancing four goals. In Section 3, we use an approximate variance formula that relies on a design effect to determine the optimal allocation for each goal. We determine the value of the design effect empirically in Section 4. Section 5 explains how a final allocation can balance individual allocations for separate goals. Finally, the subprovincial allocation is addressed in Section 6. The sample allocation for the three territories is not discussed in this paper.

2. Impact of Population Estimates on Equalization Payments

Statistics Canada is responsible for producing population estimates. One important use of those population estimates is in computing the equalization payments made by the federal Department of Finance. Although Statistics Canada is not directly concerned with the formula for equalization payments, it is still relevant to examine how the precision of the population estimates affects the precision of the equalization payments. The impact that the sample allocation has on the precision of the population estimates has been studied for many years; in this paper, we will also examine how the sample allocation affects the precision of the equalization payments.

The RRC is the survey used to measure the rate of persons missed by the Census. Traditionally, the RRC's sample allocation has been designed to achieve a compromise between having a minimum variance for the national estimated undercoverage rate (goal I) and having equally low variances for the estimated undercoverage rates of the provinces (goal II). Two more goals will be added as we examine the impact that the sample allocation has on the precision of the equalization payments.

The formula used to calculate the equalization payments, before any smoothing based on moving average, is

$$E_p = \sum_{j=1}^{33} \frac{R_{tj}}{T_{tj}} \left(\frac{T_{\text{std } j}}{P_{\text{std}}} - \frac{T_{pj}}{P_p} \right) P_p \,, \qquad (2.1)$$

where E_p is the equalization payment for beneficiary province p (at the time of writing, all provinces except Ontario and Alberta), R_{tj} is the total revenue (all provinces) from revenue source j, T_{tj} is the total tax base for revenue source j, $T_{\text{std } j}$ is the tax base of the standard provinces (all provinces except the Atlantic provinces and Alberta) for revenue source j, P_{std} is the population of the standard provinces, T_{pj} is the tax base of beneficiary province p for revenue source j, and P_p is the population of beneficiary province p.

To measure the influence that population estimates have on the equalization payments, we will rewrite equation (2.1) as

$$E_p = \left(\frac{P_p}{P_{\text{std}}} \right) C_{\text{std}} - K_p \,, \qquad (2.2)$$

where

$$C_{\text{std}} = \sum_{j=1}^{33} \frac{R_{tj} T_{\text{std } j}}{T_{tj}}$$

and

$$K_p = \sum_{j=1}^{33} \frac{R_{tj} T_{pj}}{T_{tj}} \,.$$

We note that the population of Alberta has no impact on the equalization payment of any beneficiary province. The population of Ontario only affects the equalization payment through P_{std}. For the Atlantic provinces, their equalization payment varies linearly with their population, since their population does not affect P_{std}. If we assume P_{std} is known, we can say that an error of one person in a beneficiary province's population has an impact of $C_{\text{std}} / P_{\text{std}}$ dollars on its equalization payment, for any beneficiary province. This does not mean that the equalization payment of a beneficiary province only depends on its population and not on the population of the standard provinces. However, as we will see, most of the sampling error in the equalization payment comes from the sampling error in the estimate of the beneficiary province's population, and relatively little comes from the sampling error in the estimate of the standard provinces' population.

If symbols with hats represent estimates, then from (2.2),

$$V(\hat{E}_p) \simeq C_{\text{std}}^2 \frac{1}{P_{\text{std}}^2} \begin{pmatrix} V(\hat{P}_p) + \left(\frac{P_p}{P_{\text{std}}} \right)^2 V(\hat{P}_{\text{std}}) \\ -2 \frac{P_p}{P_{\text{std}}} \text{Cov}(\hat{P}_p, \hat{P}_{\text{std}}) \end{pmatrix}. \qquad (2.3)$$

Because stratification is done separately for each province, for a beneficiary province p, which is not one of the standard provinces, we have, ignoring interprovincial migration, $\mathrm{Cov}(\hat{P}_p, \hat{P}_{\mathrm{std}}) = 0$, whereas $\mathrm{Cov}(\hat{P}_p, \hat{P}_{\mathrm{std}}) = V(\hat{P}_p)$ for any of the standard provinces. We can compute an approximation by leaving out the last two terms of (2.3):

$$V(\hat{E}_p) \simeq \left(\frac{C_{\mathrm{std}}}{P_{\mathrm{std}}}\right)^2 V(\hat{P}_p). \qquad (2.4)$$

Using data from the 2001 RRC, we can verify that the standard deviation of the equalization payment derived from (2.4) differs from that derived from (2.3) by no more than 7%, except for two beneficiary provinces: Newfoundland and Labrador, for which the approximation underestimates the standard deviation by 11%, and Quebec, for which the approximation underestimates the standard deviation by 12%.

As we can see from equation (2.4), a sample allocation that produces equal variances for beneficiary provinces' population estimates also produces equal variances for beneficiary provinces' equalization payment estimates. However, having equal CVs for the beneficiary provinces' population estimates does not guarantee equal CVs for the beneficiary provinces' equalization payment estimates, since from equation (2.2), E_p is not directly proportional to P_p, because K_p is not zero. Having equal CVs for the beneficiary provinces' population estimates is still a goal worth pursuing, since it ensures confidence intervals of equal length for the equalization payment per person. Indeed, because of the use of the approximation (2.4), if the 2001 situation recurs in 2006, the confidence interval for Newfoundland and Labrador will be 11% too short (that is, the precision for the equalization payment per person will be poorer than for other beneficiary provinces), while the confidence interval for Quebec will be 12% too long (that is, the precision for the equalization payment per person will be greater than for other beneficiary provinces). Also, if we ignore interprovincial migration, then the provincial population estimates are independent and the variance of the total equalization payment is minimized if and only if the variance of the total population of beneficiary provinces is minimized.

We are attempting to find a provincial sample allocation that minimizes the variance of the total equalization payment or, equivalently, the variance of the total population of beneficiary provinces (goal III). We also want to find a provincial sample allocation that produces equal CVs for each beneficiary province's population estimate (goal IV), in order to achieve equally good precision for the equalization payment per person.

Most of the variance in population estimates is due to the variance in the undercoverage estimates. If we ignore the contribution that overcoverage makes to the variance of the population estimate, then it is easily verified that the standard error of the estimated undercoverage rate equals the CV of the population estimate. Goals I and II can then be restated as follows: minimize the CV of the national population estimate, and produce provincial population estimates with equal CVs. The difference between goals III and I, and between goals IV and II, is that one applies to beneficiary provinces, and the other to all provinces. In what follows, we will indeed assume that the variance of the population estimates equals the variance of the undercoverage estimates.

The goals of the provincial sample allocation are summarized in Table 2.1.

Table 2.1
The Four Goals of the Provincial Sample Allocation

Goal	Description (equivalent description)
I	Minimize the variance of the estimated national undercoverage rate. (Minimize the CV of the national population estimate.)
II	Produce equal variances for the provinces' estimated undercoverage rates. (Produce provincial population estimates with equal CVs.)
III	Minimize the variance of the total equalization payment (Minimize the variance of the estimated total population of beneficiary provinces).
IV	Produce equal variances for the equalization payment per person for each beneficiary province (Produce equal CVs for the population estimate for each beneficiary province, or produce equal variances for the estimated undercoverage rates of the beneficiary provinces).

3. Optimal Provincial Sample Allocation

In this section, we will first describe the notation we plan to use, and then we will discuss approximate variance formulae for population estimates and estimated undercoverage rates. We will explore the issue of optimality with respect to the four goals mentioned above.

Five sample frames are used for the RRC in the provinces: the census frame (people enumerated in the previous census), the birth frame (intercensal births), the immigrant frame (intercensal immigrants), the non-permanent resident frame and the "missed" frame. The "missed" frame is made up of the sampled persons of the previous RRC who were missed by the previous census. With their weights, they represent the subpopulation of enumerable persons not covered by any of the other four frames. Each frame within each province is stratified separately. A stratified random sample is selected in each frame. All persons from the "missed" frame are included in the sample.

Let U_{hp} be the number of undercovered persons in stratum h who are classified in province (of classification) p. Similarly, let E_{hp} and O_{hp} be, respectively, the number of enumerated and overcovered persons in stratum h who are classified in province p, and $P_{hp} = U_{hp} + E_{hp} - O_{hp}$. The undercoverage rate for province p can then be written as

$$R_{.p} = U_{.p} / P_{.p}, \tag{3.1}$$

where $U_{.p} = \sum_h U_{hp}$ and $P_{.p} = \sum_h P_{hp}$. We see that $P_{.p}$ equals P_p as defined in the preceding section.

An estimator of the undercoverage rate for province p is

$$\hat{R}_{.p} = \hat{U}_{.p} / \hat{P}_{.p}, \tag{3.2}$$

where $\hat{U}_{.p}$ and $\hat{P}_{.p}$ are estimators of $U_{.p}$ and $P_{.p}$ respectively. Linearization gives

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2}\left[V(\hat{U}_{.p}) + \frac{U_{.p}^2}{P_{.p}^2} V(\hat{P}_{.p})\right]. \tag{3.3}$$

The second term in brackets is negligible in comparison to the first; therefore,

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_h \frac{U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.4}$$

where N_h is the size of stratum h, and n_h is the sample size in stratum h. This ignores the finite population correction factor. In what follows, we will assume that there is no non-response and that there is only one stratum per province of selection (no stratification by frame, age, sex, *etc.*). This assumption will of course be dropped in Section 6, which deals with sample allocation to subprovincial strata. To compensate for the effects of subprovincial stratification and non-response, we introduce a design effect, D_h. We assume that this design effect varies only with stratum h; in particular, the same design effect is used to represent the variance of the estimated number of persons selected in stratum h who are undercovered in province p, for all p. The variance (3.4) can be approximated by

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.5}$$

and

$$V(\hat{U}_{.p}) \cong \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.6}$$

where the summation this time is over the provinces of selection.

Goal I:

From (3.5), we have

$$V(\hat{R}_{..}) \cong \frac{1}{P_{..}^2} \sum_{h=1}^{10} \frac{D_h U_{h.}(N_h - U_{h.})}{n_h}, \tag{3.7}$$

where $P_{..} = \sum_{p=1}^{10} P_{.p}$, $\hat{R}_{..} = \hat{U}_{..} / \hat{P}_{..}$, $U_{..} = \sum_{p=1}^{10} U_{.p}$ and $U_{h.} = \sum_{p=1}^{10} U_{hp}$. This variance of the national estimated undercoverage rate will be minimized if n_h is proportional to $\sqrt{D_h U_{h.}(N_h - U_{h.})} = N_h \sqrt{D_h R_{h.}(1 - R_{h.})}$, where $R_{h.} = U_{h.} / N_h$. Therefore, the optimal allocation for goal I of a sample of total size n_I is

$$n_{p\mathrm{I}} = n_{\mathrm{I}} \left[\frac{N_p \sqrt{D_p R_{p.}(1 - R_{p.})}}{\sum_{p=1}^{10} N_p \sqrt{D_p R_{p.}(1 - R_{p.})}} \right] \quad p = 1, \ldots, 10. \tag{3.8}$$

This is an improvement over the formula used for the 2001 RRC (see Clark 2000), where no design effect was applied to the part of the sample allocated to provide the best Canada-level estimate. In addition, for the 2001 RRC, n_p was proportional to the projected population in province p. It makes sense for n_p to depend on the size of the provincial frames; it should also depend on the provincial distribution of the undercoverage.

Goal II:

We can use equation (3.5) to compute the values of n_h that yield the same variance for the estimated provincial undercoverage rates. That problem has 10 equations in 10 unknowns. There is also another difficulty: obtaining sufficiently precise estimates of the U_{hp} for $p \neq h$, especially if p is a small province. Although in many cases it is reasonable to assume that the rate of undercovered persons in a small province p, $R_{.p} = U_{.p} / P_{.p}$, that was observed in one census, is a good predictor of the rate in the next census, the individual values of the U_{hp} for $p \neq h$ are harder to estimate and still harder to predict. Instead, we will assume that $U_{hp} = 0$ for $p \neq h$ and that $U_{pp} = U_{.p}$, which will mitigate the effect that outliers have on the expected variances. The provincial estimates of the undercoverage rate will then be of equal variance, if n_h, for $h = p$, is proportional to $(1/P_{.p}^2) D_p U_{.p}(N_p - U_{.p}) = D_p R_{.p}(N_p / P_{.p} - R_{.p})$. Therefore, the optimal allocation for goal II of a sample of total size n_{II} is

$$n_{p\mathrm{II}} = n_{\mathrm{II}} \left[\frac{D_p R_{.p}(N_p / P_{.p} - R_{.p})}{\sum_{p=1}^{10} D_p R_{.p}(N_p / P_{.p} - R_{.p})} \right] \quad p = 1, \ldots, 10. \tag{3.9}$$

Note that in the 2001 RRC, for the part of the sample allocated to ensure equal precision of the provincial estimates, the sample sizes were set proportional to $D_p \hat{R}_{.p}(1 - \hat{R}_{.p})$ (see Clark 2000). Using $N_p / P_{.p}$ instead of

1 takes into account not only those units which are in the province's frame and leave the province's population but also those units of the province's population that are not in the province's frame, leaving the design effect to account only for non-response and the sample design. In 2001, adjustment for frame units leaving the population was made through the design effect, and no adjustment was made for population units not in the frame.

Goal III:

The estimate of the total population of beneficiary provinces has a variance equal to

$$V(\hat{P}_{.\text{ben}}) = V(\hat{U}_{.\text{ben}}) \cong \sum_{h=1}^{10} \frac{D_h\, U_{h\text{ben}}(N_h - U_{h\text{ben}})}{n_h}, \quad (3.10)$$

where $P_{.\text{ben}} = \sum_{p=1}^{8} P_{.p}$, $U_{h\text{ben}} = \sum_{p=1}^{8} U_{hp}$ and $U_{.\text{ben}} = \sum_{p=1}^{8} U_{.p}$ are sums over the eight beneficiary provinces (we assume that the beneficiary provinces are numbered $p = 1, \ldots, 8$, and the non-beneficiary provinces are numbered $p = 9$, 10). Equation (3.10) is minimized if n_h, for $h = 1, \ldots, 10$, is proportional to $\sqrt{D_h\, U_{h\text{ben}}(N_h - U_{h\text{ben}})} = N_h\sqrt{D_h\, R_{h\text{ben}}(1 - R_{h\text{ben}})}$, where $R_{h\text{ben}} = U_{h\text{ben}} / N_h$. Therefore, the optimal allocation for goal III of a sample of total size n_{III} is

$$n_{p\text{III}} = n_{\text{III}} \left[\frac{N_p\sqrt{D_p\, R_{p\text{ben}}(1 - R_{p\text{ben}})}}{\sum_{p=1}^{10} N_p\sqrt{D_p\, R_{p\text{ben}}(1 - R_{p\text{ben}})}} \right]$$

$$p = 1, \ldots, 10. \quad (3.11)$$

Note that because units selected in one province can be classified in another province, $R_{p\text{ben}}$, and $n_{p\text{III}}$, are not necessarily zero when p is a non-beneficiary province.

Goal IV:

From equation (3.6), we have

$$CV(\hat{P}_{.p}) \cong \frac{1}{P_{.p}}\sqrt{\sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}}. \quad (3.12)$$

We can use this equation to compute the values of n_h that yield the same coefficient of variation for the beneficiary provinces' population estimates. That problem has eight equations in eight unknowns. Again here, we have a second difficulty: obtaining sufficiently precise estimates of the U_{hp} for $p \neq h$, especially if p is a small province. As we did for goal II, we will assume instead that $U_{hp} = 0$ for $p \neq h$ and that $U_{pp} = U_{.p}$. Beneficiary provinces' population estimates will then have equal coefficients of variation if n_h, for $h = p$, is proportional to $(1/P_{.p}^2)$ $D_p U_{.p}(N_p - U_{.p}) = D_p R_{.p}(N_p / P_{.p} - R_{.p})$. Therefore, the optimal allocation for goal IV of a sample of total size n_{IV} is

$$n_{p\text{IV}} = n_{\text{IV}} \left[\frac{D_p R_{.p}(N_p / P_{.p} - R_{.p})}{\sum_{p=1}^{8} D_p R_{.p}(N_p / P_{.p} - R_{.p})} \right] \quad p = 1, \ldots, 8 \ (3.13)$$

with the two non-beneficiary provinces having $n_{p\text{IV}} = 0$, $p = 9.10$.

It is worth noting that $n_{p\text{II}} / n_{p\text{IV}}$ is constant for all eight beneficiary provinces. This shows that goal II (equal precision of the estimated provincial undercoverage rates), which is a traditional goal of the RRC sample allocation, largely overlaps with goal IV (equal precision of the beneficiary provinces' equalization payments per person). We will see in Section 5 that $n_{p\text{I}} / n_{p\text{III}}$, for the eight beneficiary provinces, is nearly constant as well. This shows that goal I (maximum precision of the estimated national undercoverage rate), which is a traditional goal of the RRC sample allocation, largely overlaps with goal III (maximum precision of the total equalization payments).

4. Design Effect

Standard errors for the 2001 RRC estimates were computed using the Generalized Estimation System. Those standard errors take into account the RRC's sampling plan and non-response by assuming that the respondents are selected with a multi-stage sampling plan. A comparison of the standard error derived from (3.6) and the standard error computed by the Generalized Estimation System is presented in Table 4.1. A design effect equal to the inverse of the cube of the response rate for the province of selection was used for this comparison.

The table shows that the standard error for Prince Edward Island derived from (3.6) is 39% higher than the standard error computed by the GES; this is due to an outlier which affects the equation (3.6) estimate more than it affects the GES estimate. For most provinces, the equation (3.6) standard error is close to the GES standard error. These empirical results show that the design effect in equations (3.5) and (3.6) is approximately equal to the inverse response rate cubed. This suggests that a sample size of "n" units with response rate "r" yields the equivalent of $n \times r^3$ units rather than the expected $n \times r$, because non-respondents are concentrated among persons missed by the Census. The GES takes into account the fact that undercovered persons are less likely to respond. This

decline in precision due to non-response occurs even though the actual sampling plan is more efficiently stratified than the assumed sampling plan of one stratum per province.

Table 4.1
Comparison of Standard Errors

Province	Response rate	D = (response rate)$^{-3}$	Standard error of under-coverage estimate from (3.6)	Standard error of under-coverage estimate from GES	(3.6) SE / GES SE
N.L.	0.97	1.08	1,783	1,689	1.06
P.E.I.	0.97	1.09	1,021	734	1.39
N.S.	0.95	1.15	3,903	3,955	0.99
N.B.	0.96	1.13	3,272	3,229	1.01
Que.	0.95	1.17	19,915	19,664	1.01
Ont.	0.92	1.28	31,502	31,602	1.00
Man.	0.95	1.15	4,762	5,115	0.93
Sask.	0.96	1.12	3,921	3,840	1.02
Alta.	0.93	1.25	10,493	10,505	1.00
B.C.	0.91	1.34	14,619	14,763	0.99
Can.	0.94	1.20	42,074	42,041	1.00

There have been no similar studies comparing the design effect and the non-response rate in previous RRCs. The weight adjustment method used to compensate for non-response is different, and the nature of non-response is significantly different from what it was before 2001.

5. Final Provincial Sample Allocation and Example

Table 5.1 shows the parameter values that will be used in the example. The values of $\hat{N}_p$ are projections of RRC frame size for 2006; the other parameters are based on 2001 RRC data.

As we might expect, the values of $\hat{R}_{p\,\text{ben}}$ in Ontario and Alberta show that few units selected in those two provinces are classified as missed by the Census in beneficiary provinces.

The final sample size allocated to province p is simply

$$n_p = \max(n_{p\text{I}}, n_{p\text{II}}, n_{p\text{III}}, n_{p\text{IV}}) \quad p = 1, \ldots, 10. \quad (5.1)$$

Whether we use the maximum of the four sizes as in (5.1), a weighted arithmetic mean, or a weighted geometric mean, each method uses four arbitrary parameters (three if the total sample size is fixed). For the maximum method, higher relative values of n_{I} (or of n_{II}, n_{III} or n_{IV}) make goal I (II, III or IV respectively) more important.

Table 5.2 presents an example with $n_{\text{I}} = 30,000$, $n_{\text{II}} = 64,000$, $n_{\text{III}} = 25,000$ and $n_{\text{IV}} = 48,078$.

The resulting total sample size is 70,028. Figures in bold represent the maximum for the four allocations, n_p. Small changes in n_{III} would affect only the final allocation for Quebec. This suggests that with the sample sizes n_{I}, n_{II}, n_{III} and n_{IV} as chosen above, the final sample size allocated to Quebec is dictated by goal III (a precise estimate of the total equalization payment). Similarly, the final sample size allocated to Ontario is dictated by goal I (a precise estimate of the national undercoverage rate). The final sample size allocated to Alberta is dictated by goal II (equal variances for the provinces' estimated undercoverage rates). The final sample sizes of the other provinces are dictated both by goal II and by goal IV (equal precision of the estimated equalization payment per person). As noted in Section 3, $n_{p\text{II}} / n_{p\text{IV}}$ is constant for all eight beneficiary provinces. In the example above, because of the "judicious" choice of n_{IV}, the constant is 1. Lowering n_{IV} would decrease Alberta's final sample size, but not that of other provinces. We note also that $n_{p\text{I}} / n_{p\text{III}}$ does not vary much for the eight beneficiary provinces. The addition of goals III and IV (relating to equalization payments) allows us to control Quebec's sample size and Alberta's sample size separately. When only goals I and II were used, Quebec's sample size tended to be closely tied to Ontario's, while Alberta's sample size was closely tied to that of the other provinces.

Table 5.1
Parameter Values

Province	$\hat{N}_p$	D_p	$\hat{P}_{.p}$	$\hat{R}_{.p}$	$\hat{R}_{p.}$	$\hat{R}_{p\text{ben}}$
N.L.	551,987	1.0804	524,722	0.0339	0.0464	0.0368
P.E.I.	145,173	1.0882	132,473	0.0334	0.0334	0.0307
N.S.	995,651	1.1527	947,099	0.0492	0.0464	0.0440
N.B.	797,488	1.1345	736,129	0.0493	0.0466	0.0440
Que.	8,079,167	1.1740	7,381,352	0.0510	0.0471	0.0460
Ont.	13,423,132	1.2752	11,702,797	0.0653	0.0565	0.0017
Man.	1,262,547	1.1558	1,136,146	0.0466	0.0437	0.0392
Sask.	1,082,238	1.1223	996,562	0.0437	0.0430	0.0402
Alta.	3,373,128	1.2478	3,010,105	0.0490	0.0403	0.0028
B.C.	4,570,444	1.3369	4,014,502	0.0761	0.0669	0.0620
Can.	34,280,955	1.2039	30,581,887	0.0587	0.0524	0.0258

Table 5.2
Provincial Sample Allocation with
$n_{\mathrm{I}} = 30{,}000,\ n_{\mathrm{II}} = 64{,}000,\ n_{\mathrm{III}} = 25{,}000,$ and $n_{\mathrm{IV}} = 48{,}078$

Province	$n_{p\mathrm{I}}$	$n_{p\mathrm{II}}$	$n_{p\mathrm{III}}$	$n_{p\mathrm{IV}}$	n_p	$n_{p\mathrm{I}} / n_{p\mathrm{III}}$
N.L.	427	**3,816**	546	3,816	3,816	0.78
P.E.I.	96	**3,956**	132	3,956	3,956	0.73
N.S.	796	**5,822**	1,107	5,822	5,822	0.72
N.B.	634	**5,921**	881	5,921	5,921	0.72
Que.	6 562	6,399	**9,262**	6,399	9,262	0.71
Ont.	**12,385**	9,220	3,148	0	12,385	3.93
Man.	982	**5,867**	1,331	5,867	5,867	0.74
Sask.	823	**5,234**	1,139	5,234	5,234	0.72
Alta.	2,622	**6,702**	1,015	0	6,702	2.58
B.C.	4,673	**11,063**	6,440	11,063	11,063	0.73
Total	30,000	64,000	25,000	48,078	70,028	

An allocation method that uses equation (5.1) and a table such as Table 5.2 makes it clear why a province's sample has to be a certain size. For example, if we look at the final sample allocation in Table 5.2 and decide that 5,867 observations in Manitoba is insufficient, then we have to specify the goal for which they are insufficient. If we want to improve on the results for goal II (or goal IV), we also have to increase the sample size in all Atlantic provinces and all western provinces (or in all Atlantic provinces and all western provinces except Alberta).

6. Subprovincial Sample Allocation

Although it is evident from equation (3.5) that the subprovincial sample allocation in one province of selection affects the variances of other provinces' estimates, we will try to optimize the allocation in one province only for that province's estimate. In other words, our problem for each province p is to minimize

$$\sum_{h \in \left\{\begin{subarray}{l}\text{strata of province}\\ \text{of selection } p\end{subarray}\right\}} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h} \qquad (6.1)$$

subject to the constraint

$$\sum_{h \in \left\{\begin{subarray}{l}\text{strata of province}\\ \text{of selection } p\end{subarray}\right\}} n_h = n_p ,$$

where n_p is a previously determined total sample size for province p. Note that the sample size allocated to the "missed" frame is fixed, which means that in what follows, the "missed" frame strata are ignored, and n_p excludes the "missed" frame sample size. The solution to that minimization problem is

$$n_{h^*} = n_p \frac{\sqrt{D_{h^*} U_{h^* p}(N_{h^*} - U_{h^* p})}}{\sum_{h \in \left\{\begin{subarray}{l}\text{strata of province}\\ \text{of selection } p\end{subarray}\right\}} \sqrt{D_h U_{hp}(N_h - U_{hp})}} \qquad (6.2)$$

for each stratum h^* in province of selection p.

As we saw in Section 4, there is empirical evidence at the provincial level that the factor D_h is inversely proportional to the cube of the RRC response rate. For the 2001 sample allocation, it was assumed that D_h varied with the inverse of the response rate. To limit the shift of sample, relative to 2001, from strata with a high response rate, such as census frame or birth frame strata, to strata with a low response rate, such as immigrant frame or non-permanent resident frame strata, we will make D_h proportional to the inverse of the square of the response rate in stratum h. Note that here, in contrast to the assumption we made in Section 3, factor D_h compensates only for non-response; it does not compensate for the stratification since it is defined at the stratum level. This is another reason for choosing a factor smaller than the inverse of the cube of the response rate.

As was the case in the 2001 sample allocation, we are faced with the problem of reliably projecting the 2006 values of U_{hp} and D_h for every stratum h. Since the birth frame, the immigrant frame and the non-permanent resident frame each have only one stratum per province, we plan to use the 2006 sizes for those strata and the 2001 undercoverage rates and response rates, along with some ad hoc adjustments for the less populous provinces if necessary. A similar procedure can be used for the Indian reserve strata of the census frame. The other census frame strata are based on sex, marital status (married, not married) and age group. For these strata, using the same age groups for each sex and each marital status, it would be possible, for each province, to rake the *national* projections to margins of *provincial* projections, and use the raked values in equation (6.2). More precisely, to produce projections for U_{hp} for all strata h in province of selection p, we would first take the 2001 estimated rates and the 2006 strata sizes and compute a projection, for each cell (sex × marital status × age group), of the number of missed persons, classified in the province where they were selected. Those national figures could populate the cells of a three-dimensional

matrix. Still using the 2001 estimated rates and the 2006 strata sizes, we would then compute a projection for the number of missed persons, classified in province p, in all of the province's strata by sex, then in all of the province's strata by marital status, and finally in all of the province's strata by age group. Those figures would provide the desired marginal totals of the three-dimensional matrix. Through raking, we could obtain projections for U_{hp} that add up to the desired provincial totals by sex, by marital status and by age group. We can avoid convergence problems, simplify programming and enhance flexibility if we replace raking; we can do so by solving a calibration problem. In fact, we need the added flexibility in this case, because the age groups for married persons are not the same as the age groups for not-married persons.

Here is an example of how calibration is used. The method is based on the following result from Théberge (1999).

If we let $\mathbf{U}$ and $\mathbf{T}$ be positive diagonal matrices of dimension n and q respectively, $\mathbf{w}_0$ a vector of dimension n, $\mathbf{A}$ a $q \times n$ matrix, and $\mathbf{b}$ a vector of dimension q, then among the weight vectors $\mathbf{w}$ of dimension n that minimize $\| \mathbf{Aw} - \mathbf{b} \|_{\mathbf{T}}^2$, the unique weight vector that minimizes $\| \mathbf{w} - \mathbf{w}_0 \|_{\mathbf{U}}^2$ is given by

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{A}\mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}(\mathbf{b} - \mathbf{A}\mathbf{w}_0), \quad (6.3)$$

where $\| \mathbf{z} - \mathbf{z}_0 \|_{\mathbf{F}}^2 = (\mathbf{z} - \mathbf{z}_0)'\mathbf{F}(\mathbf{z} - \mathbf{z}_0)$ is a weighted distance measure between $\mathbf{z}$ and $\mathbf{z}_0$, and $\mathbf{G}^{\dagger}$ is the Moore-Penrose inverse of $\mathbf{G}$.

The equation $\mathbf{Aw} = \mathbf{b}$ forms the set of q calibration constraints. We will set $\mathbf{T}$ equal to the identity matrix in equation (6.3). If the constraints can be satisfied, then the matrix $\mathbf{T}$ is irrelevant; if not, then setting $\mathbf{T}$ equal to the identity matrix has the effect of giving equal importance to each of the q constraints when we minimize the distance between $\mathbf{Aw}$ and $\mathbf{b}$.

In this case, in projecting the number of missed persons in each stratum of a given province, we have $\mathbf{A} = \mathbf{MX}$, with

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{X} = \operatorname{diag} \begin{pmatrix} x_{FN0-14} \\ x_{FN15-24} \\ x_{FN25-44} \\ x_{FN45+} \\ x_{FM25-34} \\ x_{FM35+} \\ x_{MN0-14} \\ x_{MN15-24} \\ x_{MN25-44} \\ x_{MN45+} \\ x_{MM25-34} \\ x_{MM35+} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_{FN0-14} \\ w_{FN15-24} \\ w_{FN25-44} \\ w_{FN45+} \\ w_{FM25-34} \\ w_{FM35+} \\ w_{MN0-14} \\ w_{MN15-24} \\ w_{MN25-44} \\ w_{MN45+} \\ w_{MM25-34} \\ w_{MM35+} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_{F..} \\ b_{M..} \\ b_{.N.} \\ b_{.M.} \\ b_{..0-14} \\ b_{..15-24} \\ b_{..25+} \end{pmatrix},$$

where, for example, $x_{MN25-44}$ is the number of missed persons, classified in the province where they were selected, in the strata of not-married males aged 25 to 44, $w_{MN25-44}$ is the desired weight for that stratum, and $b_{.N.}$ is the number of missed persons selected and classified in the province who belong to the "not married" strata. All persons aged 0 to 24 are in "not married" strata regardless of their actual marital status. Note that in calculating both the national figures, $\mathbf{X}$, and the provincial figures, $\mathbf{b}$, we count only persons who did not move from one province to another, so as to remain consistent with the objective set out at the beginning of this section.

Continuing the parallel with raking, the matrix $\mathbf{X}$ gives the values of the three-dimensional matrix to be raked, except that the elements are arranged in a diagonal matrix; the vector $\mathbf{w}$ provides the final "raking factors" that are applied to the elements of $\mathbf{X}$ to produce the raked values, $\mathbf{Xw}$; the constraint is that sums of those raked elements, $\mathbf{MXw}$, should be as close as possible to the desired "margins" given by the vector $\mathbf{b}$; and $\mathbf{w}$ should be as close as possible to $\mathbf{w}_0$ described below.

By choosing the vector $\mathbf{w}_0$ so that every element is equal to a constant factor, we can scale the national figures down to figures that are more appropriate for the province. We can do this if we want the weighted national figures to add up to the provincial marginal totals, with weights that are as close as possible to a constant, in order to preserve the more reliable national distribution. The national distribution of missed persons may not be appropriate if the distribution of strata sizes is not the same for Canada as it is for the province. Therefore, a better alternative is to set the $\mathbf{w}_0$ element that corresponds to stratum h^* to

$$w_{0h^*} = N_{h^*} \Big/ \sum_{h \in S_{h^*}} N_h, \quad (6.4)$$

where S_{h^*} is the set of the 10 strata (one per province) similar to stratum h^* (for example, the 10 strata of not-married males aged 15 to 24).

We could remove two constraints because the corresponding rows of $\mathbf{M}$ are linear combinations of the others (for example, the fourth row and the last row), but the solution (6.3) is sufficiently general that their removal is unnecessary. With $\mathbf{A} = \mathbf{MX}$, $\mathbf{U} = \mathbf{X}$ and $\mathbf{T}$ equal to the identity matrix, (6.3) simplifies to

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{M}'(\mathbf{MXM}')^{\dagger}(\mathbf{b} - \mathbf{MXw}_0)\,. \qquad (6.5)$$

The smoothed values for each stratum are the elements of the vector $\mathbf{Xw}$.

A similar problem can arise for non-respondents when we want to smooth the sample design's effects.

7. Conclusion

There is much overlap between the two traditional goals of RRC sample allocation, which are to obtain a minimum variance for the national estimated undercoverage rate (goal I) and to obtain equal variances for the estimated provincial undercoverage rates (goal II), and the two additional goals considered in this paper, which are to minimize the variance of the total equalization payment (goal III) and to obtain equal CVs for the beneficiary provinces' population estimates (goal IV). Nevertheless, the explicit consideration of those two additional goals may allow the sample sizes for Quebec and Alberta to vary independently from those of the other provinces. The method suggested in this paper to achieve a compromise between different allocations that is optimal with respect to the various goals, is to take, for each province, the maximum sample size over each of the distinct allocations. The method provides a more direct justification for the allocation.

A comparison of the GES standard errors with the standard errors derived from the approximation formula (3.6) shows for the 2001 RRC, n sampled units with a response rate of r are equivalent to only $n \times r^3$ full-response units.

Optimal subprovincial allocation requires smoothing of provincial parameters at the age × sex × marital status level. Calibration can be a convenient method to scale more stable national age × sex × marital status values so that they add up to provincial age values, sex values and marital status values. The method's principal goal is reminiscent of the principal goal of the raking ratio method, but a solution such as the one described in Théberge (1999), which deals with the possibility that the constraints may not be satisfied, avoids convergence problems. In addition, using the Moore-Penrose inverse prevents collinearity problems.

References

Brackstone, G.J., and Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.

Clark, C. (september 2000). 2001 Reverse Record Check: Provincial and Territorial sample allocation. Document non publié. Ottawa. Statistique Canada.

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

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

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 87-96
Statistics Canada, Catalogue No. 12-001-XIE

Sample Size Calculation for Small-Area Estimation

Nicholas Tibor Longford [1]

Abstract

We describe a general approach to setting the sampling design in surveys that are planned for making inferences about small areas (sub-domains). The approach requires a specification of the inferential priorities for the areas. Sample size allocation schemes are derived first for the direct estimator and then for composite and empirical Bayes estimators. The methods are illustrated on an example of planning a survey of the population of Switzerland and estimating the mean or proportion of a variable for each of its 26 cantons.

Key Words: Efficiency; Inferential priority; Sample size allocation; Small-area estimation.

1. Introduction

Sampling design is a key device for efficient estimation and other forms of inference about a large population when the resources available do not permit collecting the relevant information from every member of the population. In this context, efficiency is interpreted as the optimal combination of a sampling design and an estimator of a population quantity θ. By optimum we understand minimum mean squared error, although the development presented in this paper can be adapted for other criteria. The pool of the possible sampling designs is delimited by the resources, and these are usually expressed in terms of a fixed sample size. This is not always appropriate because the designs may not entail identical average costs per subject. However, within a limited range of designs, this issue can be ignored.

The problem of setting the sampling design for the purpose of efficient estimation of a single quantity is well understood, and solutions are available for many commonly encountered settings. Most of them involve a univariate constrained optimisation problem. Setting the sampling design for estimating several quantities represents a quantum leap in complexity, because the problem involves several factors, typically one for each quantity. It is essential to optimise the design simultaneously for all the factors, because the goals of efficient inference about the target quantities may be in conflict. For example, in small-area estimation, a more generous allocation of the sample size to one area has to be compensated by a less generous allocation to one or several other areas.

Small-area statistics have become an important research topic in survey methods in the last few decades (Fay and Herriot 1979; Platek, Rao, Särndal and Singh 1987; Ghosh and Rao (1994), Longford 1999; and Rao 2003), stimulated by increasing interest of government agencies, the advertising and marketing industry and the financial and insurance sector. At present, many large-scale surveys are designed for estimating national quantities but, sometimes almost as an afterthought, are used for inferences about small areas. This would be appropriate if the sampling designs optimal for small-area and national inferences were similar. We illustrate in this paper that this is not the case and that sampling design can be effectively targeted for small-area estimation, taking into account the goal of efficient estimation of national quantities. To avoid the trivial case, we assume that the areas have unequal population sizes. We apply the methods to the problem of planning inferences about the 26 cantons of Switzerland; their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million.

Literature on the subject of planning surveys for small-area estimation is rather sparse. An important contribution is Singh, Gambino and Mantel (1994). In one of the approaches they discuss, the planned sample size for the Canadian Labor Force Survey is split into two parts. One part is allocated optimally for the purpose of national (domain) estimation and the remainder optimally for small-area estimation. For the latter goal, equal subsample sizes are allocated to each area when the areas have equal within-area variances, the finite population correction can be ignored and the areas have equal survey costs per subject, but also when the targets of inference are area-level means. When the targets are population totals, equal allocation to the areas is not efficient, because it handicaps estimation for more populous areas. Even when proportions or rates (percentages) are estimated, the within-area variances depend on the population proportion, although the dependence is weak when all the proportions are distant from zero and unity. For more recent developments in sampling design for small-area estimation, see Marker (2001).

The next section describes the proposed approach based on minimising the weighted sum of the sampling variances (mean squared errors) of the planned estimators, with the

1. Nicholas Tibor Longford, Departament d'Económia i Empresa, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: NTL@SNTL.co.uk.

weights specified to reflect the inferential priorities. It is applied first to direct estimation of the area-level quantities. Then it is extended to incorporate the goal of national estimation, and, finally, to composite estimation in section 3. The concluding section 4 contains a discussion.

The remainder of this section introduces the notation used in the rest of the paper. We assume that area-level population quantities θ_d, $d = 1, ..., D$, are estimated by $\hat{\theta}_d$ with respective mean squared errors (MSE) v_d that are functions of the within-area subsample sizes n_d; $v_d = v_d(n_d)$. The overall sample size is denoted by n, and is assumed to be fixed. The population sizes are denoted by N (overall) and N_d (for area d). For brevity, we denote $\mathbf{n} = (n_1, ..., n_D)^\top$. Most population quantities θ are functions of a single variable, such as its mean, total, and the like. The variable may be recorded in the survey directly, or constructed from one or several such variables. Although our development is not restricted to such quantities, the motivation is more straightforward with them. An estimator of θ_d is said to be *direct* if it is a function of only the variable concerned on subjects in area d.

We assume that each direct estimator considered is unbiased. This is not particularly restrictive, as most direct estimators are naive estimators or are closely related to them. We assume that the sample sizes for the areas are under the control of the survey designer. This is the case in stratified sampling designs in which the strata coincide with the areas. In section 4, we discuss sampling designs in which such control cannot be exercised; they are particularly relevant for divisions of the country into many (hundreds of) areas.

2. Optimal Design for Direct Estimation

We resolve the conflict between the goals of efficient estimation of the area-level quantities θ_d by choosing the area-level sampling design that minimises the weighted sum of the sampling variances (MSEs),

$$\min_{\mathbf{n}} \sum_{d=1}^{D} P_d v_d, \qquad (1)$$

subject to the constraint of fixed overall sample size $n = \mathbf{n}^\top \mathbf{1}_D$; $\mathbf{1}_D$ is the vector of unities of length D. The coefficients P_d are called *inferential priorities*. Greater value of P_d (in relation to the values $P_{d'}$, $d' \neq d$) implies a greater urgency to reduce v_d, because the contribution of area d to the sum in (1) in magnified more than for the other areas.

The optimisation problem in (1) is solved by the method of Lagrange multipliers, or simply by substituting $n_1 = n - n_2 - ... - n_D$, so that the problem then involves $D-1$ functionally unrelated variables. The solution satisfies the condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

An analytical expression for the optimal subsample sizes n_d cannot be obtained in general, but when $v_d = \sigma_d^2 / n_d$, as in simple random sampling within areas, the solution is proportional to $\sigma_d \sqrt{P_d}$, that is,

$$n_d^\dagger = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + ... + \sigma_D \sqrt{P_D}}.$$

When the within-area variances σ_d^2 coincide, $\sigma_1^2 = ... = \sigma_D^2 = \sigma^2$, this simplifies further; the optimal sample sizes are proportional to $\sqrt{P_d}$ and do not depend on σ^2.

In most contexts, it is difficult to elicit a suitable set of priorities P_d, and so it is more constructive to propose a convenient parametric class of priorities $\mathbf{P} = (P_1, ..., P_D)^\top$ and illustrate their impact on the sample size allocation. We propose the priorities $P_d = N_d^q$ for $0 \le q \le 2$. For $q = 0$, inference is equally important for every area. With increasing q, relatively greater importance is ascribed to more populous areas. When $v_d = \sigma^2 / n_d$, the optimal sample size allocation for $q = 2$, $n_d^\dagger = n N_d / N$, is proportional to the population sizes in the areas, and so the same sampling design is optimal for national and area-level inferences. For $q > 2$ the sample size allocation is even more generous to the most populous areas, at the expense of less populous areas. As this is counterintuitive in the context of small-area estimation, the choice of an exponent $q > 2$ is probably never appropriate. A negative priority exponent q would be suitable for a survey that aims to focus on the least populous areas. Of course, such a design is very inefficient for estimating the national quantity θ, especially when the areas have widely dispersed population sizes.

The inferential priorities P_d may be functions of quantities other than N_d. For example, the sizes of certain subpopulations of focal interest, such as an ethnic minority in the area, may be used instead of N_d, P_d may be defined differently in the country's regions, or the formula for them may be overriden for one or a few areas.

In some publications of survey analyses, an estimate is reported only when it is based on a sufficiently large sample size or its coefficient of variation (the ratio of the estimated standard error and the estimate) is smaller than a specified threshold. If a 'penalty' for not reporting a quantity is specified, it can be incorporated in the definition of the inferential priorities. The difficulty that may arise is that the objective function in (1) is discontinuous and the standard approaches to its optimisation are no longer applicable. The penalty has to be set with care. If it is too low it is ineffective; if it is set too high the solution will prefer reporting estimates for as many areas as possible, but each with sample size or precision that narrowly exceeds the set

threshold. See Marker (2001) for an alternative approach to this problem.

Figure 1 illustrates the impact of the priority exponent q on the sample size allocation for a survey planned in Switzerland, with the aim of estimating the population means of a variable in its 26 cantons, assuming a common within-canton variance σ^2. The planned overall sample size is $n = 10,000$. The curves in either panel connect the optimal sample sizes for each exponent q; they are drawn on the linear scale (on the left) and on the log scale (on the right). The population sizes are marked on the horizontal bar at the bottom of each plot. On the log scale, the curves are linear. The log scale is useful also because the population sizes of the cantons are more evenly distributed on it.

For $q = 0$, each canton is allocated the same sample size, $10,000/26 = 385$, and for $q = 2$ the allocation is proportional to the canton's population size. For intermediate values of q, sample sizes of the least populous cantons are boosted in relation to proportional allocation $(q = 2)$, at the expense of reduced allocation to the most populous cantons. The subsample sizes depend very little on q for cantons with population of about 250,000, approximately 3% of the national population size.

2.1 The Priority for National Estimation

As the canton-level subsample sizes differ from the proportional allocation for priority exponent $q < 2$, optimal canton-level estimation is accompanied by a loss of efficiency of the national estimator. Consider the stratified estimator

$$\hat{\theta} = \frac{1}{N}\sum_{d=1}^{D} N_d \hat{\theta}_d$$

of the national mean θ of a variable, where $\hat{\theta}_d$ are unbiased estimators of the within-canton means of the same variable. Assuming stratified sampling with simple random sampling within strata (cantons), with $\hat{\theta}_d$ set to the within-stratum sample means,

$$\mathrm{var}(\hat{\theta}) = \frac{1}{N^2}\sum_{d=1}^{D} \frac{N_d^2}{n_d}(1 - f_d)\sigma_d^2,$$

where $f_d = n_d / N_d$ is the finite population correction.

Figure 2 displays the function that relates the standard error $\sqrt{\mathrm{var}(\hat{\theta})}$ to the priority exponent q, calculated assuming $\sigma^2 = 100$. The standard error is a decreasing function of q; it decreases more steeply at $q = 0$ than at $q = 2$, where it is quite flat. For $q = 2$, the goals of canton-level and national estimation are in accord, and $\sqrt{\mathrm{var}(\hat{\theta})} = 0.100$. For $q = 0$, $\sqrt{\mathrm{var}(\hat{\theta})} = 0.143$; in this setting, optimality of the small-area estimation exerts a considerable toll on national estimation, equivalent to halving the sample size $(0.143/0.100 \doteq \sqrt{2})$. For negative q, the toll is even greater.

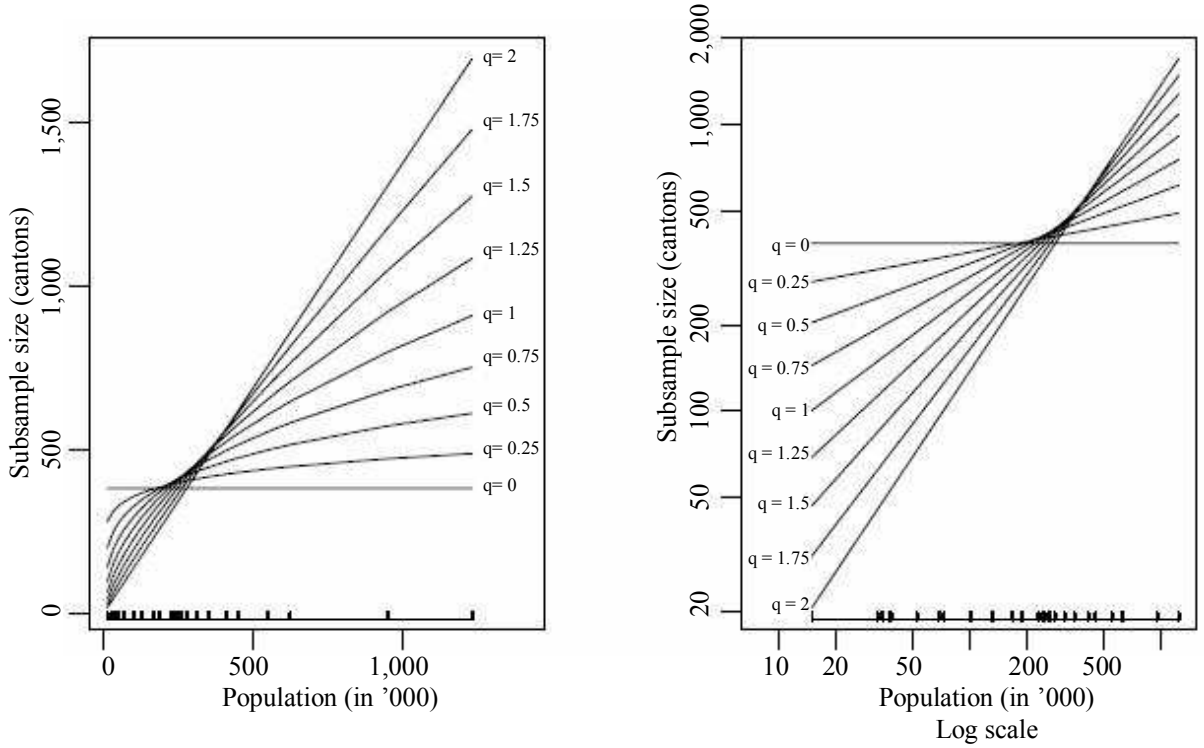


Figure 1. The sample size allocation to the Swiss cantons for a range of priority exponents q. The population sizes of the cantons are marked on the horizontal bar at the bottom of each plot.

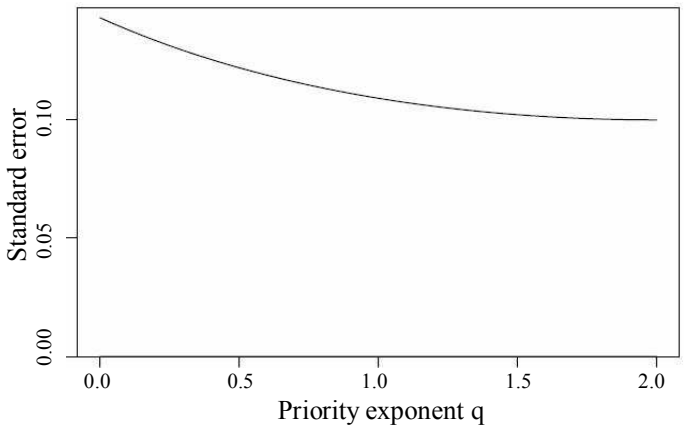


Figure 2. The standard error of the national estimator $\hat{\theta}$ of the mean of a variable, as a function of the exponent q for priorities of the canton-level estimation.

Thus, the need for efficiency of the national estimator can be addressed by increasing the priority exponent. For example, the parties with rival inferential interests may negotiate about how much loss in efficiency of $\hat{\theta}$ can be afforded, and the priority exponent would then be set to match this loss. Alternatively, this loss may be considered by applying the optimal design for area-level estimation. If it is regarded as excessive, q is increased until a balance is struck between the losses of efficiency for national and small-area estimation.

An unsatisfactory feature of these approaches is that they compromise the original purpose of the priorities $\mathbf{P}-$ to reflect the relative importance of the inferences about the distinct small areas. This drawback is addressed by associating $\hat{\theta}$ with a priority, denoted by G, relative to small-area estimation, and considering optimal estimation of the set of D area-level targets θ_d together with the national target θ. Thus, we minimise the objective function

$$\sum_{d=1}^{D} P_d v_d(n_d) + GP_+ v(\mathbf{n}),$$

where $v = \mathrm{var}(\hat{\theta})$ and $P_+ = \mathbf{P}^{\mathsf{T}}\mathbf{1}_D$. The factor P_+ is introduced to ameliorate the effect of the absolute sizes of P_d and the number of areas on the relative priority G. The priorities P_d can be interpreted only by their relative sizes, as, for any constant $c > 0$, P_d and cP_d correspond to identical sets of priorities for small-area estimation in (1).

When the sampling design within each area is simple random and $\hat{\theta}$ is the standard stratified estimator, the minimum is attained when

$$\sigma_d^2 \frac{P'_d}{n_d^2} = \text{const},$$

where $P'_d = P_d + GP_+ N_d^2 / N^2$. The optimal sample sizes for the areas are

$$n_d^* = n \frac{\sigma_d \sqrt{P'_d}}{\sigma_1 \sqrt{P'_1} + \ldots + \sigma_D \sqrt{P'_D}}.$$

This corresponds to an adjustment of the priorities P_d by $GP_+ N_d^2 / N^2$. Note that this adjustment is neither additive nor multiplicative. The priority is boosted more for the more populous areas. As a consequence, the area-level subsample sizes are dispersed more when the relative priority for national estimation is incorporated and the area-level priorities are unchanged. The finite population correction has no impact on n_d^* because it reduces each sampling variance v_d and v by a quantity that does not depend on $\mathbf{n}$.

The priority G can be set by insisting that the loss of efficiency in estimating the national quantity θ does not exceed a given percentage or that at most a few (or none) of the absolute differences $|P'_d - P_d|$ or log-ratios $|\log(P'_d / P_d)|$ are very large. However, the analytical problem is simple to solve, so the survey management can be presented by the sampling designs that are optimal for a range of values of G.

The dependence of the subsample size on the exponent q and relative priority G is plotted in Figure 3 for the least and most populous cantons, Appenzell-Innerrhoden and Zürich, in the respective panels A and C. Panels B and D plot the same curves as A and C, respectively, on the log scale. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small-area estimation to very large values of G. Throughout, we assume that $n = 10,000$ and $\sigma^2 = 100$, common to all cantons.

For each exponent $q < 2$, the sample-size curve $n_d(G)$ decreases for the less populous and increases for the more populous cantons toward the proportional representation $n_d = nN_d / N$, which corresponds to $q = 2$. On the linear scale, the increase is quite rapid for Zürich for small q and G, whereas the reduction for Appenzell-Innerrhoden is more gradual. As the relative priority G is reduced, the excess sample size is re-distributed from Zürich (and a few other populous cantons) to several less populous cantons.

Figure 4 plots the 'national' standard error $\sqrt{\mathrm{var}(\hat{\theta})}$ under the optimal sample allocation for an array of values of q and G. The diagram shows that the standard error of $\hat{\theta}$ is reduced radically by a small increase of G in the vicinity of $G = 0$, whereas for larger values of G it is affected only slightly. For each G, higher priority exponent q is associated with higher precision of $\hat{\theta}$.

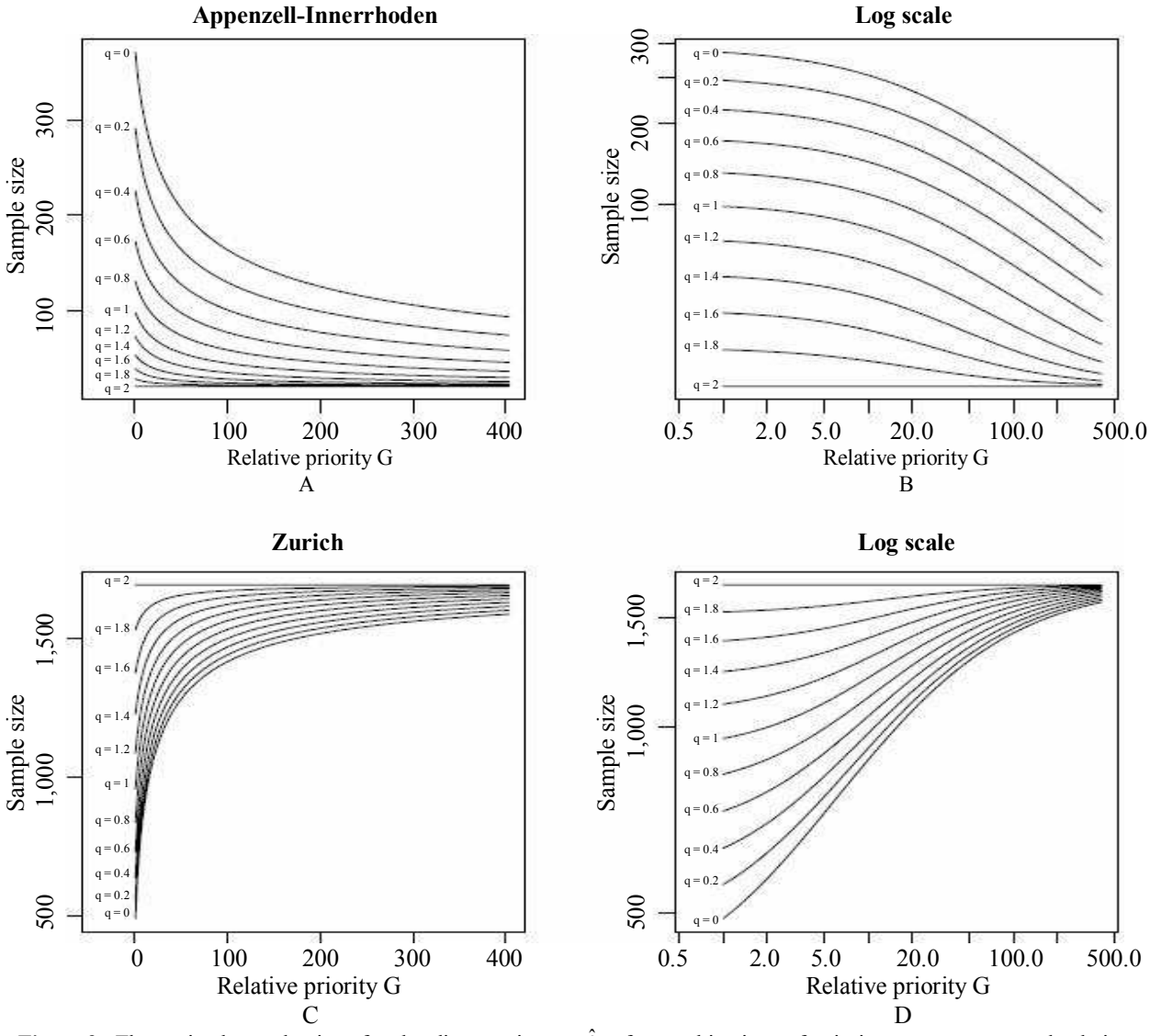


Figure 3. The optimal sample sizes for the direct estimator $\hat{\theta}_d$ for combinations of priority exponents q and relative priorities G for the least and most populous cantons.

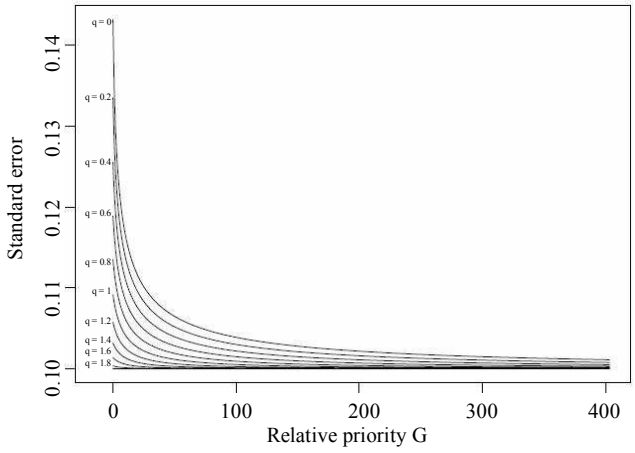


Figure 4. The standard error of the national estimator for the allocation that is optimal under an array of priorities given by q and G.

3. Composite Estimation

The resources available for the conduct of a survey are used most effectively by the optimal combination of a sampling design and estimator(s), and so the sampling design and (the selection of) the estimator should be, in ideal circumstances, optimised simultaneously. This problem is difficult to solve formally in most settings, although some estimators are more efficient than their competitors in a wide range of designs. Composite estimators (Longford 1999, 2004) are one such class. They are convex combinations of the direct small-area and national estimators,

$$\tilde{\theta}_d = (1 - b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}, \qquad (2)$$

with area-specific coefficients b_d that are estimates of the optimum. The composition $\tilde{\theta}_d$ exploits the similarity of the areas; it is particularly effective when the areas have a small between-area variance $\sigma_B^2 = D^{-1}\sum_d (\theta_d - \bar{\theta})^2$, where $\bar{\theta} = D^{-1}\sum_d \theta_d$. This variance is defined over the D population quantities θ_d and is unaffected by the sampling design. In practice, σ_B^2 has to be estimated. When planning a survey, estimates from other surveys of the same or a related population have to be used, and the uncertainty about σ_B^2 addressed. This can be done by sensitivity analysis, exploring the optimal designs for a range of plausible values of σ_B^2.

If the deviations $\Delta_d = \theta_d - \bar{\theta}$ were known the optimal coefficient b_d in (2) would be, approximately, $b_d^* = \sigma_d^2 / (\sigma_d^2 + n_d \Delta_d^2)$. As Δ_d is not known (otherwise θ_d would be estimated with high precision by $\bar{\theta} + \Delta_d$), we replace Δ_d^2 by its average over the areas, equal to σ_B^2, yielding the coefficient $b_d = 1/(1 + n_d \omega_d)$, where $\omega_d = \sigma_B^2 / \sigma_d^2$ is the variance ratio. The variance σ_B^2 also has to be estimated, but when there are many areas it is estimated with precision much higher than most Δ_d^2 are.

If the coefficients b_d are estimated with sufficient precision the composite estimator $\tilde{\theta}_d$ is more efficient than the two constituent estimators $\hat{\theta}_d$ and $\hat{\theta}$. Ignoring the uncertainty about the within- and between-area variances, as well as the national mean $\bar{\theta}$ and the correlation between the national and area-level (direct) estimators, the average MSE of $\tilde{\theta}_d$ is

$$\mathrm{aMSE}(\tilde{\theta}_d) = \frac{\sigma_B^2}{1 + n_d \omega_d}, \qquad (3)$$

where 'aMSE' denotes the MSE in which Δ_d^2 is replaced by σ_B^2, its average over the areas. The aMSE in (3) is also an approximation to the conditional variance of the EBLUP estimator of the area-level mean based on the two-level (empirical Bayes) model (Longford 1993, Goldstein 1995, Marker 1999, and Rao 2003). See Ghosh and Rao (1994) for an authoritative review of application of these models to small-area estimation.

For the composite estimators of the area-level means, we search for the sample allocation that minimises the objective function

$$\sum_{d=1}^{D} P_d \, \mathrm{aMSE}(\tilde{\theta}_d) + G P_+ v.$$

The solution satisfies the condition

$$\frac{N_d^q \sigma_B^2 \omega_d}{(1 + n_d \omega_d)^2} + G P_+ \frac{N_d^2}{N^2} \frac{\sigma_d^2}{n_d^2} = \mathrm{const.} \qquad (4)$$

This equation does not have a convenient closed-form solution, but iterative schemes can be applied to solve it. The value of n_1 determines the remaining sample sizes n_d, and so optimisation corresponds to a one-dimensional search. If the provisional sample sizes $\mathbf{n}$ based on a set value of n_1 are too large, $\mathbf{n}^\top \mathbf{1}_D > n$, n_1 is reduced and the other sample sizes n_d are calculated by solving (4). Note that the solution depends on the variances σ_d^2 and σ_B^2. The problem is simplified somewhat when the areas have a common variance $\sigma^2 = \sigma_1^2 = \ldots = \sigma_D^2$. Then the solution of (4) depends on the variances only through the ratio $\omega = \sigma_B^2 / \sigma^2$ because σ^2 is a multiplicative factor and has no impact on the optimisation.

By way of an example, suppose $q = 1$ and $G = 10$ in planning a survey of the population of Switzerland with $n = 10{,}000$, and $\omega = 0.10$ is assumed. As the initial solution, we use the allocation optimal for direct estimation with the same values of q and G. One iteration updates the sample size for each canton and, within it, the updating for all but the arbitrarily selected reference canton $d = 1$ is also iterative. The reference canton's provisional subsample size determines the current value of the constant on the right-hand side of (4). Then equation (4) is solved, iteratively, for each canton $d = 2, \ldots, D$, using the Newton method. In the application, the number of these iterations was in single digits for each canton. Finally, the subsample size for the reference canton is adjusted by the $1/D$-multiple of the difference between the current total of the subsample sizes and the target total n. The updating of the cantons is itself iterated, but only a few iterations are required to achieve convergence; for example, all the changes in the subsample sizes were smaller than 1.0 after three iterations, and smaller than 0.01 after eight iterations. The convergence is fast because the starting solution is close to the optimum; the largest difference between the two subsample sizes is for Zürich, 20.0 (from 1199.5 at the start to 1219.5 after eight iterations). For Appenzell-Innerrhoden, the sample size is reduced from 81.6 to 73.4. Change by less than unity takes place for five cantons with population sizes in the range 228,000–278,000. Note that the subsample sizes would in practice be rounded, and possibly adjusted further to conform with various survey management constraints.

No priority for national estimation

If national estimation has no priority, $G = 0$, equation (4) has the explicit solution

$$n_d^* = \frac{n\omega + D}{\omega} \frac{N_d^{q/2}}{U^{(q)}} - \frac{1}{\omega},$$

where $U^{(q)} = N_1^{q/2} + ... + N_D^{q/2}$. This allocation is related to the allocation $n_d^{\dagger}$, $d = 1, ..., D$, that is optimal for direct estimation of θ_d by the identity

$$n_d^* = n_d^{\dagger} + \frac{1}{\omega}\left(\frac{DN_d^{q/2}}{U^{(q)}} - 1\right).$$

Hence, when $q > 0$, the allocation optimal for composite estimation is more dispersed than for direct estimation. The break-even population size is $N_{\mathrm{T}} = (U^{(q)}/D)^{2/q}$; areas with population sizes $N_d < N_{\mathrm{T}}$ have smaller subsample sizes for composite than for direct estimation, and areas with greater population sizes have greater subsample sizes. (For $q = 0$, $n_d^* \equiv n/D$). The amount of extra dispersion is inversely proportional to ω.

For $\omega = 0$, the equations for the optimal sampling design lead to a singularity. In this case, each θ_d is estimated efficiently by the national estimator $\hat{\theta}$, and so the design optimal for composite estimation coincides with the design that is optimal for the national estimator $(n_d^* = nN_d/N)$. For $q > 0$, the optimal allocation yields negative sample sizes n_d^* when

$$N_d < \left\{\frac{U^{(q)}}{n\omega + D}\right\}^{2/q}. \qquad (5)$$

This (formal) solution is not meaningful. A negative solution should come as no surprise because the aMSE in (3) is an analytical function for $n_d > -1/\omega_d$. For small $\omega > 0$, the aMSE is a shallow decreasing function of the sample size n_d. A negative n_d^* indicates that a (small) canton is not worth sampling because of its low inferential priority P_d. Although additional sample size for a more populous canton d' may yield a smaller reduction of aMSE than it would for a small canton d, its impact is magnified by the larger priority $P_{d'}$.

Positive priority for the national mean

The aMSE in (3) ignores the uncertainty about the national mean θ, and this becomes acute when one of the cantons is not represented in the sample. This deficiency of (3) can be compensated for by setting the relative priority G to a positive value.

Figure 5 summarises the impact of the relative priority G and the priority exponent q on the optimal sample sizes of the least and most populous cantons, together with canton Thurgau which has the 13th (median) largest population size, 228,000. Each setting of q, indicated in the title, and G, using different line types, is represented for a canton by a graph of the optimal sample size as a function of the variance ratio ω. The limit of this function for $\omega \to +\infty$, equal to the sample size optimal for direct estimation, is marked by a bar at the right-hand margin of the panel. For $\omega = 0$, the sampling design optimal for estimation of the national mean θ is obtained. Panels A and B at the top are for the overall sample size $n = 10,000$ and panels C and D for $n = 1,000$.

The diagram shows that the optimal sample sizes are nearly constant in the range $\omega \in (\omega^*, +\infty)$; ω^* increases with q, G and $1/n$. This is a consequence of the relatively large sample size n, which ensures that the subsamples of most cantons are too large for any substantial borrowing of strength across the cantons to take place, unless the cantons are very similar $(\omega < \omega^*)$. Most shrinkage coefficients $b_d = 1/(1 + n_d\,\omega)$ are very small. When $n = 10,000$ is planned, for small values of ω, the optimal sample size increases steeply for the least populous canton and drops precipitously for the most populous canton. Dispersion of the optimal sample sizes increases with q and G, converging to the optimal allocation for estimating the national mean θ, which corresponds to $\omega = 0$. In contrast, the optimal sample sizes are discontinuous at $\omega = 0$ when $G = 0$; the solutions diverge to $-\infty$ for the least populous cantons.

In panels C and D, for $n = 1,000$, the dependence of the sample sizes on ω persists over a wider range of ω because there is a greater scope for borrowing strength across the cantons with the smaller sample sizes. The optimal sample sizes are not monotone functions of ω; for the least populous cantons there is a dip at small values of ω. The dip is more pronounced for small G and large q, that is, when the disparities of the cantons' priorities are greater and inference about the national mean is relatively unimportant. This phenomenon, somewhat exaggerated by the log-scale of the vertical axis, is similar to the case discussed for $G = 0$. Because of the disparity in the priorities P_d, a small reduction of aMSE for a more populous canton is preferred to a greater reduction for a less populous canton. The dip is present also when $n = 10,000$, but it is so shallow and narrow as to be invisible with the resolution of the graph. Note that the horizontal axes in panels C and D have three times wider range of values of ω than in panels A and B.

In the context of the planned survey, it was agreed that ω is unlikely to be smaller than 0.05. Therefore, the sample size calculations could be based on the direct estimator.

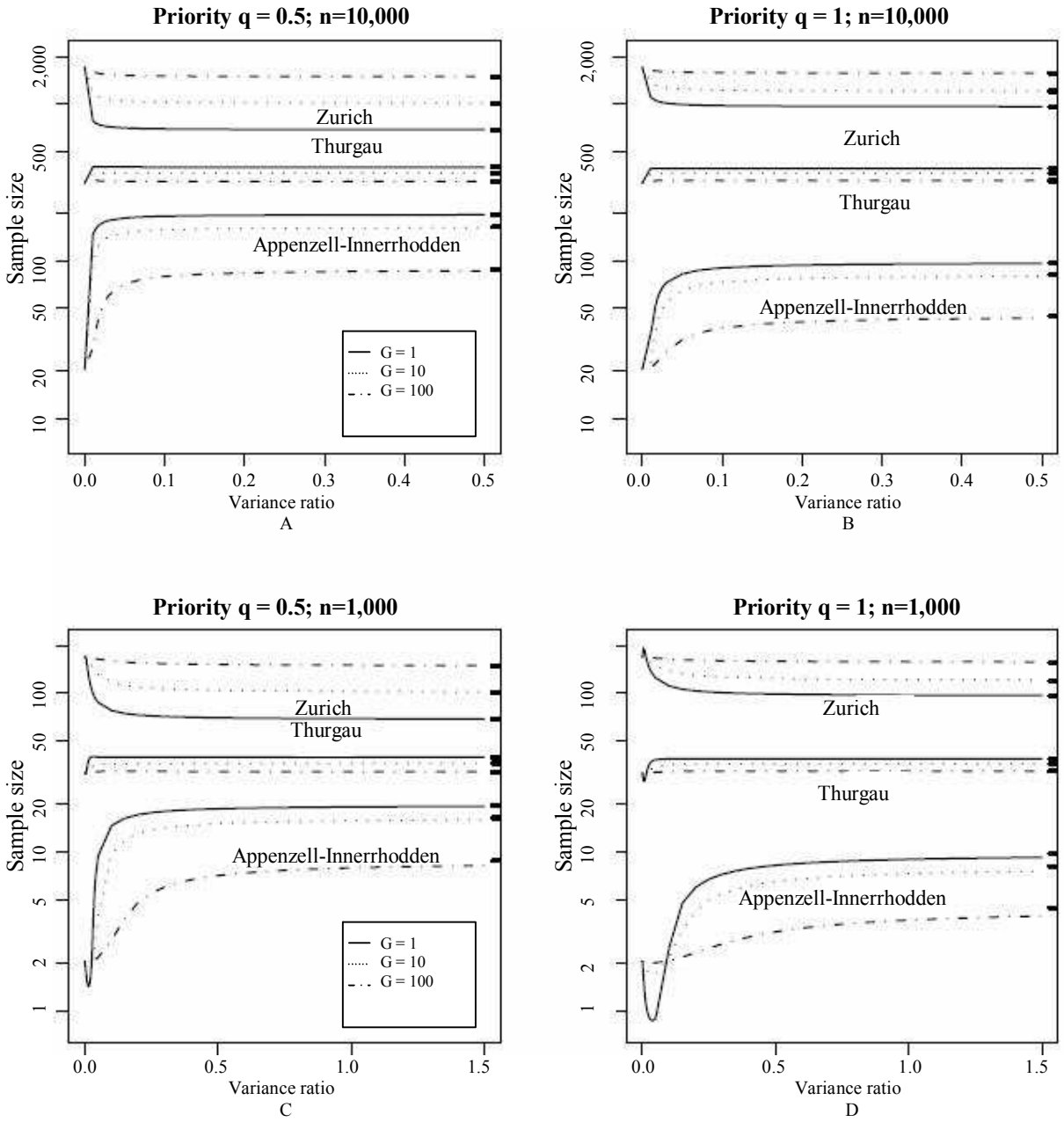


Figure 5. The sample sizes optimal for composite estimation of the population means for three cantons for a range of variance ratios ω, priority exponents $q = 0.5$ and $q = 1.0$ and relative priorities $G = 1$, 10 and 100. The overall sample sizes are 10,000 (panels A and B) and 1,000 (panels C and D).

4. Discussion

The method described in this paper identifies the optimal design for the artificial setting of stratified sampling with simple random sampling within homoscedastic strata. Specifying the priorities for small-area and national estimation is a key element of the method. In practice, the priorities may be difficult to agree on, and some of the assumptions made may be problematic, the assumptions of equal within-stratum variances and simple random sampling in particular. The method can be extended to more complex estimators, but then the values of further parameters are required. A more constructive approach regards the optimal sampling design for the simplified setting as an approximation to the sampling design that is optimal for the more realistic setting. Even if the optimal sampling design were identified, it could not be implemented literally, because of imperfections in the sampling frame and (possibly) informative and unevenly distributed nonresponse. However, the approach can be applied, in principle, to any small-area

estimator that has an analytical expression for the exact or approximate MSE. This includes all estimators based on empirical Bayes models, to which the composite estimator is closely related. Sampling weights can be incorporated in sample size calculation if they, or their within-area distributions, are known, subject to some approximation, in advance. Sample size calculation for a single (national) quantity entails the same problem.

Although the numerical solution of the problem for composite estimation with a positive priority G is simple and involves no convergence problems, it is advantageous to have an analytical solution, so that a range of scenarios can be explored. The proximity of the solutions for the direct and composite estimation suggests that the allocation optimal for direct estimation may be close to optimum also for composite estimation with realistic values of ω, say, $\omega > 0.05$.

Various management and organisational constraints are another obstacle to the literal implementation of an analytically derived sampling design. In household surveys, it is often preferable to assign an (almost) full quota of addresses to each interviewer, and so sample sizes that are multiples of the quota are preferred. These and numerous other constraints can be incorporated in the optimization problem, although they are often difficult to quantify or the designer may not be aware of them because of imperfect communication. Improvisation, after obtaining the sampling design that is optimal for a simpler setting, may be more practical. Also, priorities, or expert opinion about them, may change over time, even while the survey is being conducted and analysed. Estimates that are associated with standard errors or coefficients of variation greater than a specified threshold are often excluded from analysis reports. Intention to do this can be reflected in sample size calculation by regarding $\hat{\theta}$ as the estimator of θ_d, that is, by setting the associated MSE to the corresponding aMSE $\sigma_B^2 + \mathrm{var}(\hat{\theta})$ or to another (large) constant.

Although we propose a particular class of priorities for the small areas, no conceptual difficulties arise when another class is used instead. It may depend on several population quantities, not only the population size. In principle, the priorities can also be set for the areas individually, although this is practical only when the number of areas is small. The formula-based and individually set priorities can be combined by adjusting the priorities, such as $P_d = N_d^q$, for a few areas to reflect their exceptional role in the analysis.

Sensitivity analysis, exploring how the sampling design is changed as a result of altered input, is essential for understanding the impact of uncertainty about the estimated parameters (the variance ratio ω in particular) and the arbitrariness, however limited, in how the priorities are set. For this, an analytically simple solution that can be executed many times, for a range of settings, is preferred to a more complex solution, the properties of which are more difficult to explore.

Multivariate composite estimators exploit the similarity not only across areas, but also across (auxiliary) variables, time, subpopulations, and the like (Longford 1999 and 2005). The aMSEs of these estimators depend on the scaled variance matrix $\boldsymbol{\Omega}$, the multivariate counterpart of ω. Sample size calculation for this method is difficult to implement directly because both variances and covariances in $\boldsymbol{\Omega}$ are essential to the efficiency of the estimators. A more constructive approach matches the matrix $\boldsymbol{\Omega}$ with a ratio ω that can be interpreted as the similarity of the areas after adjusting for the auxiliary information, as in empirical Bayes methods.

When control over the sample sizes allocated to the areas is not possible sample size calculation is still meaningful as a guide for how the sample sizes should be allocated *on average*. In general, a unit reduction of the sample size is associated with greater loss of precision than a unit increase. Therefore, designs in which the sampling (replication) variance of the subsample sizes n_d (d fixed) is smaller are better suited for small-area estimation. In designs with large clusters, such variances are large because, at an extreme, an area may not be represented in the survey in some replications and may be over-represented several times in others. Using smaller clusters is in general preferable for small-area estimation if this does not inflate the survey costs and a fixed overall sample size can be maintained.

Acknowledgements

I am grateful to the Deputy Editor and referees for suggesting several improvements but mainly for leading me to discover an error in an earlier version of the manuscript. Discussions with the Polish team in the EURAREA project are acknowledged.

References

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

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Goldstein, H. (1995). *Multilevel Statistical Models.* Second Edition. Edward Arnold, London, UK.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society*, Series A, 162, 227-245.

Longford, N.T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society*, Series A, 167, 341-373.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.

Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Eds.) (1987). S*mall Area Statistics*. New York: John Wiley & Sons.

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

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 97-103
Statistics Canada, Catalogue No. 12-001-XIE

Small Area Estimation Using Area Level Models and Estimated Sampling Variances

Yong You and Beatrice Chapman [1]

Abstract

In small area estimation, area level models such as the Fay–Herriot model (Fay and Herriot 1979) are widely used to obtain efficient model-based estimators for small areas. The sampling error variances are customarily assumed to be known in the model. In this paper we consider the situation where the sampling error variances are estimated individually by direct estimators. A full hierarchical Bayes (HB) model is constructed for the direct survey estimators and the sampling error variances estimators. The Gibbs sampling method is employed to obtain the small area HB estimators. The proposed HB approach automatically takes account of the extra uncertainty of estimating the sampling error variances, especially when the area-specific sample sizes are small. We compare the proposed HB model with the Fay–Herriot model through analysis of two survey data sets. Our results have shown that the proposed HB estimators perform quite well compared to the direct estimates. We also discussed the problem of priors on the variance components.

Key Words: Gibbs sampling; Hierarchical Bayes; Prior sensitivity; Sample size; Variance components.

1. Introduction

Sample surveys, for most purposes, are usually designed to provide reliable direct estimates for total populations and large areas by using area-specific sample data. These direct estimates frequently fail to provide reliable estimates for small areas due to very small sample sizes in the areas. Since small area estimates often have unsuitably large standard errors, to gain precision and reliability it is necessary to "borrow strength" from related areas thus increasing the effective sample size to construct indirect estimates for the small areas (Rao 1999). Explicit model-based methods that use supplementary data such as census and administrative data associated with the small areas in explicit models to link the small areas have been widely used in practice to obtain reliable model-based estimators. There are two broad classifications for these models: area level models and unit level models. Area level models are based on area direct survey estimators and unit level models are based on individual observations in the areas. For overviews and appraisals of models for small area estimation, see Rao (1999, 2003). In this paper we study area level models.

To obtain a basic area level model we assume that the small area parameter of interest θ_i is related to area-specific auxiliary data $x_i = (x_{i1}, \ldots, x_{ip})'$ through a linear model

$$\theta_i = x_i'\beta + v_i,\ i = 1, \ldots, m, \qquad (1)$$

where m is the number of small areas, $\beta = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients, and the v_i's are area-specific random effects assumed to be independent and identically distributed (iid) with $\mathrm{E}(v_i) = 0$ and $\mathrm{var}(v_i) = \sigma_v^2$. The assumption of normality may also be included. This model is referred to as a linking model for θ_i.

The basic area level model also assumes that given the area-specific sample size $n_i > 1$, there exists a direct survey estimator y_i (usually design unbiased) for the small area parameter θ_i such that

$$y_i = \theta_i + e_i,\ i = 1, \ldots, m, \qquad (2)$$

where the e_i is the sampling error associated with the direct estimator y_i. We also assume are that the e_i's are independent normal random variables with mean $E(e_i \mid \theta_i) = 0$ and sampling variance $\mathrm{var}(e_i \mid \theta_i) = \sigma_i^2$. Combining models (1) and (2) lead to a linear mixed area level model

$$y_i = x_i'\beta + v_i + e_i,\ i = 1, \ldots, m. \qquad (3)$$

The well-known Fay–Herriot model (Fay and Herriot 1979) in small area estimation has the form of model (3) with the sampling variance σ_i^2 assumed to be known in the model. This is a very strong assumption. Usually a smoothed estimator of σ_i^2 is used in the model and then treated as known. In this paper, we consider the situation where the sampling variances σ_i^2 are unknown and are estimated by unbiased estimators s_i^2. Following Rivest and Vandal (2002) and Wang and Fuller (2003), we assume that the estimators s_i^2 are independent of the direct survey estimators y_i and s_i^2 has a sampling distribution $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the i^{th} area. For example, suppose we have n_i observations from small area i and these observations are iid $N(\mu_i, \sigma^2)$. Let y_i be the sample mean of the n_i observations. Then $y_i \sim N(\mu_i, \sigma_i^2)$ and $\sigma_i^2 = \sigma^2 / n_i$. Then we can obtain a direct estimator of σ_i^2 as $s_i^2 = \tau_i^2 / n_i$, where τ_i^2 is the sample

1. Yong You and Beatrice Chapman, Household Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6. E-mail: yongyou@statcan.ca.

variance of the n_i observations. Also y_i and s_i^2 are independent and $(n_i - 1)s_i^2 \sim \sigma_i^2 \chi^2_{n_i-1}$.

We are interested in estimating the small area parameters θ_i. Rivest and Vandal (2002) and Wang and Fuller (2003) obtained the empirical best linear unbiased prediction (EBLUP) estimators of θ_i and the associated mean square error (MSE) approximations assuming that m and n_i are relatively large. In this paper, we consider a hierarchical Bayes (HB) approach using the Gibbs sampling method. An advantage of the HB approach is that it is straightforward, and the inferences for parameters θ_i are "exact" unlike the EBLUP approach. The small area parameter θ_i is estimated by its posterior mean and its precision is measured by its posterior variance. The HB approach automatically takes account of the uncertainties associated with unknown parameters in the model. Section 2 presents the HB area level models and related Gibbs sampling inferences. Section 3 presents two survey data analysis and sensitivity analysis. And finally in section 4, we offer some conclusions and future work directions.

2. Hierarchical Bayes Approach

We now present the area level model (3) and the estimated sampling variances s_i^2 in a HB framework as follows:

Model 1

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2),\ i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi^2_{d_i},\ d_i = n_i - 1,\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Priors for the parameters: $\pi(\beta) \propto 1,\ \pi(\sigma_i^2) \sim \text{IG}(a_i, b_i), i = 1, \ldots, m,\ \pi(\sigma_v^2) \sim \text{IG}(a_0, b_0),$ where $a_i, b_i\ (0 \le i \le m)$ are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2. IG denotes the inverse gamma distribution.

In Model 1, the sampling variances σ_i^2 are unknown. In practice however, we may have a simpler model by replacing σ_i^2 by its estimate s_i^2 (here s_i^2 is treated as a constant) and obtain the following model:

Model 2

- $y_i \mid \theta_i \sim \text{ind } N(\theta_i, \sigma_i^2 = s_i^2),\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Priors: $\pi(\beta) \propto 1,\ \pi(\sigma_v^2) \sim \text{IG}(a_0, b_0).$

Model 2 is actually the Fay-Herriot model with sampling variances known as s_i^2. If area-specific sample sizes n_i are small, using s_i^2 in Model 2 may lead to underestimation of the MSE under the EBLUP approach or the posterior variance under the HB approach. We are interested in evaluating the effects of using s_i^2 for σ_i^2 in the model. We will obtain the HB estimates of θ_i under both Model 1 and Model 2 and compare the HB estimates through real survey data analysis.

Under the HB approach, we use the posterior mean $E(\theta_i \mid y)$ as a point estimate for θ_i and the posterior variance $V(\theta_i \mid y)$ as a measure of variability, where $y = (y_1, \ldots, y_m)'$. To estimate $E(\theta_i \mid y)$ and $V(\theta_i \mid y)$, we employ the Gibbs sampling method (Gelfand and Smith 1990). From Model 1, we obtain the following full conditional distributions for the Gibbs sampler:

- $[\theta_i \mid y, \beta, \sigma_i^2, \sigma_v^2] \sim N(\gamma_i y_i + (1-\gamma_i)x_i'\beta, \gamma_i \sigma_i^2),$ where $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2},\ i = 1, \ldots, m;$
- $[\beta \mid y, \theta, \sigma_i^2, \sigma_v^2] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i \theta_i\right), \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right);$
- $[\sigma_i^2 \mid y, \theta, \beta, \sigma_v^2] \sim \text{IG}\left(a_i + \dfrac{d_i + 1}{2}, b_i + \dfrac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right),$ where $d_i = n_i - 1,\ i = 1, \ldots, m;$
- $[\sigma_v^2 \mid y, \theta, \beta, \sigma_i^2] \sim \text{IG}\left(a_0 + \dfrac{m}{2}, b_0 + \dfrac{1}{2}\sum_{i=1}^m (\theta_i - x_i'\beta)^2\right).$

It is straight forward to draw samples from these full conditional distributions. For implementations, we use $L = 5$ parallel runs each with a "burn-in" length of $B = 1{,}000$ and Gibbs sampling size of $G = 5{,}000$. The prior parameters a_i, b_i and a_0, b_0 are chosen to 0.0001. The HB estimator of θ_i under Model 1 is thus obtained as

$$\hat{\theta}_i^{\text{HB}} = (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})})\, x_i'\, \beta^{(\text{lg})}), \qquad (4)$$

where $\gamma_i^{(\text{lg})} = \sigma_v^{2(\text{lg})} / (\sigma_v^{2(\text{lg})} + \sigma_i^{2(\text{lg})})$, and the posterior variance of θ_i can be estimated by

$$\begin{aligned} \hat{V}(\theta_i) = {} & (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} \sigma_i^{2(\text{lg})}) \\ & + (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})})^2 \\ & - \left\{ (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})}) \right\}^2, \end{aligned} \qquad (5)$$

where $\{\beta^{(lg)}, \sigma_i^{2(lg)}, \sigma_v^{2(lg)}; g = 1, \ldots, G;\ l = 1, \ldots, L\}$ is the sample generated from the Gibbs sampler. The estimators (4) and (5) are the so-called Rao-Blackwellized HB estimators. The Rao–Blackwellized estimators are more stable in terms of simulation errors as shown, for example, in Gelfand and Smith (1991) and You and Rao (2000).

Now we consider Model 2. The full conditional distributions for the Gibbs sampler under Model 2 are

- $[\theta_i \mid y, \beta, \sigma_v^2] \sim N(\gamma_i y_i + (1-\gamma_i)x_i'\beta, \gamma_i s_i^2)$, where $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + s_i^2}$, $i = 1, \ldots, m$;
- $[\beta \mid y, \theta, \sigma_v^2] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right)$;
- $[\sigma_v^2 \mid y, \theta, \beta] \sim \text{IG}\left(a_0 + \dfrac{m}{2}, b_0 + \dfrac{1}{2}\sum_{i=1}^m (\theta_i - x_i'\beta)^2\right)$.

Under Model 2, the HB estimator of θ_i and the corresponding posterior variance estimator are given by (4) and (5) respectively with $\sigma_i^{2(lg)}$ replaced by s_i^2. Note that using s_i^2 instead of $\sigma_i^{2(lg)}$ may lead to severe underestimation of the posterior variance of θ_i for some areas with small sample sizes n_i. We will compare the HB estimators and evaluate the effects of using s_i^2 in Model 2 through data analysis in the following section.

3. Data Analysis

3.1 The Data Sets

We consider two interesting data sets in our analysis. The first data set is corn and soybean data with only 8 areas and small sample sizes in each area. The second data set is milk data with 43 areas and relatively large sample sizes in each area. We will compare the HB models and estimates based on these two data sets.

Corn and Soybean Data: The corn and soybean data comes from the U.S. Department of Agriculture and was first studied by Battese, Harter and Fuller (1988). The data contains reported crop hectares and LANDSAT satellite data for corn and soybeans in sample segments of 12 Iowa counties. The reported number of hectares for each crop comprise the direct survey estimates. Used as auxiliary data are the population means of number of pixels of a given crop per segment. The sample sizes are small for these areas, ranging from 1–5. For our purposes only the counties with a sample size of 3 and greater are used (8 areas meet the criteria). Therefore of the included counties the sample sizes range from 3–5. The original data is unit level data. In order to have area level data the sample mean and the sample standard error are calculated for each county. The sample standard errors for the corn and soybean data are quite large in general (yielding some CVs in the 0.3–0.4 range and one CV of 0.532) but by chance there are also some small values in some instances (for corn data, Franklin has standard error 5.704 and CV 0.036). Because the sample sizes are so small, these sample standard errors cannot be trusted to approximate the true standard errors. Table 1 presents the modified area level data for corn and soybeans from the unit level data of Battese *et al*. (1988).

Table 1
Modified Crop Area Level Data, from Battese, Harter and Fuller (1988)

		Corn			Soybeans		
County	n_i	y_i	SD	CV	y_i	SD	CV
Franklin	3	158.623	5.704	0.036	52.473	16.425	0.313
Pocahontas	3	102.523	43.406	0.423	118.697	50.290	0.424
Winnebago	3	112.773	30.547	0.271	88.573	10.453	0.118
Wright	3	144.297	53.999	0.374	97.800	52.034	0.532
Webster	4	117.595	21.298	0.181	112.980	23.531	0.208
Hancock	5	109.382	15.661	0.143	117.478	17.209	0.146
Kossuth	5	110.252	12.112	0.110	117.844	20.954	0.178
Hardin	5	120.054	36.807	0.307	101.834	26.790	0.263

Milk Data: The milk data, used in an article by Arora and Lahiri (1997), comes from the U.S. Bureau of Labor Statistics. The estimated values are the average expenditure on fresh milk for the year 1989. There is data for 43 areas with sample sizes ranging from 95 to 633. The CVs range from 0.074 to 0.341 over the 43 areas. A more detailed description of the data can be found in Arora and Lahiri (1997). For completeness, we give the data in Table 2. Following Arora and Lahiri (1997), we use $x_i'\beta = \beta_j$ if $i \in j^{\text{th}}$ major area, a collection of similar publication areas. Arora and Lahiri (1997) used eight major areas. Since this division of the eight major areas is not given in their paper, after noting trends in the data we used the Fay-Herriot model to test two new divisions of 6 and 4 major areas that combine similar survey estimates. These major areas produced large CV reduction in general. Where the 6 groups had yielded an average CV reduction of about 20% the 4 groups gave approximately an average 25% CV reduction over the direct estimates. Comparison of the point estimates and CVs have shown that the 4 major areas perform better than the 6 major areas. The 4 major areas are 1–7, 8–14, 15–25 and 26–43. In this paper, we will use these 4 groups as auxiliary variables for illustration purpose only.

Table 2
Milk Data, from Arora and Lahiri (1997)

Small Area	n_i	y_i	SD	CV
1	191	1.099	0.163	0.148
2	633	1.075	0.080	0.074
3	597	1.105	0.083	0.075
4	221	0.628	0.109	0.174
5	195	0.753	0.119	0.158
6	191	0.981	0.141	0.144
7	183	1.257	0.202	0.161
8	188	1.095	0.127	0.116
9	204	1.405	0.168	0.120
10	188	1.356	0.178	0.131
11	149	0.615	0.100	0.163
12	290	1.460	0.201	0.138
13	250	1.338	0.148	0.111
14	194	0.854	0.143	0.167
15	184	1.176	0.149	0.127
16	193	1.111	0.145	0.131
17	218	1.257	0.135	0.107
18	266	1.430	0.172	0.120
19	214	1.278	0.137	0.107
20	213	1.292	0.163	0.126
21	196	1.002	0.125	0.125
22	95	1.183	0.247	0.209
23	195	1.044	0.140	0.134
24	187	1.267	0.171	0.135
25	479	1.193	0.106	0.089
26	230	0.791	0.121	0.153
27	186	0.795	0.121	0.152
28	199	0.759	0.259	0.341
29	238	0.796	0.106	0.133
30	207	0.565	0.089	0.158
31	165	0.886	0.225	0.254
32	153	0.952	0.205	0.215
33	210	0.807	0.119	0.147
34	383	0.582	0.067	0.115
35	255	0.684	0.106	0.155
36	226	0.787	0.126	0.160
37	224	0.440	0.092	0.209
38	212	0.759	0.132	0.174
39	211	0.770	0.100	0.130
40	179	0.800	0.113	0.141
41	312	0.756	0.083	0.110
42	241	0.865	0.121	0.140
43	205	0.640	0.129	0.202

3.2 Analysis of Results

Corn and Soybean Data: First we consider the effect of our treatment of σ_i^2 using the HB approach. Table 3 presents the HB estimates $\hat{\theta}_i^{HB}$ and the associated standard errors (SDs) and CVs for the small area corn and soybean data sets. The SD is the square root of the posterior variance. Under Model 1 (σ_i^2 unknown), the SDs and CVs are consistently larger than the corresponding SDs and CVs under Model 2 ($\sigma_i^2 = s_i^2$ known). The increased SDs and CVs of Model 1 are expected since this model takes into account the added variability of estimating σ_i^2. On average there is about 20% increase in SDs and CVs (this calculation excludes Franklin for corn data). The results support the fact that letting $\sigma_i^2 = s_i^2$, the known direct estimate of σ_i^2, leads to underestimation of the SD and CV of $\hat{\theta}_i$. Inspection of small areas Franklin and Webster for the corn data and county Winnebago for the soybean data establish in some cases where the sampling errors by chance are quite small this under estimation is severe.

Comparison of the HB estimates under Model 1 and Model 2 to the direct estimates can be made using the CVs in Table 1 and Table 3. Under Model 2 the HB estimates have smaller CVs than the direct estimates in 6 of the 8 counties for the corn data and similarly for the soybean data, 6 out of 8 counties. Of the remaining 2 counties for each crop, the CVs under Model 2 are the same as the direct survey CVs or only slightly larger. Estimators from Model 2 therefore seem to have gained efficiency compared to the direct survey estimators. Now examining the HB estimates under Model 1 and the direct survey estimates lead to mixed results for the corn and soybean data sets. Model 1 accounts for the added uncertainty of estimating the sampling variances and so in only 4 of the 8 counties the HB estimates show improvements in efficiency for the corn data. For the soybean data 5 out of 8 counties demonstrate the HB estimates as improvements on the direct survey CVs. For the remaining counties the direct estimates exhibit lower CVs and even substantially lower CVs in some cases. For the corn data, counties Franklin and Webster have CV increases with Model 1 of more than 0.09 and 0.12 respectively. As well for the soybean data, county Winnebago has a CV increase of almost 0.10 from the direct survey estimate, using Model 1. Areas where the direct estimates demonstrate smaller CVs compared to the HB estimates include a number of those areas where the CVs are by chance atypically small. So the increased model-based CVs may reflect more appropriate CVs for those areas. Of the 7 cases where the direct CVs are smaller compared to the HB CVs under Model 1, the 3 cases noted above have severe differences and the remaining 4 instances show only slight reduction in efficiency with use of Model 1. Since direct survey estimates quite often have unacceptably large CVs and yet still by chance may have CVs grossly and inexplicably small, HB estimation under Model 1 may be more reliable and reasonable by taking into consideration the uncertainty of estimating σ_i^2.

Milk data: Table 4 contains the HB estimates for the milk data. As expected, over the 43 areas the treatment of σ_i^2 as known or unknown shows negligible differences in terms of point estimates, SDs and CVs due to the large sample sizes in the 43 areas. Therefore the substitution of $\sigma_i^2 = s_i^2$ in the model is reasonable when the area-specific sample sizes are large, as clearly shown in this example. Also the HB estimates give reduced SDs and CVs when compared to the direct survey estimates in Table 2. As would be expected, the HB estimation approach is thus an improvement on the direct survey estimates.

Table 3
Comparison of HB Estimates for Crop Data

County	σ_i^2 known ($\sigma_i^2 = s_i^2$) $\hat{\theta}_i^{HB}$	SD	CV	σ_i^2 unknown $\hat{\theta}_i^{HB}$	SD	CV
			Corn			
Franklin	155.788	6.061	0.039	142.862	18.408	0.129
Pocahontas	100.813	28.297	0.281	91.560	32.420	0.356
Winnebago	115.337	28.406	0.246	113.130	35.207	0.311
Wright	131.630	28.345	0.215	123.547	30.764	0.250
Webster	109.030	20.634	0.189	97.856	29.834	0.307
Hancock	121.682	15.656	0.129	123.478	17.857	0.145
Kossuth	115.710	11.180	0.097	114.910	12.510	0.109
Hardin	135.626	23.228	0.171	135.178	23.804	0.176
			Soybean			
Franklin	75.375	16.272	0.216	88.186	21.067	0.239
Pocahontas	116.943	27.031	0.231	109.052	30.098	0.276
Winnebago	87.525	10.304	0.118	88.053	18.854	0.214
Wright	104.184	23.671	0.227	105.825	24.497	0.232
Webster	115.510	20.789	0.180	109.455	25.801	0.236
Hancock	101.368	15.741	0.155	102.876	17.311	0.169
Kossuth	102.388	14.948	0.146	101.862	15.019	0.148
Hardin	87.455	17.774	0.203	93.397	20.251	0.217

Table 4
Comparison of HB Estimates for Milk Data

Small area	σ_i^2 known ($\sigma_i^2 = s_i^2$) $\hat{\theta}_i^{HB}$	SD	CV	σ_i^2 unknown $\hat{\theta}_i^{HB}$	SD	CV
1	1.020	0.113	0.111	1.021	0.111	0.109
2	1.045	0.072	0.069	1.045	0.071	0.068
3	1.065	0.073	0.069	1.065	0.074	0.069
4	0.767	0.095	0.124	0.770	0.096	0.125
5	0.849	0.096	0.113	0.852	0.096	0.113
6	0.975	0.103	0.106	0.975	0.102	0.105
7	1.058	0.125	0.118	1.055	0.125	0.118
8	1.097	0.099	0.090	1.096	0.099	0.090
9	1.219	0.121	0.099	1.215	0.121	0.100
10	1.192	0.122	0.102	1.190	0.122	0.102
11	0.793	0.094	0.119	0.799	0.097	0.122
12	1.213	0.131	0.108	1.209	0.130	0.107
13	1.206	0.112	0.093	1.203	0.112	0.093
14	0.984	0.107	0.109	0.987	0.107	0.109
15	1.187	0.105	0.088	1.187	0.104	0.087
16	1.156	0.104	0.090	1.156	0.102	0.089
17	1.225	0.101	0.083	1.225	0.100	0.081
18	1.284	0.115	0.089	1.281	0.113	0.088
19	1.234	0.101	0.082	1.235	0.100	0.081
20	1.233	0.110	0.089	1.233	0.110	0.089
21	1.092	0.097	0.089	1.095	0.098	0.089
22	1.192	0.128	0.107	1.193	0.127	0.106
23	1.122	0.103	0.092	1.125	0.103	0.091
24	1.221	0.113	0.092	1.220	0.111	0.091
25	1.193	0.086	0.072	1.193	0.086	0.072
26	0.761	0.091	0.120	0.762	0.091	0.120
27	0.763	0.092	0.120	0.762	0.091	0.119
28	0.734	0.125	0.170	0.732	0.123	0.169
29	0.768	0.085	0.110	0.767	0.085	0.110
30	0.615	0.076	0.124	0.618	0.076	0.123
31	0.769	0.122	0.158	0.767	0.120	0.156
32	0.795	0.119	0.150	0.792	0.118	0.148
33	0.771	0.091	0.118	0.770	0.090	0.117
34	0.612	0.060	0.099	0.613	0.062	0.100
35	0.701	0.085	0.121	0.701	0.084	0.120
36	0.757	0.094	0.123	0.759	0.093	0.123
37	0.534	0.080	0.150	0.538	0.081	0.151
38	0.744	0.096	0.129	0.743	0.095	0.128
39	0.754	0.082	0.108	0.753	0.082	0.108
40	0.768	0.088	0.115	0.768	0.088	0.115
41	0.747	0.071	0.095	0.747	0.070	0.094
42	0.801	0.093	0.116	0.800	0.092	0.116
43	0.682	0.094	0.139	0.682	0.094	0.138

3.3 Priors and Sensitivity Analysis

In Model 1, the sampling variances σ_i^2 are assumed to be independent with inverse gamma prior distribution $\text{IG}(a_i, b_i)$, and the model variance σ_v^2 also has inverse gamma prior distribution $\text{IG}(a_0, b_0)$, where a_i, b_i $(0 \le i \le m)$ are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2. So we have used proper priors to avoid the problem of any improper posteriors. One may consider using flat priors for σ_i^2 and σ_v^2, *i.e.*, $\pi(\sigma_i^2) \propto 1$, and $\pi(\sigma_v^2) \propto 1$, similar to the flat prior on β. With the flat priors on σ_i^2 and σ_v^2, the full conditional distributions for σ_i^2 and σ_v^2 are given as

$$[\sigma_i^2 \mid y, \theta, \beta, \sigma_v^2] \sim \text{IG}\left(\frac{d_i - 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right),$$

and

$$[\sigma_v^2 \mid y, \theta, \beta, \sigma_i^2] \sim \text{IG}\left(\frac{m-2}{2}, \frac{1}{2}\sum_{i=1}^{m}(\theta_i - x_i'\beta)^2\right).$$

The implementation of the Gibbs sampler under the flat priors is also straightforward. However, the flat priors on σ_i^2 and σ_v^2 may lead to improper posteriors if the sample sizes and the number of small areas are small. In order to see the problem on σ_i^2 more clearly, we can study the Model 1 in two steps. First, we can obtain the posterior of σ_i^2 given its direct estimate s_i^2 as

$$\begin{aligned}\pi(\sigma_i^2 \mid s_i^2) &\propto f(s_i^2 \mid \sigma_i^2) \cdot \pi(\sigma_i^2) \\ &\propto (\sigma_i^2)^{-d_i/2} \cdot \exp\{-\sigma_i^{-2} d_i s_i^2 / 2\} \cdot \pi(\sigma_i^2).\end{aligned}$$

By assuming a flat prior $\pi(\sigma_i^2) \propto 1$, we can obtain

$$\pi(\sigma_i^2 \mid s_i^2) \sim \text{IG}\left(\frac{d_i}{2} - 1, \frac{d_i s_i^2}{2}\right),$$

provided that $d_i > 2$, or $n_i > 3$. Then we can use this proper IG posterior $\pi(\sigma_i^2 \mid s_i^2)$ as an informative prior for σ_i^2 in the sampling model $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2)$. This will ensure to have proper posterior inference. For the modified corn and soybean data, using flat priors on σ_i^2 will lead to improper posterior due to the small sample sizes $(n_i = 3)$ for some areas. Thus, proper inverse gamma priors are used in the data analysis to ensure that all the posteriors are proper, as commonly used in the HB small area estimation in practice (*e.g.*, Arora and Lahiri 1997; Datta, Lahiri, Maiti and Lu 1999; You and Rao 2000; Rao 2003). Hence we do not face the problem of some posteriors being improper, since correct HB inference should be based on proper posteriors. Under Model 2 with the sampling variance known as $\sigma_i^2 = s_i^2$, using a flat prior $\pi(\sigma_v^2) \propto 1$ for σ_v^2, the posterior of σ_v^2 will be proper provided that

$m > p + 2$, where m is the number of small areas and p is the size of regression parameters β (Rao 2003, page 238). Since the number of small areas is usually relatively large, this condition is generally satisfied in practice.

For the sensitivity analysis of vague proper priors, we can test the sensitivity of the posterior estimates to the choice of prior parameters $a_i, b_i (0 \le i \le m)$. Under Model 1, we set $a_i = b_i$ at four different values, *i.e.*, 0.0001, 0.001, 0.01 and 0.1. Table 5 presents the estimated posterior means for the corn and soybean data, and Table 6 presents the corresponding CVs.

Table 5
Comparison of Posterior Mean Estimates for Crop Data

	IG (a_i, b_i), $a_i = b_i$			
County	0.0001	0.001	0.01	0.1
		Corn		
Franklin	142.862	142.593	143.155	144.311
Pocahontas	91.560	91.912	91.422	91.974
Winnebago	113.130	113.068	121.578	114.430
Wright	123.547	124.170	125.103	125.351
Webster	97.856	98.231	99.132	98.511
Hancock	123.478	123.858	124.395	124.138
Kossuth	114.910	115.281	115.316	115.528
Hardin	135.178	134.157	135.223	136.001
		Soybean		
Franklin	88.186	89.368	89.145	89.513
Pocahontas	109.052	109.571	107.745	108.176
Winnebago	88.053	87.478	86.267	87.302
Wright	105.825	106.712	105.142	104.676
Webster	109.455	108.392	109.835	110.252
Hancock	102.876	103.413	102.240	101.808
Kossuth	101.862	101.159	101.379	100.808
Hardin	93.397	94.713	93.576	94.767

Table 6
Comparison of Posterior CVs for Crop Data

	IG (a_i, b_i), $a_i = b_i$			
County	0.0001	0.001	0.01	0.1
		Corn		
Franklin	0.129	0.124	0.128	0.125
Pocahontas	0.356	0.351	0.347	0.341
Winnebago	0.311	0.314	0.321	0.324
Wright	0.250	0.246	0.235	0.236
Webster	0.307	0.292	0.285	0.280
Hancock	0.145	0.148	0.148	0.142
Kossuth	0.109	0.110	0.107	0.104
Hardin	0.176	0.173	0.178	0.168
		Soybean		
Franklin	0.239	0.233	0.231	0.227
Pocahontas	0.276	0.281	0.271	0.296
Winnebago	0.214	0.193	0.196	0.198
Wright	0.232	0.223	0.231	0.226
Webster	0.236	0.231	0.237	0.228
Hancock	0.169	0.165	0.168	0.161
Kossuth	0.148	0.145	0.142	0.135
Hardin	0.217	0.215	0.213	0.213

It is clear from Table 5 and Table 6 that the posterior estimates and the corresponding CVs are about the same and stable, which indicates that the HB estimates are not sensitive to the choice of vague proper priors. For the milk data, the HB estimates are very stable to these proper vague priors (results are not provided here). Since the milk data has large sample sizes, flat priors on variance components can also be used to analyze the milk data under Model 1. We thus obtained the HB estimates based on the flat priors and compared them with the HB estimates based on the vague IG priors. These HB estimates are almost identical and stable with relative difference ranging from 0.07% to 2.23%, an average value of 0.69% over 43 areas, which indicates that the posterior estimates of small area means based on Model 1 are very stable and not sensitive to the choice of flat priors or vague IG priors, provided that the sample sizes and number of small areas are relatively large.

4. Conclusion and Future Work

In this paper we have studied the well-known Fay-Herriot model with the situations where σ_i^2, the sampling error variances, are assumed unknown and where they are estimated by unbiased estimators s_i^2, using the HB approach. The full HB approach with the Gibbs sampling method automatically takes into account the extra uncertainty associated with the estimation of σ_i^2. We applied the HB approach in two survey data analysis. Our results have shown that the proposed HB approach under Model 1 works quite well no matter the area-specific sample sizes are small or large. For future work, the proposed HB modeling approach can be extended to the general area level models studied by You and Rao (2002). Application of the new HB modeling approach includes the census undercoverage estimation as in You, Rao and Dick (2004). Under Model 1, the HB estimators of the sampling variances σ_i^2 can be obtained. These HB estimators of σ_i^2 can be used as alternative smoothed estimators for σ_i^2 in the sampling models. Application and evaluation of the HB estimators of the sampling variances include the census undercoverage estimation and the Canadian Labour Force Survey (LFS) unemployment rate estimation (You, Rao and Gambino 2003). We also plan to compare the HB approach with the EBLUP approach as studied by Rivest and Vandal (2002) and Wang and Fuller (2003).

Acknowledgements

The authors would like to thank two referees, an Associate Editor, the Deputy Editor and the Editor, Dr. M.P. Singh, for their constructive comments and suggestions. The authors also would like to thank J.N.K. Rao of Carleton University for his useful suggestion and Jack Gambino and Eric Rancourt of Statistics Canada

for their comments on the early version of the paper. This work was supported by Statistics Canada Methodology Branch Research Block Fund.

References

Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

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

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999) Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Gelfand, A.E., and Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Gelfand, A.E., and Smith, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statisitcs – Theory and Methods*, 20, 1747-1766.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

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

Rivest, L.P., and Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 1, 3-15.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based un-employment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 105-114
Statistics Canada, Catalogue No. 12-001-XIE

A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach

Ali-Reza Khoshgooyanfard and Mohammad Taheri Monazzah [1]

Abstract

This paper primarily aims at proposing a cost-effective strategy to estimate the intercensal unemployment rate at the provincial level in Iran. Taking advantage of the small area estimation (SAE) methods, this strategy is based on a single sampling at the national level. Three methods of synthetic, composite, and empirical Bayes estimators are used to find the indirect estimates of interest for the year 1996. Findings not only confirm the adequacy of the suggested strategy, but they also indicate that the composite and empirical Bayes estimators perform well and similarly.

Key Words: Composite estimator; Design-based estimator; Empirical Bayes estimator; Indirect estimator; Non-sampling error; Synthetic estimator; Post-strata.

1. Introduction

Each year, sample surveys are conducted in Iran to obtain statistical information required for decision and policy making. However, these surveys cannot fulfill all statistical requirements because of two factors. The first one is related to the governmental and non-governmental sectors' demand for comprehensive statistical information not only at national and regional but also at small area levels. Further, they need the information at shorter periods of time per year, say monthly or quarterly. The second factor is that the main source of statistical data in Iran is surveys, and there are financial limitations for conducting surveys several times per year at small area levels. These two factors challenge statistical agencies to find efficient strategies to balance both cost and statistical information quality. The work presented here is an endeavor to overcome this challenge by using small area estimation (SAE) methods.

The purpose of SAE methods is to provide acceptable estimates for some subpopulations in a sample design planned for the "whole" population regardless of the subpopulations. For example, a sample design is planned for estimating population parameters for the "country" and after data collection the parameters are estimated by the national sample data. If simultaneously "provincial estimates" of the parameters are needed, it is not possible to conduct separate provincial sample surveys. The provinces are unplanned subpopulations in the sense that the available sample design has been planned just for estimating the parameters for the country without considering the provincial level. In the nationwide sample, few or no sample units may be available for some provinces. Hence, acceptable estimates for such provinces (subpopulations) cannot be produced.

Before the availability of SAE methods, such subpopulation estimates were obtained by direct design-based estimation. If there were data from a given subpopulation in the nationwide sample, an estimate would be directly calculated according to the nationwide sample design by using "the available data". The direct estimate may differ substantially from the actual subpopulation parameter due to large sampling errors owing to small sample size.

Statisticians and demographers have developed ways of estimating for such subpopulations. Indirect estimators have been suggested and applications have been increasing over the last twenty years. However, the SAE methods are still an active topic of study. See Purcell and Kish (1979, 1980), Ghosh and Rao (1994), Schaible (1995), Marker (1999), Pfeffermann (2002) and especially Rao (2003a) for problem definition and a review of the SAE methods.

For a number of years, the Statistical Center of Iran (SCI) implemented annually a national one-stage cluster sample in order to estimate the intercensal unemployment rate at the country level. For sixteen years, separate one-stage cluster samples for all provinces have been conducted to estimate provincial unemployment rates. A weighted combination of provincial estimates then yields the unemployment rate for the total country. The increasing need for estimation of the unemployment rate at a provincial level on a monthly, or at least seasonal basis, and the lack of administrative records in Iran at both small and national levels persuaded SCI to try the SAE methods as the core of a revised strategy to meet the provincial need.

The revised strategy consists of designing a sample survey only at the national level and producing the provincial estimates by SAE methods. A province in the strategy is a small area. This strategy demands a smaller sample size than that for aggregating provincial samples. If the revised

1. Ali-Reza Khoshgooyanfard, The Center for Research, Studies and Program Assessments of IRIB. E-mail: khosh_ar@yahoo.com; Mohammad Taheri Monazzah, The Central Bank of Iran. E-mail: Taheri53@yahoo.com.

strategy proves practicable, time and cost of the data collection can be reduced, and produce provincial estimates on a monthly basis. The smaller sample is easier to control in the field, and estimates are less affected by nonsampling errors.

This paper is intended to answer the following questions:

1. Can a nationwide sample substitute for separate provincial samples for making estimates of the provincial unemployment rates?
2. From the three SAE methods – synthetic, composite and empirical Bayes estimators – which one produces the best estimates?

To answer empirically these two questions, estimates were produced for the year 1996 when the actual values of the provincial unemployment rates are available from the 1996 Census. Consequently, the actual bias of each provincial estimate can be computed.

The process includes the following three stages. First, a sample of size 13,000 from the whole country is selected (the 1996 Census data file). The sample size is determined at the national level, and is allocated to all provinces proportionally to population. The allocation provides sample from each province enabling direct estimates of the unemployment rate for each province. Direct estimates are not necessarily acceptable for all provinces because of the large sampling errors due to small sample sizes in some provinces. Second, applying three SAE methods, indirect estimates are produced for each province. Third, the indirect estimates are evaluated by comparing them with corresponding actual values, computing MSEs, mean of absolute errors (MAE), and mean of errors (ME).

In addition to this introduction, the paper takes in three more sections. Section 2 offers a short review of the three estimators used in this paper, including the estimation methods, their corresponding MSEs, and properties of the estimators. The estimates and corresponding computational aspects are presented in section 3, where performances of the estimators are tentatively appraised. Section 4 is devoted to final remarks and recommendations about the estimators and the merit of the SAE strategy.

2. A Glance Over the Estimators

Indirect estimators used in the study are introduced briefly. However, an excellent discussion of the SAE methodology is in Rao (2003a). First, the synthetic estimator is considered, and then the composite estimator. The empirical Bayes (EB) estimator as a model-based estimator is also considered.

2.1 Synthetic Estimator

There is a family of small area estimators characterized as synthetic, see Rao (2003a, chapter 4). The traditional and simplest is discussed here. For this estimator,

1. The country is partitioned into six post-strata on the basis of six age groups, see Table (1).
2. Next, the number of unemployed persons is estimated in each province, providing the numerator in expression (1).
3. The synthetic estimate of the i^{th} province is obtained by dividing the estimated number of unemployed persons in province i by its Economically Active Population (EAP), namely

$$\hat{P}_i^S = \left(\sum_{j=1}^{6} N_{ij} \hat{P}_j \right) \Big/ N_i \qquad (1)$$

where $\hat{P}_j$ is a direct design-based estimate of the unemployment rate in post-stratum j, N_i is the EAP in province i, and N_{ij} is the EAP in the intersection of province i and post-stratum j, cell (i, j) The synthetic estimate of the i^{th} province is according to the official definition of the unemployment rate in Iran.

The synthetic estimate shares all national sample data by using national direct estimates of the unemployment rate from the post-strata. It uses the six estimated "post-strata" unemployment rates computed over all provinces rather than specific estimates of the six "cells". This process thus **borrows strength** because each province contributes to the national sample by pooling provincial sample units to overcome small sample sizes in each province.

This estimator has three limitations:

1. The smaller the inter-post-stratum variation is, the better synthetic estimator performs. It means that all provinces should have a rather equal unemployment rate in each age group. Using the national post-strata direct estimates equally for all provinces is allowable only under this assumption. If the homogeneity assumption is not satisfied, the synthetic estimator cannot reflect specific small area variation, and the estimates could be severely biased.
2. If there are several variables that are important in post-stratification, the synthetic estimator cannot often use all of them because post-strata (after cross-classification of the several variables) have sample sizes that are too small and yield unacceptable direct estimates of the post-strata. Generally speaking, many post-strata give rise to poor direct estimates for some of the post-strata. This can create serious problems for synthetic estimation when a poor direct estimate receives a large EAP for a cell.

3. Quality of the EAPs can affect the synthetic estimates. Owing to lack of timely data sources such as administrative records, out of date EAPs from the 1986 Census data are used here in order to produce the synthetic estimates for the year 1996.

2.2 Composite Estimator

The composite estimator of the i^{th} province combines the synthetic and direct estimators of that province, namely

$$\hat{P}_i^C = W_i \hat{P}_i^D + (1-W_i)\hat{P}_i^S \tag{2}$$

where $\hat{P}_i^D$ is the direct design-based estimator for the i^{th} province, and $0 \le W_i \le 1$. Expression (2) improves upon (1) by exploiting both estimators. That is, provincial differences may take into account in the composite estimator via the provincial unbiased direct estimates and instability of the direct estimator may be reduced via the synthetic estimator.

The weight W_i can be specified so as to minimize mean square error of $\hat{P}_i^C$, $\text{MSE}(\hat{P}_i^C)$. Assuming $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, the weight is simplified as

$$W_i^{\text{opt}} = \frac{1}{\left(V(\hat{P}_i^D)/\text{MSE}(\hat{P}_i^S)\right)+1} \tag{3}$$

where $V(\hat{P}_i^D)$ and $\text{MSE}(\hat{P}_i^S)$ are the variance of $\hat{P}_i^D$ and the mean square error of $\hat{P}_i^S$, respectively. In expression (3), the weights of the direct and synthetic estimators in (2) are proportional to the MSEs of the two estimators. See Schaible (1978) and Rao (2003a, page 58) for properties of the estimator and weight.

In practice, we should estimate $\text{MSE}(\hat{P}_i^S)$ and $V(\hat{P}_i^D)$ to generate an estimate of the weight (3). If there are some sample data from the i^{th} province, according to the sample design, an unbiased design-based estimator of $V(\hat{P}_i^D)$ can be computed by using only the sample data. Therefore, only an estimator for $\text{MSE}(\hat{P}_i^S)$ is required. Under the assumption that $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, Ghosh and Rao (1994) proposed the unbiased estimator

$$\hat{\text{MSE}}(\hat{P}_i^S) = (\hat{P}_i^S - \hat{P}_i^D)^2 - \hat{V}(\hat{P}_i^D). \tag{4}$$

Under the same assumption, one can easily show that

$$\text{MSE}(\hat{P}_i^C) = W_i^2 V(\hat{P}_i^D) + (1-W_i)^2 \text{MSE}(\hat{P}_i^S). \tag{5}$$

The estimator (4) may result in negative estimates for some provinces, and the weight in expression (3) is no longer computable. In this case, instead of (3) and (4), we have used respectively the combined weight in (6) and $\hat{\text{AMSE}} = (1/I')\sum_{i=1}^{I'} \hat{\text{MSE}}(\hat{P}_i^S)$ where I' is the number of small areas having positive estimated MSE (see Gonzalez and Waksberg (1973) for more details):

$$W^C = \frac{1}{\left(\sum_i \hat{V}(\hat{P}_i^D) \Big/ \sum_i \hat{\text{MSE}}(\hat{P}_i^S)\right)+1}. \tag{6}$$

In addition to expressions (3) and (6), Copas (1972), Ghosh and Rao (1994), and Thompsen and Holmoy (1998) suggest alternative weights.

2.3 Empirical Bayes (EB) Estimator

Model based SAE methods have received more attention than the synthetic and composite esitmators. Marker (1999) regarded the SAE methods as having a common element expressed through regression models. The EB method is of the regression type. Consider the following mixed model (see Rao (2003a, page 76)):

$$\boldsymbol{g} = X\boldsymbol{\beta} + \boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \tag{7}$$

where

$$\boldsymbol{g}' = (Ln\frac{\hat{P}_1^D}{1-\hat{P}_1^D}, \ ..., \ Ln\frac{\hat{P}_I^D}{1-\hat{P}_I^D}),$$

X is an $I \times k$ design matrix of supplementary variables, $\underline{\beta}$ is a $k \times 1$ vector of unknown parameters, and $\underline{\upsilon}$ and $\underline{\varepsilon}$ are $I \times 1$ random vectors (I is the number of provinces). Assume that:

1. $\boldsymbol{\upsilon}$ and $\boldsymbol{\varepsilon}$ are independent.
2. $E(\boldsymbol{\varepsilon}) = 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \text{Diag}(d_1^2, ..., d_I^2)$.
3. $\boldsymbol{\upsilon} \sim N(0, \Sigma)$ where $\boldsymbol{\Sigma} = \text{Diag}(t^2, \ ..., \ t^2)$.

Ghosh and Meeden (1997) show that the EB estimate of the i^{th} element of $\boldsymbol{g}$ is:

$$\hat{g}_i^{\text{EB}} = \hat{W}_i \boldsymbol{x}_i' \hat{\beta} + (1-\hat{W}_i) g_i \tag{8}$$

where $\boldsymbol{x}_i'$ and g_i are the i^{th} row and the i^{th} component of X and $\boldsymbol{g}$ respectively, and $\hat{W}_i$ is an estimate of

$$W_i = \frac{d_i^2}{d_i^2 + t^2}. \tag{9}$$

Consequently, the EB estimate of the i^{th} rate is:

$$\hat{P}_i^{\text{EB}} = \frac{\exp(\hat{W}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + (1-\hat{W}_i) g_i)}{1 + \exp(\hat{W}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + (1-\hat{W}_i) g_i)}. \tag{10}$$

It is obvious that (10) needs two estimates for $\boldsymbol{\beta}$ and the weight in (9). On the other hand, the weight in (9) relies on the estimates of t^2 and d_i^2. By applying the delta method, $(g_i')^2 \hat{V}(\hat{P}_i^D)$ generates an estimate of d_i^2 where g_i' through the first derivative of $g_i = Ln(\hat{P}_i^D / 1 - \hat{P}_i^D)$. Based on Chand and Alexander (1995), estimates of $\boldsymbol{\beta}$ and t^2 are found by simultaneously solving

$$\begin{cases} t^2 = (\boldsymbol{g} - X\boldsymbol{\beta})' V^{-1} (\boldsymbol{g} - X\boldsymbol{\beta}) / (I - k) \\ \boldsymbol{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} \boldsymbol{g} \end{cases} \quad (11)$$

where $V = \text{Diag}(d_1^2 + t^2, \ldots, d_I^2 + t^2)$. Note that the equations in (11) are solved by numerical iteration with an initial value for t^2.

The EB and composite estimators have similarities although they arise from different approaches. Both estimators have two components; a direct component ($\hat{P}_i^D$ in (2) and g_i in (8)) computed from the provincial sample data, and an indirect component ($\hat{P}_i^S$ in (2) and $\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$ in (8)) constructed from the national sample data and supplementary information. Both estimators (2) and (8) give more weight to the indirect component when it is reliable. Otherwise the direct component receives more weight. Additional details are given in Cressie (1989), Ghosh *et al.* (1998) and Rao (2003 a, b).

3. Estimation for Iran

Estimates were produced for the year 1996 because the 1996 actual unemployment rate of each province is known from the 1996 Census. As a result, the actual bias of each estimate can be computed.

In 1996 the country consisted of 26 provinces. However, 21 provinces are studied here because supplementary information from the 1986 Census was available for 21 geographically unchanged provinces between the years 1986 and 1996. To make the three indirect estimates, at the national level, a sample was planned and its sample size was determined for estimating the unemployment rate of the country as a whole. Each province is a small area. The national sample was allocated among the 26 provinces proportional to population in order to have sample data from each province (a top-down approach). This enabled direct design-based estimates for each province and its corresponding variance required for both the EB and composite estimators. The sample design is able to produce good estimates for the country and for some provinces.

3.1 Computational Aspects

To construct synthetic estimates, six age groups formed the post-strata. The estimated unemployment rate of each group based on the national sample and its corresponding actual value based on the 1996 Census are presented in Table (1), which also contains absolute errors of the estimates.

The estimates for the first two groups have very large error. Therefore, if a province in expression (1) gives large EAPs to these age groups, its synthetic estimate may not perform well. The 1986 Census data were used in computing the EAPs for all provinces and cells (N_i and N_{ij} in expression (1)) because, in the absence of administrative records, the nearest census to the year 1996 is the main source of data at any level.

Table 1
Post-strata Characteristics

Age Group	Estimated Rate ($\hat{P}_i$)	Actual Value	Absolute Error
10–15	0.3240	0.2826	0.0414
16–20	0.2402	0.2629	0.0227
21–25	0.1868	0.1856	0.0012
26–30	0.0811	0.0802	0.0009
31–50	0.0363	0.0366	0.0003
More than 50	0.0653	0.0648	0.0005

To construct the composite estimates, provinces were divided into two groups. The first consists of 14 provinces that used the weight in expression (3), and the second of seven provinces used the common weight $W^C = 0.873184$ based on (6). Because the estimator in expression (4) produces negative estimates for $\text{MSE}(\hat{P}_i^S)$ for these seven provinces, the AMSE was used.

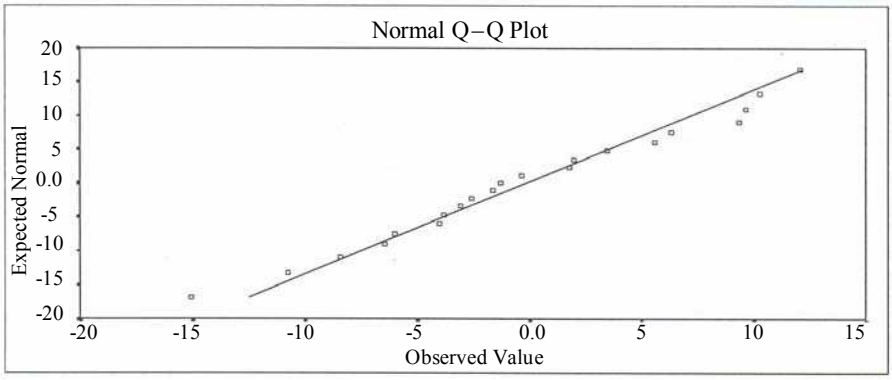
To construct the EB estimates, d_i^2 was estimated by using the delta method and then t^2 was estimated following Prasad and Rao (1990) by using a SAS/IML program (the program is available from the authors). An initial estimate for t^2 is required in this program, and was calculated by the moment estimation method as $t^2 = 0.3117194$. To solve the equations in (11), the following 21×2 design matrix was used, whose first and second columns are 1s and EAPs, respectively:

$$X = \begin{bmatrix} 1 & 133,449 \\ 1 & 141,124 \\ 1 & 883,653 \\ 1 & 795,714 \\ \vdots & \vdots \\ 1 & 522,976 \\ 1 & 162,892 \end{bmatrix}$$

The estimated t^2 and $\underline{\beta}$ are

$$\hat{t}^2 = 0.5596389, \hat{\boldsymbol{\beta}} = \begin{pmatrix} -2.066874 \\ -1.273 \times 10^{-7} \end{pmatrix}$$

To test normality, a normal Q–Q plot and a Shapiro-Wilk's test for standardized residuals of the fitted model were examined. The points in the Q–Q plot are close to a straight line, and the test did not reject the null hypothesis of normality (p–value = 0.851).



3.2 Results

The results are organized into four parts. First, bias in the forms of error and absolute error is examined using two criteria, ME and MAE. Second, MSEs are compared among methods. Third, efficiencies of the indirect estimators relative to the direct estimator are evaluated. Finally, the weights of the direct components in expressions (2) and (8) are analyzed. All the results are depicted in appropriate figures, however, details can be found in Table (2).

Suppose S_a is the allocated sample size from the national sample to a given province and S_r the required sample size if the sample size is separately determined for the province. In other words, if there is a sample of size S_r from the province, an acceptable direct estimate can be then computed for the province. Therefore, $(S_a/S_r)\times 100$ measures how much the available sample size (S_a) is adequate for a given province. This measure is used on horizontal axes of all plots as a basis for comparison sample size effects.

The synthetic estimator has the highest MAE, which was even larger than that of the direct estimator (see Figure 1). Conversely, MAEs of the composite and EB estimators are the lowest, and very similar to one another. Based on ME, there is a slight overestimation of the actual value for all estimators. The direct estimator has the lowest ME because it is unbiased. MEs of the composite and EB estimators are close and the synthetic estimator has the highest ME.

For the direct, composite, and EB estimators, all provinces with $S_a/S_r \geq 10\%$ have absolute errors less than 0.02. The highest absolute errors belong to Ilam and Kohgiluyeh & Boyerahmad which have the smallest populations and very small S_a/S_r. Plots of these three estimators have relatively similar patterns. The story is different for the synthetic estimator because the "national" sample data are only used in making synthetic estimates through the post-strata direct estimates and then the "national" sample size (not S_a/S_r) affects the synthetic estimate of a province through the cell EAPs. In other words, if a post-stratum does not have "enough" national sample data to yield acceptable direct estimates, and a province gives large EAP to the post-stratum direct estimate, the province has a poor synthetic estimate. This is the case for Sistan & Baluchestan, Bushehr, Tehran and Lorestan because of poor direct estimates for the first two post-strata (the age groups of 10–15 and 16–20) and large young populations of these provinces.

The lowest MSE always belongs to the composite or EB estimator (see Figure 2). However, MSE of the composite estimator is often lower than that of the EB estimator. The MSE of the synthetic estimator is always higher than those of the other estimators, even the direct estimator.

As the S_a/S_r increases, the MSE decreases for the direct, composite and EB estimators (see the descending trend in Figure 2). This effect is very drastic for Tehran (S_a/S_r= 36%). Again, there are two exceptions for the three estimators, Ilam and Kohgiluyeh & Boyerahmad, both having the smallest populations and very small S_a/S_r. The pattern of Figure 2 for the synthetic estimator may be misleading because seven provinces used AMSE. However, the four previous provinces (Sistan & Baluchestan, Bushehr, Tehran and Lorestan) also do not conform to the pattern. As a general rule for an estimator, the greater the dependency on the provincial direct estimates the stronger the relationship between the MSE and the ratio S_a/S_r.

The relative efficiencies (RE) of the three indirect estimators compared to the direct estimator for all provinces are often smaller than or equal to one for the composite and EB estimators and greater than one for the synthetic estimator. Some composite estimates have good REs: Semnan (0.34), West Azarbayejan (0.46), Khorasan (0.70), Kermanshah (0.75) and Hamadan (0.77). Means of REs ($\overline{RE}^{S} = 13.6$, $\overline{RE}^{C} = 0.8595$ and $\overline{RE}^{EB} = 0.9951$) indicate that the composite estimator is the most efficient estimator among the three indirect estimators. Further, in Figure 3 as S_a/S_r increases RE^{EB} approaches one. Figure 3 as well as Figure 2 may be misleading for the synthetic estimator.

Table 2
Provincial and Estimator Characteristics

Province	EAP	S_a	S_r	S_a/S_r	RE^C	RE^{EB}	RE^S	AE^C	AE^{EB}	AE^S	AE^D	MSE^C	MSE^{EB}	MSE^S	MSE^D
Bushehr	133,449	146	4,550	3.2%	0.96	1.17	25.57	0.03300	0.01687	0.06501	0.02644	**0.0003030**	0.000368	0.0080483	0.0003148
Chaharmahal & Bakhtiyari*	141,124	203	4,063	5%	0.87	0.95	6.52	0.02136	0.02135	0.03644	0.02031	**0.0003813**	0.000417	0.0028670	0.0004397
Esfahan	883,653	1032	5,850	17.6%	0.90	1.00	9.56	0.01268	0.01421	0.00990	0.01504	**0.0000533**	0.000059	0.0005631	0.0000589
Fars	795,714	925	6,175	15%	0.91	0.99	9.69	0.00610	0.00886	0.02235	0.00904	**0.0000836**	0.000091	0.0008931	0.0000922
Gilan*	734,196	683	5,364	12.7%	1.04	0.97	17.25	0.00484	0.00460	0.01107	0.00393	0.0001728	**0.000162**	0.0028670	0.0001662
Hamadan	387,517	439	4,550	9.6%	0.77	1.00	3.36	0.01294	0.01701	0.00675	0.01880	**0.0001155**	0.000150	0.0005030	0.0001498
Hormozgan*	168,268	198	4,063	4.9%	0.84	0.93	5.12	0.01984	0.01734	0.02821	0.01862	**0.0004731**	0.000519	0.0028670	0.0005600
Ilam	84,210	111	4,063	2.7%	0.83	0.87	4.94	0.04901	0.05201	0.03395	0.06579	**0.0013919**	0.001450	0.0082747	0.0016734
Kerman*	312,768	450	5,200	8.7%	0.96	0.97	12.00	0.03615	0.03672	0.02864	0.03724	**0.0002283**	0.000231	0.0028670	0.0002389
Kermanshah	357,096	436	3,575	12.2%	0.75	0.96	3.07	0.00265	0.00928	0.02641	0.01210	**0.0002747**	0.000349	0.0011190	0.0003640
Khorasan	1,410,863	1,587	8,125	19.5%	0.70	0.99	2.36	0.00515	0.00193	0.01353	0.00160	**0.0000298**	0.000042	0.0000999	0.0000424
Khuzestan*	609,044	786	4,225	18.6%	1.03	0.97	16.83	0.01034	0.01247	0.00308	0.01140	0.0001760	**0.000166**	0.0028670	0.0001704
Kohgiluyeh & Boyerahmad	90,655	105	3,575	2.9%	0.83	0.86	4.83	0.05486	0.05932	0.02630	0.07165	**0.0013629**	0.001408	0.0079493	0.0016449
Kordestan*	276,575	341	5,200	6.6%	0.91	0.95	9.22	0.03105	0.02814	0.03641	0.03027	**0.0002833**	0.000297	0.0028670	0.0003111
Lorestan	310,918	341	3,575	9.5%	0.86	0.95	6.22	0.00943	0.01383	0.04101	0.01754	**0.0004090**	0.000451	0.0029534	0.0004747
Mazandaran*	917,259	1,043	6,013	17.3%	1.30	0.98	33.57	0.00199	0.00188	0.00310	0.00183	0.0001112	**0.000084**	0.0028670	0.0000854
Semnan	110,166	121	4,713	2.6%	0.34	1.08	0.51	0.02776	0.01929	0.03661	0.01042	**0.0001534**	0.000491	0.0002317	0.0004542
Sistan & Baluchestan	272,752	318	4,875	6.5%	0.96	0.97	26.53	0.00431	0.00228	0.08606	0.00123	**0.0002519**	0.000254	0.0069347	0.0002614
Tehran	2,343,290	2,913	8,125	35.9%	0.99	1.00	83.08	0.00605	0.00573	0.04767	0.00555	**0.0000209**	0.000021	0.0017530	0.0000211
West Azarbayejan	522,976	654	6,500	10.1%	0.46	0.98	0.85	0.00505	0.01247	0.00182	0.01309	**0.0000552**	0.000118	0.0001024	0.0001199
Yazd	162,892	207	5,038	4.1%	0.82	1.36	4.52	0.01414	0.00968	0.01008	0.01950	**0.0001299**	0.000215	0.0007164	0.0001586

*Denote provinces for which expression (3) produces negative estimates for MSEs.
EAP: Economically Active Population
S_a: Allocated Sample Size
S_r: Required Sample Size
RE: Relative Efficiency
AE: Absolute Error
MSE: Mean Squared Error (the lowest MSE is bold for each province)
C, EB, S and D stand for Composite, Empirical Bayes, Synthetic and Direct estimators, respectively.

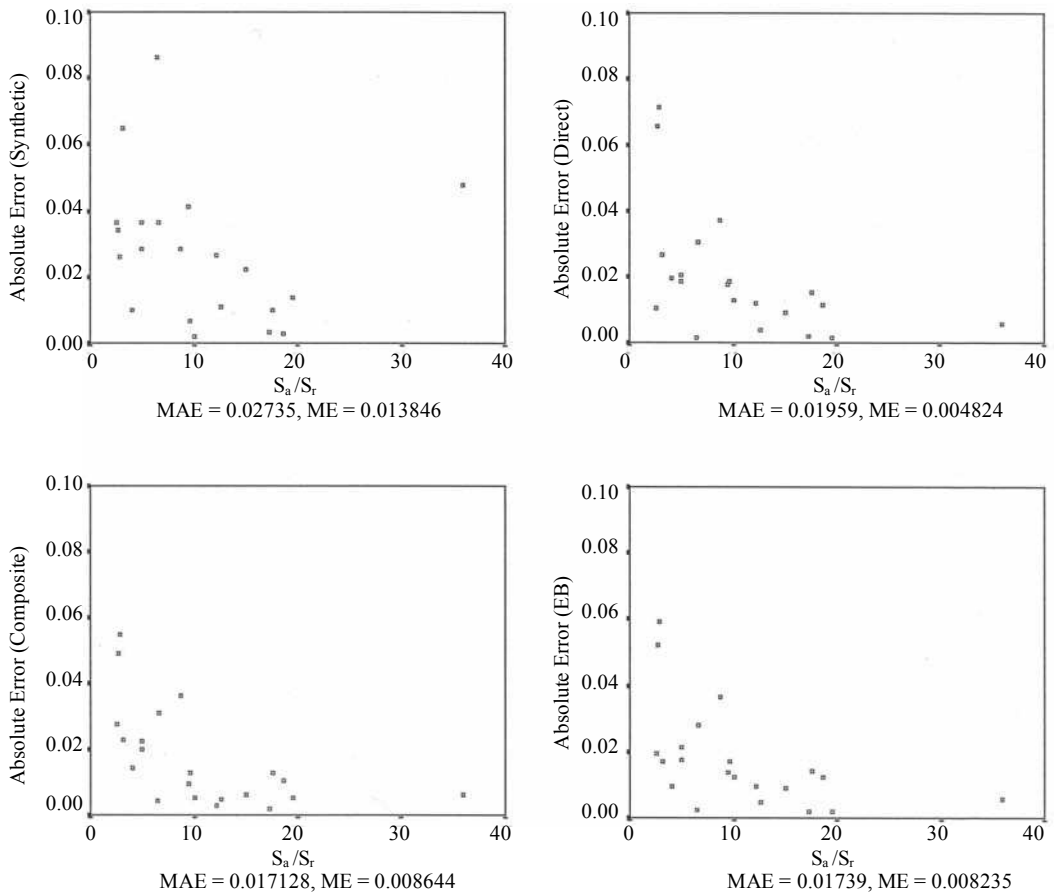


Figure 1. Absolute errors of the estimates against S_a/S_r.

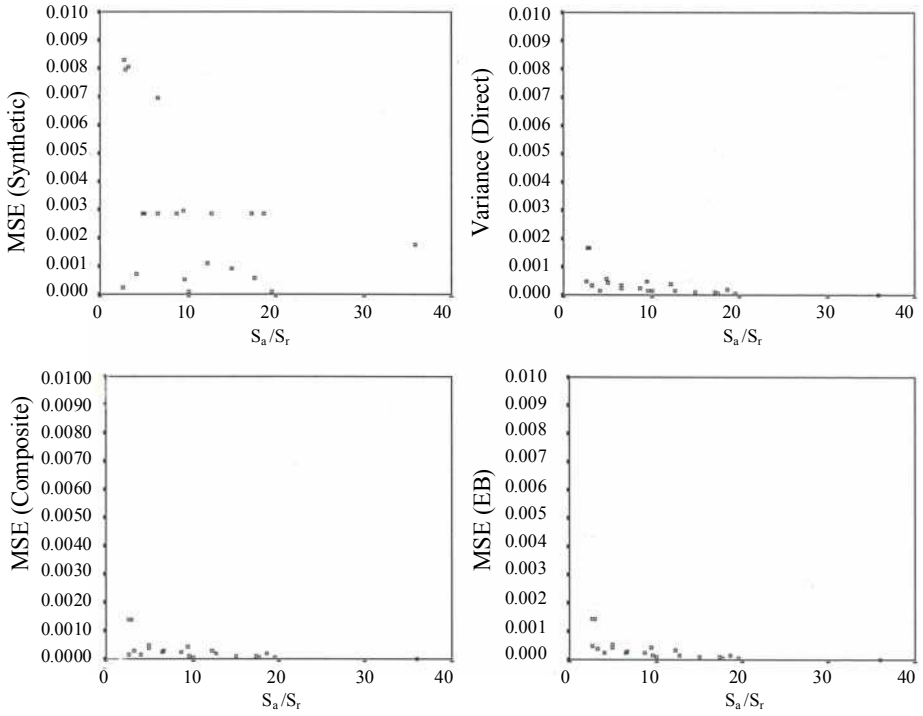


Figure 2. MSEs of the estimates against S_a/S_r.

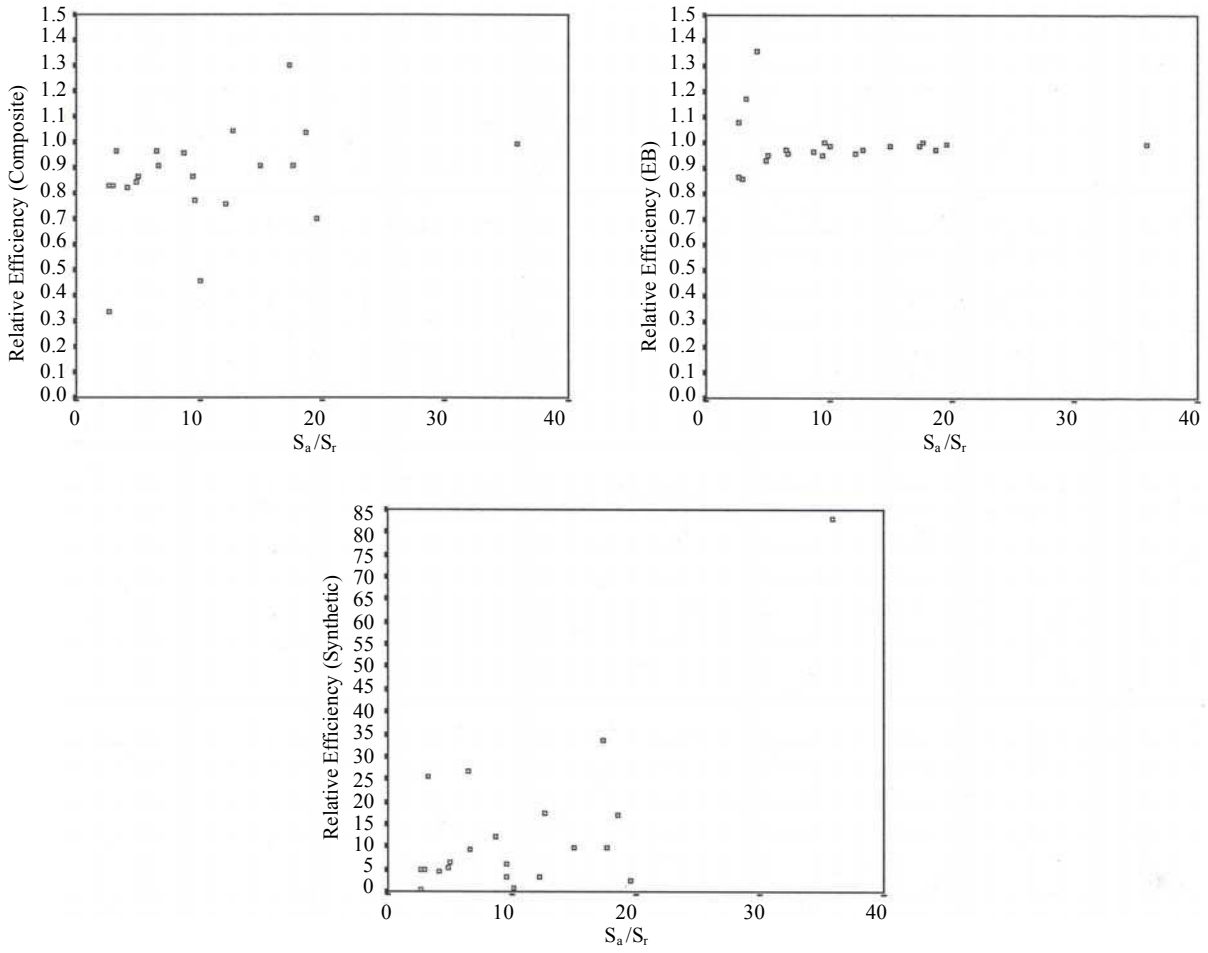


Figure 3. Relative efficiency (Estimated MSE of indirect estimator/Estimated variance of direct estimator) against S_a/S_r (the plot of synthetic estimator has a different scale on the vertical axis for legibility).

The direct component of the estimator in expression (8), g_i, always receives more weight than the indirect component. This is the case for the composite estimator, except for two provinces of Semnan and West Azarbayejan. For the composite estimator, Rao (2003a, page 58) states that "the optimal weight W_i^{opt} will be close to zero or one when one of the component estimators has a much larger MSE than the other, that is, when $f_i = MSE(\hat{P}_i^C)/MSE(\hat{P}_i^S)$ is either large or small. In this case, the estimator with larger MSE adds little information and therefore it is better to use the component with smaller MSE." This comment is clearly illustrated for Bushehr ($W = 0.962355$, $RE^S = 25.27$), Sistan & Baluchestan ($W = 0.963670$, $RE^S = 26.53$) and Tehran ($W = 0.988083$, $RE^S = 83.08$), because the direct estimates of these provinces have smaller MSEs than the synthetic estimates. Figure 4 clearly shows an ascendant relationship between the weight and S_a/S_r for the EB estimator. For the composite estimator, the lowest and highest weights pertain to the provinces with the lowest S_a/S_r and the highest S_a/S_r respectively.

In general, the synthetic estimator performs poorly based on the MAE, ME, MSE and RE criteria, even though the synthetic estimates of some provinces are individually closer to actual values than other estimates. However, the synthetic estimates have been computed under the most disadvantageous conditions. The EAPs applied to construct the synthetic estimates are based on the 1986 Census (ten years before the year when the estimates were produced). In addition, the direct estimates of the first two post-strata are quite different from the other post-strata, causing poor synthetic estimates.

To address the first problem, administrative records should be developed; for the second, post-strata estimation should be handled in the sample design in advance. If not only post-strata estimation but also provincial classifications in planning the sample design are considered in advance, good direct estimates for the post-strata can be expected. Consequently, good synthetic estimates for the provinces can be expected. The provincial classifications can increase homogeneity by putting similar provinces in classes together and using only sample data of the classified provinces to make the direct post-strata estimates to construct the synthetic estimates of those provinces.

The composite and EB estimators usually perform well when S_a/S_r is 10% or larger for a province because the direct components of the estimators (2) and (8) are relatively stable and receive a larger weight, especially for the EB estimator. Tehran, Khorasan, Khuzestan and Esfahan are of this type, while Bushehr, Ilam, Kohgiluyeh & Boyerahmad and Semnan are not.

4. Final Remarks

In developing countries like Iran, administrative records are not often available both at small and large area levels. Surveys may yield satisfactory estimates for large areas but not for small areas. Periodic censuses do not meet all demands for effective policies and planning. These limitations lead to deficiencies in official statistics. Therefore, the statistical planning activities of SCI are directed towards compensating these deficiencies by using new methods and strategies. The present study proposes a cost-effective strategy to overcome some of the limitations.

In this study, the findings support the idea that a nationwide sample design can be used instead of the separate provincial sample designs by applying suitable SAE methods. The nationwide sample design consists of nearly 13,000 persons, whereas the twenty-one separate provincial sample designs totally consist of almost 100,000 persons.

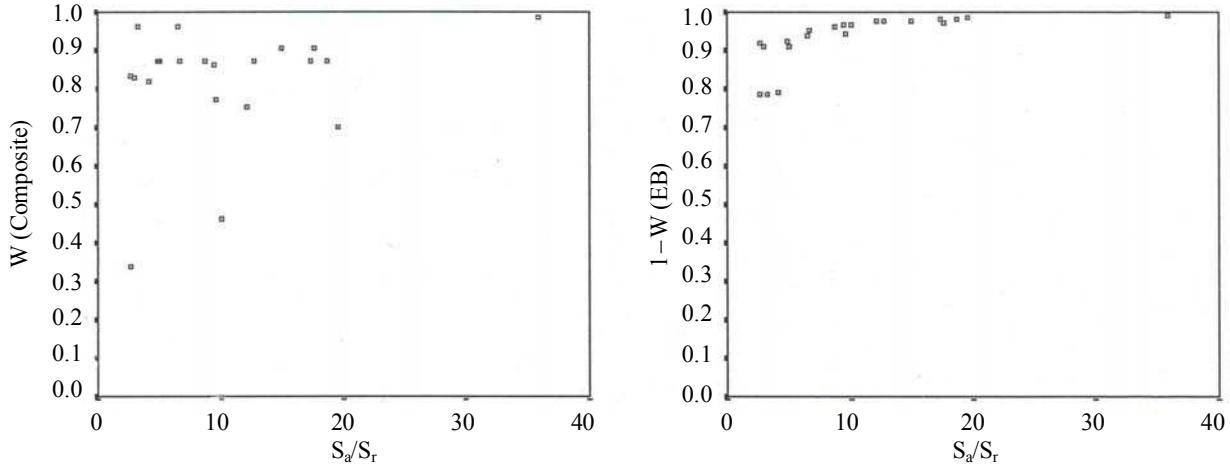


Figure 4. The weights of the direct components of composite and EB estimates against S_a/S_r.

The provincial design is the method currently used to produce provincial estimates by SCI. Using the national sample design decreases costs more than 80 percent. In addition, note that:

1. Although some SAE methods do not rely on existing sample data from all small areas, the strategy for producing provincial estimates is more appropriate when the small areas of interest are pre-specified. Therefore, the nationwide sample can be allocated to all small areas of interest to produce direct design-based estimates. It is important to adjust the sample design to accommodate the SAE methods before data collection begins. As Singh, Gambino and Mantel (1994, page 3) note

   > "small area needs should be recognized at the early stages of planning for large scale surveys. The sample design should include special features that enable production of reliable small area data using design or model estimators".

   Therefore, SCI must re-plan sample designs to reflect small area needs.

2. The SAE estimators usually perform well as the sample size increases. To improve provincial estimates, the nationwide sample size can be enlarged to have larger sample sizes from each province. Also, the provinces can be classified into groups with similar characteristics, such as unemployment rates, socio-demographic variables, and so on. Separate sample size would then be determined for each group.

3. Appropriate supplementary variables, which are related to the variable of interest, play a central role in improving the estimators.
   - The synthetic estimator used only one variable (age) for partitioning but it may be possible to use another variable or a combination of variables for partitioning. The post-strata in the synthetic estimator should be formed by variables that reduce variation in each post-stratum. These variables can indirectly affect the composite estimator as well.
   - The EB model can be improved with better supplementary information. Therefore it is important to try different supplementary variables to find the best model. In this work only EPA was used as the independent variable in the model, but there may be other variables that produce better estimates.

4. The composite estimator performs relatively better than the synthetic and EB estimators. However, the results are only meant to provide a first impression of the utility of SAE methods. More research is needed to develop a generic SAE methodology in Iran. Further, the SAE methods should be applied not only in estimating unemployment rates but also in estimating other parameters, and SAE methods should be compared with the estimates coming from separate sample designs.

Acknowledgements

Research for this paper was partially supported by the Statistical Research Center of Iran. The authors are grateful for many useful and helpful comments from the referees and the Associate Editor. My heartfelt thanks should go to Jim Lepkowski for his friendly helps. The views expressed are the authors' and do not necessarily reflect those of SCI.

References

Chand, N., and Alexander, C.H. (1995). Indirect estimation of rates and rates for small areas with continuous measurement. In *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 549-554.

Copas, J.B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika*, 59, 349-360.

Cressie, N. (1989). Empirical bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 65-93.

Ghosh, M., and Meeden, G. (1997*). Bayesian Methods for Finite Population Sampling.* Chapman & Hall, London.

Gonzalez, M.F., and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimation. *Journal of the American Statistical Association*, 73, 7-15.

Gonzalez, M.F., and Waksberg, J. (1973). Estimation of the errors of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statistician, Vienna, Austria, 18-25 August.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. In *Proceedings of the American Statistical Association*, *Social Statistics Section*, 328-331.

Marker, D.A. (1995). *Small area estimation*: *A Bayesian perspective.* Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Pfeffermann, D. (2002). Small area estimation-New developments and directions. *International Statistical Review*, 70, 125-143.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean square error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.

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

Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 2, 145-169.

Schaible, W.L. (1978). Choosing weight for composite estimators for small area statistics. In *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 741-746.

Schaible, W.L. (1995). Ed. *Lecture Notes in Statistics*: *Indirect Estimators in U.S. Federal Programs*, New York: Springer.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.

Thompsen, I., and Holmoy A.M.K. (1998). Combining data from surveys and administrative record system: The Norwegian experience. *International Statistical Review*, 66, 201-221.

Survey Methodology, June 2006
Vol. 32, No. 1, pp. 115-120
Statistics Canada, Catalogue No. 12-001-XIE

Design Effects for Multiple Design Samples

Siegfried Gabler, Sabine Häder and Peter Lynn [1]

Abstract

In some situations the sample design of a survey is rather complex, consisting of fundamentally different designs in different domains. The design effect for estimates based upon the total sample is a weighted sum of the domain-specific design effects. We derive these weights under an appropriate model and illustrate their use with data from the European Social Survey (ESS).

Key Words: Stratification; Clustering; Variance component model; Intraclass correlation coefficient; Selection probabilities.

1. Introduction

In survey research complex sample designs are often applied. These designs have features such as stratification, clustering and/or unequal inclusion probabilities, that lead to "design effects". The design effect is a measure that shows the effect of the design on the variance of an estimate. Design-based it is defined as follows (see Lohr 1999, page 239):

$$deff\,(plan, statistic) = \frac{\mathrm{V}(\text{estimate from sampling plan})}{\mathrm{V}\left(\begin{array}{c}\text{estimate from an srs with same number}\\ \text{of observation units}\end{array}\right)}$$

where srs indicates a simple random sample. The use of clustering and/or unequal inclusion probabilities typically leads to design effects greater than 1.0; in other words the variance of an estimate is increased compared to the variance of the estimate from a simple random sample with the same number of observations. The consideration of design effects is very important when deciding upon the sample size of a survey in advance. For example, if a comparative survey with different countries is planned it is very useful to have estimates of the design effects for the different countries. Then it is possible to choose the net sample sizes in a way that the precision of the estimates will be approximately equal. For this, for a certain degree of precision the sample size that would be needed under srs (effective sample size) has to be multiplied by the predicted design effect.

The European Social Survey (ESS, see www.europeansocialsurvey.com) is a survey program where design effects are taken into consideration for calculating net sample sizes – aiming at the same effective sample size for each country ($n_{\text{eff}} = 1{,}500$). 22 countries participated in the first round of the ESS, only three of them with unclustered, equal probability designs (srs): Denmark, Finland and Sweden. For all other countries there was the need to predict the design effect in advance of the study. For this, a model based approach (see Gabler, Häder and Lahiri 1999) can be used which distinguishes between a design effect due to unequal inclusion probabilities (term 1) and a design effect due to clustering (term 2):

$$deff = m\frac{\sum_{i=1}^{I} m_i w_i^2}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \times [1 + (b^* - 1)\rho] = deff_p \times deff_c \qquad (1)$$

where m_i are respondents in the i^{th} selection probability class, each receiving a weight of w_i, ρ is the intracluster correlation coefficient and

$$b^* = \frac{\sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2}$$

where b_c is the number of observations in cluster $c(c = 1, \ldots, C)$ and w_{cj} is the design weight for sample element *j* in cluster *c*. (This is of course a simplification that assumes no association between *y* and w_i or between w_i and b^* and ignores any effects of stratification, that will tend to be beneficial and modest. See Lynn, Gabler, Häder and Laaksonen (2007, forthcoming) and Park and Lee (2004) for discussion of the sensitivity of *deff* predictions to these assumptions; see Lynn and Gabler (2005) for discussion of alternative ways to predict $deff_c$).

In some countries the applied designs were even more complicated, consisting of fundamentally different designs in each of two independent domains. In the UK, *e.g.*, the design was a mixture of a clustered design with unequal inclusion probabilities (in Great Britain) and an unclustered

1. Siegfried Gabler and Sabine Häder, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Germany. E-mail: gabler@zuma-mannheim.de; Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom. E-mail: plynn@essex.ac.uk.

sample (in Northern Ireland). In Poland, simple random samples were selected in one domain (cities and large towns), while a two-stage clustered design was applied in the second domain (all other areas). In Germany, a clustered equal-probability sample was selected in each domain (West Germany including West Berlin; East Germany), but the sampling fractions differed between the domains.

The question arose how to predict design effects for these dual design samples. As we show below, it is not simply a convex combination of the design effects for the different domains – apart from in some special cases. A general solution for multiple design samples will be presented in section 2, with illustrations of the application of this solution to prediction of design effects prior to field work (section 3) and to estimation of design effects post-field work (section 4). Section 5 concludes with discussion.

2. Design Effects for Multiple Design Samples

Let $\{C_1,\ldots,C_K\}$ be a partition of the clusters into K domains. Then $C\bar{b}=\sum_{c=1}^{C} b_c=\sum_{k=1}^{K}\sum_{c\in C_k} b_c=\sum_{k=1}^{K} m_k=m$, where $m_k=\sum_{c\in C_k} b_c$ is the number of observations in the k^{th} domain of clusters. Let y_{cj} be the observation for sample element j in cluster $c (c=1,\ldots,C;\ \ j=1,\ldots,b_c)$. The usual design-based estimator for the population mean is

$$\bar{y}_w=\frac{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}}=\sum_{k=1}^{K}\frac{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}}\bar{y}_w^{(k)}$$

where

$$\bar{y}_w^{(k)}=\frac{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}\,y_{cj}}{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}}.$$

We assume the following model M1:

$$\left.\begin{aligned}E(y_{cj})&=\mu\\ \mathrm{Var}(y_{cj})&=\sigma^2\end{aligned}\right\}\quad \text{for}\quad c=1,\ldots,C;\ j=1,\ldots,b_c$$
$$\mathrm{Cov}(y_{cj},y_{c'j'})=\begin{cases}\rho_k\sigma^2 & \text{if } c=c'\in C_k;\ j\neq j'\\ 0 & \text{otherwise}\end{cases}\quad k=1,\ldots,K. \tag{2}$$

Model M1 is appropriate to account for the cluster effect with different kinds of clusters and generalises an earlier approach (see, *e.g.*, Gabler *et al.* 1999). More general models can be found in Rao and Kleffe (1988, page 62). We define the (model) design effect as $deff=\mathrm{Var}_{M1}(\bar{y}_w)/\mathrm{Var}_{M2}(\bar{y})$, where $\mathrm{Var}_{M1}(\bar{y}_w)$ is the variance of $\bar{y}_w$ under model M1 and $\mathrm{Var}_{M2}(\bar{y})$ is the variance of the overall sample mean $\bar{y}$, defined as $\sum_{c=1}^{C}\sum_{j=1}^{b_c} y_{cj}/m$, computed under the following model M2:

$$\left.\begin{aligned}E(y_{cj})&=\mu\\ \mathrm{Var}(y_{cj})&=\sigma^2\end{aligned}\right\}\quad \text{for}\quad c=1,\ldots,C;\ j=1,\ldots,b_c \tag{3}$$
$$\mathrm{Cov}(y_{cj},y_{c'j'})=0\quad \text{for all}\quad (c,j)\neq(c',j').$$

Note that model M2 is appropriate under simple random sampling and provides the usual expression, $\mathrm{Var}_{M2}(\bar{y})=\sigma^2/m$.

Quite analogous to Gabler *et al.* (1999) we note that

$$\mathrm{Var}_{M1}\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}\,y_{cj}\right)=$$
$$\sigma^2\sum_{k=1}^{K}\sum_{c\in C_k}\left\{\sum_{j=1}^{b_c} w_{cj}^2+\rho_k\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}\right\}. \tag{4}$$

Thus

$$deff=\sum_{k=1}^{K}\left(\frac{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}}\right)^2\frac{m}{m_k}deff_k \tag{5}$$

where

$$deff_k=m_k\frac{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}^2}{\left(\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}\right)^2}\times[1+(b_k^*-1)\rho_k]=deff_{pk}\times deff_{ck},$$

and

$$b_k^*=\frac{\sum_{c\in C_k}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{c\in C_k}\sum_{j=1}^{b_c} w_{cj}^2}.$$

It can be seen that *deff* is not a convex combination of the specific $\{deff_k\}$ except in some special cases. We consider here four realistic scenarios, each representing a simplification of the general case. Only in two of these scenarios (scenarios 1 and 4) does the combination become convex:

Scenario 1: *Equal weights for all units*

If $w_{cj}=1$ for all c, j, then expression (5) simplifies to:

$$deff=\sum_{k=1}^{K}\frac{m_k}{m}deff_k. \tag{6}$$

Scenario 2: *Equal weights within each domain*

If $w_{cj}=w_k$ for all $c\in C_k, j$, then expression (5) becomes:

$$deff = \sum_{k=1}^{K} \left(\frac{m_k w_k}{\sum_{k=1}^{K} m_k w_k} \right)^2 \frac{m}{m_k} deff_k. \qquad (7)$$

Scenario 3: *Weighted sample size proportional to domain population size*

If

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{N_k}{N},$$

where N_k is population size in domain $k; N = \sum_{k=1}^{K} N_k$, then expression (5) becomes:

$$deff = \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \frac{m}{m_k} deff_k. \qquad (8)$$

Scenario 4: *Unweighted sample size proportional to domain population size*

If

$$\frac{m}{m_k} = \frac{N}{N_k},$$

then expression (8) becomes:

$$deff = \sum_{k=1}^{K} \frac{N_k}{N} deff_k. \qquad (9)$$

3. Application to Prediction of *Deff*

In round 1 of the ESS, the sample design was a combination of two different sample designs for 5 out of 22 countries: United Kingdom, Poland, Belgium, Norway and Germany. We can apply the general formula (5) for design effects for multiple design samples to each of these cases, where $K = 2$. In some cases, we can equivalently use one of the simplified expressions (6) to (9). Here we illustrate how the formulae would be used in the prediction of design effects prior to fieldwork, for the purpose of establishing the required net (respondent) sample size to achieve a prescribed precision of estimation. In each case, the approach is to predict $\{deff_k\}$ using (1) for each k and then use (5) to predict *deff*. To predict $\{deff_k\}$, the observed values of $\{w_{cj}\}$ from the ESS round 1 respondent sample are used to estimate, b^*, m_i and w_i. In other words, these could be thought of as predictions for a future survey using the same design (*e.g.*, a future round of ESS). For illustration, we assume $\rho_k = 0.02 \; \forall \, k$ with a clustered design and $\rho_k = 0.00 \; \forall \, k$ with an unclustered design (0.02 is in fact the default value that was used for predicted design effects for clustered samples on the ESS in cases where estimates from previous surveys were not available). Our focus here is on the application of (5). For a more detailed description of the sample designs see Häder, Gabler, Laaksonen and Lynn (2003). We use three of the ESS countries – Poland, UK and Germany – as illustrations as these designs differ between the domains in different ways. The designs of Norway and Belgium were similar to that of Poland, with equal probabilities for all units but one domain clustered and one unclustered.

3.1 Poland

In Poland, the first domain covered the population living in towns of 100,000 inhabitants or more. Within this domain, a srs of persons was selected from the population register (PESEL data base) in each region, with slight variation between regions in the sampling fraction, reflecting anticipated differences in response rate. There were 42 towns in this domain and they accounted for about 31% of the target population.

The second domain corresponded to the rest of the population – people living in towns of 99,999 inhabitants or fewer and people living in rural areas. This part of the sample was stratified and clustered (158 clusters). The sampling of this second part was based on a two-stage design: PSUs were selected with probability proportional to size. The definition of a PSU was different for urban *vs*. rural areas. For urban areas, a PSU was equivalent to a town, whereas for rural areas, it was equivalent to a village. In the second stage, a cluster of 12 respondents was selected in each PSU by srs.

In the first domain, $\rho_1 = 0$ and $deff_{c1} = 1$. The modest variation in selection probabilities leads to $deff_{p1} = 1.005$ and, therefore, $deff_1 = deff_{c1} \cdot deff_{p1} = 1.005$. In the second domain, the design effect due to clustering is anticipated to be $deff_{c2} = 1.18$ (based on a prediction of $b^* = 10.07$) and $deff_{p2} = 1.01$ which results in $deff_2 = deff_{c2} \cdot deff_{p2} = 1.19$. Substituting these values of $deff_k$ in (5) leads to a prediction of $deff = 1.17$.

The design for Poland differs only slightly from scenario 2 and it can be seen that in this case the simpler expression, (7), provides a reasonable prediction if we approximate the weights as follows. Domain 1 contains 37.3% of the gross sample and 31% of the target population. Thus

$$w_1 = \frac{N_1 / N}{n_1 / n} = \frac{0.310}{0.373} = 0.831$$

and

$$w_2 = \frac{N_2 / N}{n_2 / n} = \frac{0.690}{0.627} = 1.100,$$

respectively, where n_k is selected sample size in domain $k; \sum_{k=1}^{K} n_k$.

Now, we can apply expression (7) to find the predicted design effect for estimates for Poland: $deff = (0.194 \cdot 1.005) + (0.821 \cdot 1.19) = 1.17$.

3.2. United Kingdom

In the UK, the ESS sample design differed between Great Britain (England, Wales, Scotland) and Northern Ireland. In Great Britain a stratified three-stage design with unequal probabilities was applied. At the first stage 162 small areas known as "postcode sectors" were selected systematically with probability proportional to the number of addresses in the sector, after implicit stratification by region and population density. At stage 2, 24 addresses were selected in each sector, leading to an equal-probability sample of addresses. At the third stage, one person aged 15+ was selected at the selected address using a Kish grid.

For Northern Ireland a simple random sample of 125 addresses was drawn from the Valuation and Land Agency's list of domestic properties. One person aged 15+ was selected at the selected address using a Kish grid. Thus, the UK sample is clustered in one domain but not in the other. In both domains, there are unequal selection probabilities.

In Great Britain we predicted $deff_{c1} = 1.20$ (based on a prediction of $b^* = 11.11$) and $deff_{p1} = 1.22$, so $deff_1 = 1.46$. In Northern Ireland we have predictions of $deff_{c2} = 1$ (by definition) and $deff_{p2} = 1.27$, so $deff_2 = 1.27$. From expression (5), $deff = 0.978 \cdot 1.46 + 0.023 \cdot 1.27 = 1.460$. It should also be noted that the selected sample sizes in the two domains were chosen to result in net sample sizes that would be approximately in proportion to the population sizes. In other words, the simplification of scenario 4 approximately holds. If we use expression (9), we get $deff = N_1/N \; deff_1 + N_2/N \quad deff_2 = 0.97 \cdot 1.46 + 0.03 \cdot 1.27 = 1.457$, demonstrating that this provides a reasonable approximation to (5) in this case.

3.3. Germany

In Germany independent samples were selected in two domains, West Germany incl. West Berlin, and East Germany incl. East Berlin. In both domains, the sample was clustered and equal-probability, but a higher sampling fraction was used in East Germany.

At the first stage 100 communities (clusters) for West Germany, and 50 for East Germany were selected with probability proportional to the population size of the community (aged 15 years or older). The number of communities selected from each stratum was determined by a controlled rounding procedure. The number of sample points was 108 in the West, and 55 in the East (some larger communities have more than one sample point). At the second stage in each sample point there was drawn an equal number of individuals by a systematic random selection process. This was done using the local registers of residents' registration offices.

Since the sampling design is self-weighting for both East and West Germany, but with disproportional allocation, scenario 2 applies and we can use expression (7), where

$$w_1 = w_{\text{EAST}} = \frac{N_{\text{EAST}}}{N} \frac{n}{n_{\text{EAST}}} = 0.567$$

and

$$w_2 = w_{\text{WEST}} = \frac{N_{\text{WEST}}}{N} \frac{n}{n_{\text{WEST}}} = 1.257.$$

(we note that common practice on some surveys is to scale the weights so that they sum to population sizes. This would make no difference to the application here as expression (5) involves only ratios of sums of weights).

The design effect due to clustering for each domain was predicted as $deff_{c1} = 1.39$ and $deff_{c2} = 1.35$, respectively (*via* predictions of $b^* = 20.56$ and 18.65 respectively), so from (7) we have

$$deff = 0.120 \cdot 1.39 + 0.991 \cdot 1.35 = 1.51.$$

It should be noted that in this case any convex combination of the domain-specific design effects will lead to a prediction of *deff* between 1.35 and 1.39. For example, (6) would give $deff = 1.36$. This fails to take into account the differences in selection probabilities *between* the domains. With this particular design – where the *only* difference in design between domains is the difference in selection probabilities – *deff* might alternatively be predicted by taking the convex combination and multiplying it by the prediction of $deff_p$ from the first term in expression (1), *viz.* $deff = 1.36 \cdot 1.09 = 1.49$. But this method is equivalent only in the special case where $\{deff_k\}$ are equal – and approximately equivalent in this case, where the variation is small.

4. Application to Estimation of *Deff*

Here we illustrate the use of expression (5) in the estimation of design effects post-fieldwork. We present estimates for 5 demographic/behavioural variables and a set of 24 attitude measures from round 1 of the European Social Survey, for the same three countries as in section 3. For comparison, we present also the estimates that would be obtained using the simpler expressions (6), (8) and (9). It can be seen that the estimates of *deff* differ greatly between variables. This is to be expected, reflecting variation in the association of y with clusters and with selection probabilities. But here we are more interested in differences between estimation methods for the same variable.

For Germany, we see that estimators (6) and (9), which ignore variation in weights and in sampling rates between the two domains respectively, under-estimates *deff* for all variables. Estimator (8), which assumes only equal response rates in each domain, produces estimates very similar to (5). For Poland, all three simplified estimators under-estimate *deff*, though (6) perhaps performs marginally better than the other two. For UK, we observe the remarkable result that all four estimators produce almost identical estimates for every variable. The assumption in (9) (and therefore also that in (8)) holds for UK and while weights are by no means equal, the distribution of weights is very similar in each domain. It can be noted that (6) holds under a weaker assumption that

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{m_k}{m},$$

i.e., that the share of the weights in each stratum equals the share of sample units. It is striking that these relationships between the estimators are consistent across all the variables considered.

5. Discussion and Conclusion

Expression (5) provides an appropriate means of combining design effects for domains with fundamentally different designs. It can be applied in estimation by estimating *deff*s in the usual way for each domain and then combining them using knowledge of the weight and domain membership of sample units. Use of (5) in the prediction of *deff*s before a survey is carried out is only slightly more demanding, requiring prediction of the share of the weights in the responding sample in each domain in addition to a method of predicting design-specific *deff*s.

Table 1
Estimates of *Deff* for Means Under 4 Estimators for 3 Countries

	DE				GB				PL			
Estimator:	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)
Demographic/behavioural												
Persons in household	1.87	1.85	1.87	1.74	1.66	1.66	1.66	1.66	1.51	1.43	1.41	1.42
Years of education	3.25	2.80	3.25	2.88	2.81	2.79	2.80	2.79	1.77	1.66	1.63	1.64
Net household income	2.46	2.15	2.46	2.19	2.82	2.80	2.80	2.80	2.16	2.00	1.95	1.98
Time watching TV	2.08	1.86	2.08	1.87	2.04	2.03	2.03	2.03	1.31	1.26	1.25	1.25
Time reading newspaper	1.79	1.62	1.79	1.61	1.35	1.35	1.35	1.35	1.73	1.63	1.60	1.61
Attitude measures												
Discriminated by race	1.16	1.03	1.16	1.04	1.92	1.92	1.92	1.92	1.02	1.01	1.01	1.01
Discriminated by religion	1.22	1.05	1.22	1.08	1.26	1.26	1.26	1.26	1.07	1.05	1.05	1.05
General happiness	2.56	2.11	2.55	2.23	1.56	1.55	1.56	1.55	1.49	1.42	1.40	1.41
Trust in others	2.20	1.96	2.20	1.98	1.85	1.84	1.84	1.84	1.66	1.57	1.54	1.55
Trust in Euro Parliament	1.83	1.59	1.83	1.62	1.50	1.50	1.50	1.50	1.43	1.37	1.35	1.36
Trust in legal system	2.07	1.72	2.07	1.81	1.37	1.37	1.37	1.37	1.42	1.36	1.34	1.35
Trust in police	1.92	1.63	1.92	1.69	1.24	1.24	1.24	1.24	1.24	1.20	1.19	1.19
Trust in politicians	1.75	1.62	1.75	1.59	1.38	1.38	1.38	1.38	1.63	1.54	1.51	1.53
Trust in parliament	1.64	1.48	1.64	1.48	1.45	1.45	1.45	1.45	1.13	1.10	1.10	1.10
Left-right scale	1.70	1.65	1.70	1.58	1.48	1.47	1.48	1.48	1.31	1.26	1.25	1.25
Satisfaction with life	2.06	1.74	2.06	1.81	1.68	1.67	1.67	1.67	1.30	1.25	1.24	1.25
Satisfaction with education system	3.03	2.89	3.03	2.79	1.37	1.37	1.37	1.37	1.40	1.34	1.32	1.33
Satisfaction with health system	3.76	3.21	3.76	3.32	1.65	1.64	1.64	1.64	1.65	1.56	1.53	1.54
Religiosity	1.94	1.75	1.94	1.75	1.57	1.56	1.56	1.56	1.73	1.63	1.60	1.61
Attitudes to immigrants	2.77	2.68	2.77	2.57	1.92	1.92	1.92	1.92	1.89	1.76	1.73	1.74
Supports law against ethnic discrimination	2.82	2.85	2.82	2.66	1.73	1.72	1.72	1.72	2.57	2.36	2.29	2.33
Importance of family	2.17	1.99	2.17	1.97	1.19	1.19	1.19	1.19	1.21	1.17	1.17	1.17
Importance of friends	2.31	2.09	2.31	2.08	1.34	1.34	1.34	1.34	1.54	1.46	1.44	1.45
Importance of work	2.20	2.16	2.20	2.05	1.90	1.89	1.89	1.89	1.69	1.59	1.57	1.58
Support people worse off	2.70	2.47	2.70	2.45	1.35	1.35	1.35	1.35	1.78	1.67	1.64	1.66
Always obey law	2.43	2.21	2.43	2.20	1.53	1.52	1.52	1.52	2.11	1.96	1.91	1.93
Political activism	3.26	2.83	3.26	2.89	1.94	1.94	1.94	1.94	2.16	2.00	1.96	1.98
Liberalism	2.28	2.18	2.28	2.10	1.78	1.77	1.78	1.78	1.75	1.64	1.61	1.63
Participation in groups	3.75	3.04	3.75	3.24	2.26	2.25	2.25	2.25	1.82	1.71	1.68	1.69

We have shown in section 4 above that use of alternative, simpler, methods of combining the domain-specific *deff*s does not always result in good estimates. In particular, the use of a convex combination will tend to result in an under-estimation, the extent of which depends on the extent of departure from the assumptions underlying the simplified expressions. In our empirical illustration, departures were modest, but it is easy to imagine designs with greater variation between domains in mean selection probabilities or in the distribution of design weights. We would therefore recommend that estimators (6)–(9) are used only if the assumptions genuinely hold, or if the sample design data necessary to calculate (5) is not available, in which case the analyst should at least make arbitrary allowance for under-estimation based on his or her knowledge of the design.

An important issue that is outside the scope of this article is how to deal with non-response when predicting or estimating design effects for multiple design samples. The expressions throughout section 2 of this article refer to the number of observations, *i.e*., respondent sample units, in each domain, m_k, and the calculations in sections 3 and 4 are based on predicted numbers of observations and actual numbers of observations respectively. But the natural interpretation of the differences between the four scenarios in section 2 may be in terms of sample design, where the weights are design weights. Thus, scenario 2, for example, would refer to a design that is *epsem* within domains, but where the sampling fraction is permitted to differ between domains. However, in most realistic applications non-response will occur and may well be differential both between and within domains. This is often reflected in an adjustment to the design weight. Thus, the simplification of scenario 2 would only apply if the non-response adjustment were constant within domains, in addition to the design being *epsem* within domains.

Scenario 3, if interpreted with respect to design alone, should always hold for any well-specified design in which the domains form explicit strata. Expression (8) is therefore equivalent to expression (5) in the absence of non-response. In the presence of non-response, scenario 3 requires that the (design-weighted) response rates are equal in each domain. Similarly, scenario 4 requires that the net inclusion rate (the product of coverage rate, sampling fraction and response rate) is equal in each domain, whereas a design interpretation would not consider the response rate component.

Appropriate ways to incorporate non-response adjustment into design effect estimation and, in particular, how that might effect estimation for multiple design samples, would appear to be an area worthy of further research.

Acknowledgement

The third author is grateful to ZUMA for a guest professorship which provided the time and stimulation to write this paper and for the support of the UK Longitudinal Studies Centre at the University of Essex, which is funded by grant number H562255004 of the UK Economic and Social Research Council.

References

Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.

Häder, S., Gabler, S., Laaksonen, S. and Lynn, P. (2003). The sample. Chapter 2 in *ESS 2002/2003: Technical Report*. http://www.europeansocialsurvey.com.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lynn, P., and Gabler, S. (2005). Approximations to *b** in the prediction of design effects due to clustering. *Survey Methodology*, 31, 101-104.

Lynn, P., Gabler, S., Häder, S. and Laaksonen, S. (2007, forthcoming). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics,* accepted.

Park, I., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 183-193.

Rao, C.R., and Kleffe, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

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

Contents
Volume 21, No. 4, 2005

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

CONTENTS TABLE DES MATIÈRES

Volume 33, No. 4, December/décembre 2005

Volume 34, No. 1, March/mars 2006

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

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

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", *etc*.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω ; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

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