Catalogue no. 12-001-XWE
ISSN: 1492-0921

Survey Methodology

January 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

email at infostats@statcan.gc.ca,

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

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

Depository Services Program

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

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

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

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

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

Standard symbols

The following symbols are used in Statistics Canada publications:

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

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 39, Number 2, December 2013

Contents

Waksberg Invited Paper Series

Regular Papers

Survey Methodology, December 2013
Vol. 39, No. 2, p. 247
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

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

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2015 Waksberg Award.

This issue of *Survey Methodology* opens with the thirteenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Mary E. Thompson (Chair), Steve Heeringa, Cynthia Clark and J.N.K. Rao for having selected Ken Brewer as the author of this year's Waksberg Award paper.

2013 Waksberg Invited Paper

Author: Kenneth R.W. Brewer

Ken Brewer was a practicing survey methodologist in the Australian Bureau of Statistics for about twenty years. He joined the Sampling Section in 1954 two years after its founding, and eventually became director of the sampling and methodology division. He "retired" in 1992 to do full-time research at the Australian National University. Ken always had a fascination for the foundations of statistical inference, both in survey sampling and in general. He is well known for his 1983 book "Sampling with Unequal Probabilities", written with M. Hanif, but he also made many other important contributions. Among them, he was a pioneer in the use of modeling for inference and analysis using survey data. He made important contributions to the discussion of model-based versus design-based inference in sampling, and he introduced to concept of "cosmetic calibration" in an attempt to reconcile the two view-points.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 249-262
Statistics Canada, Catalogue No. 12-001-X

Three controversies in the history of survey sampling

Ken Brewer[1]

Abstract

The history of survey sampling, dating from the writings of A.N. Kiaer (1897), has been remarkably controversial. First Kiaer himself had to struggle to convince his contemporaries that survey sampling itself was a legitimate procedure. He spent several decades in the attempt, and was an old man before survey sampling became a reputable activity. The first person to provide both a theoretical justification of survey sampling (in 1906) and a practical demonstration of its feasibility (in a survey conducted in Reading which was published in 1912) was A.L. Bowley. In 1925, the ISI meeting in Rome adopted a resolution giving acceptance to the use of both randomization and purposive sampling. Bowley used both. However the next two decades saw a steady tendency for randomization to become mandatory. In 1934, Jerzy Neyman used the relatively recent failure of a large purposive survey to ensure that subsequent sample surveys would need to employ random sampling only. He found apt pupils in M.H. Hansen, W.N. Hurwitz and W.G. Madow, who together published a definitive sampling textbook in 1953. This went effectively unchallenged for nearly two decades. In the 1970s, however, R.M. Royall and his coauthors did challenge the use of random sampling inference, and advocated that of model-based sampling instead. That in turn gave rise to the third major controversy within little more than a century. The present author, however, with several others, believes that both design-based and model-based inference have a useful part to play.

Key Words: Rule of three; Representative method; *p*-statistic; Prediction; Randomization; Model, Horvitz-Thompson.

1 Introduction

One of the most difficult problems I struck in writing this paper was in knowing where to begin. Initially I had intended to start with Laplace as I had in an earlier paper (Brewer and Gregoire 2009), which incorrectly described him as being unable to fulfill his ambition to estimate the population of France by using what we would now describe as survey sampling. He had, in fact, achieved that using a sample of the small administrative districts known as communes as early as September 22, 1802 (Cochran 1978). In later accounts, I had read of him struggling to repeat that performance while the boundaries of France were in a constant state of flux, and I had jumped to the incorrect conclusion that he had never achieved it at all.

However, I soon found myself being pulled further back into history. No, Laplace had not been the first person to use a ratio estimator, not even the first Frenchman (Stephan 1948). The Englishman John Graunt had used the ratio estimator in his estimation of the population of London (Graunt 1662). Well, perhaps he had not really used the ratio estimator (he probably hadn't used anything that would be recognized as a ratio estimator today, certainly not by a finicky survey statistician like me!), but he had admittedly used the Rule of Three.

I had not come across that Rule before, but apparently it was well-known to be this: "If AB = CD and D is unknown, then D = AB/C." Obviously the present-day ratio estimator was a particular case of that Rule of Three. In fact, the Rule of Three must have predated the 17th Century by a considerable margin, so it might genuinely be of interest when searching for a survey-sampling start date.

1. Ken Brewer, School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Australia. E-mail: ken.brewer@anu.edu.au.

It soon occurred to me that the Rule of Three was bound to have been known to Hammurabi's astronomers, getting on for 4,000 years ago, because they were very arithmetically minded, having invented a sexagesimal system of counting that still survives today in the using of "hours" (and also of "degrees") "minutes" and "seconds", and also of 30^0-60^0-90^0 ["30-60-90"] triangles.

That realization encouraged me to start to look for a more recent starting point for this paper, and I eventually concluded that a good choice would be to start with "modern survey sampling", a topic that had been suggested to me before. The paper is structured as follows. Section 2 discusses the first controversy which is Anders Kiaer's "Representative Method." Section 3 provides a discussion on the second controversy, which is the exclusive use of randomization as a means for selecting samples, as advocated by Neyman (1934). The arguments for using the model-assisted or the model-based approach as a means for inference in survey sampling is described in Section 4. Section 5 provides a middle ground that incorporates both procedures. The paper ends with a summary given in Section 6.

2 The first controversy: Anders Kiaer and the "Representative method"

Anders Kiaer (1838-1919), was the founder and first director of Statistics Norway. Although many now claim him to be the first modern survey statistician, his contribution to statistics did not go unchallenged at the time. It was claimed, for instance, that his approaches to sampling lacked a theoretical description. In addition, there was also a serious lack of references in Kiaer's papers. Most of the charges made against him by his contemporaries have merit, but it is also true that with the first publication of his ideas in 1895 he started a process that ended in the development of modern survey sampling theory. Kiaer was also the first to use a sample survey on its own, as opposed to a by-product from a full enumeration.

By 1895, Kiaer had been conducting sample surveys successfully in his own country for fifteen years or more, finding to his own satisfaction that it was not always necessary to enumerate an entire population to obtain useful information about it. He decided that it was time to convince his peers of this fact, and he attempted to do so at the session of the International Statistical Institute (ISI) being held in Berne that year. Kaier there argued that what he called a "partial investigation", based on a subset of the population units, could indeed provide such information, provided only that the subset in question had been carefully chosen to reflect the whole of that population in miniature. He described this process as his "representative method", and he was able to gain some support for it, notably from his Scandinavian colleagues. Unfortunately, his idea of "representation" was too subjective and (in hindsight) too lacking in probabilistic rigour, to make headway against the then universally held belief that only complete enumerations, "censuses", could provide any useful information (Wright 2001, Lie 2002).

Moreover, all Kiaer's innovations, and in particular his idea of a sample being "representative", were controversial enough to create serious opposition to his ideas among his contemporaries, and this was particularly evident in the seriously unfavourable reactions to the paper that he presented at that 1895 meeting. However, he persisted and continued to present papers about his surveys and the methods he used in them at later ISI meetings.

Eight years later, at the ISI's Berlin meeting in 1903, Lucien March suggested that randomization might provide an objective basis for the use of "partial investigations" (Wright 2001, Lie 2002).

This idea was further developed by Sir Arthur Lyon Bowley, first in a theoretical paper (Bowley 1906) and later by a practical demonstration of its feasibility in a survey conducted in Reading, England (Bowley 1912).

By 1925, the ISI at its Rome meeting was sufficiently convinced (largely by the report of a study that it had itself commissioned!) to adopt a resolution giving acceptance to the idea of sampling. However it was left to the discretion of the investigators whether they should use randomized or purposive sampling. With the advantage of hindsight we may conjecture that, however vague their awareness of the fact, the writers of that report were intuiting that while purposive sampling was sometimes capable of presenting useful estimates, the underpinning of randomization was also desirable.

In the following year, Bowley himself published a substantial monograph (Bowley 1926) in which he presented what was then known concerning the purposive and randomizing approaches to sample selection, and also made suggestions for further developments in both of them. These included the notion of collecting similar units into groups called "strata," including the same proportions of units from each stratum in the sample. Furthermore, there was an attempt to make purposive sampling more rigorous by taking into account the correlations between the variables of interest for the survey and any other auxiliary variables that might be helpful in the estimation process.

3 The second controversy: Neyman advocates the exclusive use of randomization

By the 1920s the situation was clear, though hardly ideal. Sampling was no longer regarded as off the agenda, but there was little or no guidance as to whether the sample should be chosen randomly or purposefully. The next two decades saw a slow but steady tendency for the randomization approach to become mandatory. And there was a good reason behind that tendency, for there were no other attractive models available to cause sampling statisticians to want to use them.

A particularly influential paper advocating the exclusive use of randomization was Jerzy Neyman's (1934) 68-page attack on a survey conducted by Gini and Galvani (1929). Those two authors had selected a "purposive" sample of 29 out of 214 districts (circondari) from the 1921 Italian Population Census. Their sample was chosen in such a way as to reflect almost exactly the whole-of-Italy average values for seven variables chosen for their importance; but Neyman showed that it exhibited substantial differences for other important variables. He then went on to attack this study with a three-pronged argument.

1) Because randomization had not been used, the investigators had not been able to invoke the Central Limit Theorem. Consequently they had been unable to use the normality of the estimates to construct the "confidence intervals" that Neyman himself had recently invented. That idea appeared in English for the first time in this paper.
2) On Gini's and Galvani's own admissions, the difficulty of their achieving their "purposive" requirement (that the sample match the population closely on seven variables) had caused them to limit their attention to the 214 districts rather than to the 8,354 communes into which Italy had also been divided. In consequence, their 15% sample consisted of only 29 districts (instead of perhaps 1,200 or 1,300 communes). Neyman further showed that a considerably more

accurate set of estimates could have been expected had the sample consisted of a much larger number of those (order of magnitude smaller) communes.

3) Crucially, the population model used by the investigators was unrealistic and inappropriate. (Neyman was convinced that models by their very nature were always liable to represent the actual situation inadequately.) Furthermore, randomization obviated the need for such population modelling. Using randomization-based inference, the statistical properties of an estimator could be established by using the distribution of its estimates from all the samples that could possibly be drawn. Moreover, when using randomisation, that same estimator under different designs could have different statistical properties. (A good example of this, though not one of Neyman's, is that an estimator that is biased under an equal probability design might well be unbiased under an unequal probability design.)

These three arguments were not all equally valid or convincing, but even Gini and Galvani were ready to admit that something was seriously wrong with their approach. Moreover, the second argument (that the sample size of 29 was too small) was an easy one for Neyman to argue. It was incontrovertible. The third argument, that the population modelling was inadequate, was also one that the survey designers were ready to acknowledge. The first argument (about confidence intervals) seems to have been accepted for no better reason than that Neyman was saying it, and that since he was certainly right on the other two points, he was probably right on that one as well.

3.1 Bowley's opposition to Neyman's first argument and the outcomes

One statistician who was not prepared to accept Neyman's way of thinking was Bowley, who moved the vote of thanks to him for his 1934 presentation. We are, in consequence, able to quote the actual words used by both the disputants. Bowley actually started the argument by wondering aloud whether confidence intervals were just "a confidence trick"!

He asked, "Does [a confidence interval] really lead us to what we need—the chance that within the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event had occurred *or* that the proportion in the population is within these limits… The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity."

In his reply, Neyman asserted that Bowley's question (about the confidence interval being a confidence trick) "contain[ed] the statement of the problem in the form of Bayes" and that in consequence its solution "*must* depend upon the probability law *a priori*." He added, "In so far as we keep to the old form of the problem, any further progress is impossible." He thus concluded that there was a need to stop asking Bowley's "Bayesian" question and instead adopt the stance that Neyman's own "*either…or*" statement [that *either* an improbable event had occurred *or* the proportion of the population was within the stated limits] "form[ed] a basis for the practical work of a statistician concerned with problems of estimation…"

However, the fact remains that confidence intervals are not easy to understand. A confidence interval is in fact a sample-specific range of potentially true values of the parameter being estimated, which has been constructed so as to have a particular property. This property is that, over a large number of sample observations, the proportion of times that the true parameter falls inside that range (constructed for each

sample separately) is equal to a predetermined value known as the confidence level. This confidence level is conventionally written as $p = 1 - \alpha$, where α is small compared with unity. Conventional values for α are 0.05, 0.01, and sometimes 0.001. Thus, if many samples of size n are drawn independently from a normal distribution, the proportion of times that the true parameter value will lie within any given sample's own confidence interval will, before that sample is selected, be $[1 - \alpha]$.

"It is not the case, however, that the probability of this true parameter value lying within the confidence interval as calculated for any individual sample of size n will be $[1 - \alpha]$. The confidence interval calculated for any individual sample of size n will, in general, be wider or narrower than average and might be centred well away from the true parameter value, especially if n is small. It is also sometimes possible to recognise when a sample is atypical and, hence, make the informed guess that in this particular case, the probability of the true value lying in a particular 95% confidence interval differs substantially from 0.95."

Let us then consider, in particular, the most commonly used of all 95% confidence intervals, namely that between $p = 0.05$ and $p = 1.00$. (Fisher (1925) had actually suggested using the interval between $p = 1/22$ and $p = 1$.) Editors of publications in a great variety of fields (most of them not themselves statisticians) feel this definition of "significance" to be the one that very conveniently gives them leave to publish *p*-values that fall outside that range and reject those that do not. I believe the time is long overdue for looking at that suggestion of Fisher's very carefully.

What Fisher claimed (using $p = 1/22$ rather than $p = 0.05$) was that "Using this criterion we should be led to follow up a false indication only once in 22 trials". But what did he (and what do we now) mean by "following up a false indication"? What we *should* mean is this: that if the null hypothesis (H_0) is true, a "false indication", that is to say, "a misleadingly significant observation," will be observed , on average, once in 22 (or 20) times. But this is not what many non-statistical users of the *p*-statistic imagine that it means. Such users seem to think it means that only one in 20 of their "significant observations" (*i.e.*, that only one in 20 of all their observations with *p*-values less than 0.05) will be *misleadingly* significant.

That is the notorious *p*-statistic fallacy! (See Berger and Sellke (1987) for details.) To say "If H_0 is true, observations will be misleadingly described as 'significant' only once in 20 (or 22) times", is correct but unhelpful, for if H_0 is true, it follows that *every* observation described as "significant", for whatever reason, must also have been described that way misleadingly. But simply to say "Whether H_0 is true or not, $p < 0.05$", is also misleading. A meaningful false discovery rate (FDR) in these circumstances is (in fact) something that approximates to $p < 0.0025$ or $p < 0.05^2$.

This is a subject on which I have expended some thought of late. In particular, I co-authored a four-part article on it.

Part 1 (Brewer and Hayes 2011a) discusses how the notoriously parsimonious Bayesian Information Criterion (BIC) can be remedied by adding certain obviously needed penalty terms. The resulting Augmented Bayesian Information Criterion (ABIC) is nearly always intermediate between the original BIC and the (equally notoriously *lacking* in parsimony) Akaike Information Criterion (AIC). Another useful feature of the ABIC is that in its univariate case it is a simple function of T (the large sample limiting case of Student's t).

In Part 2 (Brewer and Hayes 2011b), a reference Bayesian hypothesis test is derived that is fully compatible with the ABIC of Part 1. An important role is played here by an obvious generalisation of Benford's (purely empirical) Law of Numbers, in providing an objective (though not flat) Bayesian prior distribution over the entire range from zero (or minus infinity) to plus infinity for the relevant hypothesis test. (The problem that characteristically arises with zero prior probabilities is avoided here by the use of Lebesgue-type measures instead.) Importantly, when $T = 1$, the relevant Bayesian hypothesis test yields a posterior measure that is indifferent between the null and alternative hypotheses. Furthermore, when the ABIC is generalised to small samples, as a function of the *t*-statistic, Fisher's p sets an upper bound to the false discovery rate (FDR), regardless of the number of degrees of freedom involved.

In Part 3 (Brewer, Hayes, and Gillison 2012), a set of some 1,300 regression slopes from a biodiversity sample survey of tropical landscape mosaics is used to provide empirical support for the ABIC, and the earlier theoretical findings are thereby confirmed.

In Part 4 (Hayes and Brewer 2012), the approximate results derived in Parts 1 to 3 are supplemented by exact results that can be obtained using a somewhat similar approach, but one that requires no explicit null hypothesis. Finally we suggest some likely consequences of the recognition that, when the implied null hypothesis is precise, much smaller values of $|p|$ (typically of the order of 0.0025 rather than 0.05) are needed to provide any useful FDR.

3.2 The acceptance of Neyman's second and third arguments

The second and third ideas that Neyman had advocated in his paper (namely the inefficiency of Gini and Galvani's (1929) selection procedure and the need to use only randomized sampling) though both relevant for their time and well presented, caught on only gradually over the course of the next decade. W. Edwards Deming heard Neyman in London in 1936. He was impressed and arranged for Neyman to lecture, and for his approach to be taught to U.S. government statisticians. A crucial event in its acceptance was the use in the 1940 U.S. Population and Housing Census of a one-in-twenty sample, designed by Deming along with Morris Hansen and others, to obtain answers to additional questions. Once fully accepted, however, Neyman's second and third arguments swept all other considerations aside for at least two decades.

Those twenty-odd years were a time of great progress. In the terms introduced by Kuhn (1962), finite population sampling had found a universally accepted "paradigm" in randomization-based inference, and an unusually long period of "normal science" based on "probability sampling" had ensued. ("Probability sampling" requires that all the elements in the population have known and positive probabilities of inclusion in sample.)

3.3 The appearance of relevant textbooks

This agreed consensus made it possible for several influential sampling textbooks to be published. Kish's (1995) historical article mentions five that appeared in quick succession: Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurwitz and Madow ("HH&M") (1953) and Sukhatme (1954).

In my estimation the two most important of these were those by Cochran and by HH&M, but for quite opposite reasons. HH&M seem not to have wanted any truck at all with population modelling. (I doubt

whether the word "model" is even mentioned in either of their two volumes. It does not appear in either index.) Cochran (1953), on the other hand found several uses for such models, even as early as 1953.

Re-reading Cochran (1953) recently, I had the distinct impression that the more he wrote, the more he was at ease in using population models. So I started to count them. This first edition had 316 pages of text. The words "model" and "models" were used on 23 occasions. In the first half of the book, the word "model" appeared only once (on page 123) and "models" not at all. But Cochran used those words again three times in the third quarter and 19 times in the last quarter. (Numbers sometimes speak louder than words!)

Another strange thing was that although HH&M's two-volume book on *Sample Survey Methods and Theory* appears not to have used the word "model" at all, each of its two volumes included a chapter on "regression estimation". I don't see how one can have a regression estimator without a regression model, at least in the back of one's mind.

HH&M also defined four "estimates" in Chapter 11 of their Volume 1: the *difference estimate*, the *regression estimate*, the *ratio estimate* and the *simple unbiased estimate*. In Chapter 11 of Volume 2 only the *difference estimate* and the *regression estimate* are defined, but of course the other two would have been well known to anyone who was already familiar with Volume 1.

The question still remains as to whether HH&M would have regarded the regression estimate as implying a model. My guess is that they would have been reluctant to do so!

3.4 My fifteen months in the USA

In 1966-67, I was privileged to spend over a year in the USA, visiting (in order) the U.S. Bureau of the Census in Washington DC, and then Harvard and Princeton Universities. At the Bureau of the Census I had hoped to be able to spend some time with Morris Hansen, and was looking forward to suggesting to him that there were actually some useful things that could be done with population models, but when the first opportunity occurred, he cut me off short, saying "We don't need *models*," and immediately changed the subject!

Conversely, when I went to Harvard, where I spent a considerable time with Cochran, we were able to look at the topic rationally together and agree that models had a useful if limited role to play. At Princeton, I attempted to interest several well-known statisticians at the university about the topic, but without any serious success.

Quite a different challenge to Hansen's model-free orthodoxy had been voiced by Godambe (1955), with his proof of the non-existence of any uniformly best randomization-based estimator of the population mean. A new notation and class of estimators were required for the argument, and this framework in its earliest form met with some resistance. In Section 5 of that paper, citing Yates' (1949) textbook and Cochran's (1939) paper as antecedents, Godambe suggested an alternative optimality criterion, the minimization of the expected sampling variance under what was later called a superpopulation model.

At that time few others working in this excitingly innovative field of survey sampling seemed to be concerned by this result. I must confess that I wasn't myself concerned at the time, but I now think that perhaps I should have been!

4 The third controversy: "Sampling inference: Model-assisted or model-based?"

It came as a considerable shock to the finite population sampling establishment when Royall (1970) issued his highly readable call to arms for the reinstatement of purposive sampling and prediction-based inference. To read this paper was to read Neyman (1934) being stood on its head. The identical issues were being considered but the opposite conclusions were being drawn.

By 1973, however, Royall had withdrawn the most extreme of his recommendations. This was that the best sample to select would be the one that was optimal in terms of a model represented by the following Equations:

$$Y_i = \beta X_i + U_i \tag{4.1}$$

$$E(U_i) = 0 \tag{4.2}$$

$$E(U_i^2) = \sigma^2 X_i \tag{4.3}$$

and

$$E(U_i U_j) = 0. \tag{4.4}$$

Such a sample would typically have consisted of the n largest units in the population as measured by their realized x_i values, asking for trouble if the parameter β had not been close to constant over the entire range of the sizes of the population units.

In later articles (Royal and Herson 1973a, Royal and Herson 1973b, Cumberland and Royall 1981), Royall suggested that the chosen sample be "balanced," in other words, that the moments of the sample x_i should be as close as possible to the corresponding moments of the whole population. This formalized the much earlier notion that samples should be chosen purposively to resemble the population in miniature. The samples of Gini and Galvani had been chosen in something of the same way – meaning here "something of the same way in intention", but certainly not anything like the same success in execution.

For the most part, Royall's original stand remained unshaken. The business of a sampling statistician was to make a realistic model of the relevant population, design a sample to estimate its parameters, and make all inferences regarding that population in terms of those parameter estimates. The randomization-based concept of defining the variance of an estimator in terms of the variability of its estimates over all possible samples was to be discarded in favour of the prediction-based variance, which was sample-specific, and based on averaging all possible realizations of the chosen prediction model.

Regardless of what sample was drawn, Royall's estimator for a population total $T_y = \sum_U y_i$ had this prediction form:

$$t_y = \sum_s y_i + \sum_{U-s} x_i \hat{\beta}_{\text{BLUE}},$$

where $\hat{\beta}_{\text{BLUE}} = \sum_s y_i \Big/ \sum_s x_i$ was the best linear unbiased estimator for β based on the sample under model in equation (4.1). This is in prediction form since the *y*-values of $U - s$ are predicted by the model.

Sampling statisticians had at no stage been slow to take sides in this debate. Now the battle-lines were drawn. The heat of the argument appears to have been exacerbated by language-blocks; for instance the words "expectation" and "variance" carried one set of connotations for randomization-based inference and quite a different set for prediction-based inference. So assertions made on one side appeared to those on the other side to be unintelligible nonsense.

A major establishment counter-attack was launched with an article by Hansen, Madow and Tepping (1983). A small (and by most standards undetectable) divergence from Royall's model was shown nevertheless to be capable of distorting the sample inferences substantially. The obvious counter would have been "But this distortion would not have occurred if the sample had been drawn in a balanced fashion."

5 A third alternative, "Use them both together"

Eventually, a third position was also offered, the one held by the present author, namely that since there were merits in both the design-based (or randomization-based) and the model-based (or prediction-based) approaches, and that since it was possible to combine them, the two should be used together. I had actually foreshadowed this possibility in Brewer (1963), a paper that provoked little interest at the time, but was later spotted and accorded recognition by J.N.K. Rao, at least to the extent that he invited me to visit him in Ottawa for six weeks in 1974.

To combine these two approaches was relatively simple. In each of them there was a variable y which was of central interest and a related or auxiliary variable x, about which something additional was known that could be of assistance in estimating the value of that y variable. That "something additional" was typically the known population total of all the x values, denoted by T_x. Consequently the *relationship* of central interest, was that which linked the crucial parameter β in equation (4.1) to its *cosmetic* estimator $\hat{\beta}_{\text{COS}}$, namely

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (\pi_i^{-1} - 1) y_i}{\sum_s (\pi_i^{-1} - 1) x_i}, \tag{5.1}$$

where π_i is the probability that unit i is selected in the sample, or in the notation used by Särndal (2011),

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (d_k - 1) y_i}{\sum_s (d_k - 1) x_i}, \tag{5.2}$$

where his d_k is identical to my π_i^{-1}. The resulting estimator of the total $Y = \sum_U y_k$ is

$$\hat{Y}_{\mathrm{COS}} = \sum_s d_k y_k + \left(\sum_U x_k - \sum_s d_k x_k \right) \frac{\sum_s (d_k - 1) y_k}{\sum_s (d_k - 1) x_k}. \qquad (5.3)$$

Särndal (2011) also shows that these x and y values can be related to each other in several different ways, but also shows that there is a common theme that runs through all of those ways. That common theme is that y increases linearly as x increases, and that the extent of that linearity is measured by the parameter β in equation (4.1). Importantly, however, when $\hat{\beta}_{\mathrm{COS}}$ replaces $\hat{\beta}_{\mathrm{BLUE}}$ in Royall's prediction estimator, the estimator can be shown to be nearly unbiased under the design regardless of the validity of the assumed model.

Equation (5.2) can also be found explicitly on page 569 of Brewer (2011), immediately following its more general formula in matrix notation, namely

$$\hat{\beta}_{\mathrm{COS}} = \left[X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) X_s \right]^{-1} X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) y_s . \qquad (5.4)$$

When, the question arises as to how many explanatory variables should be used in the relevant model, Särndal (2011) makes an apparently disparaging distinction between "explanatory rich" and "explanatory poor" countries. He certainly treats those "explanatory poor" countries as being at a substantial disadvantage as a result of having relatively few "explanators".

There is at least one "explanatory rich" country (Australia) that appears to have made a deliberate decision to ignore whatever advantages might be available to those that are "explanatory rich". The current Australian procedure (the one used primarily to produce seasonally adjusted series) is to use only a single auxiliary variable, namely the latest available Census total, as the single "explanator".

Earlier, Brewer (1999a) had also presented a case that it might be preferable to use a cosmetic regression estimator to compensate for any lack of balance, rather than go to the trouble of selecting balanced samples. However, those who prefer to use balanced sampling directly can now select randomly from among many balanced or nearly balanced samples using the "cube method" (Deville and Tillé 2004). That paper also contains several references to earlier methods of selecting balanced samples, but regardless of how the relevant balanced sample is arrived at, the ways in which it needs to be used are identical.

In Brewer and Gregoire (2009) all three of the relevant approaches to estimation (randomization alone, prediction alone, and the two together) are examined. At this point, it is convenient to quote from yet another paper of mine (Brewer 2005, pages 390-391) which sets out the reasons why I was, and still am, concerned to use both methods simultaneously, and how readily it can be done.

"Each approach has its merits, and there are advantages in using both together. Consider how each of these inferences works.

First, design-based inference. Consider the general case where the inclusion probabilities π_i are known but may differ from unit to unit. In that case we can imagine the sampling statistician constructing a model of the population by looking at each of the sample units in turn and saying, *Oh yes, you (the first unit) were included with one chance in 10, so my model of the population includes you and nine other non-sample units with the same Y_k value as you. But you (the second unit) you were included with only one chance in two, so my model includes you and only one other unit like you.*"

The consequence of using this procedure here was therefore that the model of the population in the sampler's mind would consist of two real sample units (one from each sample stratum) plus ten imaginary units, (nine from the stratum with a sample fraction of one in ten, plus one from the stratum with a sample fraction of one in two) and finally plus all the units from the completely enumerated stratum.

Brewer (2005, page 391) continues as follows: "So even design-based estimation can be thought of as being based on a model, but on a model quite different from the prediction models… that are favoured by the so-called *model-based* school. More accurately that school should be described as *prediction-based* and the *design-based* school should be described as *randomization-based*. Each school uses a model, but one uses a prediction model and the other a randomization model."

The randomization-based approach described above is the one that was used for the selection of two sample units (one from each sampled stratum) plus all the units in the completely enumerated stratum. It also gave rise to the well-known Horvitz-Thompson estimator, which may be written

$$\hat{T}_{\mathrm{HT}} = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} \delta_i \frac{Y_i}{\pi_i} \tag{5.5}$$

where δ_i is an inclusion indicator taking the value "one" if the i^{th} unit is either in the sample or in the completely enumerated sector, and the value "zero" otherwise. In this particular case it is defined over both the two sampled units and also all the units in the completely enumerated sector. [This last sentence corrects the error mentioned above.]

Statisticians of the prediction-based school ridicule the use of randomization-based inference because the inclusion probabilities are chosen arbitrarily by the sample designer, and are therefore unable (they say) to tell us anything meaningful about the population! They prefer instead to use the Best Linear Unbiased Estimator (BLUE) of the regression parameter β as a step towards arriving at the Best Linear Unbiased Predictor (BLUP) of T. It is a predictor, because T is a random variable under the model, not a parameter.

Which is then the better estimator of T, the HT or the BLUP? The BLUP is the better if the prediction model holds exactly, and is much the better if both the sample and the population are small. However there will always be some sample size beyond which the HT is the more efficient estimator unless the model holds exactly.

6 Summary

In conclusion, we can see that survey sampling, over its relatively short history, has been remarkably vulnerable to controversies. In the first instance there was opposition to the notion that there should be any sampling at all. The only valid source of statistical information was taken to be the complete collection. It took the determination of Kiaer, a person already in a senior position of authority, to break down the opposition to what was eventually demonstrated to be a valuable tool.

The second controversy was also due to the determination of just a few people. Neyman took the lead, but this time there were others who were involved. Bowley was certainly involved to start with, but Neyman seems to have had the more convincing arguments at the crucial time. They were controversial,

even to begin with, and I am certainly not impressed with them now, but at the time he found a ready disciple in Hansen, who dominated the sampling fraternity for decades, at least until the mid-1970s.

The third controversy is still in progress and it is not altogether clear as to how it will turn out, but my current preference (at least for middling-sized samples) would be to use the prediction and randomization estimators combined.

In summary, both the HT and the BLUP can be useful in different situations. The BLUP makes sense to use when the sample size is small, and a model is desperately needed. The HT provides protection against prediction-model failure as the sample grows large. A prudent statistician would combine the principles of both.

References

Berger, J.O., and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p*-values and evidence (with discussion). *Journal of the American Statistical Association*, 82, 112-139.

Bowley, A.L. (1906). Address to the economic and statistics section of the British association for the advancement of science. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1912). Working class households in reading. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1926). Measurement of the precision obtained in sampling. *Bulletin of the International Statistical Institute*, 22, 11-62 (supplement).

Brewer, K.R.W. (2011). Remarks on the paper on "Combined inference in survey sampling" by Carl-Erik Särndal. *Pakistan Journal of Statistics*, 27, 4, 567-572.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 10, 213-233.

Brewer, K.R.W. (1999a). Design-based or model-based inference? Stratified random *vs* stratified balanced sampling. *International Statistical Review*, 67, 35-47.

Brewer, K.R.W. (1999b). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 205-212.

Brewer, K.R.W. (2005). Anomalies, probings, insights: Ken Foreman's role in the sampling inference controversy of the late 20th century. *Australian and New Zealand Journal of Statistics*, 47, 4, 385-399.

Brewer, K.R.W., and Gregoire, T.G. (2009). Introduction to survey sampling. Chapter 1 of *Handbook of Statistics* 29A, *Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfefferman and C.R. Rao), Elsevier.

Brewer, K.R.W., and Hayes, G. (2011a). Understanding and using Fisher's *p*: Part 1: Countering the *p*-statistic Fallacy. *Mathematical Scientist*, 36, 107-116.

Brewer, K.R.W., and Hayes, G. (2011b). Understanding and using Fisher's *p*: Part 2: A Reference Bayesian Hypothesis Test. *Mathematical Scientist*, 36, 117-125.

Brewer, K.R.W., Hayes, G. and Gillison, A.N. (2012). Understanding and using Fisher's *p*: Part 3: Examining an Empirical Data Set. *Mathematical Scientist*, 37, 20-26.

Cochran, W.G. (1953). *Sampling Techniques.* First Edition, Wiley.

Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492- 510.

Cochran, W.G. (1978). Laplace's ratio estimator. In *Contributions to Survey Sampling and Applied Statistics*; papers in honor of H.O. Hartley; H.A. David (Editor), 3-10.

Deming, W.E. (1950). *Some theory of sampling.* Dover books on mathematics.

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

Fisher, R.A. (1925). *Statistical methods for research workers.* 14th Edition (1970) Oliver and Boyd.

Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentativo all' ultimo censimento italiano della popolazione (1 dicembre 1921). *Annali di statistica* VI 4, 1-107.

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

Graunt, J. (1661/2). Natural and political observations made upon the Bills of Mortality. Reprinted (1939) Baltimore: The John Hopkins Press.

Hansen, M.H., Hurwitz W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory (2 vols.)* (Republished 1993) Wiley, New York.

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

Hayes, G., and Brewer, K.R.W. (2012). Understanding and using Fisher's *p*: Part 4: Do we even need to specify a prior measure at H_0? *Mathematical Scientist*, 37, 27-33. Sons, New York.

Kiaer, A.N. (1897). The representative method of statistical surveys. Papers from the Norwegian Academy of Science and Letters, II The Historical, philosophical Section, 1897 No. 4.

Kish, L. (2003). Selected Papers. Graham Kalton (Editor) Steven Heeringa (Editor) Wiley.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2 (5), 813- 830.

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Lie, E. (2002). The rise and fall of sample surveys in Norway, 1875-1906. *Science in Context*, 15 (3), 385-1906.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

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

Royall, R.M., and Herson, J. (1973a). Robust estimation in finite population I. *Journal of the American Statistical Association*, 68, 880-889.

Royall, R.M., and Herson, J. (1973b). Robust estimation in finite population II: Stratification on a size variable. *Journal of the American Statistical Association*, 68, 890-893.

Särndal, C.-E. (2011). Combined inference in survey sampling. *Pakistan Journal of Statistics*, 27 (4) 359-370.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.

Sukhatme, P.V. (1954). Sampling theory of surveys: With applications. Asia Publishing House.

Wright, T. (2001). Selected moments in the development of probability sampling: Theory and practice. *Survey research methods section newsletter*, American Statistical Association, Alexandria, VA. Issue 13, 1-6.

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

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 263-282
Statistics Canada, Catalogue No. 12-001-X

A weighted composite likelihood approach to inference for two-level models from survey data

J.N.K. Rao, François Verret and Mike A. Hidiroglou[1]

Abstract

Multi-level models are extensively used for analyzing survey data with the design hierarchy matching the model hierarchy. We propose a unified approach, based on a design-weighted log composite likelihood, for two-level models that leads to design-model consistent estimators of the model parameters even when the within cluster sample sizes are small provided the number of sample clusters is large. This method can handle both linear and generalized linear two-level models and it requires level 2 and level 1 inclusion probabilities and level 1 joint inclusion probabilities, where level 2 represents a cluster and level 1 an element within a cluster. Results of a simulation study demonstrating superior performance of the proposed method relative to existing methods under informative sampling are also reported.

Key Words: Composite likelihood; Inclusion probabilities; Informative sampling; Multi-level models.

1 Introduction

Data collected from large-scale socio-economic, health and other surveys are extensively used for analysis purposes, such as inference on the regression parameters of linear and logistic linear regression population models. Ignoring the survey design features (such as stratification, clustering and unequal selection probabilities) can lead to erroneous inferences on model parameters because of sample selection bias caused by informative sampling. It is tempting to expand the models by including among the auxiliary variables all the design variables that define the selection process at the various levels and then ignore the design and apply standard methods to the expanded model. The main difficulties with this approach are the following (Pfeffermann and Sverchkov 2003): (1) Not all design variables may be known or accessible to the analyst; (2) Too many design variables can lead to difficulties in making inference from the expanded model; (3) The expanded model may no longer be of scientific interest to the analyst. On the other hand, the design-based approach can provide asymptotically valid repeated sampling inferences without changing the analyst's model. A unified approach based on the survey weighted estimating equations leads to design-consistent estimators of the "census" or finite population parameters which in turn estimate the associated model parameters. Further, re-sampling methods, such as the jackknife and the bootstrap for survey data, can provide valid variance estimators and associated inferences on the census parameters. The same methods may also be applicable to inference on the model parameters, in many cases of large-scale surveys. In other cases, it is necessary to estimate the model variance of the census parameters from the sample. The estimator of the total variance is then given by the sum of this estimator and the re-sampling variance estimator. Beaumont and Charest (2010) extended the bootstrap to estimate the total variance associated with the model parameters. We refer the reader to Rao *et al.* (2010) for an overview of methods for making inference on regression parameters from complex survey data.

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: jrao@math.carleton.ca; François Verret, Statistics Canada, 15 B, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: francois.verret@statcan.gc.ca; Mike A. Hidiroglou, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: mike.hidiroglou@statcan.gc.ca.

In this paper, our focus is on making design-based inference on the variance component parameters and regression parameters of multi-level models from data obtained from multi-stage sampling designs corresponding to the levels of the model. For example, in an education study of students, schools (first-stage sampling units) may be selected with probabilities proportional to school size and students (second-stage units) within selected schools by stratified random sampling. Again, ignoring the survey design and using traditional methods for multi-level models can lead to erroneous inferences in the presence of sample selection bias. In the design-based approach, estimation of variance component parameters of the model is more difficult than that of regression parameters. Past work on multi-level models for survey data is summarized in Section 2. Our main purpose is to present a unified approach to making inference for general multi-level models from survey data, based on a weighted log composite likelihood approach (Section 4). The proposed methods lead to asymptotically valid inferences on the variance component parameters even when the within-cluster sample sizes are small, provided the number of sample clusters is large, unlike some of the existing methods summarized in Section 2. Limited simulation results are presented in Section 5.

2 Two-level models: Past work

2.1 Two-level models

Multi-level (or hierarchical) models are extensively used in social sciences, education, health and other areas to analyze survey data with a hierarchical structure. Here we focus on two-level models associated with two-stage sampling of clusters (level 2): a sample, s, of level 2 units, i, is selected according to a specified design and then a sample, $s(i)$, of elements (or level 1 units), j, is selected from each sampled level 2 unit i according to another specified design. We assume, following the literature on multi-level models for survey data, that the model matches the design hierarchy, as in the example of an educational survey of students. However, in some multipurpose surveys, the design hierarchical structure could be quite different from the model hierarchy. For example, the Canadian National Longitudinal Survey of Children and Youth uses a multi-stage design where the stages are geographical areas, households within an area and students within a household, whereas an educational multilevel model may include as levels students, classes, schools and school boards (Rao and Roberts 1998). Since the design clusters cut across the model clusters for such surveys, it is difficult to develop a suitable design-weighted method of inference on the model parameters that can handle informative sampling of clusters and or elements within sampled clusters. Under informative sampling, the assumed model for the population may not hold for the sample.

Let N be the number of level 2 units in the population and M_i be the number of level 1 units in the level 2 unit i. A two-level super-population model is given by

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right), \ \mathbf{v}_i \sim_{iid} f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right), i = 1,...,N \ ; \ j = 1,...,M_i, \tag{2.1}$$

where y_{ij} and $\mathbf{x}_{ij} = \left(x_{ij0},...,x_{ij,p-1}\right)^T$ are the response and a p-vector of covariate values associated with element j within cluster i and $x_{ij0} = 1$, $\mathbf{v}_i$ denotes a level 2 random effect, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote the

parameters associated with the two stages of the assumed model. Here $f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)$ and $f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right)$ are specified density functions of y_{ij} given $\mathbf{x}_{ij}$ and $\mathbf{v}_i$ and of $\mathbf{v}_i$, respectively. Note that in model (2.1), the responses y_{ij} for a given i are assumed to be conditionally independent given the random effect $\mathbf{v}_i$ but they are correlated marginally due to the common $\mathbf{v}_i$. The model formulation (2.1) covers both linear two-level models and generalized linear two-level models. Under informative sampling of clusters and or elements within sampled clusters, standard methods for multi-level models that ignore the design and assume that model (2.1) holds for the sample can lead to asymptotically biased estimators of model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (Pfeffermann *et al.* 1998).

Special cases

(1) A simple nested error mean model that is often used in simulation studies related to two-level models is given by

$$y_{ij} = \mu + v_i + e_{ij}, e_{ij} \sim_{iid} N\left(0, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \tag{2.2}$$

where $i = 1,...,N; j = 1,...,M_i$. Model (2.2) may be written in the form (2.1) as

$$y_{ij} \mid v_i \sim_{ind} N\left(\mu + v_i, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \boldsymbol{\theta}_1 = \left(\mu, \sigma_e^2\right), \theta_2 = \sigma_v^2.$$

Marginally, $y_{ij} \sim N\left(\mu, \sigma_v^2 + \sigma_e^2\right)$ but y_{ij} and y_{ik} $(j \neq k)$ are correlated: $\operatorname{corr}\left(y_{ij}, y_{ik}\right) = \rho = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2\right), j \neq k$.

(2) A linear two-level model, often used in practice, is given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_i + e_{ij}, i = 1,...,N; j = 1,...,M_i, \tag{2.3}$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim_{iid} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}_v\right), i = 1,...,N$ and $e_{ij} \sim_{iid} N\left(0, \sigma_e^2\right)$. This model may also be expressed in the form (2.1) as

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} N\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2\right), \mathbf{v}_i \sim_{iid} N_p\left(0, \boldsymbol{\Sigma}_v\right) \tag{2.4}$$

where $\boldsymbol{\theta}_1 = \left(\boldsymbol{\beta}^T, \sigma_e^2\right)^T$ and $\boldsymbol{\theta}_2$ is the vector of $p(p+1)/2$ distinct elements of $\boldsymbol{\Sigma}_v$. Marginally, $y_{ij} \sim N\left(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} + \sigma_e^2\right)$, but y_{ij} and y_{ik} $(j \neq k)$ are correlated through the common random effect $\mathbf{v}_i$. However, in the case of a generalized linear two-level model, the marginal distribution of y_{ij} generally does not yield a closed-form expression; for example, in the case of a logistic linear two-level model for binary responses.

2.2 Point estimation

The "census" or population log-likelihood under the assumed two-level model (2.1) is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log L_i(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) = l(\boldsymbol{\theta}), \tag{2.5}$$

where $\boldsymbol{\theta}$ is the vector with elements $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and

$$L_i(\boldsymbol{\theta}) = \int \exp\left[\sum_{j=1}^{M_i} \log f\left(y_{ij} \middle| \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)\right] f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right) d\mathbf{v}_i \tag{2.6}$$

see Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). The census score function $\mathbf{U}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ satisfies $E_m\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$, where E_m denotes the model expectation. The census parameter $\boldsymbol{\theta}_N$ is defined as the unique solution to $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ and $\boldsymbol{\theta}_N$ is model consistent for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}_N$ is the vector with elements $\boldsymbol{\theta}_{1N}$ and $\boldsymbol{\theta}_{2N}$.

Let the sample consist of n clusters with m_i elements from sample cluster i. Let π_i and $\pi_{j|i}$ respectively denote the level 2 and level 1 inclusion probabilities associated with cluster i and element j within cluster i. Then the level 2 and level 1 weights are given by $w_i = \pi_i^{-1}$ and $w_{j|i} = \pi_{j|i}^{-1}$ respectively. Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) proposed a weighted sample pseudo log-likelihood obtained by replacing $\sum_{j=1}^{M_i}(.)$ in (2.6) by $\sum_{j \in s_i} w_{j|i}(.)$ and $\sum_{i=1}^{N}(.)$ in (2.5) by $\sum_{i \in s} w_i(.)$, where s denotes the sample of clusters and $s(i)$ denotes the sample of elements within clusters $i \in s$. It is given by

$$\tilde{l}_w(\boldsymbol{\theta}) = \sum_{i \in s} w_i \tilde{l}_{wi}(\boldsymbol{\theta}) \tag{2.7}$$

where $\tilde{l}_{wi}(\boldsymbol{\theta}) = \log \tilde{L}_{wi}(\boldsymbol{\theta})$ and

$$\tilde{L}_{wi}(\boldsymbol{\theta}) = \int \exp\left[\sum_{j \in s(i)} w_{j|i} \log f(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1)\right] f(\mathbf{v}_i \mid \boldsymbol{\theta}_2) d\mathbf{v}_i. \tag{2.8}$$

Maximizing the pseudo log-likelihood $\tilde{l}_w(\boldsymbol{\theta})$, given by (2.7), we get a pseudo maximum likelihood (PML) estimator $\tilde{\boldsymbol{\theta}}_w$. Computational details are discussed in Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). In the special case of linear two-level models, Pfeffermann *et al.* (1998) used an iterative generalized least squares method proposed by Goldstein (1986). Note that we need both level 1 and level 2 weights to compute $\tilde{\boldsymbol{\theta}}_w$, unlike in the case of marginal models that require only the unconditional element weights $w_{ij} = w_i w_{j|i}$.

Design consistency of the PML estimator $\tilde{\boldsymbol{\theta}}_{2w}$ of the census parameter $\boldsymbol{\theta}_{2N}$ or design-model consistency of $\tilde{\boldsymbol{\theta}}_{2w}$ as an estimator of the model parameter $\boldsymbol{\theta}_2$ requires that both the number of sample clusters, n, and the within cluster sample sizes, m_i, tend to infinity, even in the linear case. Also, the relative bias of the estimators will be considerable when m_i are small. To remedy this problem, several weight-scaling methods have been proposed in the literature. In particular, level 1 weights $w_{j|i}$ in (2.8) are scaled by a factor k_{1i} before maximizing the pseudo log-likelihood (2.7). We consider only two weight-scaling methods here, denoted A and A1 (Asparouhov 2006). Method A uses

$$k_{1i} = m_i \Big/ \sum_{j \in s(i)} w_{j|i} \tag{2.9}$$

In method A1, k_{1i} is the same as in method A but level 2 weights w_i are also scaled by the factor $k_{2i} = 1/k_{1i}$ to offset level 1 weight scaling. Asparouhov (2006) mentioned the use of accelerated EM algorithm for calculating the PML estimator $\tilde{\boldsymbol{\theta}}_w$ with M plus 3: www. Statmodel.com: Muthén and Muthén, 1998-2005.

2.3 Variance estimation

Turning to variance estimation, Asparouhov (2006) proposed a Taylor linearization "sandwich" variance estimator of $\tilde{\boldsymbol{\theta}}_w$. It is given by

$$v_L\left(\tilde{\boldsymbol{\theta}}_w\right) = \left(\tilde{\mathbf{I}}''_w\right)^{-1}\left[\sum_{i \in s}\left(k_{2i}w_i\right)^2 \tilde{\mathbf{I}}'_{wi}\left(\tilde{\mathbf{I}}'_{wi}\right)^T\right]\left(\tilde{\mathbf{I}}''_w\right)^{-1}, \tag{2.10}$$

where $\tilde{\mathbf{I}}'_w$ and $\tilde{\mathbf{I}}''_w$ respectively denote the first derivative vector and the second derivative matrix of $\tilde{l}_w(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$, and $\tilde{\mathbf{I}}'_{wi}$ is the first derivative of $\tilde{l}_{wi}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$. If the level 2 sampling fraction is small, then $v_L\left(\tilde{\boldsymbol{\theta}}_w\right)$ tracks the variance of $\tilde{\boldsymbol{\theta}}_w$ well, but not the MSE of $\tilde{\boldsymbol{\theta}}_w$ if the relative bias of $\tilde{\boldsymbol{\theta}}_w$ is large.

Kovacevic *et al.* (2006) studied bootstrap variance estimators for $\tilde{\boldsymbol{\theta}}_w$. They considered two options: options 1 and 2. In option 1, level 2 bootstrap weights $w_i(b)$, based on the Rao, Wu and Yue (1992) method, are used and level 1 weights are not changed, *i.e.*, $w_{j|i}(b) = w_{j|i}$, where $b = 1,...,B$ denote the B bootstrap samples. For option 2, the Rao, Wu and Yue (1992) bootstrap method is applied to both level 1 and level 2, and the level 1 bootstrap weights are rescaled. Replacing the weights w_i and $w_{j|i}$ by $w_i(b)$ and $w_{j|i}(b)$ in (2.7) and (2.8), bootstrap PML estimators $\tilde{\boldsymbol{\theta}}_w(b), b = 1,...,B$ are obtained and the resulting bootstrap variance estimator is given by

$$v_{Boot}\left(\tilde{\boldsymbol{\theta}}_w\right) = \frac{1}{B}\sum_{b=1}^{B}\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]^T. \tag{2.11}$$

A simulation study of (2.11), based on the simple mean model (2.2), showed that option 1 may lead to underestimation of the variance of $\tilde{\sigma}^2_{ew}$. Option 2 performed better than option 1. Grilli and Pratesi (2004) studied an alternative bootstrap method for variance estimation.

3 Design-weighted estimating equations

In Sections 3 and 4 we study methods of generating design-weighted estimating equations for the model parameters of multi-level models that lead to design-model consistent estimators, even in the case of small within-cluster sample sizes. The proposed methods depend only on the first order inclusion probabilities π_i and $\pi_{j|i}$ and the joint inclusion probabilities $\pi_{jk|i}$ within clusters. Section 3 introduces a

simple moment-based weighted estimating equations approach applicable to linear nested error regression models. A unified method, based on weighted log composite likelihoods, is proposed in Section 4. This method can handle linear and generalized linear multi-level methods, unlike the moment-based method, and it leads to design-model consistent estimators. It also depends only on π_i, $\pi_{j|i}$ and $\pi_{jk|i}$.

3.1 Point estimation

We first illustrate the weighted estimating equations approach, using the simple mean model (2.2). Here our interest is to estimate $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ from a two-stage cluster sampling design matching the model hierarchy. We have chosen the following three estimating functions (EF) for this purpose:

$$u_1\left(y_{ij}, \boldsymbol{\theta}\right) = y_{ij} - \mu, \tag{3.1}$$

$$u_2\left(y_{ij}, \boldsymbol{\theta}\right) = \left(y_{ij} - \mu\right)^2 - \left(\sigma_v^2 + \sigma_e^2\right) \tag{3.2}$$

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[\left(y_{ij} - \mu\right) - \left(y_{ik} - \mu\right)\right]^2 - 2\sigma_e^2 = z_{ijk}{}^2 - 2\sigma_e^2, j \neq k, \tag{3.3}$$

where $z_{ijk} = y_{ij} - y_{ik}$. The corresponding census estimating equations are given by

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) = 0, U_2(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) = 0 \tag{3.4}$$

$$U_3(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j<k=1}^{M_i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = 0. \tag{3.5}$$

The resulting census parameter, $\tilde{\boldsymbol{\theta}}_N$, is model-consistent for $\boldsymbol{\theta}$ because the model expectations of the three estimating functions (3.1) – (3.3) are zero. It follows from (3.4) and (3.5) that the design-weighted estimating equations (WEE) are given by

$$\hat{U}_{w1}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w1i}(\boldsymbol{\theta}) = 0 \tag{3.6}$$

$$\hat{U}_{w2}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w2i}(\boldsymbol{\theta}) = 0 \tag{3.7}$$

$$\hat{U}_{w3}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w3i}(\boldsymbol{\theta}) = 0, \tag{3.8}$$

where $w_{jk|i} = \pi_{jk|i}^{-1}$. The WEE estimator, $\hat{\boldsymbol{\theta}}_w$, is obtained by solving (3.6) – (3.8). For the mean model, we obtain explicit solutions to WEE as

$$\hat{\mu}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij} y_{ij}\right) \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} \equiv \bar{y}_w \tag{3.9}$$

$$\hat{\sigma}^2_{vw} = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \bar{y}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}^2_{ew} \tag{3.10}$$

$$\hat{\sigma}^2_{ew} = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} z^2_{ijk} \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right), \tag{3.11}$$

where $w_{ij} = w_i w_{j|i}$. Note that the above moment method is distribution free.

We note that $\hat{U}_{wt}(\boldsymbol{\theta}), t = 1,2,3$ are estimating functions with zero expectation with respect to the design and the model, *i.e.*, $E_m E_p\left\{\hat{U}_{wt}(\boldsymbol{\theta})\right\} = 0$. Using this result, it can be shown that the WEE estimator $\hat{\boldsymbol{\theta}}_w = \left(\hat{\mu}_w, \hat{\sigma}^2_{vw}, \hat{\sigma}^2_{ew}\right)^T$ is design-model consistent for $\boldsymbol{\theta}$ as the number of level 2 units in the sample, n, increases, even when the within cluster sample sizes, m_i, are small. This property does not necessarily hold for the estimators presented in Section 2. The proposed method, however, requires the within-cluster joint inclusion probabilities $\pi_{jk|i}$. The latter are readily available for simple random or stratified random sampling within clusters, or when the within cluster sampling fraction is small. Also several good approximations to $\pi_{jk|i}$ when sampling within clusters is based on unequal probability sampling are also available, and those approximations depend only on the marginal inclusion probabilities $\pi_{j|i}$ (Haziza, Mecatti and Rao 2008). The WEE estimator $\hat{\boldsymbol{\theta}}_w$ is also design-consistent for $\tilde{\boldsymbol{\theta}}_N$, noting that $E_p\left\{\hat{U}_{wt}(\tilde{\boldsymbol{\theta}}_N)\right\} = 0$, $t = 1,2,3$.

The choice of estimating functions (3.1) – (3.3) is not necessarily unique. For example, we could replace the previous $u_2\left(y_{ij}, \boldsymbol{\theta}\right)$ by $\tilde{u}_2\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left(y_{ij} - \mu\right)\left(y_{ik} - \mu\right) - \sigma^2_v$ in (3.7) and retain (3.6) and (3.8). The resulting WEE estimator is also design-model consistent for $\boldsymbol{\theta}$ as the number of level 2 units increases. The weighted pairwise composite likelihood approach of Section 4 provides a unified method of generating the estimating functions.

Korn and Graubard (2003) used an alternative approach for the mean model which has some similarities with the proposed approach. Under this approach, "census parameters", S^2_e and S^2_v are first obtained by assuming that the model holds for the finite population. Survey weighted estimators $\hat{S}^2_{ew}$ and $\hat{S}^2_{vw}$ of the census parameters are then obtained, assuming M_i is known for the sampled clusters. The estimator $\hat{S}^2_{ew}$ is given by

$$\hat{S}^2_{ew} = \left\{\frac{1}{2}\sum_{i\in s}\left(M_i - 1\right) w_i \left[\sum_{j<k\in s(i)} w_{jk|i}\left(y_{ij} - y_{ik}\right)^2 \Big/ \sum_{j<k\in s(i)} w_{jk|i}\right]\right\}\left[\sum_{i\in s}\left(M_i - 1\right) w_i\right]^{-1}, \tag{3.12}$$

assuming $m_i > 1$ for all sampled clusters. Note that (3.12) requires the joint inclusion probabilities $\pi_{jk|i}$ as in the proposed method, but it induces within-cluster ratio bias when the within-cluster sample sizes are small unlike our method. The expression for $\hat{S}^2_{vw}$ is more complicated and we refer the reader to Korn and Graubard (2003) for the relevant formula.

The WEE method readily extends to the nested error linear regression model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij};\ e_{ij} \sim_{iid} N\left(0, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right). \tag{3.13}$$

In this case, the estimating function (3.1) is changed to

$$u_1\left(y_{ij}, \boldsymbol{\theta}\right) = \mathbf{x}_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right), \tag{3.14}$$

(3.2) to

$$u_2\left(y_{ij}, \boldsymbol{\theta}\right) = \left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2 - \left(\sigma_v^2 + \sigma_e^2\right) \tag{3.15}$$

and (3.3) to

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T\boldsymbol{\beta}\right]^2 - 2\sigma_e^2,\ j \neq k, \tag{3.16}$$

where $\boldsymbol{\theta}$ is the vector with elements $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 and $z_{ijk} = y_{ij} - y_{ik}$. Explicit solutions to $\hat{U}_{wt}(\boldsymbol{\theta}) = 0,\ t = 1,2,3$ corresponding to (3.14) – (3.16) are obtained as

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}y_{ij}\right), \tag{3.17}$$

$$\hat{\sigma}_{vw}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.18}$$

and

$$\hat{\sigma}_{ew}^2 = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T\hat{\boldsymbol{\beta}}_w\right]^2 \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right). \tag{3.19}$$

3.2 Variance estimation

A Taylor linearization sandwich variance estimator of the WEE estimator $\hat{\boldsymbol{\theta}}_w$ can be obtained along the lines of the variance estimator (2.10), provided the level 2 sampling fraction is small. Let $\hat{\mathbf{U}}_w(\boldsymbol{\theta})$ be the column vector with components $\hat{U}_{w1}(\boldsymbol{\theta})$, $\hat{U}_{w2}(\boldsymbol{\theta})$ and $\hat{U}_{w3}(\boldsymbol{\theta})$ and similarly $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ be the column vector with components $\hat{U}_{w1i}(\boldsymbol{\theta})$, $\hat{U}_{w2i}(\boldsymbol{\theta})$ and $\hat{U}_{w3i}(\boldsymbol{\theta})$. Then the linearization variance estimator is given by

$$v_L\left(\hat{\boldsymbol{\theta}}_w\right) = \left(\hat{\mathbf{U}}_w'\right)^{-1}\left(\sum_{i\in s} w_i^2\,\hat{\mathbf{U}}_{wi}\hat{\mathbf{U}}_{wi}^T\right)\left[\left(\hat{\mathbf{U}}_w'\right)^{-1}\right]^T, \tag{3.20}$$

where $\hat{\mathbf{U}}_{wi}$ and $\hat{\mathbf{U}}_w'$ denote $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$ and the first derivative $\hat{\mathbf{U}}_w'(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$, respectively. Properties of the variance estimator (3.20) are studied through simulation in Section 5.2.

4 Weighted log composite likelihood: A unified approach

In this section we propose a unified approach applicable to both linear and generalized linear multi-level models. This approach is based on the concept of composite likelihood which has become popular in the non-survey literature to handle clustered or spatial data (see *e.g.*, Lindsay 1988, Lele and Taper 2002 and Varin, Reid and Firth 2011). A pairwise marginal composite likelihood is obtained by multiplying the likelihood contributions from all the distinct pairs within clusters. Note that the composite likelihood is obtained by pretending the sub-models are independent. When the super-population model holds for the sample, then we can obtain parameter estimators by maximizing the pairwise composite likelihood. Here we extend this approach to handle informative designs by obtaining weighted estimating equations that require only the marginal weights w_i and $w_{j|i}$ and the pairwise weights $w_{jk|i}$, as in Section 3.

The census log pairwise composite likelihood is given by

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j<k=1}^{M_i} \log f\left(y_{ij},\, y_{ik} \middle| \boldsymbol{\theta}\right), \tag{4.1}$$

where $f\left(y_{ij},\, y_{ik} \middle| \boldsymbol{\theta}\right)$ is the marginal joint density of y_{ij} and y_{ik}. We estimate (4.1) by the design-weighted log pairwise composite likelihood

$$l_{wC}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} \log f\left(y_{ij},\, y_{ik} \middle| \boldsymbol{\theta}\right) \tag{4.2}$$

which depends only on the first order level 1 and level 2 inclusion probabilities and the second order level 1 probabilities. We then solve the weighted composite score equations

$$\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta}) = \partial l_{wC}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}, \tag{4.3}$$

obtained from (4.2) to get a weighted composite likelihood estimator, $\hat{\boldsymbol{\theta}}_{wC}$, of $\boldsymbol{\theta}$. The proposed method is applicable to linear and generalized linear two-level models.

We note that $\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})$, given by (4.3), is a vector of estimating functions with zero expectation with respect to the design and the model, *i.e.*, $E_m E_p \left\{\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})\right\} = \mathbf{0}$. Using this result, it can be shown that the weighted composite likelihood (WCL) estimator $\hat{\boldsymbol{\theta}}_{wC}$ of $\boldsymbol{\theta}$ is design-model consistent as the number of level 2 units in the sample, n, increases, even when the within cluster sample sizes, m_i, are small. Details of the proof are given in Yi, Rao and Li (2012). In the non-survey context, we have limited theoretical and empirical evidence that the composite likelihood approach leads to efficient estimators (*e.g.*, Bellio and Varin 2005, Lindsay *et al.* 2011). Our simulation study (Section 5) indicates that the weighted composite likelihood approach performs well in terms of efficiency, even for small within-cluster sample sizes.

In the case of the nested error model (3.13), following Lele and Taper (2002) we can simplify the pairwise composite likelihood approach by replacing the bivariate density function $f\left(y_{ij}, y_{ik} \middle| \boldsymbol{\theta}\right)$ by the univariate density functions of y_{ij} and the difference $z_{ijk} = y_{ij} - y_{ik}$. For the mean model (2.2), we have $y_{ij} \sim N\left(\mu,\, \sigma_v^2 + \sigma_e^2\right)$ and $z_{ijk} \sim N\left(0,\, 2\sigma_e^2\right)$. By reparametrizing $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ as $\boldsymbol{\phi} = \left(\mu, \sigma^2, \sigma_e^2\right)^T$

where $\sigma^2 = \sigma_v^2 + \sigma_e^2$, we see that the parameters of the two univariate density functions are distinct and the log composite likelihoods corresponding to y_{ij} and z_{ijk} are given by

$$l_{wCy}\left(\mu,\sigma^2\right) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} \log f\left(y_{ij}\middle|\mu,\sigma^2\right)$$

and

$$l_{wCz}\left(\sigma_e^2\right) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \log f\left(z_{ijk}\middle|\sigma_e^2\right).$$

We then solve the resulting weighted composite score equations

$$\hat{U}_{wCy1}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\mu = \sum_{i\in s} w_i \sum_{j\in s_i} w_{j|i}\left(y_{ij}-\mu\right)\Big/\sigma^2 = 0,$$

$$\hat{U}_{wCy2}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\sigma^2 = \frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[-\frac{1}{\sigma^2} + \frac{\left(y_{ij}-\mu\right)^2}{\sigma^4}\right] = 0$$

$$\hat{U}_{wCz}\left(\sigma_e^2\right) = \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 = \frac{1}{2}\sum_{i} w_i \sum_{j<k\in s(i)} w_{jk|i}\left(-\frac{1}{\sigma_e^2} + \frac{z_{ijk}^2}{2\sigma_e^4}\right) = 0$$

to get the weighted composite likelihood (WCL) estimators $\hat{\mu}_{wC}, \hat{\sigma}_{vwC}^2$ and $\hat{\sigma}_{ewC}^2$. The WCL estimators are identical to (3.9) – (3.11) obtained by the weighted estimating equations approach of Section 3.

We now turn to the nested error linear regression model (3.13). We first note that $y_{ij} \sim N\left(\mathbf{x}_{ij}^T\boldsymbol{\beta},\ \sigma^2\right)$ where $\sigma^2 = \sigma_v^2 + \sigma_e^2$, and $z_{ijk} = y_{ij} - y_{ik} \sim N\left\{\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta},\ 2\sigma_e^2\right\}$. It follows that the weighted composite score equations are given by

$$\begin{aligned}\hat{\mathbf{U}}_{wCy1}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\boldsymbol{\beta} \\ &= \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\mathbf{x}_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right) = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{U}}_{wCy2}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\sigma^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[\frac{1}{\sigma^2} - \frac{\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2}{\sigma^4}\right] = 0\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{U}}_{wCz}(\sigma_e^2) &= \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left\{\frac{1}{\sigma_e^2} - \frac{\left[z_{ijk}-\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta}\right]^2}{2\sigma_e^4}\right\} = 0.\end{aligned}$$

The resulting WCL estimators of $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 are given by

$$\hat{\boldsymbol{\beta}}_{wC} = \left(\sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right)^{-1} \left(\sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} y_{ij} \right),$$

$$\hat{\sigma}_{wC}^2 = \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \left(y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{wC} \right)^2 \Big/ \sum_{i \in s} \sum_{j \in s(i)} w_{ij},$$

and

$$\hat{\sigma}_{ewC}^2 = \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} \left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik} \right)^T \hat{\boldsymbol{\beta}}_{wC} \right]^2 \Big/ \left(2 \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} \right).$$

The estimator of σ_v^2 is given by $\hat{\sigma}_{vwC}^2 = \hat{\sigma}_{wC}^2 - \hat{\sigma}_{ewC}^2$. Again, the WCL estimators $\hat{\boldsymbol{\beta}}_{WC}$, $\hat{\sigma}_{vWC}^2$ and $\hat{\sigma}_{eWC}^2$ are identical to (3.17) – (3.19) obtained from the weighted estimating equations approach of Section 3.

The above composite likelihood approach, based on y_{ij} and $z_{ijk} = y_{ij} - y_{ik}$, is not applicable to the linear two-level model given by (2.4) because the parameter vector, $\boldsymbol{\theta}$, is not identifiable under the composite likelihood obtained from the y_{ij} and z_{ijk}. We need the pairwise method to handle model (2.4).

Marginally, $\left(y_{ij}, y_{ik} \right)^T$ is bivariate normal with means $\mathbf{x}_{ij}^T \boldsymbol{\beta}$ and $\mathbf{x}_{ik}^T \boldsymbol{\beta}$ and 2×2 covariance matrix

$$\boldsymbol{\Sigma}_{i(jk)} = \begin{bmatrix} \sigma_e^2 + \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} & \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ik} \\ \mathbf{x}_{ik}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} & \sigma_e^2 + \mathbf{x}_{ik}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ik} \end{bmatrix}.$$

It now follows from (4.3) that the weighted composite score equations are given by

$$\boldsymbol{\beta}: \qquad \hat{\mathbf{U}}_{wC\boldsymbol{\beta}} = \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} \mathbf{X}_{i(jk)}^T \boldsymbol{\Sigma}_{i(jk)}^{-1} \left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \boldsymbol{\beta} \right) = \mathbf{0} \tag{4.4}$$

and

$$\begin{aligned} \boldsymbol{\tau}: \qquad \hat{\mathbf{U}}_{wCl} = \frac{1}{2} \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} \Bigg[& \left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \boldsymbol{\beta} \right)^T \boldsymbol{\Sigma}_{i(jk)}^{-1} \frac{\partial \boldsymbol{\Sigma}_{i(jk)}}{\partial \tau_l} \boldsymbol{\Sigma}_{i(jk)}^{-1} \left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \boldsymbol{\beta} \right) \\ & - \operatorname{tr} \left(\boldsymbol{\Sigma}_{i(jk)}^{-1} \frac{\partial \boldsymbol{\Sigma}_{i(jk)}}{\partial \tau_l} \right) \Bigg] = \mathbf{0}, \qquad l = 1, \ldots, p(p+1)/2 + 1 = P \end{aligned} \tag{4.5}$$

where $\mathbf{X}_{i(jk)}$ is the $2 \times p$ matrix with rows $\mathbf{x}_{ij}^T$ and $\mathbf{x}_{ik}^T$, $\mathbf{y}_{i(jk)} = \left(y_{ij}, y_{ik} \right)^T$ and $\boldsymbol{\tau}$ is the P-vector with elements $\tau_1 = \sigma_e^2$ and the $p(p+1)/2$ distinct elements of $\boldsymbol{\Sigma}_v$ denoted by $\tau_2, \ldots, \tau_P$. We can solve the weighted composite score equations (4.4) and (4.5) iteratively using the Newton-Raphson method or some other iterative method to obtain the WCL estimators $\hat{\boldsymbol{\beta}}_{wC}$ and $\hat{\boldsymbol{\tau}}_{wC}$.

In the special case of the nested error linear regression model (3.13), the census score equations, based on the full census log-likelihood $l(\boldsymbol{\theta})$ given by (2.5), can be written in a closed form. The corresponding sample weighted score equations depend only on the level 1 weights $w_{j|i}$ and $w_{jk|i}$ and the level 2 weights

w_i, similar to the weighted composite score equations (see the Appendix). The resulting estimators are design-model consistent for $\boldsymbol{\theta}$, unlike the estimators based on the weighted pseudo log-likelihood $l_w(\boldsymbol{\theta})$ given by (2.7) and (2.8). However, for more complex models, such as two level models with random slopes, the sample weighted score equations will depend on third order and fourth order level 1 inclusion probabilities, unlike the weighted composite score equations (4.3) that depend only on the first order and second order level 1 inclusion probabilities, even for complex multi-level models. We have therefore not included the weighted score equations approach, based on the full census log-likelihood, in the simulation study.

5 Simulation study

We conducted a small simulation study on the performance of the proposed WEE estimators under the simple nested error mean model, using $\mu = 0.5$, $\sigma_v^2 = 0.5$ and $\sigma_e^2 = 2.0$. The population consists of $N = 1{,}000$ clusters, each containing $M_i = M = 100$ elements. A two-stage sampling design with $n = 50$ sample clusters and $m_i = m = 5$ sample elements from each sample cluster is used. Clusters are selected by simple random sampling, and the elements within clusters by the Rao-Sampford probability proportional to size (PPS) sampling method (Rao 1965 and Sampford 1967) with specified size measures z_{ij}. The size measures are chosen to reflect different levels of informativeness.

Following Asparouhov (2006), we considered both invariant and non-invariant selections. For invariant selection, the size measure z_{ij} depends only on the level 1 errors and is invariant across clusters. In particular, we let

$$z_{ij} = \left(1 + \exp\left\{-0.5\left[e_{ij}/\alpha + e_{ij}^{*}\left(1 - \alpha^{-2}\right)^{1/2}\right]\right\}\right)^{-1}, \tag{5.1}$$

where e_{ij}^{*} is independent of e_{ij} but with the same distribution, $N\left(0, \sigma_e^2 = 2.0\right)$. For non-invariant selection, the size measure z_{ij} depends on both level 1 and level 2 errors and hence non-invariant across clusters. In particular, we replace e_{ij} and e_{ij}^{*} in (3.7) by $v_i + e_{ij}$ and $v_i^{*} + e_{ij}^{*}$ respectively, where v_i^{*} is independent of v_i but with the same distribution $N\left(0, \sigma_v^2 = 0.5\right)$. We considered four values of α in (5.1): $\alpha = 1, 2, 3, \infty$, where $\alpha = \infty$ corresponds to non-informative sampling within each cluster, $\alpha = 1$ corresponds to the most informative sampling and informativeness decreases as α increases.

We used the design-model (*pm*) approach to simulate $R = 1{,}000$ samples for each specified α and separately for invariant and non-invariant selections. Under this approach, we generated a population with $N = 1{,}000$ and $M_i = M = 100$ from the model and then selected a two-stage sample of elements as specified above. The two-step process was repeated $R = 1{,}000$ times to simulate $1{,}000$ samples.

5.1 Performance of estimators

From each sample, we computed the estimates of μ, σ_v^2 and σ_e^2 using REML, weighted scaling methods A and A1, the proposed WEE method and the alternative method of Korn and Graubard

(abbreviated KG). Biases and variances of the estimators were computed from the 1,000 estimates. Performance of alternative estimators is judged using two performance measures: Bias ratio = BR = (Bias)/(square root of variance) and relative root mean squared error = RRMSE = (square root of MSE)/(true parameter value). Tables 5.1, 5.2 and 5.3 respectively report the BR values of the estimators of μ, σ_v^2 and σ_e^2. RRMSE values of the estimators of μ, σ_v^2 and σ_e^2 are reported in Tables 5.4, 5.5 and 5.6 respectively.

Table 5.1
Bias ratio (%) of estimators of μ

α	Invariant			Non-invariant		
	REML	A	A1/WEE/KG	REML	A	A1/WEE/KG
1	346.5	80.2	2.2	370.9	83.9	3.0
2	167.7	40.1	0.3	172.3	45.3	6.1
3	114.3	30.7	4.5	114.9	30.8	4.8
∞	2.0	2.5	2.1	-1.5	-2.4	-2.2

Table 5.1 reports bias ratio (%) of the estimators of μ based on REML, weight-scaling methods A and A1, KG and WEE. Note that in the case of μ, estimators A1, KG and WEE (WCL) are identical. Results in Table 5.1 show that BR is similar for invariant and non-invariant selections and that BR of REML and A decrease as α increases. Further, REML leads to large bias under informative sampling, even for $\alpha = 3$; for example, BR for REML ranges from 114% to 346% under invariant selection. Method A also leads to significant BR under informative sampling; for example BR for A ranges from 30.8% to 83.9% under non-invariant selection. On the other hand, BR of WEE, A1 and KG does not depend on α and it is small ($|BR| < 6\%$). Under non-informative sampling, REML performs well as expected ($|BR| < 3\%$).

Turning to the estimation of σ_v^2, we first note that the proportion of times the estimate of σ_v^2 is negative is zero in the simulations for all four values of α and for all the estimation methods (REML, A, A1, WEE and KG). Table 5.2 reports BR values of the estimators of σ_v^2. It shows that the BR of REML is not affected by α under invariant selection, but is affected under non-invariant selection. In the latter case, REML leads to serious underestimation for $\alpha = 1$ (BR = -49%) but |BR| decreases as α increases. Table 5.2 also shows that methods A and A1 do not perform well under informative sampling (BR ranging from 16% to 60%). KG did not perform well for $\alpha = 1$ (BR=33% under invariant selection and BR = 24% under non-invariant selection). On the other hand, WEE performs well for all values of α (BR ranging from -4% to -13%) although underestimation is consistent across values of α.

Table 5.3 reports BR values of the estimators of σ_e^2. It shows that BR values are similar for invariant and non-invariant selections, as in the case of μ. REML and KG lead to serious underestimation when $\alpha = 1$ (BR = -107% for REML and BR = -71% for KG under invariant selection), but |BR| decreases as α increases and becomes negligible for $\alpha = \infty$. Estimators A and A1 perform poorly for all values of α

including $\alpha = \infty$. On the other hand, WEE performs well for all values of α with |BR|<8%. It appears that the instability introduced by the scale factor (2.9) might have contributed to the large |BR| for methods A and A1 even for the case of non-informative sampling $(\alpha = \infty)$.

Table 5.2
Bias ratio (%) of estimators of σ_v^2

α	REML	A	A1	WEE	KG
			Invariant Selection		
1	0.6	59.5	59.3	-8.5	33.2
2	0.5	24.5	26.3	-10.0	8.0
3	-3.4	16.1	18.2	-13.6	0.4
∞	-0.1	14.8	17.1	-8.9	0.6
			Non-invariant Selection		
1	-49.0	50.1	58.9	-4.4	24.0
2	-10.9	24.6	28.7	-7.0	7.1
3	-4.0	20.0	22.7	-7.8	4.6
∞	-1.3	12.8	13.9	-13.3	-1.6

Table 5.3
Bias ratio (%) of estimators of σ_e^2

α	REML	A	A1	WEE	KG
			Invariant Selection		
1	-106.9	-118.4	-66.9	2.4	-71.2
2	-22.7	-43.6	-34.3	2.1	-16.5
3	-9.4	-31.7	-28.4	2.9	-6.5
∞	-0.4	-21.8	-23.8	0.3	0.4
			Non-invariant Selection		
1	-115.3	-131.3	-79.6	-6.9	-82.6
2	-30.4	-51.1	-43.3	-7.6	-23.9
3	-12.5	-34.9	-32.2	-2.3	-10.3
∞	1.1	-20.2	-21.8	2.6	1.6

Table 5.4
Relative root mean squared error (%) of estimators of μ

α	Invariant			Non-invariant		
	REML	A	A1/WEE/KG	REML	A	A1/WEE/KG
1	93.3	35.9	29.4	92.5	35.4	29.2
2	51.6	29.3	27.8	52.8	30.4	28.9
3	40.5	28.2	27.5	40.8	28.7	28.1
∞	25.8	26.1	26.5	26.6	27.3	27.7

Relative root mean squared error

Table 5.4 shows that the RRMSE (%) values for estimators of μ are similar for invariant and non-invariant selections and that RRMSE of REML and A decrease as α increases. For informative sampling with $\alpha = 1$, RRMSE for REML is large relative to RRMSE for WEE (A1 and KG) due to large BR. For example, RRMSE=93% for REML compared to RRMSE=29% for WEE. As expected, REML has the smallest RRMSE under non-informative sampling, but the increase in RRMSE for the other methods is quite small. Also, RRMSE of WEE (A1 and KG) depends on α.

Table 5.5
Relative root mean squared error (%) of estimators of σ_v^2

α	REML	A	A1	WEE	KG
			Invariant Selection		
1	36.5	47.3	51.1	43.6	43.8
2	37.1	39.7	41.1	40.5	39.5
3	36.3	37.3	38.7	39.5	37.8
∞	35.8	36.9	38.1	38.7	37.2
			Non-invariant Selection		
1	36.7	44.6	52.6	43.4	41.5
2	35.6	37.9	40.4	39.3	37.7
3	37.0	38.7	40.4	40.2	38.8
∞	36.6	37.2	38.0	39.0	37.8

Turning to RRMSE of estimators of σ_v^2, Table 5.5 shows that REML performs well for all α under invariant selection due to small BR in this case. We also note that KG and WEE are comparable in terms of RRMSE for all values of α. Table 5.5 also shows that A and A1 lead to somewhat larger RRMSE for $\alpha = 1$: 51% for A1 and 47% for A under invariant selection compared to 44% for WEE.

Table 5.6
Relative root mean squared error (%) of estimators of σ_e^2

α	REML	A	A1	WEE	KG
			Invariant Selection		
1	13.5	14.5	12.8	13.9	12.9
2	9.7	10.4	10.4	11.0	10.0
3	9.5	10.0	10.1	10.7	9.8
∞	10.1	10.3	10.5	11.1	10.3
			Non-invariant Selection		
1	13.7	14.8	12.9	13.2	13.0
2	10.0	10.9	10.9	11.3	10.3
3	9.7	10.4	10.7	11.2	10.2
∞	10.3	10.6	10.8	11.4	10.7

Table 5.6 gives RRMSE values of the estimators of σ_e^2 and we note that the values are similar for invariant and non- invariant selections. It also shows that RRMSE values are comparable for methods WEE, A, Al and KG even though in terms of bias ratio A, Al and KG performed poorly relative to WEE. This is due to larger variance for WEE compared to other methods. For example, in the case of invariant selection and $\alpha = 1$ we have the following variances for WEE, KG and REML: 0.0771, 0.0438 and 0.0339 with corresponding bias ratios (%) from Table 5.3: 2.4, -71.2, and -106.9.

5.2 Performance of variance estimator

We now report some simulation results on the relative bias of the linearization variance estimator (3.12) of the WEE (WCL) estimator $\hat{\boldsymbol{\theta}}_w$. We first repeated the two-step process $R_1 = 2{,}000$ times and computed $v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ from each two-stage sample $r = 1,..., 2{,}000$. The averages of the diagonal elements of $E\left\{v_L\left(\hat{\boldsymbol{\theta}}_w\right)\right\} \approx v_L\left(\hat{\boldsymbol{\theta}}_w\right) = R_1^{-1}\sum_{r=1}^{R_1} v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ are denoted by $\bar{v}_L\left(\hat{\mu}_w\right)$, $\bar{v}_L\left(\hat{\sigma}_{vw}^2\right)$ and $\bar{v}_L\left(\hat{\sigma}_{ew}^2\right)$ respectively. We then generated $R_2 = 10{,}000$ independent samples and computed the empirical mean squared error

(MSE) of the three estimators $\hat{\mu}_w$, $\hat{\sigma}^2_{vw}$ and $\hat{\sigma}^2_{ew}$. We have $\mathrm{MSE}(\hat{\mu}_w) \approx R_2^{-1}\sum_{r=1}^{R_2}(\hat{\mu}_w^{(r)} - \mu)^2$ where $\hat{\mu}_w^{(r)}$ is the estimate of μ from the *r*-th simulated sample, and similar expressions for $\mathrm{MSE}(\hat{\sigma}^2_{vw})$ and $\mathrm{MSE}(\hat{\sigma}^2_{ew})$.

The relative bias of $v_L(\hat{\mu}_w)$ is calculated as

$$\mathrm{RB}\{v_L(\hat{\mu}_w)\} = [\bar{v}_L(\hat{\mu}_w)/\mathrm{MSE}(\hat{\mu}_w)] - 1$$

and similarly $\mathrm{RB}\{v_L(\hat{\sigma}^2_{vw})\}$ and $\mathrm{RB}\{v_L(\hat{\sigma}^2_{ew})\}$ were calculated. Table 5.7 reports the RB values of the three variance estimators for invariant and non-invariant selections and $\alpha = 1, 2, 3, \infty$. It is clear from Table 5.7 that the linearization variance estimator performs well over all combinations with $|\mathrm{RB}| < 10\%$.

Table 5.7
Relative bias (%) of variance estimators

α	$v_L(\hat{\mu}_w)$	$v_L(\hat{\sigma}^2_{vw})$	$v_L(\hat{\sigma}^2_{ew})$
	Invariant Selection		
1	-3.0	-6.2	-7.5
2	-5.2	-4.5	-3.1
3	-1.3	-3.8	-1.8
∞	-0.9	-2.5	-2.0
	Non-invariant Selection		
1	-3.8	-8.3	-4.2
2	-4.5	-5.8	-7.3
3	-4.3	-4.6	-5.7
∞	-2.4	-2.7	-2.9

6 Concluding remarks

In this paper, we have proposed a unified weighted composite likelihood (WCL) approach for two-level models to make inferences from complex survey data. The proposed WCL methods are asymptotically valid even when the sample sizes within sampled clusters (level 1 units) are small, unlike some of the existing methods, but knowledge of the joint inclusion probabilities within sampled clusters is required. Often it may be possible to treat the sample within clusters as drawn with replacement because of small sampling fractions within clusters. Also, excellent approximations to joint inclusion probabilities, depending only on the marginal inclusion probabilities, are also available when the sampling fractions are not small (Haziza *et al.* 2008). We plan to study the accuracy of such approximations in a future study.

Simulation studies on the performance of the WCL estimators (4.5) and (4.6) for two-level models (2.3), based on the pairwise method, will also be conducted.

Composite likelihood methods are mostly used when the full likelihood is complex. Our development in the survey sampling context demonstrates that the full likelihood method with weights is not feasible for multi-level models whereas the weighted composite likelihood method facilitates valid inferences even when the cluster sample sizes are small.

7 Acknowledgements

We thank two referees and the associate editor for constructive comments and suggestions.

Appendix

Weighted score equations: nested error linear regression model

For the nested error linear regression model (2.3), an explicit form for the census full log-likelihood is obtained using the explicit form for the covariance matrix $\mathbf{V}_i$ of $\mathbf{y}_i = \left(y_{i1},...,y_{iM_i}\right)^T$. We have $\mathbf{V}_i^{-1} = \sigma_e^{-2}\left[\mathbf{I}_i - \sigma_v^2 / \lambda_i \mathbf{1}_i \mathbf{1}_i^T\right]$, where $\lambda_i = \sigma_e^2 + M_i\sigma_v^2$, $\mathbf{I}_i$ is the $M_i \times M_i$ identity matrix and $\mathbf{1}_i$ is the $M_i \times 1$ unit vector. Using the expression for $\mathbf{V}_i^{-1}$, the census score equations are obtained as

$$\boldsymbol{\beta}: \quad \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}y_{ij} - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}y_{ik}\right)\right] - \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ik}^T\right)\right]\boldsymbol{\beta} = 0 \quad \text{(A.1)}$$

$$\sigma_v^2: \quad \sum_{i=1}^{N}\lambda_i^{-2}\left[\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik} - \mathbf{x}_{ik}^T\boldsymbol{\beta}\right)\right] - \sum_{i=1}^{N}\lambda_i^{-1}M_i = 0 \quad \text{(A.2)}$$

$$\begin{aligned}\sigma_e^2: \quad & \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2 + \sum_{i=1}^{N}\left(M_i\sigma_v^4\lambda_i^{-2} - 2\sigma_v^2\lambda_i^{-1}\right)\sum_{j,k=1}^{M_i}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik} - \mathbf{x}_{ik}^T\boldsymbol{\beta}\right) \\ & -\sigma_e^2\sum_{i=1}^{N}\left(1 - \sigma_v^2\lambda_i^{-1}\right)M_i = 0\end{aligned} \quad \text{(A.3)}$$

From (A.1), we obtain weighted score equations

$$\begin{aligned}\boldsymbol{\beta}: \quad & \sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}y_{ij} - \sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}y_{ik}\right) \\ & -\left[\sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}\mathbf{x}_{ik}\right)\right]\boldsymbol{\beta} = 0\end{aligned} \quad \text{(A.4)}$$

where $w_{jj|i} = w_{j|i}$. Note that the cluster sizes M_i for $i \in s$ are assumed to be known. One should not replace M_i by its estimate $\sum_{j \in s(i)} w_{j|i}$ because it includes ratio bias due to small within cluster sample sizes. The estimating equation (A.4) is design-unbiased for the census equation (A.1).

Turning to the weighted score equation for σ_v^2, we obtain from (A.2)

$$\sigma_v^2: \quad \sum_{i \in s} w_i \lambda_i^{-2} \left[\sum_{j \in s(i)} \sum_{k \in s(i)} w_{jk|i} \left(y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta} \right) \left(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta} \right) \right] - \sum_{i \in s} w_i \lambda_i^{-1} \sum_{j \in s(i)} w_{j|i} = 0 \tag{A.5}$$

The estimating equation (A.5) is unbiased for (A.2). Finally, the weighted score equation for σ_e^2 is obtained from (A.3) as

$$\begin{aligned} \sigma_e^2: \quad & \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} \left(y_{ij} - x_{ij}^T \beta \right)^2 + \sum_{i \in s} w_i \left(M_i \sigma_v^4 \lambda_i^{-2} - 2\sigma_v^2 \lambda_i^{-1} \right) \sum_{j,k \in s(i)} w_{jk|i} \left(y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta} \right) \left(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta} \right) \\ & - \sigma_e^2 \sum_{i \in s} w_i \left(1 - \sigma_v^2 \lambda_i^{-1} \right) \sum_{j \in s(i)} w_{j|i} = 0 \end{aligned} \tag{A.6}$$

It follows from (A.4) – (A.6) that the weighted score equations depend only on the first order weights w_i and $w_{j|i}$ and the second order weights $w_{jk|i}$ in the special case of a nested error linear regression model.

References

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35, 439-460.

Beaumont, J.-F., and Charest, A.-S. (2010). Bootstrap variance estimation with survey data when estimating model parameters. Unpublished report (courtesy of the authors).

Bellio, R., and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 3, 217-227.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Grilli, L., and Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, 30, 93-103.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society B*, 65, 175-190.

Kovacevic, M.S., Rong, H. and You, Y. (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. *Proceedings of ASA Section on Survey Research Methods*, American Statistical Association, 3260-3269.

Lele, S., and Taper, M.L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103, 117-125.

Lindsay, B.G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, (Ed. N.U. Prabhu), Providence: American Mathematical Society, 221-239.

Lindsay, B.G., Yi, G.Y. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71-105.

Muthén, L.K., and Muthén, B.O. (1998-2005). *Mplus User's Guide. 3rd ed. Los Angeles, CA: Muthén & Muthén*.

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, (Eds. R. Chambers and C.J. Skinner) 175-196, Wiley, Chichester.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B*, 60, 23-56.

Rabe-Hesketh, S., and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society A*, 169, 805-827.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J.N.K., and Roberts, G. (1998). Discussion on the papers by Firth and Bennett and Pfeffermann *et al*. *Journal of the Royal Statistical Society B*, 60, 50-51.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Rao, J.N.K., Hidiroglou, M., Yung, W. and Kovacevic, M. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of the Indian Society of Agricultural Statistics*, 64, 129-135.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Yi, G.Y., Rao, J.N.K. and Li, H. (2012). A weighted composite likelihood approach for analysis of survey data under two level models. Available on request to jrao@math.carleton.ca.

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 283-301
Statistics Canada, Catalogue No. 12-001-X

Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption

Hervé Cardot, Alain Dessertaine, Camelia Goga, Étienne Josserand and Pauline Lardin[1]

Abstract

When the study variables are functional and storage capacities are limited or transmission costs are high, using survey techniques to select a portion of the observations of the population is an interesting alternative to using signal compression techniques. In this context of functional data, our focus in this study is on estimating the mean electricity consumption curve over a one-week period. We compare different estimation strategies that take account of a piece of auxiliary information such as the mean consumption for the previous period. The first strategy consists in using a simple random sampling design without replacement, then incorporating the auxiliary information into the estimator by introducing a functional linear model. The second approach consists in incorporating the auxiliary information into the sampling designs by considering unequal probability designs, such as stratified and πps designs. We then address the issue of constructing confidence bands for these estimators of the mean. When effective estimators of the covariance function are available and the mean estimator satisfies a functional central limit theorem, it is possible to use a fast technique for constructing confidence bands, based on the simulation of Gaussian processes. This approach is compared with bootstrap techniques that have been adapted to take account of the functional nature of the data.

Key Words: Bonferroni; Bootstrap; Horvitz-Thompson estimator; Covariance function; Model-assisted estimator; Functional linear model; Hájek formula.

1 Introduction

With the development of automated data acquisition processes at fine time scales, it is no longer unusual to have very large databases on phenomena that change over time. For example, in the coming years in France, approximately 30 million electric meters will be replaced by smart meters. These will be able to measure the consumption of each household and business at potentially very fine time scales (by the second or minute) and send the measurements once a day to a central server. Another example is measuring the viewership of different television channels. Boxes measure in continuous time information on whether the television is on and what channel is being viewed.

The statistical unit studied is accordingly a function (of time or space), which calls for the introduction of functional analysis tools. Although this branch of statistics has existed since the 1970s (Deville 1974), Dauxois and Pousse (1976), it truly developed during the 1990s with advances in computer technology. It has applications to various fields such as climatology, economics, remote sensing, medicine and quantitative chemistry. Readers may consult the recent references Ramsay and Silverman (2005) and Ferraty and Romain (2011) for a panorama of the different techniques and examples of applications.

1. Hervé Cardot, Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 av. Alain Savary, 21078 DIJON, FRANCE; Alain Dessertaine, LA POSTE - DIRECTION DU COURRIER - DFI – DCPES, 2 Boulevard Newton 77543 MARNE LA VALLEE CEDEX 2 and EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle, 92141 CLAMART, France; Camelia Goga, Université de Bourgogne, Institut de Mathématiques de Bourgogne; Étienne Josserand, Université de Bourgogne, Institut de Mathématiques de Bourgogne. E-mail: camelia.goga@u-bourgogne.fr; Pauline Lardin, Université de Bourgogne, Institut de Mathématiques de Bourgogne and EDF, R&D, ICAME-SOAD.

When the potential databases are very large, it can be difficult and costly to collect, save and analyze the entire data set. Moreover, if one is interested in simple indicators such as the mean curve under constraints of memory space or the cost of transmission, the use of survey techniques to extract a sample can provide a precise estimate at a reasonable cost (Dessertaine 2008).

In the statistical literature, there are as yet few studies that combine functional data analysis and sampling theory. Cardot, Chaouch, Goga and Labruère (2010) are interested in using principal component analysis to reduce the dimension of the data, while Cardot and Josserand (2011) examine the uniform convergence properties of Horvitz-Thompson estimators of mean curves. Chaouch and Goga (2012) provide a robust estimator of central curves.

The objective of this study is to compare different sampling strategies in a functional context, using a real example. These real data concern the electricity consumption, measured every half hour for two weeks, of a test population of $N = 15,069$ electric meters. The time profile of individuals' electricity use depends on covariables such as weather conditions (temperature, *etc.*) or geographic characteristics (altitude, latitude or longitude). Unfortunately, those variables are not available for this study, and we use only one variable as auxiliary information: the mean consumption from each meter during the previous week. This information can easily be transmitted by all the meters in the population.

Extending estimation methods that use auxiliary information to the functional framework is not always straightforward. Cardot and Josserand (2011) propose stratifying the population of curves to improve the estimate of the mean curve. Chaouch and Goga (2012), who are interested in the median curve, suggest using PPS (probability proportional to size) sampling with replacement as well as a post-stratified estimator. In this article, we propose to compare several strategies that take auxiliary information into account. The first strategy uses auxiliary information in selection of the sample: sampling with an unequal probabilities design (stratified, πps) and estimation with the Horvitz-Thompson estimator. The second strategy introduces this information at the estimation stage: a simple random sample is drawn without replacement and estimation is performed using a linear regression model (Särndal, Swensson and Wretman 1992) adapted to the functional framework (Faraway 1997).

A new question, related to the functional nature of the data, naturally arises: how to quantify sampling uncertainty? The construction of confidence intervals—a central concern for survey methodologists—has received little attention in the field of functional data statistics, where it is a matter of constructing confidence bands. Drawing on techniques based on estimation of the covariance function of the estimator (see Faraway (1997), Cuevas, Febrero and Fraiman (2006) or more recently Degras (2011)), we first propose to construct confidence bands by simulating Gaussian processes. An asymptotic justification of the validity of these techniques is given in Cardot, Degras and Josserand (2013) when the hypotheses of the central limit theorem are verified and there is a precise estimator of the covariance function. A second method of construction, which is based on bootstrap techniques, is also applied. It basically consists of three bootstrap techniques for use in a finite population: the bootstrap without replacement proposed by Gross (1980), the rescaling bootstrap (Rao and Wu 1988) and the mirror-match bootstrap (Sitter 1992). In this study, we use the bootstrap without replacement, which is based on adaptations for the stratified and PPS designs proposed by Chauvet (2007).

In Section 2, we introduce notations, estimators of the mean curve where there is auxiliary information, and estimators of their covariance function. The algorithms for constructing confidence bands, based on the bootstrap or simulation of Gaussian processes, are described in Section 3. Section 4 then compares the

different strategies—in terms of precision of the estimators, width and coverage of the confidence bands and computation time—for purposes of estimating the consumption curves of the French electricity company EDF (Électricité de France). For this we use samples of size $n = 1,500$ in our test population consisting of $N = 15,069$ curves. To finish, we present several perspectives of research in Section 5.

2 Functional data in a finite population

Consider a finite population $U = \{1, ..., N\}$ of size N and assume that for each unit k in the population U, we can observe the deterministic curve $Y_k = \left(Y_k(t)\right)_{t \in [0,T]}$. The objective is to estimate the mean curve of the population, which is defined for any instant $t \in [0, T]$, by

$$\mu(t) = \frac{1}{N} \sum_{k \in U} Y_k(t).$$

Let s be a sample of fixed size n, selected randomly in U, according to a sampling design $p(.)$. Let $\pi_k = \Pr(k \in s)$ and $\pi_{kl} = \Pr(k \,\&\, l \in s)$ be the first- and second- order inclusion probabilities respectively. Assume that $\pi_k > 0$ for any unit k in population U.

The mean curve μ is estimated using the Horvitz-Thompson estimator (Cardot *et al.* 2010) as follows:

$$\hat{\mu}(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_k(t)}{\pi_k} = \frac{1}{N} \sum_{k \in U} \frac{Y_k(t)}{\pi_k} 1_{k \in s}, \quad t \in [0, T], \tag{2.1}$$

where $1_{k \in s}$ is the indicator that unit k belongs to the sample s. For each instant $t \in [0, T]$, the estimator $\hat{\mu}(t)$ is unbiased for $\mu(t)$, meaning that $E\left(\hat{\mu}(t)\right) = \mu(t)$ where the expectation is considered in relation to the sampling design.

Generally, the trajectories $Y_k(t)$ are not observed continuously for $t \in [0, T]$ but only for a set of D measurement instants $0 = t_1 < t_2 < \ldots < t_D = T$. In functional data analysis, a classical strategy is to interpolate or smooth discretized trajectories to obtain objects that are truly functions (Ramsay and Silverman 2005). This also makes it possible to deal with curves whose measurement instants are not identical. In the context of surveys, Cardot and Josserand (2011) studied linear interpolation where there is no measurement error at the discretized points, while Cardot *et al.* (2013) examined smoothing procedures. If there are enough discretization points and the trajectories are fairly regular (but not necessarily derivable), the approximation error due to smoothing or interpolation is negligible in relation to the sampling error. We subsequently assume that the trajectories are observed at any point t of the interval $[0, T]$.

The Horvitz-Thompson covariance function $\gamma(r, t) = \operatorname{cov}\left(\hat{\mu}(r), \hat{\mu}(t)\right)$ is given by

$$\gamma(r, t) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l}$$

for any $(r,t) \in [0,T] \times [0,T]$ and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. If we assume that the second-order probabilities of inclusion satisfy $\pi_{kl} > 0$, an unbiased estimator of $\gamma(r,t)$ is given by the Horvitz-Thompson unbiased estimator of the variance,

$$\hat{\gamma}(r,t) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l} \tag{2.2}$$

for any $(r,t) \in [0,T] \times [0,T]$.

2.1 Using auxiliary information for estimating the mean trajectory

It is well known that using auxiliary information that effectively explains the variable of interest can greatly improve the precision of the Horvitz-Thompson estimator. In the case of the EDF data, the outside temperature or the type of contract could probably be useful auxiliary variables. A stratification based on geographic position would also yield estimates for different regions. In this study, we have as an auxiliary variable the total electricity consumption for the previous week. We assume that this variable (a real one) is observed for all units in the population.

In this section, we present the Horvitz-Thompson estimator for the mean curve as well as an estimate of the covariance function of this estimator, both for a stratified design using simple random sampling without replacement (SRSWOR) in each stratum, denoted hereafter as STRAT, and for PPS sampling without replacement, which will be denoted as πps. We also consider an estimator of the mean curve, assisted by a functional linear model.

2.1.1 Stratified sampling with SRSWOR in each stratum (STRAT)

The population U is assumed to be stratified into a fixed number H of strata $U_1, \ldots, U_H$ of sizes $N_1, \ldots, N_H$. Within each stratum U_h, a sample s_h of size n_h is drawn according to an SRSWOR design.

We denote $\mu_h(t) = \sum_{k \in U_h} Y_k(t) / N_h$, for $t \in [0,T]$, the mean curve in each stratum, and $\hat{\mu}_h(t) = \sum_{k \in s_h} Y_k(t) / n_h$, its estimate. The estimator of the mean curve μ is then defined by

$$\hat{\mu}_{\text{strat}}(t) = \frac{1}{N} \sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{k \in s_h} Y_k(t) \right), \; t \in [0,T]. \tag{2.3}$$

The Horvitz-Thompson estimator of the covariance function γ is then

$$\hat{\gamma}_{strat}(r,t) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Y(r)Y(t),s_h} \quad r,t \in [0,T], \tag{2.4}$$

where

$$S_{Y(r)Y(t),s_h} = \frac{1}{n_h - 1} \sum_{k \in s_h} \left(Y_k(r) - \hat{\mu}_h(r) \right) \left(Y_k(t) - \hat{\mu}_h(t) \right)$$

is the estimator of the covariance function $S_{Y(r)Y(t),U_h}$ in stratum h. For $r = t \in [0,T]$, we obtain the estimator of the variance function as follows:

$$\hat{\gamma}_{strat}(r) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{Y(r),s_h},$$

where

$$S^2_{Y(r),s_h} = \frac{1}{n_h - 1}\sum_{k\in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)^2$$

is the estimator of the variance $S^2_{Y(r),U_h}$ in stratum h. Cardot and Josserand (2011) propose an extension, in the functional framework, of Neyman's optimal allocation. When the sizes n_h of the samples s_h verify

$$n_h = n\frac{N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}}, \quad h = 1,...,H, \tag{2.5}$$

the integrated variance, $\int_0^T \hat{\gamma}_{\text{strat}}(t)\,dt$, of the stratified estimator is minimized. This allocation is similar to the one obtained in a multivariate context by Cochran (1977). By replacing the variable Y by another variable X that is known for the entire population and is highly correlated with the variable of interest, we obtain an allocation that can be described as x- optimal.

Note 2.1 *For $H = 1$, we obtain the simple random design without replacement (SRSWOR), and the mean curve $\mu(t)$ is estimated by*

$$\hat{\mu}_{\text{srswor}}(t) = \frac{1}{n}\sum_{k\in s} Y_k(t), \quad t \in [0,T]. \tag{2.6}$$

The estimator of the covariance function defined in (2.2) is then

$$\hat{\gamma}_{\text{srswor}}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}. \tag{2.7}$$

2.1.2 PPS sampling without replacement (πps)

PPS sampling designs with or without replacement are often used in practice because they are more effective than equal probability designs when the variable of interest is basically proportional to an auxiliary variable X that has strictly positive values.

In the case of samples of fixed size n drawn without replacement, it is possible to give the equivalent of the formula of Yates and Grundy (1953) and Sen (1953). The covariance function of $\hat{\mu}$ verifies

$$\gamma(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl} - \pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right),\ r,t \in [0,T]. \quad (2.8)$$

Assume that the values x_k of variable X are known for all units k in the population. It is then possible to define the inclusion probabilities as follows:

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}.$$

Methods have been proposed in the literature for the case $\pi_k > 1$ (Särndal *et al.* 1992).

Second-order inclusion probabilities are generally very difficult to calculate for πps designs, and therefore Formula (2.2) cannot be used. However, there is a simple asymptotic approximation of the variance, which was proposed by Hájek (1964) and which entails only first-order inclusion probabilities. This approximation proves to be very effective when the sample is large and the entropy of the sampling design is close to maximum entropy. To select sample s with inclusion probabilities π_k, the cube algorithm (Deville and Tillé 2004) balanced on the variable $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ can be used. Deville and Tillé (2005) show that for this particular sampling design, the Hàjek formula is highly effective for estimating the variance of a total or a mean. This formula for approximating the variance can also be used for the covariance, which is then estimated by

$$\hat{\gamma}_{\pi\text{ps}}(r,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(r)}{\pi_k} - \hat{R}(r)\right)\left(\frac{Y_k(t)}{\pi_k} - \hat{R}(t)\right),\ \ r,t \in [0,T], \quad (2.9)$$

where

$$\hat{R}(t) = \frac{\sum_{k\in s}\frac{Y_k(t)}{\pi_k}(1-\pi_k)}{\sum_{k\in s}(1-\pi_k)}.$$

We also used the systematic sampling with unequal probabilities proposed by Madow (1949), since it is simple to use. Unfortunately, it is difficult to estimate the variance for this type of design, and we will therefore not use it to construct confidence bands.

2.2 The model-assisted estimator

Consider p real auxiliary variables $X_1,\ldots,X_p$ and let x_{kj} be the value of the variable X_j for the k^{th} individual. Let $\mathbf{x}_k = (x_{k1},...,x_{kp})'$ denote the vector containing the values of p auxiliary variables measured on the k^{th} individual. We consider that the relationship between the variable of interest and the auxiliary variables is modeled by the following superpopulation model

$$\xi : Y_k(t) = \mathbf{x}_k'\boldsymbol{\beta}(t) + \varepsilon_{kt},\ \ t \in [0,T] \quad (2.10)$$

with

$$E_\xi\left(\varepsilon_{kt}\right) = 0, E_\xi\left(\varepsilon_{kt}\varepsilon_{lt'}\right) = 0 \text{ for } k \neq l \text{ and } E_\xi\left(\varepsilon_{kt}\varepsilon_{kt'}\right) = \sigma^2_{tt'} \text{ for } k = l.$$

This model is an immediate generalization of the functional linear model proposed by Faraway (1997) to several auxiliary variables.

The estimate of $\boldsymbol{\beta}$ based on the model ξ and the sampling design $p(.)$ is given by

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0,T]. \tag{2.11}$$

Note that the sampling weights do not depend on the time $t \in [0,T]$. Let $\hat{Y}_k(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)$ be the estimator based on the sampling design for the prediction of $Y_k(t)$ under the model ξ. By direct analogy with the univariate case (Särndal *et al.* 1992), we finally obtain the following estimator for the mean, for $t \in [0,T]$,

$$\begin{aligned}\hat{\mu}_{MA}(t) &= \frac{1}{N}\sum_{k\in s}\hat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\left(\hat{Y}_k(t) - Y_k(t)\right)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\frac{Y_k(t) - \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)}{\pi_k} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}(t).\end{aligned} \tag{2.12}$$

If the ξ contains the constant variable 1, then the estimator becomes

$$\hat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t), \quad t \in [0,T]. \tag{2.13}$$

For fixed r and t, the asymptotic covariance of $\hat{\mu}_{MA}(r)$ and $\hat{\mu}_{MA}(t)$ can be calculated according to the classical residual technique (Särndal *et al.* 1992),

$$\gamma_{MA}(r,t) \simeq \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{\left(Y_k(r) - \tilde{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \tilde{Y}_l(t)\right)}{\pi_l}, \tag{2.14}$$

where $\tilde{Y}_k(r) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$ is the prediction of $Y_k(t)$ under the superpopulation model and $\tilde{\boldsymbol{\beta}}(t) = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_U \mathbf{x}_k Y_k(t)\right)$ is the estimate of $\boldsymbol{\beta}$ at the level of the population and $r,t \in [0,T]$. Cardot, Goga and Lardin (2013) show that this result remains valid uniformly in $r,t \in [0,T]$.

As an estimator of the covariance function $\gamma_{MA}(r,t)$, we propose the Horvitz-Thompson estimator of asymptotic covariance given by (2.14) where $\tilde{\boldsymbol{\beta}}(t)$ is replaced by its estimator $\hat{\boldsymbol{\beta}}(t)$ based on the sampling design,

$$\hat{\gamma}_{MA}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{\left(Y_k(r) - \hat{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \hat{Y}_l(t)\right)}{\pi_l}, \quad r,t \in [0,T]. \tag{2.15}$$

Note 2.2 *It is entirely possible to consider a superpopulation model ξ that is more general than the linear model proposed here. Estimation techniques based on smoothing by B-splines (Goga and Ruiz-Gazen 2012) can then also be considered. In our study, the relationship between consumption at instant t and the mean consumption for the previous week is almost linear (cf. Figure 4.1), which justifies not using these non-parametric approaches.*

3 Construction of confidence bands

Here we are considering confidence bands for the mean curve μ that have the form

$$\mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha \hat{\sigma}(t)\right],\ \forall t \in \left[0,T\right]\right) = 1-\alpha, \tag{3.1}$$

where the value of the coefficient c_α is unknown and depends on the desired confidence level $1-\alpha$, and $\hat{\sigma}(t)$ is an estimator of the standard deviation of $\hat{\mu}(t)$. The calculation of c_α is based on the fact that according to some hypotheses (Cardot *et al.* 2013), the process

$$Z(t) = (\hat{\mu}(t) - \mu(t)) / \hat{\sigma}(t),\quad t \in [0,T],$$

converges toward a Gaussian process in the space of continuous functions $\mathcal{C}\left(\left[0,T\right]\right)$. We then have

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left|Z(t)\right| \le c_\alpha\right) = \mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha \hat{\sigma}(t)\right], \forall t \in \left[0,T\right]\right) \tag{3.2}$$

and it is therefore sufficient to determine c_α, the quantile of order $1-\alpha$ of the real random variable $\sup_{t\in[0,T]} \left|Z(t)\right|$ to construct the confidence band completely. The distribution of the sup of Gaussian processes is known explicitly for only a few specific cases, such as the Brownian motion.

We propose two approaches to determine the value of c_α. The first is based on a direct estimate of the standard deviation and the simulation of Gaussian processes $Z(t)$. The second, which does not require having an estimator of the variance, is based on resampling techniques where both the standard deviation and the value of c_α are obtained from bootstrap replications.

3.1 Construction of confidence bands by simulation of Gaussian processes

The steps of the algorithm are as follows:

1) Draw sample s of size n using sampling design p and calculate the estimator $\hat{\mu}$ as well as the estimator $\hat{\gamma}(r,t)$ of the covariance function $\gamma(r,t)$, $r,t \in \left[0,T\right]$.
2) Simulate M curves Z_m, $m = 1,\ldots M$, of the same distribution as Z where Z is a Gaussian process of expectation 0 and of covariance function ρ where $\rho(r,t) = \hat{\gamma}(r,t) / \left(\hat{\gamma}(r)\hat{\gamma}(t)\right)^{1/2}$ $r,t \in \left[0,T\right]$.

3) Determine c_α, the quantile of order $1-\alpha$ of the variables, $\left(\sup_{t\in[0,T]}\left|Z_m(t)\right|\right)_{m=1,\ldots,M}$.

This algorithm, which is very fast and easy to implement, has already been proposed in the context of i.i.d. observations by Faraway (1997), Cuevas *et al.* (2006) and Degras (2011) to construct confidence bands. A rigorous asymptotic justification of this approach may be found in Cardot *et al.* (2013) for sampling in finite populations.

3.2 Construction of confidence bands by bootstrapping

In this work, we use the bootstrap method proposed by Gross (1980) for SRSWOR sampling and the extensions proposed by Chauvet (2007) for STRAT and πps designs. It is based on the following principle: the sample s is used to simulate a fictitious population U^* in which we select a number of bootstrapped samples. The implementation of this algorithm is not straightforward when the ratio $1/\pi_k$ is not an integer. Many variants have been proposed in the literature to deal with the general case, and we decided to adopt the one first proposed by Booth, Butler and Hall (1994) for the SRSWOR design.
Assume that sample s of size n was selected using sampling design p and let $\hat{\mu}$ be the estimator of μ calculated from s.

General bootstrap algorithm

1) Duplicate each individual $k \in s$, $\left[1/\pi_k\right]$ times, where [.] designates the integer portion. We complete the population thus obtained by selecting a sample in s with an inclusion probability $\alpha_k = 1/\pi_k - \left[1/\pi_k\right]$. Let Y_k^*, $k \in U^*$ be the value of the variable of interest in the pseudo-population.
2) Draw M samples s_m^*, $m = 1,\ldots,M$ of size n in the pseudo-population U^* using the sampling design p^* with inclusion probabilities π_k^* and calculate

$$\hat{\mu}_m^*(t) = \frac{1}{N}\sum_{k\in s_m^*}\frac{Y_k^*(t)}{\pi_k^*},\ t\in\left[0,T\right] \text{ and } m = 1,\ldots,M.$$

3) Estimate the function $\hat{\sigma}(t)$ by the corrected empirical standard deviation of $\hat{\mu}_m^*(t)$, $m = 1,\ldots,M$,

$$\hat{\sigma}^2(t) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mu}_m^*(t) - \hat{\mu}_\bullet^*(t)\right)^2,$$

where

$$\hat{\mu}_\bullet^*(t) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mu}_m^*(t) \text{ and } t\in\left[0,T\right].$$

4) Choose c_α as the quantile of order $1-\alpha$ of the variables

$$\left(\sup_{t\in[0,T]} \frac{\mid \hat{\mu}_m^*(t) - \hat{\mu}(t) \mid}{\hat{\sigma}(t)} \right)_{m=1,\ldots,M}.$$

A technique similar to the one used in step 4 of the algorithm was used by Bickel and Krieger (1989) to construct confidence bands for a distribution function.

The SRSWOR design uses the general bootstrap algorithm for $\pi_k^* = n / N$, and for the STRAT design, we apply in each stratum U_h, for $h = 1, \ldots, H$, the algorithm used for the SRSWOR design with $\pi_k^* = n_h / N_h \;\; k \in U_h$. In this case, we are back to the algorithm proposed by Booth *et al.* (1994).

The adaptation of the bootstrap algorithm to the πps design was proposed by Chauvet (2007). It consists in selecting, during step 2 of the general algorithm, the sample s^* in U^* with inclusion probabilities

$$\pi_k^* = \frac{nx_k}{\sum_{k\in U^*} x_k}.$$

This change is necessary in order to comply with the constraint of fixed size during re-sampling. The inclusion probabilities π_k^* are also used to estimate $\hat{\mu}_m^*$ in step 2 of the general algorithm. The selection of a πps sample can be carried out using the cube algorithm with the balancing variable π. In these conditions, it is desirable to perform a random sort in the population U (or U^*) before the selection of s (or s_m^*) in order to obtain a sampling design close to maximum entropy (Chauvet 2007, Tillé 2011). Chauvet (2007) also gives asymptotic results concerning the convergence of the variance estimator obtained in the case of the bootstrap for the πps design.

Finally, it is also possible to adapt this general algorithm to estimate the variance function of the estimator $\hat{\mu}_{MA}$. In step 1 of the algorithm, we also calculate the values $\mathbf{x}_k^*$ of $\mathbf{x}_k$ in the pseudo-population U^*. Using the fact that the linear-model-assisted estimator is a nonlinear function of Horvitz-Thompson estimators, we calculate the bootstrapped value $\hat{\mu}_{MA}^*$ of $\hat{\mu}_{MA}$ over sample s_m^* according to

$$\hat{\mu}_{MA}^*(t) = \frac{1}{N} \sum_{k\in s_m^*} \frac{Y_k^*(t) - \mathbf{x}_k^{*\prime}\hat{\boldsymbol{\beta}}^*(t)}{\pi_k^*} + \frac{1}{N}\left(\sum_{k\in U} \mathbf{x}_k^{\prime} \right)\hat{\boldsymbol{\beta}}^*(t)$$

where $\hat{\boldsymbol{\beta}}^*(t) = \left(\sum_{s_m^*} \mathbf{x}_k^* \mathbf{x}_k^{*\prime} \right)^{-1} \sum_{s_m^*} \mathbf{x}_k^* Y_k^*(t)$. As Canty and Davison (1999) note, using the total of the variable $\mathbf{x}_k$ over the population U instead of the pseudo-population U^* yields better results, especially when this variable has extreme values.

4 Study of the mean electricity consumption curve

We have a population U consisting of $N = 15,069$ electricity consumption curves measured every half hour during two consecutive weeks. We have $D = 336$ measurement points for each week, and we

want to estimate the mean consumption curve for the second week. We denote by $\mathbf{Y}'_k = \left(Y_k\left(t_1\right),\ldots,Y_k\left(t_D\right)\right)$, the electricity consumption of individual $k \in U$ measured in the second week and $\mathbf{X}'_k = \left(X_k\left(t_1\right),\ldots,X_k\left(t_D\right)\right)$, the individual's consumption during the first week. The mean consumption of each individual k during the first week, $x_k = \sum_{d=1}^{D} X_k\left(t_d\right) / D$, which is simple piece of information that is inexpensive to transmit, will be used as auxiliary information. This variable (a real one), which is known for all units k in the population, is strongly related to the current consumption curve. As Figure 4.1 shows, the current consumption in each t is almost proportional to the mean consumption for the previous week.

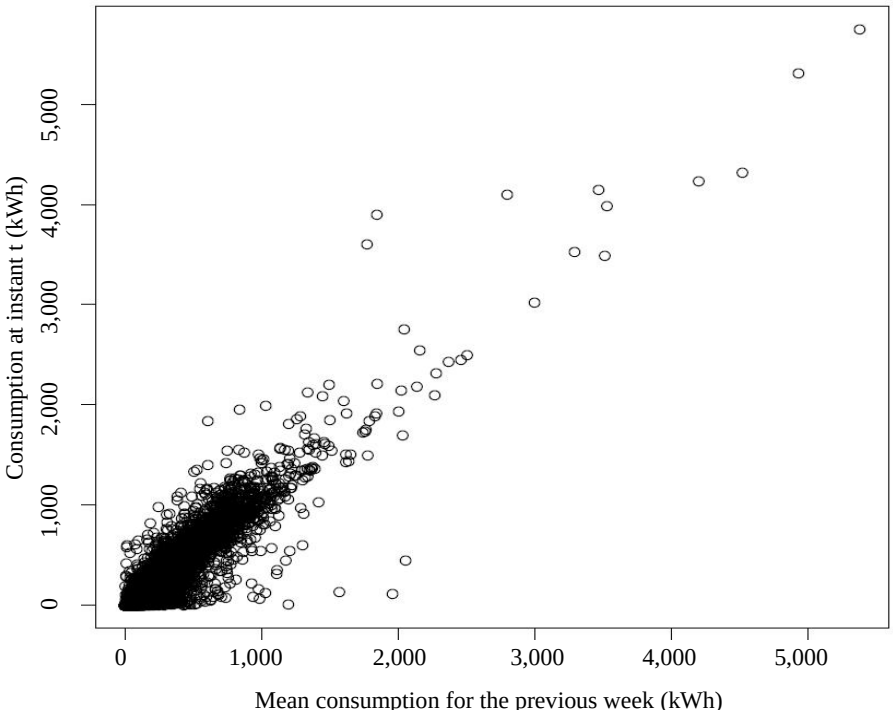


Figure 4.1 Representation of consumption at an instant t as a function of the mean consumption for the previous week

4.1 Description of strategies used

We consider samples of fixed size $n = 1,500$ obtained using different sampling designs. The strategies presented are repeated I times to evaluate and compare their performance.

1. *SRSWOR sampling and Horvitz-Thompson estimator*

This design is simple to implement; the Horvitz-Thompson estimator of the mean curve is given by (2.6) and the estimator of its covariance by (2.7).

2. *STRAT stratified design and Horvitz-Thompson estimator*

A stratified design is very effective if the strata are homogenous in relation to the variable of interest. In this study, we used the k-means algorithm to create the strata, and we considered $H = 10$ strata. A first stratification (STRAT 1) was carried out using the classification of the discretized trajectories $\mathbf{X}'_k$ for the

previous week. A second stratification, which uses only the aggregate information x_k was also considered. It is denoted by STRAT 2.

Tables 4.1 and 4.2 show the sizes of the strata N_h yielded using the two stratifications and the optimal sizes n_h, according to (2.5), of the samples to be selected in each stratum. In both cases, the strata are numbered in ascending order in relation to the mean consumption for each stratum. More specifically, stratum 1 corresponds to small consumers of electricity and stratum 10 refers to the 10 largest consumers. Note that the first stratification, which requires knowing the electricity consumption at each measurement instant t, requires more information than the second stratification. The mean curve is constructed using (2.3), and its covariance is estimated by (2.4).

Table 4.1
STRAT 1: Stratification based on curves. The strata are constructed using the curves for week 1. The optimal allocation n_h is calculated using the curves for week 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3,866	4,769	623	2,690	664	1,251	806	328	62	10
n_h	212	345	87	242	117	179	172	101	35	10

Table 4.2
STRAT 2: Stratification based on the mean consumption x_k. The optimal allocation n_h is calculated using the mean consumption for week 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3,257	4,236	3,139	1,937	1,189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

3. *πps sampling and Horvitz-Thompson estimator*

We used the cube algorithm proposed by Deville and Tillé (2004) and Chauvet and Tillé (2006), where the inclusion probabilities are proportional to $x_k, k \in U$. To have a sampling design close to maximum entropy, a random sort of the population is performed before selection of the sample s. The covariance of the estimator of the mean is estimated using Formula (2.9). The cube algorithm is available in R in the *sampling* package, *samplecube* function, and a SAS macro is available on the INSEE website (Institut National de Statistique et des Études Économiques).

4. *SRSWOR sampling and MA estimator*

The estimator $\hat{\mu}_{MA}$ assisted by the model ξ is constructed using the auxiliary information given by $\mathbf{x}'_k = (1, x_k)$, where x_k is the mean consumption for the previous week. In these conditions, $\hat{\mu}_{MA}$ is the sum over any population U of the values $\hat{Y}_k$ estimated by the model (*cf.* Formula (2.13)). The covariance of the estimator of the mean is estimated using Formula (2.15).

4.2 Error of estimation of the mean curve

The error of estimation of the mean curve μ at instants $t_1, ..., t_{336}$, is evaluated according to the following criterion:

$$R_2\left(\hat{\mu}\right) = \frac{1}{336}\sum_{i=1}^{336}\left(\hat{\mu}\left(t_i\right) - \mu\left(t_i\right)\right)^2 \approx \frac{1}{T}\int_0^T \left(\hat{\mu}\left(t\right) - \mu\left(t\right)\right)^2 dt.$$

The results are shown in Table 4.3 for $I = 10,000$ simulations (replications). They clearly show that for this study, taking account of total consumption for the previous week substantially improves the precision of the estimate of the mean compared with simple random sampling without replacement, dividing the mean square error R_2 by 5. Among the different strategies, the best appear to be those that take account of auxiliary information via inclusion probabilities (STRAT, πps and systematic PPS).

Table 4.3
Square error R_2 of estimation of the mean μ, with $I = 10,000$ replications

Strategy	Mean	1st quartile	Median	3rd quartile
SRSWOR	40.53	10.82	22.16	51.09
STRAT (1)	5.78	3.68	5.08	7.07
STRAT (2)	6.49	4.03	5.48	7.88
πps	7.06	3.99	5.52	8.16
Systematic $\pi - ps$	6.73	3.85	5.20	8.07
MA	8.29	5.24	7.14	10.06

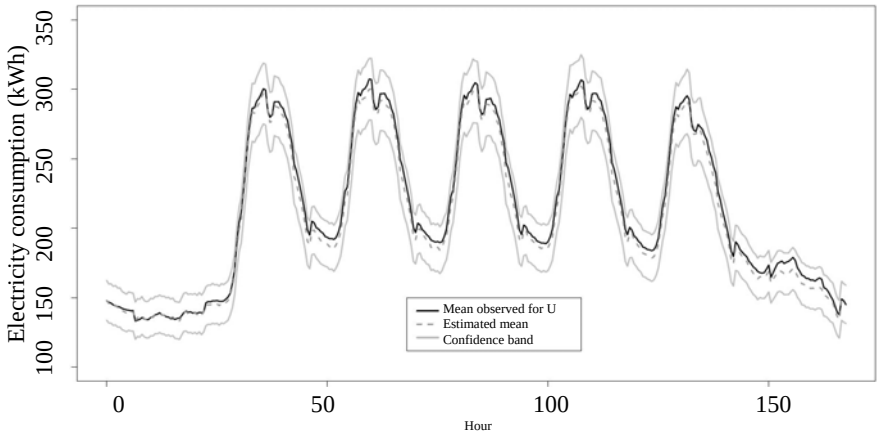
4.3 Coverage rate and width of confidence bands

The construction of confidence bands of level $1 - \alpha$ requires calculating quantiles of order $1 - \alpha$ of the supremum of Gaussian processes.

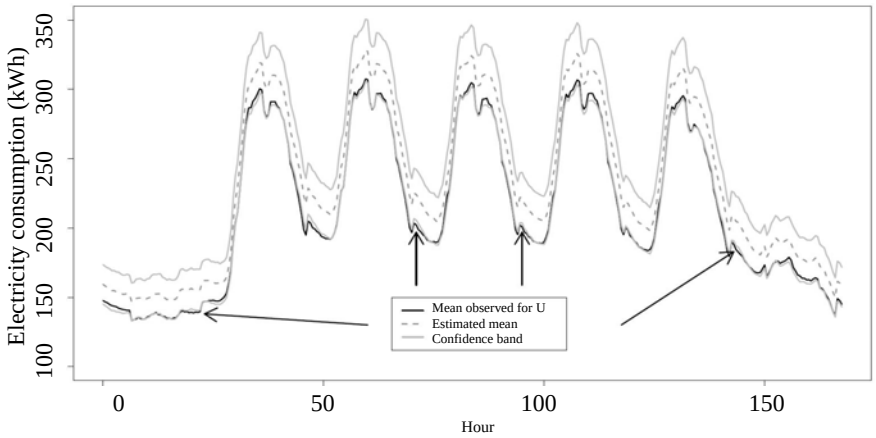
So as not to favour one method of constructing confidence bands over the other, we applied the two algorithms to the same sample s and we considered the same number M of processes. This number M varies from one estimator to the other owing to the computation time needed for the bootstrap approaches (see Section 4.4).

The empirical coverage rate is the proportion of times, among the $I = 2,000$ replications, where the true mean curve μ appears, for all instants t, within the confidence band constructed using an estimate $\hat{\mu}$. Figure 4.2 shows two examples of confidence bands (continuous grey curves) constructed from estimated curves (dotted grey curves). Figure 4.2(A) shows that the true mean curve for the population (continuous black curve) is within the confidence band at every instant. Conversely, Figure 4.2(B) shows that mean curve for the population is generally overestimated and there are a few instants (indicated by arrows) where the curve shown is outside the confidence band. Empirical coverage rates are shown in Table 4.4.

The two methods of constructing confidence bands yield coverage rates that are similar and fairly close to the desired nominal rates (95% and 99%). However, the results seem slightly less satisfactory for the πps designs and for the MA approach, for which the variance of the estimator is complex and more difficult to estimate precisely.



(A) The mean curve shown falls within the confidence band.



(B) The mean observed does not fall within the confidence band at the instants shown by the arrows.

Figure 4.2 Examples of confidence bands

Table 4.4
Empirical coverage rate (in %), for $I = 2,000$ replications

Method	Number M of processes	Bootstrap		Gaussian process	
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
SRSWOR	5,000	94.95	98.85	94.80	98.70
STRAT (1)	5,000	93.92	98.34	94.09	98.43
STRAT (2)	5,000	94.3	98.45	94	98.55
πps	1,000	94.73	98.77	93.87	98.61
MA	5,000	94.3	98.5	92.85	98.15

Another useful indicator is the mean width of the confidence band,

$$\frac{1}{336}\sum_{i=1}^{336} 2c_{\alpha}\hat{\sigma}\left(t_i\right) \approx \frac{1}{T}\int_0^T 2c_{\alpha}\hat{\sigma}\left(t\right)dt$$

the values of which are shown in Table 4.5. The two methods provide confidence bands of largely similar width. Also note that the use of the auxiliary variable considerably reduces the mean band width, which is cut in half if one of the stratified designs is used rather than a SRSWOR design.

Table 4.5
Mean width of confidence bands, for $I = 2,000$ replications

Method	Number M of processes	Bootstrap		Gaussian process	
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
SRSWOR	5,000	35.98	43.35	35.99	43.19
STRAT (1)	5,000	16.64	18.92	16.62	18.88
STRAT (2)	5,000	17.58	19.99	17.55	19.94
πps	1,000	17.85	20.31	17.62	19.93
MA	5,000	19.88	22.65	19.75	22.44

Figures 4.3 and 4.4 show the widths of the confidence bands for a level $\alpha = 0.05$, for each instant, depending on whether they are pointwise ($c_\alpha = 1.96$), estimated by simulations of Gaussian processes or obtained using the approach based on the Bonferroni inequality applied to each measurement point. We then have, in the latter case, $c_\alpha = 3.793048$, the quantile of order $1 - 0.05 / (336 \times 2)$ of a distribution $N(0,1)$. The bands obtained by Bonferroni are conservative, and they cover what might be considered the worst case in terms of information, the case of independence of the pointwise intervals. Note that the simulation approach substantially reduces the mean width of the bands in comparison with Bonferroni when the design does not allow all temporal information on the data to be taken into account (Figure 4.3). Conversely, for the stratified design (Figure 4.4), which provides a precise estimate of the mean curve, the confidence band constructed by simulation is close to that of Bonferroni, which can intuitively be interpreted as meaning that almost all the information was captured by the sampling design.

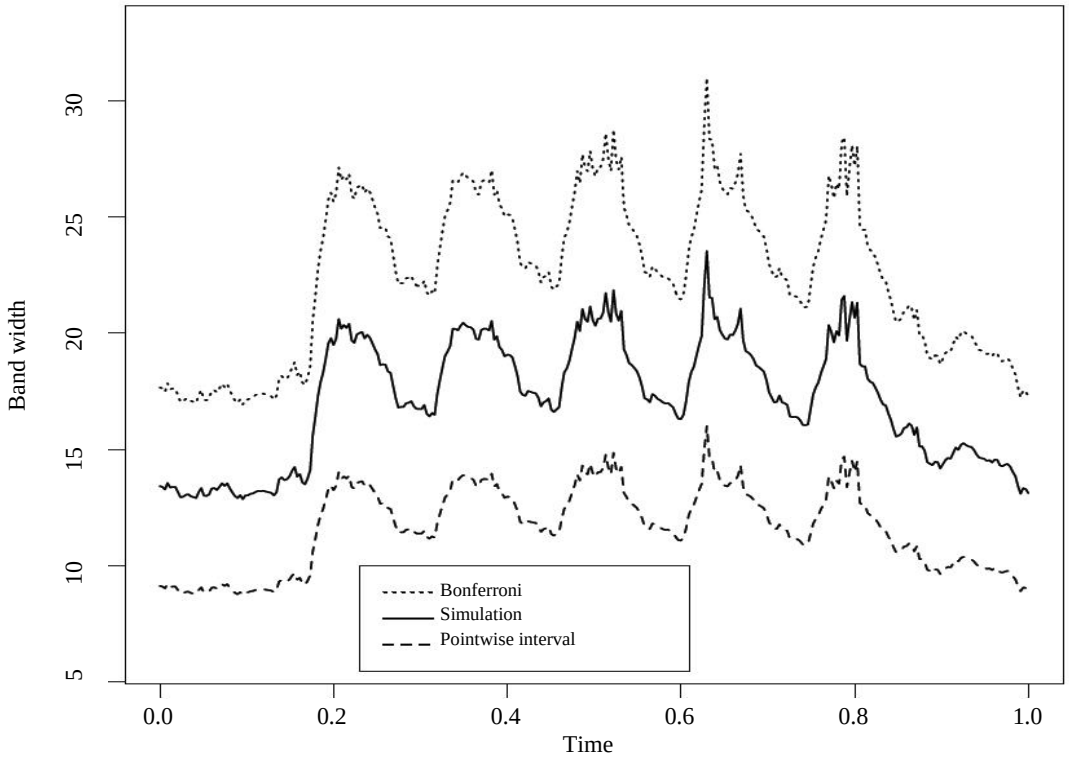


Figure 4.3 Simple random sampling without replacement. Width of confidence bands—pointwise, overall by process simulations, and with Bonferroni ($\alpha = 0.05$)

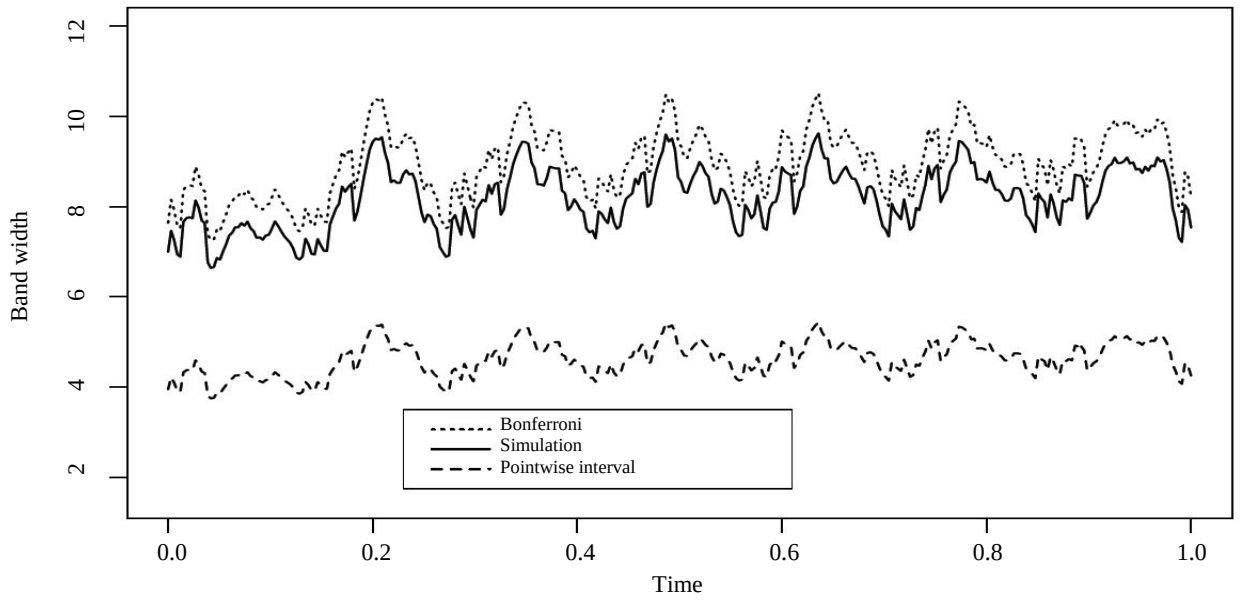


Figure 4.4 Stratified sampling (STRAT 1). Width of confidence bands—pointwise, overall by process simulations, and with Bonferroni (with $\alpha = 0.05$)

4.4 Computation time

Computation times with the bootstrap method are much greater—by a factor of approximately 1 to 1,000—than those with the Gaussian processes simulation method (*cf.* Table 4.6). The reason for this major difference is that in the bootstrap methods, the entire estimation process (construction of the fictitious population, drawing of a new sample, calculation of the estimator) must be repeated for each bootstrapped sample. Also, the designs that introduce auxiliary information are slower than SRSWOR, even though if used individually their computation time is entirely reasonable.

Table 4.6
Run time of a simulation in seconds for $M = 5,000$ replications. The SRSWOR, MA and STRAT strategies were programmed with R and πps with SAS.

Strategy	Bootstrap	Gaussian processes
SRSWOR	1,170.6	1.0
STRAT	1,839.5	1.4
πps	5,020.0	7.3
MA	3,156	1.4

5 Conclusion and perspectives for research

In this study, we have implemented and compared different strategies for using auxiliary information for estimating, and constructing confidence bands for, the mean of data in the form of curves. This information can be taken into consideration at the time of sampling by using unequal probability designs

or during estimation with simple random sampling without replacement, assisted by a functional-response regression model. It seems clear from our example of electricity consumption curves that when total consumption for the previous week is known, the precision of estimators of the mean can be greatly improved compared with an SRSWOR-type sampling.

Also, in this context of large samples and high-dimensional data, it also seems possible to construct, for these different strategies, confidence bands that have empirical coverage rates close to the desired rates. The two considered approaches—estimation of the covariance function and simulation of Gaussian or bootstrap processes—seem to perform comparably in terms of the width of the confidence bands; the main difference is in the computation time. The bootstrap, which seems more general because it does not require having a good estimator of the covariance function, proves to be much slower in practice.

Sometimes, in these flows of large-scale data, there are losses of information owing to signal transmission problems. The end result is that the utility has incomplete records of some trajectories. This issue, of partial non-response, can probably be dealt with by considering adaptations of classical non-response techniques (Haziza 2009) in the functional context. A fundamental question, then, is how to construct good estimators of the covariance function.

Acknowledgements

We wish to thank the anonymous referees as well as Guillaume Chauvet and Jean-Claude Deville for their helpful comments, which led to improvements in this study.

References

Bickel, P., and Krieger, A. (1989). Confidence bands for a distribution function using the bootstrap. *Journal of the American Statistical Association*, 84, 95-100.

Booth, J., Butler, R. and Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, 89, 1282-1289.

Canty, A.J., and Davison, A.C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48, 379-391.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140, 75-91.

Cardot, H., Degras, D. and Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19, 2067-2097.

Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic J. of Statistics*, 7, 562-596.

Cardot, H., and Josserand, E. (2011). Horvitz-thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98, 107-118.

Chaouch, M., and Goga, C. (2012). Using complex surveys to estimate the l1-median of a functional variable: Application to electricity load curves. *International Statistical Review*, 80, 40-59.

Chauvet, G. (2007). Méthodes de bootstrap en population finie. Ph.D. thesis, Université de Rennes II.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Cochran, W. (1977). Sampling techniques. New York: John Wiley & sons, Inc., 3rd Edition.

Cuevas, A., Febrero, M. and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51, 1063-1074.

Dauxois, J., and Pousse, A. (1976). Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Ph.D. thesis, Université Paul Sabatier, Toulouse.

Degras, D. (2011). Simultaneous confidence bands for parametric regression with functional data. *Statistica Sinica*, 21(4), 1735-1765.

Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir des mesures synchrones. In *Méthodes de Sondages* (Eds., P. Guibert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Dunod, France, 353-357.

Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Ann. Insee, 15, 3-104.

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

Deville, J., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Faraway, J. (1997). *Regression analysis for a functional response*. Technometrics, 39(3), 254-261.

Ferraty, F., and Romain, Y., editors (2011). *Oxford Handbook of Functional Data Analysis*. Oxford University Press.

Goga, C., and Ruiz-Gazen, A. (2013). Efficient estimation of nonlinear finite population parameters using nonparametrics, to appear in the *Journal of the Royal Statistical Society*, Series B, DOI: 10.1111/rssb.12024.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

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

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Sample Surveys: Theory Methods and Inference*, volume 29 of Handbook of Statistics, (Eds., C. Rao and D. Pfeffermann), North-Holland, 215-246.

Madow, W. (1949). On the theory of systematic sampling, ii. *Annals of Mathematical Statistics*, 19, 535-545.

Ramsay, J., and Silverman, B. (2005). *Functional data analysis*. Springer, New York, second edition.

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

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

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

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: An appraisal. *Survey Methodology*, 37, 215-226.

Yates, F., and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 303-321
Statistics Canada, Catalogue No. 12-001-X

Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data

Chen Xu, Jiahua Chen and Harold Mantel[1]

Abstract

Regression models are routinely used in the analysis of survey data, where one common issue of interest is to identify influential factors that are associated with certain behavioral, social, or economic indices within a target population. When data are collected through complex surveys, the properties of classical variable selection approaches developed in i.i.d. non-survey settings need to be re-examined. In this paper, we derive a pseudo-likelihood-based BIC criterion for variable selection in the analysis of survey data and suggest a sample-based penalized likelihood approach for its implementation. The sampling weights are appropriately assigned to correct the biased selection result caused by the distortion between the sample and the target population. Under a joint randomization framework, we establish the consistency of the proposed selection procedure. The finite-sample performance of the approach is assessed through analysis and computer simulations based on data from the hypertension component of the 2009 Survey on Living with Chronic Diseases in Canada.

Key Words: Variable selection; Sampling weights; Model-design-based inference; BIC; Penalized likelihood; Selection consistency.

1 Introduction

In many areas of scientific research, one common interest is to identify the influential factors associated with certain behavioral, social, or economic indices within a target population. For example, sociologists would like to identify important factors that affect the unemployment rate in a specific region, and epidemiologists are interested in finding risk behavior for diseases. In such studies, researchers often start with a survey of the target population (*e.g.*, Rahiala and Teräsvirta 1993; Korn and Graubard 1999; Wolfson 2004). A representative sample is then selected and measurements of the variables of interest for the sampled units are collected. A regression model is routinely employed to summarize the information contained in the data. It explains variations in the response variable through a simple function of explanatory variables (covariates). When they lack prior knowledge, researchers may collect information on many potential explanatory variables. The goal of identifying influential factors can be achieved through a variable selection procedure.

Variable selection is fundamental in statistical modeling. In non-survey settings, classical selection criteria have been developed to assess and select candidate variables. Examples include Mallow's C_p statistic (Mallows 1973), the (generalized) cross-validation (CV/GCV; Stone 1974; Craven and Wahba 1979), the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (BIC; Schwarz 1978). All these criteria are very useful and can provide meaningful inferences in practice.

Despite the abundance of the literature on variable selection, it has received little attention in the context of survey sampling. When variable selection methods are applied to survey data, many potential complications arise. We focus on issues related to special features of surveys. First, data collected through

1. Chen Xu and Jiahua Chen, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4. E-mail: chen.xu@stat.ubc.ca and jhchen.stat.ubc.ca; Harold Mantel, Statistical Research and Innovation Division, Statistics Canada, Ottawa, ON, Canada, K1A 0T6. E-mail: Harold.Mantel@statcan.gc.ca.

survey sampling are usually obtained from a finite population without replacement, and hence they have an intrinsic dependence structure. Second, in complex survey designs, the inclusion probabilities of sampling units often vary over the target population. Consequently, the correlation between the response and the covariates reflected in the sample can be distorted from the population. This is potentially the case when some parts of the population are sampled more intensively than the others. Ignoring survey designs in the selecting process may result in biased selection results for the target population.

In the literature, sampling weights are often utilized in estimating parameters in regression models based on survey data. The weighted estimates of regression coefficients are helpful to avoid the biased inference from informative sampling (Pfeffermann 1993; Fuller 2009, Section 6.3; Skinner 2012). Although model estimation and selection serve for their own purposes, they often have coherent linkage in a modeling process. It is natural to conjecture that using sampling weights is beneficial for the variable selection.

In this spirit, we investigate the use of pseudo-likelihood to take account of the sampling weights, and derive a pseudo-likelihood-based BIC criterion for variable selection of survey data. A penalized pseudo-likelihood-based procedure (PPL) is further proposed for numerical implementation of the proposed criterion. Under a joint randomization framework, we prove that the new procedure consistently identifies the influential variables. The weighted selection method is assessed through simulation studies and using data from the 2009 Survey on Living with Chronic Diseases in Canada.

The paper is organized as follows. In Section 2, we introduce the joint randomization mechanism and the super-population model. In Section 3, we derive the pseudo-likelihood-based BIC for the analysis of survey data and propose its implementation via the PPL procedure. In Section 4, we investigate the asymptotic behavior of the proposed BIC procedure. We use numerical studies in Section 5 to further assess the performance of our approach and provide concluding remarks in Section 6. We provide the proofs of theorems in a separate technical supplement: Xu and Chen (2012), where the derivation of proposed BIC can also be found.

2 Joint inference and super-population

The random behavior of an inference procedure is mostly inherited from the randomness in the data. In the context of surveys, the set of sampled units is random because of the probabilistic sampling design. At the same time, the value of each sampling unit may be regarded as a random outcome from some conceptual infinite super-population (Royall 1976).

In a design-based analysis, the finite population is regarded as nonrandom and all measurements of sampling units are constants. The parameters of interest are finite population quantities such as the population total or the population median. The statistical inference is evaluated based on the randomness from the probability design.

One may also regard the design-induced randomness as an artifact. The measurements of sampled units are independent realizations of a random variable from a probability model for the postulated super-population. The parameters of interest are related to the assumed model and model-based inferences are evaluated solely based on the randomization introduced from the model.

A third approach is called model-design-based inference; it incorporates the randomization from both design and model. In such a joint randomization mechanism, the finite population is regarded as a random

sample from a super-population. The survey sample is considered as a second-phase sampling from the super-population. The parameters of interest can be either model or finite-population parameters. In this mechanism, inferences on the finite-population parameters are motivated from the super-population model. Model-design-based inference can be more efficient than pure design-based approaches when the finite population is well described by the super-population model. Compared with pure model-based approaches, it protects against model violation and is therefore more robust in general (see, *e.g.*, Binder and Roberts 2003; Kalton 1983).

We study the variable selection problem under the joint randomization mechanism. Let $\mathcal{D} = \{1, \ldots, N\}$ be a finite population consisting of N sampled units. The measurements on the i^{th} unit are denoted $(y_i, \mathbf{x}_i)$, where y_i is the response of interest and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a p-dimensional explanatory vector (covariate vector). These are regarded as independent realizations of $(Y, \mathbf{X})$ from a super-population. We postulate a generalized linear model (GLM) on the super-population as follows. Conditioning on $\mathbf{X}$, the distribution of Y belongs to a natural exponential family, the density of which takes the form

$$f(y;\theta) = c(y)\exp\{\theta y - b(\theta)\}. \tag{2.1}$$

θ is known as the natural parameter of $f(y;\theta)$ such that $b'(\theta) = E[Y|X] \equiv \mu$ and $b''(\theta) = \text{Var}[Y|X] \equiv \sigma^2$, and $c(y)$ is a non-negative base measure. The influence of the explanatory variable $\mathbf{X}$ on Y is expressed through $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$ for some assumed linkage function $g(\cdot)$, where the vector $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ is the p-dimensional regression coefficient. If $g(\cdot)$ is the canonical link, *i.e.*, $g(\mu) = \theta$, then we have $\theta = \mathbf{X}^T\boldsymbol{\beta}$. For simplicity, we focus on the canonical link in this paper.

Based on this model, the effect of the explanatory variable is characterized through the size of the corresponding regression coefficient. In applications, a complex model with many variables often leads to over-fitting and a poor interpretive value. Hence, it is desirable to fit the data with a parsimonious model in which many regression coefficients are estimated to be zero. Explanatory variables with nonzero coefficients are then considered to be influential on the response. To this end, we assume that $\boldsymbol{\beta}$ is ideally sparse, and address the variable selection problem through identifying a sparse model formed by the covariates with nonzero coefficients.

3 Pseudo-likelihood-based selection with BIC

3.1 BIC in surveys

With the model settings described in Section 2, it is clear that, if the measurement $(y_i, \mathbf{x}_i)$ is observed for every unit in population $\mathcal{D}$, the randomness in the data introduced by the probability sampling design is completely gone. In this situation, the selection of the influential variables is based on the entire population and the classical selection criteria developed in non-survey settings (purely model-based) remain valid for model-design-based inference. In particular, let $s \subseteq \{1, \ldots, p\}$ be an arbitrary set of

$\tau(s)$ covariates, which corresponds to a candidate model in form of (2.1). The "census-based" BIC (Schwarz 1978) selects the model (covariates) that minimizes

$$\mathrm{BIC}_N(s) = -2l_N\left(\tilde{\boldsymbol{\beta}}_s\right) + \tau(s)\log N, \tag{3.1}$$

where $l_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log f\left(y_i; \mathbf{x}_i\boldsymbol{\beta}\right)$ is the census log-likelihood function and $\tilde{\boldsymbol{\beta}}_s$ is the maximizer of $l_N(\boldsymbol{\beta})$ based on s. It can be seen that the BIC (3.1) is a decreasing function of the maximized log-likelihood and an increasing function of the number of variables included in the model. Hence, a lower BIC implies either a simpler model (fewer explanatory variables), a better fit (higher maximized likelihood), or both. A model with balanced complexity and goodness of fit is preferred.

We note that the census BIC (3.1) is conceptual, because observing $\left(y_i, \mathbf{x}_i\right)$ for all units in $\mathcal{D}$ is usually not feasible in applications. Instead, a representative sample $d = \left\{i_1, \ldots, i_n\right\} \subset \left\{1, \ldots, N\right\}$ with n units is often drawn from $\mathcal{D}$ and the measurements are observed based on the sampled units. Due to the intrinsic dependence structure among the sampled units, a full likelihood on d is prohibitive to compute in general. Alternatively, for the model-design-based inference, a pseudo-log-likelihood function is frequently used, which takes the form

$$l_n(\boldsymbol{\beta}) = \sum_{i \in d} w_i \log f\left(y_i; \boldsymbol{\beta}\right) \tag{3.2}$$

with $w_i = k/P(i \in d)$ denoting the survey weight for the i^{th} unit. The scaling parameter k in w_i does not have analytical impacts on the pseudo-likelihood-based inference. For the simplicity of presentation, we choose $k = n/N$ such that $n^{-1}l_n(\boldsymbol{\beta})$ is design-unbiased to $N^{-1}l_N(\boldsymbol{\beta})$. Maximizing $l_n(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ leads to a maximum pseudo-likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, *i.e.*,

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} l_n(\boldsymbol{\beta}).$$

Under the appropriate sampling designs, $\hat{\boldsymbol{\beta}}$ is often $n^{-1/2}$ consistent for $\boldsymbol{\beta}$ under the joint randomization framework. The idea of using pseudo-likelihood for inference on model parameters has been widely adopted in the literature (see, *e.g.,* Binder 1983; Godambe and Thompson 1986; Molina and Skinner 1992).

In this paper, we aim to develop an analogue of BIC criterion based on the pseudo-likelihood. Following the super-population formulation described in Section 2, let $\boldsymbol{\beta}_s$ be the $\tau(s)$-dimensional coefficient of model s and let ν_s be the prior density of $\boldsymbol{\beta}_s$. Then a pseudo-marginal density function of the data is given by

$$P_n\left(\mathbf{y}|s\right) = \int L_n\left(\mathbf{y}; \boldsymbol{\beta}_s\right) \nu_s\left(\boldsymbol{\beta}_s\right) d\boldsymbol{\beta}_s$$

with $L_n\left(\mathbf{y}; \boldsymbol{\beta}_s\right) = \exp\left\{l_n\left(\mathbf{y}; \boldsymbol{\beta}_s\right)\right\}$. Consequently, we may regard the following expression as the pseudo-posterior probability of the model s :

$$P_n\left(s|\mathbf{y}\right) = \frac{P_n\left(\mathbf{y}|s\right)P(s)}{\sum_{s \in S} P(s)\,P_n\left(\mathbf{y}|s\right)}, \tag{3.3}$$

where S denotes the collection of all candidate models. In the spirit of Bayesian analysis, the model with the highest $P_n\left(s|\mathbf{y}\right)$ is then considered to be the one that receives the most support from the data. Since $\sum_{s\in S} P(s)\,P_n\left(\mathbf{y}|s\right)$ does not depend on any specific model, the highest $P_n\left(s|\mathbf{y}\right)$ is achieved by the model that maximizes the corresponding $P_n\left(\mathbf{y}|s\right)P(s)$. When the uniform prior $P(s)=\zeta$ is used and the weight scaling is chosen as $k=n/N$, we obtain a Laplace approximation under some regularity conditions (see Xu and Chen 2012):

$$-2\log\left\{P_n\left(\mathbf{y}|s\right)\right\} = -2l_n\left(\hat{\boldsymbol{\beta}}_s\right) + \tau(s)\log n + O_p(1).$$

Accordingly, we choose the model s that minimizes

$$\mathrm{BIC}_n(s) = -2l_n\left(\hat{\boldsymbol{\beta}}_s\right) + \tau(s)\log n. \tag{3.4}$$

Compared with the census BIC (3.1), the first term in BIC (3.4) is the maximum survey-weighted pseudo-likelihood, which is potentially helpful to avoid sampling errors that might lead to biased inferences for the target population. We refer to (3.4) as a pseudo-likelihood-based version of BIC in the context of surveys. In the joint randomization framework, we establish the selection consistency of using BIC (3.4) through a PPL-based implementation procedure, as will be seen in Section 4.

3.2 Implementing BIC via penalized pseudo-likelihood

In applications, a straightforward way to implement BIC is best-subset selection, where BIC is evaluated and compared for each candidate model. However, this procedure can be computationally impractical when the number of covariates is large. Alternatively, penalized likelihood methods have recently been used as computationally efficient procedures for implementing a selection criterion. These methods exclude variables from the model by estimating their coefficients to be zero, and shrink the other coefficients accordingly. By varying the penalty on the likelihood, we can obtain a series of models with differing sparsity. To avoid an exhaustive search of the entire model space, the selection criterion is used to pick an optimal one among these sparse models. The effectiveness of this implementation strategy has been illustrated in the non-survey context for BIC (Wang, Li and Tsai 2007; Liu, Wang and Liang 2011) and GCV (Fan and Li 2001; Xie, Pan and Shen 2008) among others.

Sharing the same spirit, we proposed a penalized pseudo-likelihood (PPL) procedure for the implementation of BIC (3.4) for survey data. Specifically, following pseudo-likelihood (3.2) with $k=n/N$, we define the survey-weighted penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$ that maximizes the penalized pseudo-likelihood function

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{p}\phi_\lambda\left(|\beta_j|\right), \tag{3.5}$$

where $\phi_\lambda(\cdot)$ is a penalty function indexed by a tuning parameter λ controlling the size of the penalty. With an appropriate choice of $\phi_\lambda(\cdot)$, $\hat{\boldsymbol{\beta}}_\lambda$ contains zero estimates for some coefficients and thus automatically produces a sparse model. The desirable sparsity of $\hat{\boldsymbol{\beta}}_\lambda$ typically requires the singularity of the corresponding $\phi_\lambda(\cdot)$ at the origin. Some popular choices of $\phi_\lambda(\cdot)$ include the L_γ penalty (Frank and

Friedman 1993; Tibshirani 1996), *i.e.*, $\phi_\lambda\left(|\beta|\right) = \lambda|\beta|^\gamma$ with $\gamma \in (0,1]$, and the SCAD penalty (Fan and Li 2001), which is defined by the following derivative:

$$\phi'_\lambda\left(|\beta|\right) = \lambda\left\{I\left(|\beta| \le \lambda\right) + \frac{\left(a\lambda - |\beta|\right)_+}{(a-1)\lambda} I\left(|\beta| > \lambda\right)\right\} \tag{3.6}$$

with $a = 3.7$ being a common choice.

With different values of λ for a properly specified $\phi_\lambda(\cdot)$, $\hat{\boldsymbol{\beta}}_\lambda$ leads to models of differing sparsity. These sparse models (with respect to λ) naturally form a collection of candidate models. BIC (3.4) can then be used to select an optimal model within this collection. To be more specific, let Ω be the range of λ and let s_λ denote the model produced by $\hat{\boldsymbol{\beta}}_\lambda$. We treat $S_\Omega = \{s_\lambda : \lambda \in \Omega\}$ as the collection of candidate models under consideration, and select the model $s^* \in S_\Omega$ such that $\mathrm{BIC}_n(s^*) = \min_{\lambda\in\Omega} \mathrm{BIC}(s_\lambda)$. We refer to this selection procedure as the penalized pseudo-likelihood-based BIC method (PPL-BIC). Compared with traditional best-subset selection, the PPL-BIC procedure focuses on the models that are produced by the survey-weighted penalized estimators, and therefore it can be much less computationally expensive.

4 Consistency of PPL-BIC

We now investigate the asymptotic behavior of the PPL-BIC procedure under the joint randomization framework. Suppose there is a sequence of finite populations, say $\mathcal{D}_r$ with $r \to \infty$. Each $\mathcal{D}_r$ is an independent and identically distributed (i.i.d.) sample of size N_r from a super-population modeled by (2.1) with random variable $\left(Y, \mathbf{X} = \{X_1, \ldots, X_p\}\right)$. Within each $\mathcal{D}_r$, a sample d_r of size n_r is drawn according to some sampling scheme. We assume that both N_r and n_r increase to infinity as $r \to \infty$, with the sampling fraction n_r/N_r bounded by some constant $C < 1$. For simplicity of notation, we will drop the index r in the following discussion.

Without loss of generality, we assume that the first q coefficients are nonzero and denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0 = \{\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}\}$ with $\boldsymbol{\beta}_{02} = 0$. Also, we use s_0 to denote the true model $\{1, \ldots, q\}$ to be identified. We establish the selection consistency of PPL-BIC in two steps. In the first step we show that, for appropriate choices of $\phi_\lambda(\cdot)$, the PPL can consistently identify the true s_0 so that $s_0 \in S_\Omega$ with probability tending to 1. In the second step, we verify that BIC (3.4) consistently selects s_0 over S_Ω.

For the asymptotic analysis, we define $\varphi_\lambda = \max\left\{\phi'_\lambda\left(|\beta_{0j}|\right) \text{ for } j \in s_0\right\}$ and associate λ with n to make φ_λ a sequence. Under the joint randomization framework, we show the claim of step 1 as the following theorem.

Theorem 1 Under regularity conditions on model (2.1) and other requirements specified in the online supplement, if $\varphi_\lambda \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_\lambda = \left(\hat{\boldsymbol{\beta}}_{\lambda 1}, \hat{\boldsymbol{\beta}}_{\lambda 2}\right)$ of the penalized pseudo-likelihood function (3.5) such that

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2} + \varphi_\lambda\right) \quad \text{and} \quad P\left\{\hat{\boldsymbol{\beta}}_{\lambda 2} = 0\right\} \to 1$$

with $\|.\|$ denoting the Euclidean norm.

The consistency result in Theorem 1 holds for popular nonconvex penalty functions. For example, for the L_γ penalty with $\gamma \in (0,1)$, consistency holds if $\lambda \to 0$; for the SCAD penalty, consistency holds if $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$. It also implies that with probability tending to 1, the true model s_0 is included in S_Ω, which serves as a prerequisite for the selection consistency of BIC over S_Ω.

We now establish the consistency of using BIC on S_Ω with a specified $\phi_\lambda(\cdot)$ that satisfies Theorem 1. Following the notation used in Section 3.2, let s_λ be the model corresponding to a PPL estimator $\hat{\boldsymbol{\beta}}_\lambda$, and let Ω be the range of λ under consideration. We define two collections of candidate models as follows:

- Over-fitted models: $S_+ = \{s : s_0 \subset s, s \neq s_0\}$;
- Under-fitted models: $S_- = \{s : s_0 \not\subset s\}$.

Notation $\not\subset$ denotes there is at least one different element between two sets, so that S_- is the collection of candidate models which does not include all variables in the true model. Then, Ω can be partitioned accordingly into

$$\Omega_+ = \{\lambda : s_\lambda \in S_+\},\ \Omega_- = \{\lambda : s_\lambda \in S_-\},\ \Omega_0 = \{\lambda : s_\lambda = s_0\}. \tag{4.1}$$

By Theorem 1, we have shown that $P(\Omega_0 \neq \varnothing) \to 1$. Therefore, the selection consistency of BIC over S_Ω is achieved if BIC is able to identify s_0 from any model s_λ with $\lambda \in \Omega_+ \cup \Omega_-$. We use the following theorem to establish this consistency result.

Theorem 2 Under the same conditions as in Theorem 1,

$$P\left\{\min_{\lambda \in \Omega_+ \cup \Omega_-} \mathrm{BIC}_n(s_\lambda) \leq \mathrm{BIC}_n(s_0)\right\} \to 0,$$

where Ω_+ and Ω_- are defined in (4.1).

5 Numerical studies

To evaluate the finite sample performance of PPL-BIC, extensive numerical studies have been conducted using data from the Survey on Living with Chronic Diseases in Canada (SLCDC; Statistics Canada 2009). In particular, we compare the proposed procedure with classic non-survey methods based on regression models postulated between SLCDC variables and hypothetical (simulated) responses. We tentatively reveal some insights for using pseudo-likelihood-based selection under two simulation scenarios. In the first scenario, populations are generated from presumed models and samples are obtained by designs that potentially create spurious correlations among SLCDC variables. In the second scenario, populations are not accurately generated from presumed models and samples are obtained by a design related to both response and candidate covariates. Also, we report the analysis of the original SLCDC 2009 data as an example for using PPL-BIC in real applications.

5.1 SLCDC data

SLCDC is a cross-sectional study sponsored by the Public Health Agency of Canada that collects information related to the experiences of Canadians with chronic health conditions. One of the main

objectives of SLCDC is to identify health behavior that influences disease outcomes, so that the government can better plan and provide health services for people with chronic diseases.

SLCDC takes place every two years, with two chronic diseases covered in each survey cycle. The 2009 survey focused on arthritis and hypertension. We restrict our attention to hypertension. The target population for the hypertension survey is Canadians aged twenty years or older from the ten provinces who have been diagnosed with hypertension and who live in private dwellings. To facilitate the survey process, the sampling units of SLCDC 2009 are people with hypertension who completed the 2008 Canadian community health survey (CCHS). For the purpose of SLCDC, the population is first stratified according to the CCHS respondents based on sex and four age groups: 20-44, 45-64, 65-75, and 75+. Therefore, the finite population formed by the CCHS respondents was divided into 8 categories, age (4 levels) by sex (2 levels). A stratified sampling plan is used for SLCDC with proportional sample size allocation. An overall sample of 9,005 was selected from the 17,437 CCHS respondents, and 6,142 respondents completed the SLCDC survey.

We identified 40 variables relevant to hypertension based on the original SLCDC data, among which 7 variables have complete information on all 6,142 respondents. The remaining 33 variables have some amount of missing values due to the non-responses in the original questionnaire (see Table 5.5 in Appendix for the list of variables and corresponding non-response rates). There was no obvious systematic reason for the item non-response. The variable with most severe missingness is INCDRPR (household income) with a 9.6% non-response rate, while the amount of missing data is relatively minor for the remaining variables. To facilitate the analysis, we used simple imputation methods for the missing data as follows. For a categorical variable, we imputed the non-response value by a random value from the response set; for a continuous variable, we imputed the non-response value by the mean value of the responses. Two exceptions for above imputation are variables BMHX_02 and CNHX_05. The former one acts as the response variable of the regression model in the later data analysis, while the later one has natural restrictions on the range of its value. Instead, we removed the 274 observations with missing values in these two variables, which results in the basic working data with 5,868 observations. The imputation/removal procedure does not have any effect on evaluating the BIC procedure based on simulated population. It could bias the analysis of the real data. Yet given the low rate of missingness, and plausibility of missing at random in the specific case, the conclusion is unlikely to be severely affected.

Since the SLCDC is a follow-up to the CCHS, the sampling weights for SLCDC were initially obtained from the weights of the CCHS data. The weights were then adjusted to ensure that the SLCDC respondents represent the target population. Consequently, the adjusted weights show considerable variation between sampled units. After scaling by $k = n/N \approx 10^{-3}$, the adjusted weights vary between 0.01 to 33.62 with an inter-quartile range of 0.76.

5.2 Scenario 1: Spurious correlation

As mentioned, in complex survey designs, the correlation structure between variables reflected in the sample can be distorted from the population. In the first simulation scenario, we assess the purposed BIC method when data are collected through designs that potentially create spurious correlations between candidate covariates. Specifically, we treat the 40 identified variables as candidate covariates for some hypothetical response Y, and index them as X_1 to X_{40} for simplicity. We consider both continuous and binary responses in our simulations. For the continuous cases, we generate the values of Y according to

- Model 1 : $Y = 0.7X_6 + 0.7X_{10} + 0.6X_{18} - 0.6X_{22} + \varepsilon,$
- Model 2 : $Y = 0.7X_6 + 0.6X_{10} + 0.6X_{18} - 0.5X_{22} + 0.3X_{30} - 0.3X_{34} + \varepsilon,$

with $\varepsilon \sim N(0,1)$. For the binary cases where $Y \in \{0,1\}$, we generate the values of Y according to the logistic models

- Model 3 : $\text{logit}\left(\Pr\{Y = 1|\mathbf{X}\}\right) = 0.7X_7 - 0.6X_8 + 0.5X_{26},$
- Model 4 : $\text{logit}\left(\Pr\{Y = 1|\mathbf{X}\}\right) = 0.8X_7 - 0.7X_8 + 0.6X_{26} - 0.5X_{28} + 0.4X_{36}.$

The specified models include one of the strata identifiers in SLCDC (*i.e.*, X_6 or X_7) with a nested structure for each modeling context.

The finite population used in the simulation was created as follows. The basic working data of 5,868 respondents was duplicated 10 times proportional to the rounded integer values of SLCDC weights, which results in a pseudo-finite population of size 55,950 with complete information on $X_1, \ldots, X_{40}$. The values of response Y were then generated based on Models 1-4 respectively. We consider the variable selection problem to be the identification of the postulated model that generates the values of Y.

We investigate the performance of proposed procedure under two stratified sampling plans. Specifically, we create 4 strata based on variables X_6 (age, 55-/55+) and X_7 (sex, Male/Female), which leads to the group (Female, 55-) of size 7,120, group (Female, 55+) of size 19,199, group (Male, 55-) of size 6,187, and group (Male, 55+) of size 23,458. In the first plan, a simple random sampling without replacement (SRSWOR) with equally allocated sample size is drawn from each stratum. The inference is made based on the four SRSWORs pooled together. In the second plan, we further construct three subgroups within each stratum based on the sum of two binary covariates of the postulated models. In particular, the subgroups are built according to $X_{18} + X_{22}$ for data generated by Models 1-2, while the subgroups are similarly construct based on $X_8 + X_{26}$ for data from Models 3-4. We then make inference based on SRSWORs drawn from each sub-group of the four strata. The overall sample size is equally allocated at the stratum level with a 2:1:2 proportion for the three subgroups within a same stratum. A simple Monte Carlo computation reveals that the sample correlation between X_{18} and X_{22} (for data from Models 1-2) can be as high as 0.5, whereas their population-based correlation is merely around 0.02. Similar phenomenon is also observed between X_8 and X_{26} (for data from Models 3-4). We therefore expect variable selection under the second sampling plan to be more challenging due to this systematic inflation. In the simulations, we set the overall sample size $n = 500$ for Models 1-2 and $n = 1{,}500$ for Models 3-4. A summary of influential variables to the response and the design variables affecting the sampling probabilities can be found in Appendix (Table A.2).

The PPL-BIC selection procedure was carried out on probability samples obtained from the finite population. In particular, we scaled the survey weights as mentioned in (3.2) and chose the SCAD penalty for the penalized pseudo-likelihood function (3.5). The corresponding maximizer of (3.5) was solved by using the thresholding-based iterative algorithm (She 2011). For comparison purpose, the ideas of AIC and GCV are also used as alternatives for the proposed BIC (3.4). Based on the discussion in Section 3, we define the pseudo-likelihood-based AIC and GCV as

$$\text{AIC}_n(s) = -2l_n\left(\hat{\boldsymbol{\beta}}_s\right) + 2\tau(s),$$

$$\mathrm{GCV}_n(s) = -\frac{1}{n}\frac{l_n\left(\hat{\boldsymbol{\beta}}_s\right)}{(1-\tau(s)/n)^2},$$

which are similarly implemented though the PPL-based procedure. Moreover, for each setup, we repeat the selection procedure with all survey weights ignored (being set as unity). The unweighted selection results are corresponding to pure model-based inferences as discussed in Section 2. In particular, the pseudo-likelihood-based BIC reduces to the classic BIC (3.1) used for non-survey situations.

In Tables 5.1-5.2, we summarize the simulation results based on 1,000 repetitions in terms of the positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR), and averaged model size (AMS). Specifically, let s_0 be the true model that generates the finite population and s'_j be the selected model based on the j^{th} sample, $j = 1,\ldots,1{,}000$. The PSR, FDR, CSR and AMS are estimated as

$$\mathrm{PSR} = \frac{\sum_{j=1}^{1{,}000}\tau\left(s_0 \cap s'_j\right)}{1{,}000\,\tau\left(s_0\right)}, \mathrm{FDR} = \frac{\sum_{j=1}^{1{,}000}\tau\left(s'_j/s_0\right)}{1{,}000\,\tau\left(s'_j\right)},$$

$$\mathrm{CSR} = \frac{\sum_{j=1}^{1{,}000} I\left(s'_j = s_0\right)}{1{,}000}, \mathrm{AMS} = \frac{\sum_{j=1}^{1{,}000}\tau\left(s'_j\right)}{1{,}000},$$

where $\tau(s)$ denotes the size of model s and $I(\cdot)$ is the indicator function. In addition, we assess the predictive accuracy of the selected model as follows. For each setup, a test sample of size 200 is generated by SRSWOR from the same finite population as that for the training sample. For Models 1-2, we use the averaged residual sum of squares (RSS) on the test data as a measurement of the predictive ability of the selected model. For Models 3-4, we compute both positive and negative prediction rates. To be specific, let π^* be a specified benchmark and $\hat{\pi}_i$ be the estimated success probability of the i^{th} test sample, $i = 1,\ldots,200$. We then predict the i^{th} response y_i by $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi^*$ and $\hat{y}_i = 0$ otherwise. The correct prediction rates are estimated by

$$\mathrm{PPR} = \frac{\sum_{i\in\{i:y_i=1\}}^{200} I\left(\hat{y}_i = 1\right)}{\sum_{i=1}^{200} I\left(y_i = 1\right)}, \mathrm{NPR} = \frac{\sum_{i\in\{i:y_i=0\}}^{200} I\left(\hat{y}_i = 0\right)}{\sum_{i=1}^{200} I\left(y_i = 0\right)}.$$

The final PPR and NPR are averaged based on 1,000 replications. Note that PPR and NPR here are similar to sensitivity and specificity in the clinical studies, which indicate the ability of a 0-1 prediction approach in terms of correct positive and negative predictions. In general, a larger π^* leads to high NPR but low PPR. The value of π^* should be cautiously specified in applications. In our simulation studies, we fix $\pi^* = 0.5$ for simplicity.

The results are encouraging for the proposed BIC method. From Tables 5.1-5.2, we observe that models selected by AIC have both high PSR and FDR, which indicates an excessive inclusion of the irrelevant variables. In comparison, the BIC significantly reduces the FDR of selected models with a slight sacrifice on PSR, and selects the model with sizes closer to the truth. Although the GCV behaves similarly

to BIC in the linear model settings, it concurs with AIC for the logistic models where less information is provided from the binary responses.

In the first sampling plan, the inclusion probabilities are related to Y only through a single covariate in the model (*i.e.*, X_6 or X_7). The sample correlation structure between the response and covariates is largely maintained from the finite population. Consequently, no substantial difference is observed between the weighted and unweighted selection procedures from Table 5.1.

The insights of using sampling weights in variable selection are tentatively revealed from the second sampling plan, where the sample correlation structure is systemically distorted. Clearly, the spurious correlation between covariates in the sampled units deteriorates the efficiency of selection methods. This is reflected from the depressed PSRs and the inflated FDRs from the unweighted procedures. Incorporating sampling weights in the selecting process is helpful to correct the biased result. In particular, noticeable improvements have been observed for the BIC-based selection. In the most impressive case (*i.e.*, Model 3 of Table 5.2), the pseudo-likelihood-based BIC substantially improves the classic BIC by increasing the PSR from 0.65 up to 0.89, while reduces the corresponding FDR from 0.62 down to 0.50. Our observation echoes the rationale of weighting as the removal of bias due to the informative sampling (Section 6.3, Fuller 2009).

Table 5.1
Selection for the design not generating strong spurious correlations (1st plan). Results are summarized in terms of positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR) and averaged model size (AMS); Prediction assessments for Models 1-2 are based on the testing residual sum of squares (RSS), while for Models 3-4 they are based on positive/negative prediction rate (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
			Model 1			
Ignored	GCV	0.96	0.19	0.28	4.9	1.04
	AIC	0.99	0.48	0.05	8.7	1.08
	BIC	0.96	0.19	0.28	4.9	1.04
Included	GCV	0.95	0.24	0.19	5.2	1.05
	AIC	0.99	0.61	0.01	11.4	1.11
	BIC	0.95	0.24	0.20	5.3	1.05
			Model 2			
Ignored	GCV	0.72	0.19	0.02	5.5	1.07
	AIC	0.89	0.44	0.01	10.3	1.09
	BIC	0.73	0.19	0.03	5.6	1.07
Included	GCV	0.74	0.24	0.02	6.1	1.08
	AIC	0.89	0.54	0.01	12.5	1.12
	BIC	0.74	0.24	0.03	6.1	1.08
			Model 3			
Ignored	GCV	0.99	0.59	0.00	7.8	(0.71, 0.45)
	AIC	0.99	0.62	0.00	8.4	(0.69, 0.49)
	BIC	0.96	0.43	0.00	5.1	(0.72, 0.44)
Included	GCV	0.99	0.67	0.00	9.9	(0.71, 0.47)
	AIC	0.99	0.70	0.00	10.7	(0.68, 0.48)
	BIC	0.94	0.45	0.00	5.3	(0.71, 0.45)
			Model 4			
Ignored	GCV	0.97	0.44	0.01	9.4	(0.66, 0.55)
	AIC	0.98	0.47	0.01	9.8	(0.65, 0.56)
	BIC	0.87	0.26	0.07	6.0	(0.69, 0.53)
Included	GCV	0.98	0.54	0.01	11.4	(0.66, 0.54)
	AIC	0.98	0.56	0.00	11.9	(0.66, 0.55)
	BIC	0.86	0.30	0.05	6.2	(0.68, 0.53)

Table 5.2
Selection for the design generating strong spurious correlations (2nd plan). Results are summarized in terms of positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR) and averaged model size (AMS); Prediction assessments for Models 1-2 are based on the testing residual sum of squares (RSS), while for Models 3-4 they are based on positive/negative prediction rate (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
			Model 1			
Ignored	GCV	0.83	0.23	0.17	4.6	1.09
	AIC	0.97	0.49	0.04	8.6	1.10
	BIC	0.83	0.23	0.17	4.6	1.09
Included	GCV	0.95	0.31	0.13	5.9	1.07
	AIC	0.99	0.65	0.00	12.5	1.12
	BIC	0.95	0.30	0.14	5.9	1.07
			Model 2			
Ignored	GCV	0.62	0.22	0.02	5.0	1.13
	AIC	0.88	0.45	0.01	10.3	1.14
	BIC	0.62	0.22	0.02	5.1	1.12
Included	GCV	0.72	0.28	0.01	6.5	1.10
	AIC	0.89	0.59	0.00	13.7	1.12
	BIC	0.72	0.27	0.01	6.5	1.10
			Model 3			
Ignored	GCV	0.87	0.62	0.00	7.3	(0.66, 0.44)
	AIC	0.88	0.63	0.00	7.6	(0.65, 0.45)
	BIC	0.65	0.62	0.00	4.5	(0.68, 0.42)
Included	GCV	0.97	0.74	0.00	11.9	(0.70, 0.46)
	AIC	0.97	0.75	0.00	12.4	(0.68, 0.46)
	BIC	0.89	0.50	0.00	5.6	(0.70, 0.44)
			Model 4			
Ignored	GCV	0.94	0.48	0.00	9.5	(0.62, 0.51)
	AIC	0.95	0.50	0.00	10.0	(0.62, 0.52)
	BIC	0.72	0.41	0.00	6.1	(0.64, 0.49)
Included	GCV	0.93	0.61	0.00	12.5	(0.64, 0.53)
	AIC	0.94	0.62	0.00	12.9	(0.64, 0.53)
	BIC	0.82	0.34	0.01	6.4	(0.67, 0.54)

5.3 Scenario 2: Model mis-specification

A well-known rationale for using sampling weights is that it provides protection against model mis-specification (Pfeffermann and Holmes 1985; Kott 1991): the inferences based on weighted estimates may remain valid for the surveyed population, even when the model fails. To gain further insights of weighting in variable selection, we further compare the proposed BIC with the classic unweighted methods in the simulation where the presumed model is misspecified from the model that generates the data. In such situations, a postulated "true" model does not exist, and the goal of variable selection is to find an optimal model that well describes the finite population. We still make use of the stratified pseudo-finite population in Section 5.2, but generate the response variable Y according to the strata. Specifically, the values of Y for units in strata (Male, 55+) and (Female, 55+) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + 0.6X_{38} + \varepsilon,$$

while the values Y for units in the strata (Male, 55-) and (Female, 55-) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + \varepsilon$$

with $\varepsilon \sim N(0,1)$ denoting a random error. In other words, we assume that variable X_{38} is influential only for people aged 55 and older, but not for people younger than 55. In addition, we further violate the presumed Model 1 by excluding X_6 from the set of candidate covariates, which mimics the situation where one important design feature is omitted in the modeling. A stratified SRSWOR of size 500 or 1,000 is drawn using the first sampling plan in Section 5.2. The weighted and unweighted procedures are then tested for the variable selection based on the sampled units.

We summarize the simulation results in Table 5.3 by estimating the selection rates of X_{18}, X_{20}, and X_{38} based on 1,000 replications. Similar to the previous simulations, the averaged model size (AMS) and the testing RSS of selected models (*i.e.*, the averaged RSS based on testing data of size 200) are also included in the summary. From Table 5.3, we see that when the model assumption is violated, the pseudo-likelihood-based BIC still achieves relatively high prediction accuracy by suggesting relevant variables with high probability. In contrast, ignoring the survey weights leads to nearly 9% relative loss on the testing RSS because of the exclusion of X_{38}. Apparently, increasing the sample size helps to improve the goodness of fit for the misspecified models, yet the improvement is at a cost by including more variables.

Table 5.3
Selection frequency of influential variables in model mis-specified case; The averaged model size (AMS) and the testing residual sum of squares (RSS) are also reported.

Weights	Criterion	X_{18}	X_{20}	X_{38}	AMS	Testing RSS
			$n = 500$			
Ignored	GCV	0.78	0.95	0.56	5.9	1.93
	AIC	0.95	0.99	0.73	12.5	1.95
	BIC	0.83	0.97	0.60	6.6	1.93
Included	GCV	0.73	0.92	0.84	6.3	1.77
	AIC	0.91	0.99	0.85	12.5	1.79
	BIC	0.78	0.94	0.83	6.9	1.77
			$n = 1{,}000$			
Ignored	GCV	0.96	1.00	0.79	7.6	1.87
	AIC	0.99	1.00	0.87	13.1	1.88
	BIC	0.97	1.00	0.80	7.9	1.87
Included	GCV	0.93	1.00	0.94	7.6	1.71
	AIC	0.98	1.00	0.96	13.0	1.72
	BIC	0.94	1.00	0.94	7.7	1.71

5.4 Analysis of SLCDC data

To illustrate the application of proposed BIC, we use it to identify health behaviors that affect the control of blood pressure using SLCDC 2009. The response variable is BMHX_02 from the working data obtained from SLCDC, which has 2 levels indicating whether or not the blood pressure of the respondent is under control, based on the latest measurement by a health professional. We treat the remaining 39 variables in the working data as candidate covariates, and our goal is to identify the influential covariates that are associated with blood-pressure control. We build a logistic regression of BMHX_02 on the candidate covariates and use PPL-BIC with SCAD penalty to select the influential ones (weights are scaled by $k = 10^{-3}$). As a preliminary step, each covariate is standardized such that the corresponding first and second weighted sample moments are zero and unity respectively. For comparison, the AIC and GCV are also used in the analysis.

In Figure 5.1, we plot the scores of criterion with respect to the degree of model sparsity. We see that the BIC selects a model with 11 covariates, while the GCV and AIC pick the same model with 24 covariates. When survey weights are ignored in the selection procedure, models with 7 or 21 covariates are suggested based on the standard BIC or GCV/AIC. The distinction between the weighted and unweighted selection results reflects the potential distortion in the correlation structure of the sampled units. Such a distinction may also be explained by model mis-specification for part of the SLCDC population (Lohr and Liu 1994). Given the potential bias for unweighted methods, the weighted selection results are more plausible in the analysis.

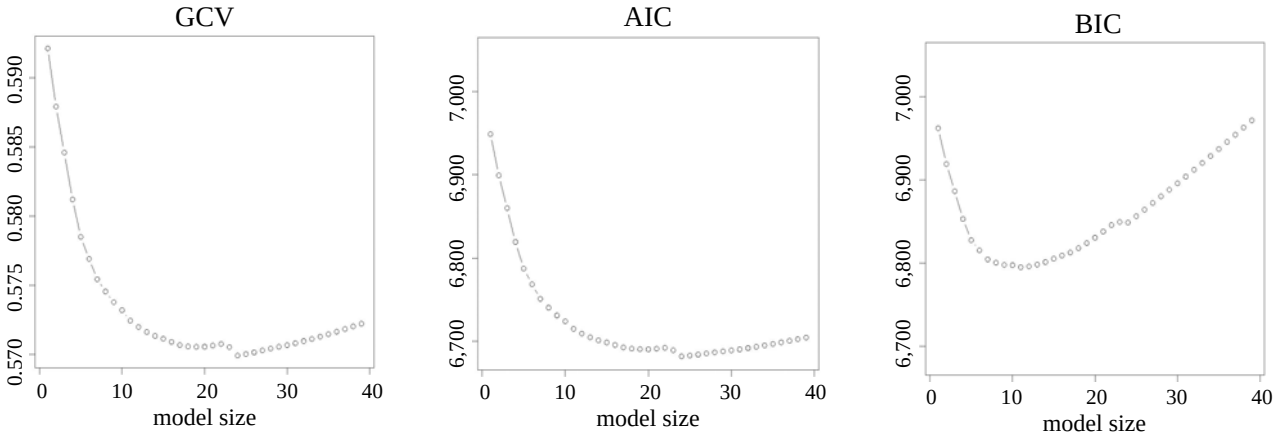


Figure 5.1 Selection criteria values based on candidate models

We further assess the selected models in terms of predictive accuracy as follows. First, we draw 500 independent sets of 5,868 bootstrap samples (with replacement) from the working data of SLCDC. For the t^{th} bootstrap sample d_t, $t = 1, \ldots, 500$, the survey weight w_i for the i^{th} unit is adjusted by $\tilde{w}_{ti} = v_{ti} w_i$ with v_{ti} denoting the number of times that the i^{th} unit is selected in d_t. We then fit the selected models to each bootstrap sample (with weights accounted accordingly), and evaluate their weighted positive and negative prediction rates (WPPR, WNPR) by

$$\text{WPPR} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 1, y_i = 1\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 1\right)}, \text{WNPR} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 0, y_i = 0\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 0\right)},$$

where y_i and $\hat{y}_i$ denote the i^{th} response in BMHX_02 and its predicted value. We summarize the averaged WPPR and WNPR based on 500 bootstrap samples in Table 5.4 according to three different benchmark values (*i.e.*, 0.25, 0.35, 0.45).

From Table 5.4, we observe that the models selected from unweighted analysis have lower WPPR in general, which provides additional support for using survey weights in the selection procedure. Compared with GCV/AIC, the BIC selects the model with a slightly conservative WPPR but a higher WNPR. Nevertheless, the difference is not significant. Noticeably, the size of BIC-selected model is much less

than the GCV/AIC selected one, which provides an easier interpretation between the response BMHX_02 and the covariates.

Table 5.4
Prediction accuracy for selected models: (WPPR, WNPR) based on different benchmarks.

Weights	Criteria	≥0.25	≥ 0.35	≥ 0.45
Ignored	AIC/GCV	(0.646, 0.525)	(0.460, 0.688)	(0.299, 0.811)
	BIC	(0.649, 0.513)	(0.445, 0.705)	(0.265, 0.818)
Included	AIC/GCV	(0.645, 0.523)	(0.488, 0.682)	(0.338, 0.790)
	BIC	(0.654, 0.532)	(0.485, 0.706)	(0.322, 0.830)

To assess the stability of selection, we repeat the weighted selection procedure based on the 500 bootstrap samples. In Table 5.5, we list the bootstrap selection rate for the seven most significant covariates according to their MLE in the original SLCDC working data. The corresponding coefficient estimates and standard errors are also included based on the bootstrap samples. From Table 5.5, we find that only four significant covariates (*i.e.*, DHHX_AGE, GENXDHMH, INHX_06, HWTDBMI) are consistently selected by BIC, while the GCV/AIC tends to pick more unreliable ones in the model. The BIC-based selection result suggests that the control of blood pressure is strongly associated with age, body weights, mental health and the medication information. Our observation echoes many hypertension studies in the literature (see, *e.g.*, Gelber, Gaziano, Manson, Buring and Sesso 2007; Yan, Liu, Matthews, Daviglus, Ferguson and Kiefe 2003).

Table 5.5
Bootstrap selection results for significant variables: (Estimated coefficient, Standard error, Selection rate).

Variable	GCV	AIC	BIC
GEO_ON	(0.14, 0.09, 0.86)	(0.16, 0.09, 0.92)	(0.09, 0.09, 0.58)
DHHX_AGE	(-0.29, 0.09, 1.0)	(-0.32, 0.09, 1.0)	(-0.27, 0.08, 1.0)
GENXDHMH	(-0.15, 0.05, 0.99)	(-0.15, 0.05, 0.99)	(-0.14, 0.06, 0.92)
SMHXDSLT	(0.11, 0.07, 0.76)	(0.12, 0.07, 0.84)	(0.08, 0.09, 0.47)
MOHXDBPM	(-0.08, 0.07, 0.67)	(-0.09, 0.06, 0.81)	(-0.05, 0.07, 0.35)
INHX_06	(0.18, 0.06, 0.97)	(0.18, 0.06, 0.99)	(0.18, 0.07, 0.91)
HWTDBMI	(0.14, 0.06, 0.95)	(0.14, 0.06, 0.97)	(0.13, 0.06, 0.91)
Ave. Model Size	23.1	27.8	10.3

6 Concluding remarks

In this paper, we have addressed the variable selection problem in the analysis of complex surveys. When units are selected through disproportionate sampling, the data correlation structure reflected in the sample can be distorted. Incorporating sampling weights in the selection process is protective against the biased selection results. In this spirit, we derived a survey-weighted BIC criterion based on the pseudo-likelihood and further proposed an efficient procedure (PPL) for its implementation. With some regularity

conditions, we showed that our criterion consistently identifies the influential variables under a joint randomization framework. The decent performances of proposed method was confirmed by numerical studies.

Acknowledgements

The authors are grateful to the associate editor and the two anonymous referees for their insightful comments and valuable suggestions. The authors are indebted to Professor J.N.K. Rao of Carleton University for his constructive comments to an earlier manuscript. This work was supported by Statistics Canada and MITACS.

Appendix

Table A.1
Variables for analysis of SLCDC data with non-response adjustments: A: allocate to other categories; D: delete from the data; M: impute by mean values; NA: no adjustment applied.

	Variable	Description	Levels	Missing	Adjust
1	BMHX_02	Blood pressure control status	2	1.6%	D
2	GEO_QB	Provinces grouped by region - QC	2	- -	NA
3	GEO_ON	Provinces grouped by region - ON	2	- -	NA
4	GEO_BC	Provinces grouped by region - BC	2	- -	NA
5	GEO_PR	Provinces grouped by region - PR	2	- -	NA
6	DHHX_AGE	Age	Cont.	- -	NA
7	DHHX_SEX	Sex	2	- -	NA
8	GENXDHMH	Perceived mental health	2	0.2%	A
9	CNHX_05	High blood pressure - age when diagnosed	Cont.	2.7%	D
10	MEHX_02	No. of medications taken	Cont.	0.3%	M
11	MEHX_03	No. of times per day medications taken	Cont.	0.1%	M
12	MEHXGMED	No. of medications for high blood pressure	Cont.	2.0%	M
13	MEHX_06	No. of times per day bp medication taken	Cont.	1.0%	M
14	MEHXDMCO	Medication compliance - overall	2	0.2%	A
15	HUHXDHP	Consulted family doctor about hbp	2	0.1%	A
16	SMHX_11A	Smoked at any time since being diagnosed	2	0.1%	A
17	SMHX_13A	Drank alcohol since being diagnosed	2	0.2%	A
18	SMHXDSLT	Daily salt intake	2	0.2%	A
19	SMHXDFDC	Dietary foods	2	0.1%	A
20	SMHXDPAC	Exercise/physical activity	2	0.1%	A
21	SMHXDBW	Body weight control	2	0.2%	A
22	MOHXDBPM	Self-monitoring of blood pressure	2	0.3%	A
23	MOHX_02	Correct use of bp measurement device	2	0.5%	A
24	INHX_01A	Info from family doctor	2	2.4%	A
25	INHX_01F	Info from family member/friend	2	2.4%	A
26	INHX_02A	Info from book, pamphlet, brochure	2	1.5%	A
27	INHX_02C	Info from package insert with medication	2	1.5%	A
28	INHX_02G	Info from media	2	1.5%	A
29	INHX_02H	Info from internet	2	1.5%	A
30	INHX_04	Info received - emotional impact of hbp	2	0.8%	A
31	INHX_06	Info received - correct use of medication	2	0.6%	A
32	INHX_07	Info received - additional information	2	0.9%	A
33	CPGFGAM	Gambling activity	2	0.5%	A
34	DHHDECF	Household type	2	0.2%	A
35	EDUDH04	Highest level of education in household	2	3.4%	A
36	FVCGTOT	Daily consumption - fruits and vegetables	2	5.2%	A
37	GEODUR2	Urban and rural areas	2	- -	NA
38	HWTDBMI	Body mass index (BMI) self-report	Cont.	2.1%	M
39	INCDRPR	Household income - provincial level	10	9.6%	A
40	SACDTOT	Total number hours - sedentary activities	Cont.	1.5%	M

Table A.2
Influential and design variables in simulation settings: ∗ - influential variable to the response; • - design variable affecting sampling probabilities in the 1st plan; ◊ - design variable affecting sampling probabilities in the 2nd plan.

	Variable	Model 1	Model 2	Model 3	Model 4
6	DHHX_AGE	∗ • ◊	∗ • ◊	• ◊	• ◊
7	DHHX_SEX	• ◊	• ◊	∗ • ◊	∗ • ◊
8	GENXDHMH			∗ ◊	∗ ◊
10	MEHX_02	∗	∗		
18	SMHXDSLT	∗ ◊	∗ ◊		
22	MOHXDBPM	∗ ◊	∗ ◊		
26	INHX_02A			∗ ◊	∗ ◊
28	INHX_02G				∗
30	INHX_04		∗		
34	DHHDECF		∗		
36	FVCGTOT				∗

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (Eds., B.N. Petrox and F. Caski), 267-281.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D., and Roberts, G. (2003). *Analysis of Survey Data*, Chapter: Design-based and model-based methods for estimating model parameters. Wiley Series in Survey Methodology, Chichester.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377-403.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Frank, I.E., and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley, Hoboken.

Gelber, R.P., Gaziano, J.M., Manson, J.E., Buring, J.E. and Sesso, H.D. (2007). A prospective study of body mass index and the risk of developing hypertension in men. *American Journal of Hypertension*, 20, 370-377.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175-188.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc.

Kott, P.S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45, 107-112.

Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21, 1225-1248.

Lohr, S.L., and Liu, J. (1994). A comparison of weighted and unweighted analyses in the NCVS. *Journal of Quantitative Criminology*, 10, 343-360.

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, 15, 661-675.

Molina, E.A., and Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis*, 13, 395-405.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Pfeffermann, D., and Holmes, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, Series A, 148, 268-278.

Rahiala, M., and Teräsvirta, T. (1993). Business survey data in forecasting the output of Swedish and Finnish metal and engineering industries: A Kalman filter approach. *Journal of Forecasting*, 12, 255-271.

Royall, M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

She, Y. (2011). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, in press.

Skinner, C. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, manuscript.

Statistics Canada (2009). Survey on living with chronic diseases in Canada 2009: User guide. Supplementary documentation.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 111-147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267-288.

Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wolfson, W.G. (2004). Analysis of labour force survey data for the information technology occupations 2000-2003. *Report for the Software Human Resource Council*, WGW Services Ltd., Ottawa, Ontario.

Xie, B., Pan, W. and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64, 921-930.

Xu, C., and Chen, J. (2012). Technical supplement to "Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data". Available from the first author.

Yan, L.L., Liu, K., Matthews, K.A., Daviglus, M., Ferguson, T.F. and Kiefe, C.I. (2003). Psychosocial factors and risk of hypertension: The coronary artery risk development in young adults (CARDIA) study. *The Journal of the American Medical Association*, 290, 2138-2148.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 323-349
Statistics Canada, Catalogue No. 12-001-X

Design-based analysis of factorial designs embedded in probability samples

Jan A. van den Brakel[1]

Abstract

At national statistical institutes experiments embedded in ongoing sample surveys are frequently conducted, for example to test the effect of modifications in the survey process on the main parameter estimates of the survey, to quantify the effect of alternative survey implementations on these estimates, or to obtain insight into the various sources of non-sampling errors. A design-based analysis procedure for factorial completely randomized designs and factorial randomized block designs embedded in probability samples is proposed in this paper. Design-based Wald statistics are developed to test whether estimated population parameters, like means, totals and ratios of two population totals, that are observed under the different treatment combinations of the experiment are significantly different. The methods are illustrated with a real life application of an experiment embedded in the Dutch Labor Force Survey.

Key Words: Completely randomized designs; Design-based inference; Embedded experiments; Measurement error models; Model-assisted inference; Randomized block designs.

1 Introduction

The fields of randomized experiments and probability sampling are traditionally two separated domains of applied statistics. Both, however, come together if experiments are embedded in ongoing sample surveys. Randomized experiments embedded in ongoing sample surveys are frequently conducted to compare and test the effect of alternative survey implementations on the outcomes of a sample survey. The purpose of such empirical research is to improve the quality and efficiency of the underlying survey processes or to obtain more quantitative insight into the various sources of non-sampling errors. Many experiments conducted in this context are small scaled or conducted with specific groups. The value of empirical research into survey methods is strengthened as conclusions can be generalized to populations larger than the sample that is included in the experiment. Selecting experimental units randomly from a larger target population, is an important tool to secure that results of an experiment can be generalized to populations larger than the group of people included in the experiment, as emphasized by Fienberg and Tanur (1987, 1988, 1989 and 1996). This naturally leads to randomized experiments embedded in ongoing sample surveys. In the survey literature, such experiments are also referred to as split-ballot designs or interpenetrating subsampling, and date back to Mahalanobis (1946).

At national statistical offices such experiments are particularly useful to quantify discontinuities in the series of repeated surveys due to adjustments to the survey process. Repeatedly conducted surveys make up series that describe the development of target parameters. Embedded experiments can be used to avoid one or more modifications in the survey process resulting in unexplained differences in the series of a survey.

1. Jan A. van den Brakel, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. E-mail: jbrl@cbs.nl.

An important issue in the analysis of this kind of experiment is to find the right mode of inference. The statistical inference in survey sampling is traditionally design based or model assisted. This implies that the inference is predominantly based on the stochastic structure induced by the sampling design. A well-known design-based estimator is the Horvitz-Thompson (HT) estimator, developed by Narain (1951) and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement. Under the model assisted approach developed by Särndal, Swensson and Wretman (1992), the accuracy of the HT estimator is improved by taking advantage of available auxiliary information about the complete target population, resulting in the generalized regression (GREG) estimator. Many national statistical institutes rely on this design-based and model-assisted approach to compile official statistics.

The statistical inference that is traditionally employed in the theory of design and analysis of randomized experiments is predominantly model-based. The observations that are obtained in the experiment are assumed to be the realization of a linear model. To test hypotheses about treatment effects, F-tests are derived under the assumption of normally and independently distributed observations. An exception is Kempthorne (1955), where a randomization approach is proposed in a way that is similar to the design-based inference approach in sampling theory. The F-test is used as an approximation of the randomization test. The model-based inference for randomized experiments is not necessarily appropriate for the analysis of embedded experiments, particularly if a design-based or model-assisted inference is used in the ongoing survey to compile official statistics.

In an embedded experiment the probability sample of the ongoing survey is randomly divided into different subsamples according to an experimental design. Each subsample can be considered as a probability sample drawn from the finite target population and can be used to estimate parameters such as means, totals and ratios, that are observed under the different survey implementations or treatments of the experiment using the estimation procedure that is applied in the regular survey to compile official statistics. The purpose of such embedded experiments is to compare the effect of alternative survey implementations on the main parameter estimates of the ongoing survey and to test whether the observed differences between these parameter estimates are statistically significant. This is obtained with a design-based approach where point and variance estimates for the population parameters, are (approximately) design-unbiased with respect to the sample design used to draw an initial probability sample from the target population, and the experimental design used to randomize this sample over the different subsamples. This analysis must also reflect the specific details of the regular estimation approach used to compile official statistics, as far as this is possible with the available sample size under the different treatments.

Previous research has proposed such a design-based theory for the analysis of single-factor experiments that are designed as completely randomized designs (CRDs) or randomized block designs (RBDs) to test the effect of one factor on $K \geq 2$ levels (van den Brakel (2008); van den Brakel and Renssen (1998, 2005); van den Brakel and van Berkel (2002)). In their approach the GREG estimator is applied to derive design-based Wald- and t-statistics to test whether the differences between finite population parameter estimates observed under the different survey implementations are significantly different. This theory is further extended to the experiments embedded in rotating panel designs by Chipperfield and Bell (2010).

From standard experimental design theory it is well known that it is efficient to test different treatment factors simultaneously in one factorial design instead of conducting separate single-factor experiments

(Hinkelmann and Kempthorne (1994); Montgomery (2001)). It can be expected that different design parameters in a survey process interact with each other, *e.g.*, when different questionnaire designs and data collection modes are compared empirically. Factorial setups are indeed appropriate if more than one factor in the survey is adjusted and tested in an embedded experiment, since fewer experimental units are required to test the main effects of the treatment factors whereas interactions between the factors can be analyzed. Another advantage of testing different treatments simultaneously in a factorial design is that the validity of the observed results is extended, since the effects are observed over a wider range of conditions (Hinkelmann and Kempthorne (1994)). Therefore the design-based theory for the analysis of embedded experiments is extended to factorial designs in this paper.

The theory for factorial designs where the effect of two factors is tested simultaneously is developed in section 2. Subsequently the methodology is extended to higher order factorial designs in section 3. In section 4, the methodology is extended to test hypotheses about ratios of population totals and designs where clusters of sampling units are randomized over the treatment combinations. In section 5 these methods are applied to a factorial experiment with advance letters in the Dutch Labor Force Survey (LFS). The paper concludes with a discussion in section 6.

2 Analysis of embedded $K \times L$ factorial experiments

2.1 Experimental designs embedded in probability samples

In a $K \times L$ factorial design, the effects of two factors are tested simultaneously. The first factor, denoted A contains $K \geq 2$ levels. The second factor, denoted B contains $L \geq 2$ levels. The purpose of the experiment is to test the main effects of the two factors and the interactions between both factors on the main parameter estimates of the ongoing survey. To this end a probability sample s of size n is drawn from a finite target population U of size N according the sample design of the regular survey. This sample design can be generally complex, and is described by its first order inclusion probabilities π_i for unit i and second order inclusion probabilities $\pi_{ii'}$ for units i and i'.

Subsequently, this sample is randomly divided into KL subsamples according to a randomized experiment. In the case of a CRD, the sample s of size n is randomly divided into KL subsamples s_{kl}, each with a size of n_{kl} sampling units. The sampling units of each subsample are assigned to one of the KL treatment combinations. Under a CRD, $n_{++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$ denotes the total number of sampling units in the sample s. The probability that sampling unit i is assigned to subsample s_{kl}, conditionally on the realization of s, equals n_{kl} / n_{++}. The unconditional probability that sampling unit i is selected in subsample s_{kl} equals $\pi_i^* = \pi_i (n_{kl} / n_{++})$.

The power of an experiment might be improved by using sampling structures such as strata, clusters or interviewers as block variables in an RBD since restricted randomization removes the variance between the blocks from the analysis of the experiment (Fienberg and Tanur (1987, 1988)). In the case of an RBD, the sampling units are deterministically grouped in B more or less homogeneous blocks s_b. Within each block, the sampling units are randomly assigned to one of the KL treatment combinations. Let n_{bkl} denote the number of sampling units in block b assigned to treatment combination kl, and

$n_{b++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{bkl}$ the number of sampling units in block b. The probability that sampling unit i is assigned to subsample s_{kl}, conditionally on the realization of s and $i \in s_b$, equals n_{bkl} / n_{b++}, $i \in s_b$. The unconditional probability that sampling unit i is selected in subsample s_{kl} equals $\pi_i^* = \pi_i (n_{bkl} / n_{b++})$.

In many practical applications one of the KL subsamples is assigned to the regular survey and serves, besides being used to produce estimates for the regular publication, as the control group in the experiment. In such situations, the size of this subsample will be substantially larger than the other subsamples.

There are a lot of issues in the planning and design stage of embedded experiments. The field staff, for example, requires special attention, since an embedded experiment can have a large impact on their daily routine of data collection, to which they are accustomed. See van den Brakel and Renssen (1998) and van den Brakel (2008) for more details about such design issues.

Although factorial designs are efficient from a statistical point of view, there might be strong practical arguments against a factorial set-up. The number of treatment combinations increases rapidly with the number of factors in full factorial designs, which might be difficult to implement in the data collection of a survey process. A general solution, known from standard experimental design theory, is to confound higher order interactions with blocks or to apply fractional factorial designs (Hinkelmann and Kempthorne (2005); Montgomery (2001)). These balanced designs, however, are generally hard to combine with the fieldwork restrictions encountered in the daily practice of survey sampling. In many applications the factors that changed in a survey redesign are therefore combined into one treatment. The total effect of these modifications is tested against the standard alternative in a two-treatment experiment. This implies that the effects of all factors in the experiment are confounded and cannot be separately estimated.

2.2 Testing hypotheses about finite population parameters

The purpose of embedded experiments is to test whether alternative survey implementations result in significantly different estimates for finite population parameters. Such differences are the result of non-sampling errors, like measurement errors and response bias. A measurement error model is required to link systematic differences between finite population parameters due to different survey implementations or treatments. Therefore the measurement error model for single-factor experiments proposed by van den Brakel and Renssen (2005) and van den Brakel (2008) is extended to factorial designs.

Let y_{iqkl} denote the observation obtained from the i^{th} individual observed under the kl^{th} treatment combination and the q^{th} interviewer. It is assumed that the observations are a realization of the measurement error model

$$y_{iqkl} = u_i + \beta_{kl} + \gamma_q + \varepsilon_{ikl}. \tag{2.1}$$

Here u_i is the true intrinsic value of the i^{th} individual, β_{kl} the effect of the kl^{th} treatment combination and ε_{ikl} an error component. The model also allows for interviewer effects, *i.e.*, $\gamma_q = \psi + \xi_q$, where ψ denotes a systematic interviewer bias and ξ_q the random effect of the q^{th} interviewer, respectively. Let E_m and cov_m denote the expectation and the covariance with respect to the measurement error model. It

is assumed that $\mathrm{E}_m(\varepsilon_{ikl}) = 0$, $\mathrm{var}_m(\varepsilon_{ikl}) = \sigma^2_{ikl}$, and that measurement errors between sampling units are independent. Furthermore it is assumed that $\mathrm{E}_m(\xi_q) = 0$, $\mathrm{var}_m(\xi_q) = \tau^2_q$ and that random interviewer effects between interviewers are independent. As a result the model allows for correlated response between sampling units that are interviewed by the same interviewer. The measurement error model allows for separate variances for measurement errors under different treatment combinations and separate variances for interviewers.

The treatment effects β_{kl} can be interpreted as the bias in the estimated population parameter if the true intrinsic population value of u is measured by means of the kl^{th} survey implementation. The treatment effect can be decomposed in the traditional way of an analysis of variance for a two-way layout:

$$\beta_{kl} = u + A_k + B_l + AB_{kl}, \tag{2.2}$$

with u the overall effect, A_k and B_l the main effects of treatment factors A and B and AB_{kl} the interactions between treatment factors A and B. If the treatment effects are defined as fixed deviations from the individuals' intrinsic value u_i, then the overall mean u equals zero. In that case A_k corresponds with the bias associated with the k^{th} level of factor A averaged over all levels of factor B, B_l the bias associated with the l^{th} level of factor B, averaged over all levels of factor A, and AB_{kl} the additional bias associated with the combination of the k^{th} level of factor A and the l^{th} level of factor B on top of A_k and B_l.

The following restrictions are required to identify model (2.2):

$$\sum_{k=1}^{K} A_k = 0, \ \sum_{l=1}^{L} B_l = 0, \tag{2.3}$$

and

$$\sum_{k=1}^{K} AB_{kl} = 0, l = 1, 2, \ldots, L, \ \sum_{l=1}^{L} AB_{kl} = 0, k = 1, 2, \ldots, K. \tag{2.4}$$

For each sampling unit, a potential response variable is defined under each of the KL treatment combinations. Therefore the measurement error model can be expressed in matrix notation as:

$$\mathbf{y}_{iq} = \mathbf{j}_{KL}\, u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \gamma_q + \boldsymbol{\varepsilon}_i, \tag{2.5}$$

where $\mathbf{y}_{iq} = (y_{iq11}, ..., y_{iqkl}, ..., y_{iqKL})^t$, $\boldsymbol{\beta} = (\beta_{11}, ..., \beta_{kl}, ..., \beta_{KL})^t$, $\mathbf{j}_{KL}$ a vector of order KL with each element equal to one and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, ..., \varepsilon_{ikl}, ..., \varepsilon_{iKL})^t$. The sampling units are assigned to one of the treatment combinations only, so only one of the responses of $\mathbf{y}_{iq}$ is actually observed. The model assumptions specified above are stated as:

$$\mathrm{E}_m\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{0}, \tag{2.6}$$

$$\mathrm{cov}_m\left(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}\right) = \begin{cases} \boldsymbol{\Sigma}_i & : \ i = i' \\ \mathbf{O} & : \ i \neq i' \end{cases}, \tag{2.7}$$

$$\mathrm{E}_m\left(\xi_q\right) = 0, \tag{2.8}$$

$$\mathrm{cov}_m\left(\xi_q, \xi_{q'}\right) = \begin{cases} \tau_q^2 & : \quad q = q' \\ 0 & : \quad q \neq q' \end{cases}, \tag{2.9}$$

$$\mathrm{cov}_m\left(\varepsilon_{ikl}, \xi_q\right) = 0, \tag{2.10}$$

where $\mathbf{0}$ is a vector of order KL with each element zero, $\boldsymbol{\Sigma}_i$ a matrix of order $KL \times KL$ containing the variances of the measurement errors σ_{ikl}^2, and $\mathbf{O}$ a matrix of order $KL \times KL$ with each element zero.

Let $\bar{\mathbf{Y}} = (\bar{Y}_{11}, ..., \bar{Y}_{1L}, ..., \bar{Y}_{kl}, ..., \bar{Y}_{K1}, ..., \bar{Y}_{KL})^t$ denote the KL dimensional vector of population means of $\mathbf{y}_{iq}$ defined by (2.5). These are the values obtained under a complete enumeration of the finite population under each of the treatment combinations and are defined as:

$$\bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \psi + \mathbf{j}_{KL} \sum_{q=1}^{Q} \frac{N_q}{N} \xi_q + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varepsilon}_i\,, \tag{2.11}$$

where Q denotes the total number of interviewers available for the data collection and N_q the number of units assigned to the q^{th} interviewer in the case of a complete enumeration.

Only systematic differences between the population parameters that are reflected by the treatment effects $\boldsymbol{\beta}$ should lead to a rejection of the null hypotheses of no treatment effects. This is accomplished by formulating hypotheses about $\bar{\mathbf{Y}}$ in expectation over the measurement error model, *i.e.*,

$$\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \psi\,. \tag{2.12}$$

Consequently, hypotheses about main effects and interactions are formulated as.

$$\begin{aligned} H_0&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{2.13}$$

where $\mathbf{C}$ denotes an appropriate contrast matrix, and $\mathbf{0}$ a vector with elements equal to one and a dimension that is equal to the number of contrasts (rows) defined by $\mathbf{C}$. The contrast matrix for the hypothesis about the main effects of factor A is defined as

$$\mathbf{C}_{\mathbf{A}} = \frac{1}{L}\left(\mathbf{j}_{(K-1)} \mid -\mathbf{I}_{(K-1)}\right) \otimes \mathbf{j}_L^t \equiv \frac{1}{L} \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}_L^t, \tag{2.14}$$

with $\mathbf{I}_{(K-1)}$ the identity matrix of order $K-1$. Matrix $\tilde{\mathbf{C}}_{\mathbf{A}}$ defines the $K-1$ contrasts between the K levels of factor A, averaged over the L levels of factor B. From (2.12) and due to restrictions (2.3) and (2.4) it follows that the contrasts between the population parameters exactly correspond to the contrasts between the main effects of the first factor:

$$\tilde{\mathbf{C}}_{\mathbf{A}} \mathrm{E}_m \bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{A}} \boldsymbol{\beta} = (A_1 - A_2, ..., A_1 - A_K)^t\,.$$

The contrast matrix for the hypothesis about the main effects of factor B is defined as

$$\mathbf{C}_{\mathbf{B}} = \frac{1}{K}\mathbf{j}_K^t \otimes \left(\mathbf{j}_{(L-1)} \mid -\mathbf{I}_{(L-1)}\right) \equiv \frac{1}{K}\mathbf{j}_K^t \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.15}$$

This matrix defines the $L-1$ contrasts between the L levels of factor B, averaged over the K levels of factor A. From (2.12) and due to restrictions (2.3) and (2.4) it follows that the contrasts between the population parameters exactly correspond to the contrasts between the main effects of the second factor:

$$\tilde{\mathbf{C}}_{\mathbf{B}}\mathrm{E}_m\bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{B}}\boldsymbol{\beta} = (B_1 - B_2, ..., B_1 - B_L)^t.$$

The contrast matrices for the main effects use the first level of factors A and B as the reference category. This implies that treatment combination $A_1 \times B_1$ is considered as the control group in the experiment.

Interactions between the two treatment factors are defined as the $L-1$ contrasts of factor B between the $K-1$ contrasts of factor A or, equivalently, as the $K-1$ contrasts of factor A between the $L-1$ contrasts of factor B, Hinkelmann and Kempthorne (1994, chapter 11). Therefore the contrast matrix for the hypothesis about the interactions between factor A and B can be defined as

$$\mathbf{C}_{\mathbf{AB}} = \left(\mathbf{j}_{(K-1)}\Big| - \mathbf{I}_{(K-1)}\right) \otimes \left(\mathbf{j}_{(L-1)}\Big| - \mathbf{I}_{(L-1)}\right) = \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.16}$$

This matrix contains the $(K-1)(L-1)$ contrasts that define the interactions between factor A and B. The contrasts between the population parameters exactly correspond to the interactions between the first and the second factor, since

$$\begin{aligned}\tilde{\mathbf{C}}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{AB}}\boldsymbol{\beta} = (&AB_{11} - AB_{12} - AB_{21} + AB_{22}, ...,\\ &AB_{11} - AB_{1L} - AB_{21} + AB_{2L}, ...,\\ &AB_{11} - AB_{12} - AB_{K1} + AB_{K2}, ...,\\ &AB_{11} - AB_{1L} - AB_{K1} + AB_{KL})^t.\end{aligned}$$

Each element of this $(K-1)(L-1)$ vector defines one of the $(K-1)(L-1)$ interactions, which neatly corresponds to the contrasts between the interaction effects defined by (2.2). The first element *e.g.*, can be interpreted as the deviation of the treatment effect of the particular combination of factor A at level 2 and factor B at level 2 from the two main effects of these factors.

2.3 Wald test

The hypotheses specified in section 2.2, can be tested with a Wald test (Wald 1943), which is frequently applied in design-based testing procedures, see for example Skinner, Holt and Smith (1989) or Chambers and Skinner (2003). If $\hat{\bar{\mathbf{Y}}}$ denotes a design-unbiased estimator for $\bar{\mathbf{Y}}$, $\mathbf{C}$ the contrast matrix $\mathbf{C}_{\mathbf{A}}$, $\mathbf{C}_{\mathbf{B}}$, or $\mathbf{C}_{\mathbf{AB}}$ defined in (2.14), (2.15) and (2.16), and $\mathrm{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}})$ the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}$, then hypotheses can be tested with the Wald statistic $W = \hat{\bar{\mathbf{Y}}}^t\mathbf{C}^t\{\mathrm{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}})\}^{-1}\mathbf{C}\hat{\bar{\mathbf{Y}}}$.

The GREG estimators, proposed by van den Brakel and Renssen (2005) and van den Brakel (2008) for single-factor experiments are extended to embedded factorial designs in this section. For notational convenience, the subscript q will be omitted in y_{iqkl}, since there is no need to sum explicitly over the interviewer subscript in most of the formulas developed in the rest of this paper.

To apply the model-assisted mode of inference to the analysis of embedded experiments, it is assumed for each unit in the population that the intrinsic value u_i in measurement error model (2.5) is an independent realization of the following linear regression model:

$$u_i = \beta^t \mathbf{x}_i + e_i, \tag{2.17}$$

where $\mathbf{x}_i$ H-vector with auxiliary information, β a H-vector with the regression coefficients and e_i the residuals, which are independent random variables with variance ω_i^2. It is required that all ω_i^2 are known up to a common scale factor, that is $\omega_i^2 = \omega^2 \nu_i$, with ν_i known. The GREG estimator for $\bar{Y}_{kl}$, based on the n_{kl} observations of subsample s_{kl}, is defined as (Särndal *et al.* 1992)

$$\hat{\bar{Y}}_{kl;greg} = \hat{\bar{Y}}_{kl} + \hat{\mathbf{b}}_{kl}^t(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}), k = 1, 2, \ldots, K, \text{ and } l = 1, 2, \ldots, L, \tag{2.18}$$

where,

$$\hat{\bar{Y}}_{kl} = \frac{1}{N}\sum_{i=1}^{n_{kl}} \frac{y_{ikl}}{\pi_i^*}, \tag{2.19}$$

denotes the HT estimator for $\bar{Y}_{kl}$, $\bar{\mathbf{X}}$ the finite population means of the auxiliary variables $\mathbf{x}$, and $\hat{\bar{\mathbf{X}}}$ the HT estimator for $\bar{\mathbf{X}}$ based on the n_{kl} sample units of subsample s_{kl}. Furthermore,

$$\hat{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*}\right)^{-1} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}, \tag{2.20}$$

denotes the HT-type estimator for the regression coefficients in (2.17) based on the n_{kl} sampling units in subsample s_{kl}. In (2.19) and (2.20), π_i^* are the first order inclusion probabilities for the sampling units in the KL different subsamples, derived in subsection 2.1. Now $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = (\hat{\bar{Y}}_{11;greg}, ..., \hat{\bar{Y}}_{KL;greg})^t$ is an approximately design-unbiased estimator for $\bar{\mathbf{Y}}$ and also for $\mathrm{E}_m \bar{\mathbf{Y}}$ by definition.

Under the null hypotheses that there are no treatment effects and no interactions, it follows that $\mathbf{b}_{kl} = \mathbf{b}_{k'l'}$. In that case, it might be efficient to substitute for $\hat{\mathbf{b}}_{kl}$ in the GREG estimator (2.18) the pooled estimator

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*}\right)^{-1} \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}. \tag{2.21}$$

Since H instead of $KL \times H$ regression coefficients have to be estimated, the pooled estimates of the regression coefficients $\hat{\mathbf{b}}$ will be more precise, particularly in the case of small subsamples. Note,

however, that many commonly used weighting schemes meet the condition that a constant vector λ exists such that $\omega_i^2 = \lambda \mathbf{x}_i$ for all $i \in U$. In this situation the GREG estimator reduces to the simplified form $\hat{\bar{Y}}_{kl;greg} = \hat{\mathbf{b}}_{kl}^t \bar{\mathbf{X}}$ (Särndal *et al.* 1992, section 6.5). Under this simplified form, the treatment effects are completely included in the regression coefficients. In case of the pooled estimator (2.21), the KL GREG estimators are exactly equal by definition, since $\hat{\bar{Y}}_{kl;greg} = \hat{\mathbf{b}}^t \bar{\mathbf{X}}$ for all k and l.

An expression for the covariance matrix of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ where the covariance is taken over the sampling design, the experimental design and the measurement error model, is given by

$$\text{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}) = \text{E}_m \text{E}_s \mathbf{C}\mathbf{D}\mathbf{C}^t, \tag{2.22}$$

where E_s denotes the expectation with respect to the sampling design, and $\mathbf{D}$ a $KL \times KL$ diagonal matrix with diagonal elements

$$d_{kl} = \frac{1}{n_{kl}\left(n_{++}-1\right)} \sum_{i=1}^{n_{++}} \left(\frac{n_{++}\left(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i\right)}{N\pi_i} - \frac{1}{n_{++}} \sum_{i'=1}^{n_{++}} \frac{n_{++}\left(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'}\right)}{N\pi_{i'}} \right)^2, \tag{2.23}$$

in the case of a CRD and

$$d_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}\left(n_{b++}-1\right)} \sum_{i=1}^{n_{b++}} \left(\frac{n_{b++}\left(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i\right)}{N\pi_i} - \frac{1}{n_{b++}} \sum_{i'=1}^{n_{b++}} \frac{n_{b++}\left(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'}\right)}{N\pi_{i'}} \right)^2, \tag{2.24}$$

in the case of an RBD. An estimator for $\mathbf{D}$ can be derived from the experimental design, conditionally on the measurement error model and the sampling design. Therefore the covariance matrix (2.22) is conveniently stated implicitly as the expectation over the measurement error model and the sampling design. A design-based estimator for this covariance matrix is given by

$$\hat{\text{cov}}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}) = \text{E}_m \text{E}_s \mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t, \tag{2.25}$$

with $\hat{\mathbf{D}}$ a $KL \times KL$ diagonal matrix with elements

$$\hat{d}_{kl} = \frac{1}{n_{kl}(n_{kl}-1)} \sum_{i=1}^{n_{kl}} \left(\frac{n_{++}(y_{ikl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{kl}} \sum_{i'=1}^{n_{kl}} \frac{n_{++}(y_{i'kl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{2.26}$$

in the case of a CRD and

$$\hat{d}_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}(n_{bkl}-1)} \sum_{i=1}^{n_{bkl}} \left(\frac{n_{b++}(y_{ikl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bkl}} \sum_{i'=1}^{n_{bkl}} \frac{n_{b++}(y_{i'kl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{2.27}$$

in the case of an RBD. Proofs for (2.22) and (2.25) are given by van den Brakel (2010) and resemble the derivation of the covariance matrix for single factor experiments, given by van den Brakel and Renssen (2005) and van den Brakel (2008).

The results for (2.22) and (2.25) are obtained under the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. This is a rather weak condition, since it implies that a weighting model is used that at least uses the size of the finite population as a priori information. See van den Brakel and Renssen (2005) or van den Brakel (2008) for a more detailed discussion.

Since the KL subsamples are drawn without replacement from a finite population, there is a nonzero design covariance between elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$. From that point of view, it is remarkable that (2.25) has a structure as if the subsamples are drawn independently through sampling with replacement using unequal selection probabilities. This gives rise to an attractive variance estimation procedure for embedded experiments, since no design covariances between the subsample estimates appear in (2.25) and no second order inclusion probabilities are required in the variance estimators (2.26) and (2.27). This result is obtained since the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ is derived instead of the covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ itself. A detailed interpretation of this result is given by van den Brakel and Renssen (2005) or van den Brakel (2008). See van den Brakel and Binder (2000) and Hidiroglou and Lavallée (2005) for approximations of the covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$.

The design-based estimators $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ and $\hat{\text{cov}}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}})$ can be used to construct a design-based Wald statistic to test the hypotheses described in section 2.2:

$$W = \hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}^t \mathbf{C}^t (\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1} \mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}. \tag{2.28}$$

Design-based inferences are generally based on normal large-sample approximations to construct confidence intervals for point estimates or *p*-values and critical regions for test statistics. Under this approach it follows under the null hypothesis that the Wald statistic is asymptotically distributed as a central chi-squared random variable, where the number of degrees of freedom equals the number of contrasts specified in the hypothesis.

The Wald statistic for the hypotheses about the main effects and interactions are given by (2.28) using the contrast matrix $\mathbf{C}_\mathbf{A}$, $\mathbf{C}_\mathbf{B}$, or $\mathbf{C}_\mathbf{AB}$. Under the null hypothesis, it follows that $W \rightarrow \chi^2_{[K-1]}$ for the test about the main effects of factor A, $W \rightarrow \chi^2_{[L-1]}$ for the test about the main effects of factor B and $W \rightarrow \chi^2_{[(K-1)(L-1)]}$ for the test about interactions, where $\chi^2_{[p]}$ denotes a central chi-squared distributed random variable with *p* degrees of freedom.

The Wald test for the main effects can be further simplified. Expressions are developed for the Wald test for the main effects for factor A. Similar expressions can be derived for the main effects of factor B. Denote

$$\begin{aligned} \hat{\bar{\mathbf{Y}}}_{\mathbf{A;GREG}} &= (\hat{\bar{Y}}_{1.;greg}, ..., \hat{\bar{Y}}_{K.;greg})^t, \text{ with } \hat{\bar{Y}}_{k.;greg} = \frac{1}{L}\sum_{l=1}^{L} \hat{\bar{Y}}_{kl;greg}, \\ \hat{\mathbf{D}}_\mathbf{A} &= \text{Diag}(\hat{d}_{1.}, ..., \hat{d}_{K.}), \text{ with } \hat{d}_{k.} = \frac{1}{L^2}\sum_{l=1}^{L} \hat{d}_{kl}. \end{aligned} \tag{2.29}$$

It follows that $\mathbf{C}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{GREG}} = \tilde{\mathbf{C}}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}$ and $\mathbf{C}_\mathbf{A}\hat{\mathbf{D}}\mathbf{C}_\mathbf{A}^t = \tilde{\mathbf{C}}_\mathbf{A}\hat{\mathbf{D}}_\mathbf{A}\tilde{\mathbf{C}}_\mathbf{A}^t$. With the matrix inversion lemma, the Wald statistic for the main effects of factor A can be simplified to:

$$\begin{aligned} W &= \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}^t\tilde{\mathbf{C}}_\mathbf{A}^t(\tilde{\mathbf{C}}_\mathbf{A}\hat{\mathbf{D}}_\mathbf{A}\tilde{\mathbf{C}}_\mathbf{A}^t)^{-1}\tilde{\mathbf{C}}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \\ &= \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}^t\left(\hat{\mathbf{D}}_\mathbf{A}^{-1} - \frac{1}{\text{Trace}(\hat{\mathbf{D}}_\mathbf{A}^{-1})}\hat{\mathbf{D}}_\mathbf{A}^{-1}\mathbf{j}_{(K-1)}\mathbf{j}_{(K-1)}^t\hat{\mathbf{D}}_\mathbf{A}^{-1}\right)\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \\ &= \sum_{k=1}^{K}\frac{\hat{\tilde{Y}}_{k.;greg}^2}{\hat{d}_{k.}} - \left(\sum_{k=1}^{K}\frac{1}{\hat{d}_{k.}}\right)^{-1}\left(\sum_{k=1}^{K}\frac{\hat{\tilde{Y}}_{k.;greg}^2}{\hat{d}_{k.}}\right)^2. \end{aligned} \tag{2.30}$$

Finally note that the HT estimator (2.19) does not meet the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. The minimum use of auxiliary information used in the GREG estimator is obtained with a weighting scheme that only uses the size of the finite population as a priori knowledge, *i.e.*, $(x_i) = 1$ and $\omega_i^2 = \omega^2$ (Särndal *et al*. 1992, section 7.4). Under this weighting scheme it follows that

$$\hat{\tilde{Y}}_{kl;greg} = \left(\sum_{i=1}^{n_{kl}}\frac{1}{\pi_i^*}\right)^{-1}\left(\sum_{i=1}^{n_{kl}}\frac{y_{ikl}}{\pi_i^*}\right) \equiv \tilde{y}_{kl}, \tag{2.31}$$

and $(\hat{\mathbf{b}}_{kl}) = \tilde{y}_{kl}$. Expression (2.31) can be recognized as Hájek's ratio estimator for a population mean (Hájek 1971). This weighting scheme satisfies the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. Therefore an approximately design-unbiased estimator for the covariance matrix of the contrasts between subsample estimates is given by (2.26) and (2.27) for a CRD and an RBD respectively, where $\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i = \tilde{y}_{kl}$. Estimator (2.31) is preferable above the HT estimator (2.19), since (2.31) is more stable and the covariance matrix of the contrasts between (2.31) always has the relatively simple form of (2.25).

2.4 Special cases

It will be shown for two special cases that the design-based Wald statistic is equal to the F-test of a standard analysis of variance. Therefore, an ANOVA-type pooled variance estimator for the diagonal elements of $\hat{\mathbf{D}}$ should be considered as an alternative for (2.26) or (2.27). Such a pooled variance estimator for a CRD is given by

$$\hat{d}_{kl}^p = \frac{1}{n_{kl}(n_{++} - KL)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_{k'l'}}\left(\frac{n_{++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{k'l'}}\sum_{i'=1}^{n_{k'l'}}\frac{n_{++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{2.32}$$

and for an RBD by

$$\hat{d}_{kl}^p = \sum_{b=1}^{B}\frac{1}{n_{bkl}(n_{b++} - KL)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_{bk'l'}}\left(\frac{n_{b++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bk'l'}}\sum_{i'=1}^{n_{bk'l'}}\frac{n_{b++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2. \tag{2.33}$$

Now consider a CRD that is embedded in a self-weighted sample, *i.e.*, $\pi_i = n_{++} / N$, with equally sized subsamples, *i.e.*, $n_{kl} = n_{k'l'} = n_s$. The inclusion probabilities for all units in the KL subsamples are given by $\pi_i^* = n_s / N$. Let $\bar{y} = (1 / n_s)\sum_{i=1}^{n_s} y_{ikl}$. Under Hájek's ratio estimator (2.31) and the pooled variance estimator (2.32) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

$$\hat{d}_{kl}^{\,p} = \frac{1}{n_s(n_{++} - KL)} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_s} \left(y_{ik'l'} - \bar{y}_{k'l'}\right)^2 \equiv \frac{\hat{S}^2_{p;CRD}}{n_s}.$$

The parameter estimates of the K levels of factor A averaged over the L levels of factor B are denoted as

$$\bar{y}_{k.} = \frac{1}{L}\sum_{l=1}^{L} \bar{y}_{kl} = \frac{1}{n_{k+}} \sum_{l=1}^{L} \sum_{i=1}^{n_s} y_{ikl}, k = 1, ..., K, \tag{2.34}$$

with $n_{k+} = \sum_{l=1}^{L} n_{kl}$. The diagonal elements of $\hat{\mathbf{D}}_{\mathbf{A}}$ are now given by

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L} \hat{d}_{kl}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L} \frac{\hat{S}^2_{p;\,\mathrm{CRD}}}{n_s} = \frac{\hat{S}^2_{p;\,\mathrm{CRD}}}{n_{k+}}, k = 1, \ldots, K. \tag{2.35}$$

Let $\bar{y}_{..} = (1 / n_{++})\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_s} y_{ikl}$. Inserting (2.34) and (2.35) into (2.30), gives rise to the following expression for the Wald statistic of the main effects of factor A

$$W = \frac{1}{\hat{S}^2_{p;CRD}} \left(\sum_{k=1}^{K} n_{k+} \bar{y}^2_{k.} - n_{++} \bar{y}^2_{..} \right). \tag{2.36}$$

Note that $W / (K - 1)$ in (2.36) corresponds with the F-statistic for the main effects of an analysis of variance for the two-way layout with interactions (Scheffé 1959, chapter 4). Under the null hypothesis and the assumption of normally and independently distributed errors, the F-statistic in the two-way layout follows an F-distribution with $(K - 1)$ and $(n_{++} - KL)$ degrees of freedom, which is denoted as $F^{[K-1]}_{[n_{++}-KL]}$. If $n_{++} \to \infty$, then $F^{[K-1]}_{[n_{++}-KL]} \to \chi^2_{[K-1]} / (K - 1)$. Consequently the F-statistic and the Wald statistic have the same limit distribution.

Now consider an RBD that is embedded in a self-weighted sampling design with equal subsample sizes, thus $\pi_i = n_{+++} / N$ and $n_{kl} = n_{k'l'} = n_s$, with $n_{+++} = \sum_{b=1}^{B} n_{b++}$. Let $\bar{y}_{bkl} = (1 / n_{bkl})\sum_{i=1}^{n_{bkl}} y_{ikl}$. Furthermore, it is assumed that the fraction of sampling units assigned to each treatment combination within each block is equal, *i.e.*, $n_{bkl} / n_{b++} = n_s / n_{+++}$, and that the block sizes are sufficiently large to assume that $n_{b++} / (n_{b++} - KL) \approx 1$. Under Hájek's ratio estimator (2.31) and the pooled variance estimator (2.33) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

$$\hat{d}_{kl}^{\,p} = \sum_{b=1}^{B} \frac{1}{n_{bkl}(n_{b++} - KL)} \left(\frac{n_{b++}}{n_{+++}} \right)^2 \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'} \right)^2$$
$$\approx \frac{1}{n_s n_{+++}} \sum_{b=1}^{B} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'} \right)^2 \equiv \frac{\hat{S}^2_{p;RBD}}{n_s}.$$

The parameter estimates of the K levels of factor A averaged over the L levels of factor B and the blocks are denoted as

$$\bar{y}_{.k.} = \frac{1}{L} \sum_{l=1}^{L} \bar{y}_{kl} = \frac{1}{n_{+k+}} \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}, k = 1, \ldots, K, \tag{2.37}$$

where $n_{+k+} = \sum_{b=1}^{B} \sum_{l=1}^{L} n_{bkl}$. The diagonal elements of $\hat{\mathbf{D}}_{\mathbf{A}}$ are given by

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2} \sum_{l=1}^{L} \hat{d}_{kl}^{\,p} = \frac{\hat{S}^2_{p;\,\mathrm{RBD}}}{n_{+k+}}, k = 1, \ldots, K. \tag{2.38}$$

Let $\bar{y}_{...} = (1 / n_{+++}) \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}$. If these results are inserted into (2.30), then the expression for the Wald statistic of the main effects of factor A can be simplified to

$$W = \frac{1}{\hat{S}^2_{p;RBD}} \left(\sum_{k=1}^{K} n_{+k+} \bar{y}^2_{.k.} - n_{+++} \bar{y}^2_{...} \right). \tag{2.39}$$

It can be recognized that $W / (K - 1)$ in (2.39) corresponds with the F-statistic for the main effects of an analysis of variance for the three-way layout with interactions, (Scheffé 1959, chapter 4). As in the case of a CRD, this Wald and F-statistic have the same limit distribution.

3 Factorial designs with more than two factors

The results developed for $K \times L$ factorial designs are extended to designs with more than two factors. A more general notation for the treatment factors is introduced first. Let A_g denote the g^{th} treatment factor in the experiment with levels $a_g = 1, ..., M_g$. In the general case there are g = 1,...,G factors included in the experiment. The population parameters observed under the $M_1 M_2 ... M_G$ treatment combinations are collected in the vector $\bar{\mathbf{Y}} = (\bar{Y}_{11...1}, ..., \bar{Y}_{a_1 a_2 ... a_G}, ..., \bar{Y}_{M_1 M_2 ... M_G})^t$. The index for the levels of a factor runs within each level of its preceding factor. Thus index a_g runs from $a_g = 1, ..., M_g$ within each level of $a_{(g-1)}$. Hypotheses about the main effects and interactions are, as motivated in section 2.2, formulated about $\bar{\mathbf{Y}}$ in expectation over the measurement error model.

The contrast matrices for the main effects and interactions in (2.13) are developed for the general case of a $M_1 \times M_2 \times ... \times M_G$ factorial design. Let $\mathcal{A} = \{1, ..., G\}$ denote the set of labels for the factors and $\tilde{\mathbf{C}}_{\mathbf{A_g}} = (\mathbf{j}_{(M_g - 1)} \mid -\mathbf{I}_{(M_g - 1)})$. The following three functions are defined first:

$$\mathbf{J}_{1_g} = \begin{cases} \mathbf{j}^t_{M_1} \otimes ... \otimes \mathbf{j}^t_{M(g-1)} & : \ g > 1 \\ 1 & : \ g = 1 \end{cases},$$

$$\mathbf{J}_{2_g} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes ... \otimes \mathbf{j}^t_{M_G} & : \ g < G \\ 1 & : \ g = G \end{cases},$$

$$\mathbf{J}_{3_{g,g'}} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes ... \otimes \mathbf{j}^t_{M(g'-1)} & : \ g' - g > 1 \\ 1 & : \ g' = g + 1 \end{cases}.$$

The main effect of factor A_g is defined as the $M_g - 1$ contrasts between the M_g levels, averaged over the levels of the other $G - 1$ factors and is given by:

$$\mathbf{C}_{\mathbf{A}_{g_1}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1\}} M_g \right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{2_{g_1}}, g_1 = 1, \ldots, G.$$

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{2_{g_1}}$ sums over the levels of the factors $A_{(g_1+1)}...A_G$ that are nested within each level of A_{g_1}.Subsequently, $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ defines the $M_{g_1} - 1$ contrasts between the levels of A_{g_1} that are nested within each combination of the levels of $A_1...A_{(g_1-1)}$. Premultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{1_{g_1}}$ adds the contrast matrices $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ that are nested within all combinations of the levels of $A_1...A_{(g_1-1)}$.

The interaction between A_{g_1} and A_{g_2} is defined as the $M_{g_2} - 1$ contrasts of factor A_{g_2} between the $M_{g_1} - 1$ contrasts of A_{g_1} averaged over the levels of the other $G - 2$ factors and is given by:

$$\mathbf{C}_{\mathbf{A}_{g_1}\mathbf{A}_{g_2}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1, g_2\}} M_g \right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{2_{g_2}},$$
$$g_1 = 1, \ldots, G-1, g_2 = 1, \ldots, G, g_1 < g_2.$$

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ by $\mathbf{J}_{2_{g_2}}$ adds the levels of the factors $A_{(g_2+1)}...A_G$ that are nested within each level of A_{g_2}. $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ defines the contrasts of the main effect of factor A_{g_2} which are nested within each combination of the levels of $A_1...A_{(g_2-1)}$. Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{3_{g_1,g_2}}$ sums the contrast matrices $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ over the levels of $A_{(g_1+1)}...A_{(g_2-1)}$ that are nested within each combination of the levels of $A_1...A_{g_1}$. Premultiplication of $\mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{2_{g_2}}$ with $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ defines the contrasts of the interactions between A_{g_1} and A_{g_2}, within each combination of the levels of $A_1...A_{(g_1-1)}$. Finally, Premultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{1_{g_1}}$ sums the contrasts of the interactions between A_{g_1} and A_{g_2} over the levels of $A_1...A_{(g_1-1)}$.

The interaction between A_{g_1}, A_{g_2} and A_{g_3} is defined as the $M_{g_3} - 1$ contrasts of factor A_{g_3} between the interactions of A_{g_1} and A_{g_2}, averaged over the levels of the other $G - 3$ factors. This process expands in a similar way to higher order interactions, which results in the following definitions of the higher order interactions:

$$\mathbf{C}_{\mathbf{A}_{g1}\mathbf{A}_{g2}\mathbf{A}_{g3}} = \left(\prod_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3\}} M_g\right)^{-1} \mathbf{J}_{1_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \mathbf{J}_{3_{g1,g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \mathbf{J}_{3_{g2,g3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \mathbf{J}_{2_{g3}},$$

$$g_1 = 1,\ldots,G-2, g_2 = 2,\ldots,G-1, g_3 = 3,\ldots,G, g_1 < g_2 < g_3,$$

$$\mathbf{C}_{\mathbf{A}_{g1}\mathbf{A}_{g2}\mathbf{A}_{g3}\mathbf{A}_{g4}} = \left(\prod_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3,g_4\}} M_g\right)^{-1} \mathbf{J}_{1_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \mathbf{J}_{3_{g1,g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \mathbf{J}_{3_{g2,g3}} \otimes$$

$$\tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \mathbf{J}_{3_{g3,g4}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g4}} \otimes \mathbf{J}_{2_{g3}},$$

$$g_1 = 1,\ldots,G-3, g_2 = 2,\ldots,G-2, g_3 = 3,\ldots,G-1,$$

$$g_4 = 4,\ldots,G, g_1 < g_2 < g_3 < g_4,$$

$$\vdots$$

$$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\ldots\mathbf{A}_G} = \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \ldots \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g4}}$$

The number of rows of each contrast matrix coincides with the number of contrasts that define the various main effects and interactions. The number of columns of these matrices equals $M_1M_2...M_G$.

These contrast matrices are inserted in (2.13) to define the various hypotheses about the main effects and interactions between the G treatment factors. The sampling units in the initial sample are randomly divided over all possible treatment combinations according to a CRD or an RBD, resulting in $M_1M_2...M_G$ different subsamples. Let $n_{a_1...a_G}$ denote the number of sampling units assigned to treatment combination $a_1...a_G$ in subsample $s_{a_1...a_G}$ and $n_{+...+}$ the size of the initial sample. In the case of a CRD, the first order inclusion probabilities for the units in subsample $s_{a_1...a_G}$ are now given by $\pi_i^* = \pi_i(n_{a_1...a_G} / n_{+...+})$. In the case of an RBD, the first order inclusion probabilities for the units in subsample $s_{a_1...a_G}$ are given by $\pi_i^* = \pi_i(n_{ba_1...a_G} / n_{b+...+})$ where $n_{ba_1...a_G}$ denotes the number of sampling units assigned to treatment combination $a_1...a_G$ in block b and $n_{b+...+}$ the total number of sampling units in block b.

Now $\hat{\bar{Y}}_{a_1...a_G;greg}$ denotes the GREG estimator for $\bar{Y}_{a_1...a_G}$ based on the observations obtained in subsample $s_{a_1...a_G}$ and is defined analogously to expression (2.18). These $M_1M_2...M_G$ GREG estimators are collected in the vector $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = (\hat{\bar{Y}}_{1...1;greg},...,\hat{\bar{Y}}_{M_1...M_G;greg})^t$ and is an approximately design-unbiased estimator for $\bar{\mathbf{Y}}$ and $\mathrm{E}_m\bar{\mathbf{Y}}$. Design-based estimators for the covariance matrices of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (2.26) in the case of a CRD or (2.27) in the case of an RBD.

Finally hypotheses about main effects and interactions are tested with the Wald statistic (2.28), which is asymptotically distributed as a chi-squared random variable where the number of degrees of freedom equals the number of contrasts specified in the various hypotheses. As an example, the contrast matrices of the main effects and interactions in a factorial design with four factors are given in Table 3.1.

Table 3.1
Contrasts in a $M_1 \times M_2 \times M_3 \times M_4$ factorial design

Contrast matrix	Number of contrasts (degrees of freedom)
$\mathbf{C}_{\mathbf{A}_1} = 1/\left(M_2M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_1 - 1$
$\mathbf{C}_{\mathbf{A}_2} = 1/\left(M_1M_3M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_2 - 1$
$\mathbf{C}_{\mathbf{A}_3} = 1/\left(M_1M_2M_4\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$M_3 - 1$
$\mathbf{C}_{\mathbf{A}_4} = 1/\left(M_1M_2M_3\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$M_4 - 1$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2} = 1/\left(M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3} = 1/\left(M_2M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_4} = 1/\left(M_2M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_1M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_1M_3\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1M_2\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_2\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = \tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$

4 Further extensions

So far, experimental designs are considered where the ultimate sampling units of the sampling design are randomized over the treatments. Owing to restrictions in the field work there might be practical reasons to randomize clusters of sampling units over the different treatments, at the cost of reduced power for testing hypotheses about treatment effects. It might for example be attractive to assign the sampling units that belong to the same household or are assigned to the same interviewer to the same treatment combination. In van den Brakel (2008) a design-based analysis procedure is developed for single-factor experiments designed as CRDs and RBDs where clusters of sampling units are randomized over the

treatments. These methods directly extend to the analysis of the factorial designs that are considered in this paper.

Consider the general case of a $M_1 \times M_2 \times ... \times M_G$ factorial design. The clusters of sampling units in the initial sample are randomized over the different treatment combinations. The conditional probability that a sampling unit is assigned to a subsample is now derived from the fractions of clusters that are assigned to the different treatment combinations within the sample or within each block. See van den Brakel (2008) for details. The GREG estimator for $\bar{Y}_{a_1 \ldots a_G}$ is defined analogously to expression (2.18). Design-based estimators for the covariance matrices of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (4.6) in van den Brakel (2008), which is based on the variance between the estimated cluster totals.

The target parameters of a survey are often defined as a ratio of two population totals. In van den Brakel (2008) a design-based analysis procedure is developed to test hypotheses about ratios in single-factor experiments designed as a CRD or an RBD. These results can be extended to the analysis factorial designs treated in this paper. Based on each subsample a ratio of two GREG estimators can be constructed for each treatment combination. Design-based estimators for the covariance matrices of the contrasts between the ratios are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (4.11) in van den Brakel (2008), which is an estimator for the variance of the ratio of two GREG estimators. Hypotheses about main effects and interactions are tested with the Wald statistic (2.28).

5 Testing new advance letters for the Dutch Labor Force Survey

In this section an experiment with different advance letters embedded in the Dutch Labor Force Survey (LFS) is described, which serves as a numerical example to illustrate the methodology developed in this paper.

5.1 Survey design

The LFS is based on a rotating panel survey. Each month a stratified two-stage cluster sample of about 6,000 addresses is drawn from a register of all known addresses in the Netherlands. Strata are formed by geographical regions, municipalities are considered as primary sampling units, and addresses as secondary sampling units. All households residing at an address, with a maximum of three, are included in the sample. In the first wave, data are collected by means of computer assisted personal interviewing. The respondents are re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing.

The weighting procedure of the LFS is based on the GREG estimator of Särndal *et al.* (1992). The inclusion probabilities reflect the sample design used to select households as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different socio-demographic categorical variables. One of the most important parameters of the LFS is the unemployed labor force, which is defined as the ratio of the total unemployment and the total labor force.

5.2 Experimental design

Advance letters are one of the design parameters of a survey that affect response rates and cooperation of respondents (De Leeuw, Callegaro, Hox, Korendijk and Lensveit-Mulders (2007)). The standard advance letter of the LFS is addressed to the occupants of the accommodation and the tone is formal and high-handed. As a result, this letter does not conform to social psychological theories regarding survey participation proposed by Groves, Cialdini and Couper (1992) and Groves and Couper (1998). In an attempt to improve the LFS response rates, Luiten, Campanelli, Klaasen and Beukenharst (2008) proposed different advance letters for the LFS that better meet these principles about survey participation. The effects of these alternative letters are investigated empirically by means of a large-scale field experiment embedded in the LFS.

The first factor considered in this experiment, say A, concerns the salutation of the respondent on two levels, *i.e.*, the standard approach where the letter is addressed to the occupants of the accommodation (A_1) versus a named letter (A_2). It is anticipated that named letters are more likely to be read and therefore increase response rates and survey participation. The second factor, say B, concerns the content of the letter on three levels, *i.e.*, the standard formal letter (B_1) versus two alternative letters (B_2 and B_3). In the first alternative, the content of the standard letter is adapted by explaining why the survey is conducted, what the respondent gains by participating and why it is important for Statistics Netherlands that the respondent participates in the survey. The second alternative attempts to improve the formal tone of the standard letter. The three versions of the advance letters can be found in van den Brakel (2010).

A new letter is only considered for implementation as a standard in the LFS, if its positive effect on response behavior has been demonstrated and if its effect on the main parameter estimates is quantified in a randomized experiment. Both factors are tested in a 2×3 factorial design resulting in six treatment combinations. This experiment is embedded in the first wave of the LFS for a period of five months (December 2007 through April 2008). During this period the monthly gross sample size is randomized over six subsamples according to an RBD with interviewers as the block variables. About 220 interviewers were available for the field work. In the analysis, adjacent interviewer regions were collapsed into 13 blocks. A fraction of 0.8 of the sample is assigned to the regular advance letter, *i.e.*, treatment combination $A_1 \times B_1$. A fraction of 0.04 of the sample is assigned to each of the other five alternative treatment combinations.

The allocation of the sampling units over the treatments is predominantly based on practical arguments. Embedding experiments in ongoing sample surveys serves two competing purposes. To estimate official figures as precisely as possible it is beneficial to allocate as many sampling units as possible to the control group, since this subsample is also used for regular publication purposes. To estimate the contrasts in the experiment as precisely as possible it is, on the other hand, beneficial to divide the total sample equally over the different treatment combinations. In this application it was decided that a loss of at most 20% of the sample size for regular publication purposes could be tolerated. This led to the aforementioned allocation over the treatment combinations. Under a response rate of 56% and a monthly sample size of 6,000 households it is expected that about 13,440 households are observed in the control group $A_1 \times B_1$ and 670 households in each of the alternative treatment combinations.

Although the allocation is based on practical considerations, it is important to have a notion of the power of the planned experiment. The target variable analyzed in this paper is the unemployed labor force,

expressed as a percentage. Ignoring the block design of this experiment, it follows that the variance of the treatments equals to $\hat{d}_{kl} = \hat{S}_{kl}^2 / n_{kl}$, where $\hat{S}_{kl}^2$ is implicitly defined by (2.26). It is assumed that $\hat{S}_{kl}^2$ is equal to say $\hat{S}^2$ for each treatment combination. With available sample data it follows for the unemployed labor force that $\hat{S}^2 = 285$. Now the minimal observable difference for a contrast that would reject the null hypothesis under a pre-specified significance and power level equals

$$\Delta = \sqrt{\text{var}(\Delta)}(Z_{(1-\alpha/2)} + Z_{(1-\beta)}), \tag{5.1}$$

where $Z_{(\gamma)}$ denotes the γ^{th} percentile point of the standard normal distribution, α the significance level of the test and $(1-\beta)$ the power. The main effect of factor A concerns one contrast $\hat{\Delta}_A = (\hat{\bar{Y}}_{1.;greg} - \hat{\bar{Y}}_{2.;greg})$. From (2.29) it follows that the variance of this contrast equals $\hat{\text{var}}(\hat{\Delta}_A) = (\hat{S}^2 / 9)\sum_{l=1}^{3}(1/n_{1l} + 1/n_{2l})$. The main effect of factor B concern two contrasts $\hat{\Delta}_{B_l} = (\hat{\bar{Y}}_{.1;greg} - \hat{\bar{Y}}_{.l;greg})$, $l = 2,3$ with variances $\hat{\text{var}}(\hat{\Delta}_{B_l}) = (\hat{S}^2 / 4)\sum_{k=1}^{2}(1/n_{k1} + 1/n_{kl})$, $l = 2,3$. The interactions between factors A and B concern the two contrasts $\hat{\Delta}_{AB_l} = (\hat{\bar{Y}}_{11;greg} - \hat{\bar{Y}}_{1l.;greg} - \hat{\bar{Y}}_{21;greg} + \hat{\bar{Y}}_{2l.;greg})$ with variances $\hat{\text{var}}(\hat{\Delta}_{AB_l}) = \hat{S}^2(1/n_{11} + 1/n_{1l} + 1/n_{21} + 1/n_{2l})$, $l = 2,3$.

Inserting the variances of the different contrasts in (5.1), gives minimum values of differences that would reject the null hypothesis for main effects and interactions for pre-specified sample sizes, significance levels and power levels. In Table 5.1 these differences for the unemployed labor force are calculated for the aforementioned applied allocation, and a balanced design where the sample size for each treatment combination is equal to 2,800. Values are given for unspecified alternative hypotheses at a 5% significance level and a power of 50%, 80% and 90%. In experimental design theory, 80% is a widely accepted power level by sample size determination. In survey sampling minimum sample size requirements are generally based on significance level requirements only, which corresponds to a power level of 50%. Differences are specified for separate tests of the contrasts. The main effect of factor B and the interaction effects both contain two contrasts. To preserve an overall significance level of 5%, differences for both tests are also calculated using Bonferroni's simultaneous comparison procedure.

Table 5.1 illustrates different aspects of embedded experiments and factorial designs. First it illustrates the cost-benefits of a factorial setup. Twice as many experimental units are required if the main effects of both factors are tested at the same precision in two separate single factor experiments. Table 5.1 also shows that the power for the test of interactions is much smaller than for the tests of the two main effects. The more treatment factors that are combined in one experiment, the smaller the sample size allocated to each treatment combination and the smaller the power for the tests of interactions. This puts the often cited advantage that factorial designs also allow testing of interactions between the different treatment factors into perspective. In practice, sample sizes are based on power calculations for the tests on the main effects. Consequently, only large interactions can be detected with sufficient power. A factorial design still has the advantage that the validity of observed main effects increases, since they are tested over a wider range of conditions.

Table 5.1
Observable difference for the unemployed labor force in percentages at 5% significance levels and different power levels

Contrast	Number of contrasts	Power separate t-test			Power Bonferroni t-test		
		50%	80%	90%	50%	80%	90%
Applied design							
Main effect A	1	0.96	1.36	1.58	0.96	1.36	1.58
Main effect B	2	1.12	1.59	1.85	1.27	1.75	2.00
Interaction	2	2.23	3.19	3.69	2.55	3.51	4.00
$A_1 \times B_1 - A_k \times B_l$	5	1.31	1.87	2.17	1.72	2.28	2.57
Balanced design							
Main effect A	1	0.51	0.73	0.84	0.51	0.73	0.84
Main effect B	2	0.63	0.89	1.03	0.71	0.98	1.12
Interaction	2	1.25	1.79	2.07	1.43	1.97	2.25
$A_1 \times B_1 - A_k \times B_l$	5	0.88	1.26	1.46	1.16	1.54	1.74

If the null hypothesis of no interactions is rejected, then main effects are difficult to interpret. In that situation it is more useful to compare the control group, *i.e.*, $A_1 \times B_1$, with the five alternative treatment combinations. The minimum observable differences of these five contrasts that reject the null hypothesis at a 5% significance level and different power levels are also included in Table 5.1.

Comparing minimum values for the differences under the applied design and the balanced design, illustrates the loss of power if an extreme skew allocation over the treatment combinations is chosen. Minimizing the risk of losing too much precision for the regular publication is the motivation behind the choice for this allocation. It clearly illustrates the duality of combining two competing purposes in an embedded experiment; estimation for the regular publication purposes versus testing contrasts of different treatment combinations.

To assess the value of the results that can be obtained with this experiment, the minimum observable differences with this experiment are related to the standard errors of the regular survey estimates. Standard errors for the survey estimates at the national level will generally be much smaller than the minimum observable differences with an experiment since the sample size allocated to the alternative treatments is generally much smaller than the regular sample size. If, however, the assumption is adopted that differences observed with an experiment at the national level also apply to the survey estimates for important domains, then the differences observable with the experiment might become comparable with the standard errors of these domain estimates. This assumes no interaction between domains and treatment effects. The standard errors for the monthly unemployed labor force figures at the national level equals 0.15 percent points. The standard errors for the domains vary between 0.3 and 1.0 percent points. Comparing these standard errors with the differences in Table 5.1 shows that the main effects are still larger than the standard errors at the national level but become comparable with the precision of the regular monthly domain estimates.

5.3 Results

Table 5.2 contains an overview of the response rates of the households in the six subsamples of the experiment. It follows that the different advance letters result in relatively small differences in the response rates. Factor A results in an increase of the response of 2.4 percent points by using a personalized letter (after correcting proportions for the unbalanced allocation of the sample over the treatment combinations). The alternative letters considered in factor B resulted in a decrease of 1.5 percent points (alternative B_2) and 1.9 percent points (alternative B_3).

Table 5.2
Response rates experiment with advance letters

Treatment	Response		Refusal		Rest		Total
$A_1 \times B_1$	13,234	56.69%	5,127	21.96%	4,985	21.35%	23,346
$A_1 \times B_2$	604	53.59%	271	24.05%	252	22.36%	1,127
$A_1 \times B_3$	635	56.34%	254	22.54%	238	21.12%	1,127
$A_2 \times B_1$	662	59.00%	256	22.82%	204	18.18%	1,122
$A_2 \times B_2$	663	59.09%	236	21.03%	223	19.88%	1,122
$A_2 \times B_3$	627	55.64%	259	22.98%	241	21.38%	1,127

Response behavior is modeled in a logistic regression model to test hypotheses about the effect of the two treatment factors. This is a typical conditional analysis that does not account for sample design features like unequal selection probabilities and clustering of households within municipalities. Clustering induced by the two-stage sample design is ignored, since households are randomized over the treatments in the experiment. In this logistic regression analysis interest is focussed on differences in the observed sample, in this case due to differences in selective non-response. This gives additional information on whether the factors increase the response across the entire target population or that specific groups react differently to the treatments. Second and higher order interactions between the two treatment factors and socio-demographic categorical variables in the logistic regression model indicate that the variation in response between different subpopulations increases and that they react differently to the treatments.

In the logistic regression model, the dependent binary variable indicates whether a household completely responded versus the remaining response categories. The response behavior is assumed to depend upon:

- a general mean,
- treatment factors A (name) and B (content),
- a block variable in 13 categories,
- auxiliary variables:
  - urbanization level at five categories,

- gender in three categories, specifying whether a household consists of men only, women only, or a mixture of men and women,
- age as a quantitative variable containing the average age of the household members,
- ethnicity in seven categories, specifying household compositions of native, western background, non-western background, and all possible mixtures,
- family composition in four categories: partners, single-parent family, single, and a remainder category.

All third order interactions between the variables are initially considered for backward model selection. The final selected model contains the terms that are given in the first column of Table 5.3. For brevity, the regression coefficients with their standard errors and test statistics for separate categories are only expressed for the treatment factors. The hypothesis that there are no interactions between the two treatment factors cannot be rejected (p -value Wald statistic equals 0.121). From Table 5.3 it follows that factor A, *i.e.*, using a letter addressed to a named individual, has a positive but non-significant effect on the response rate. Factor B, *i.e.*, two alternative letters with an improved content, has even a slightly negative but non-significant effect on the response rates. This is a remarkable result, since the two alternative letters attempt to improve the formal tone of the standard letter, but in line with the results of an earlier experiment where the response to a more informal advance letter for the LFS also resulted in significantly smaller response rates (van den Brakel 2008). Since there are no interactions between the treatment factors and the auxiliary variables, there are also no indications that the treatment factors induce the response of specific subpopulations.

Table 5.3
Logistic regression analysis for response rates

Parameter	Coefficient	Standard error	Wald statistic	D.f.	p -value
Mean	0.287	0.078	13.604	1	0.000
Block			212.425	12	0.000
Treatment A (name, A_2)	0.083	0.045	3.394	1	0.065
Treatment B (content)			2.965	2	0.227
Alternative 1 (B_2)	-0.046	0.051	0.816	1	0.366
Alternative 2 (B_3)	-0.083	0.051	2.678	1	0.102
Urbanization			16.589	4	0.002
Ethnic			127.734	6	0.000
Gender			48.076	2	0.000
Family composition			27.339	3	0.000

In the second step of this analysis it is tested whether the estimates for the unemployed labor force obtained with the six subsamples under the different advance letters are significantly different. The design-based analysis procedure developed in this paper is used to account for the sampling design, the

experimental design and the estimation procedure of the LFS. The GREG estimator is applied to obtain estimates for the unemployed labor force under the six different treatment combinations. With this unconditional analysis it is tested whether the different advance letters introduce differences in selection bias, after correcting for the differences in response rates using the design-based estimation procedure applied in the regular LFS.

With this analysis, the linear measurement error model (2.1) is applied to a binary response variable. This might appear to be ridged, since logistic models are more natural in this case. Under the model-assisted approach linear regression models, however, are frequently applied to derive a GREG estimator for binary response variables. Also in the Dutch LFS a linear regression model is assumed to derive a GREG estimator for official labor force figures. To develop a design-based analysis procedure for embedded experiments that also account for the GREG estimator used in the regular survey, a linear measurement error model is assumed in a similar way. A detailed discussion about the use and interpretation of a linear measurement error model applied to binary response variables is given by van den Brakel (2008).

The inclusion probabilities in the GREG estimator (2.18) reflect the sampling design of the LFS and the experimental design used to divide the initial sample into six subsamples. The following weighting scheme was applied to calibrate the design weights: *age* + *region* + *marital status* + *gender* + *urbanization level*, where the five variables are categorical. This is a reduced version of the regular weighting scheme of the LFS.

The estimation results for the six subsamples are summarized in Table 5.4, where the unemployed labor force is expressed in percentages. It appears that there are no systematic patterns between subsample estimates. The subsample estimates and their variance estimates indicate that there are no significant differences between the control group and the five alternative treatment combinations. Finally the main effects and the interaction effects of the two treatment factors are tested, taking into account that the experiment was designed as an RBD where adjacent interviewer regions are collapsed in 13 blocks. The analysis results are summarized in Table 5.5.

Table 5.4
Point estimates and standard errors unemployed labor force (expressed in percentages)

Treatment combination		Estimate $\hat{\bar{Y}}_{kl;greg}$	Standard error $\sqrt{\hat{d}_{kl}}$
$k\,(A_k)$	$l\,(B_l)$		
1	1	4.100%	0.145%
1	2	3.761%	0.646%
1	3	5.264%	0.753%
2	1	3.609%	0.608%
2	2	4.546%	0.666%
2	3	3.385%	0.664%

Table 5.5
Analysis main effects and interactions unemployed labor force (expressed in percentages)

Source	Estimate $C\hat{\bar{Y}}_{greg}$	Wald statistic	D.f.	p -value
Treatment A (name) $A_1 - A_2$	0.528	1.109	1	0.292
Treatment B (content)		0.732	2	0.694
$B_1 - B_2$	-0.300			
$B_1 - B_3$	-0.471			
Interaction		3.801	2	0.150
$AB_{11} - AB_{12} - AB_{21} + AB_{22}$	1.276			
$AB_{11} - AB_{13} - AB_{21} + AB_{23}$	-1.388			

From the analysis results, summarized in Table 5.5, it can be concluded that there are no indications that the different advance letters result in different parameter estimates. This is in line with the analysis results of the response rates. Since there is no empirical evidence that the different advance letters affect response rates of the entire population or a subpopulation, it might be expected that no significant differences between the parameter estimates occur.

There are no indications that the alternative letters, considered in this experiment, improve response behavior or result in systematic effects in the estimates for target variables like the unemployed labor force. Therefore it was decided not to adapt the standard advance letter of the LFS.

6 Discussion

In factorial designs the levels of two or more treatment factors are varied and all possible treatment combinations are considered simultaneously. These designs are widely used in scientific experimentation for several reasons. The main effects of the factors are averaged over the levels of the other factors. Conclusions about the various effects are therefore based on a wider range of conditions, which increases the validity of the results. Furthermore, interaction between the different treatment factors can be analyzed, although the power of these tests decreases as the number of factors that are combined in one experiment increases. Finally factorial designs are more efficient compared to single-factor experiments, since fewer experimental units are required to estimate the main effects with the same precision.

In this paper a design-based theory is developed for the analysis of factorial designs that are embedded in probability samples. This approach is particularly appropriate to quantify the effects of the different design parameters of a survey process on the parameter estimates of a sample survey. Applications can be found in total survey design, empirical research into survey practice and quantifying discontinuities in series of repeatedly conducted surveys. Design-based analysis procedures are developed to test hypotheses about population means for factorial designs where the ultimate sampling units are randomized over the different treatment combinations through a CRD or an RBD. Procedures for factorial designs where clusters of sampling units are randomized over the treatment combinations or to test hypotheses about

ratios of population totals are obtained analogously to the methods developed in van den Brakel (2008) for single-factor experiments.

The design-based variance estimator that is developed for the various treatment effects does not require joint inclusion probabilities nor design-covariances between the different subsamples. As a result a design-based analysis procedure for factorial designs embedded in complex probability samples is obtained with the attractive relatively simple structure as if the sampling units are drawn with unequal selection probabilities with replacement. The traditional advantages of factorial designs, summarized in the first paragraph of the discussion, still apply under this design-based approach. As illustrated with variance expression (2.29) fewer experimental units are required to estimate the main effects with the same precision in a factorial setup compared to separate single-factor designs.

The advantage of an RBD over a CRD is that the between block variance is removed from the estimated treatment effects. In the standard model-based theory for the analysis of randomized experiments, an F-test for the blocks as well as the treatment factors is available. Under restricted randomization of an RBD, however, it is generally argued that a F-test for the block effects is not valid. In these cases alternative measures to evaluate the efficiency of an RBD are available; see for example Montgomery (2001). In the design-based theory developed for RBDs in this paper there is an asymmetry between the block and treatment factors, as in the case of the randomization approach followed by Hinkelmann and Kempthorne (1994). Due to the restricted randomization within the blocks there is no meaningful test for the main effect of the block factor available.

Acknowledgements

The author wishes to thank the Associate Editor and the unknown referees for giving constructive comments on a former draft of this paper. The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

References

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*, Chichester: John Wiley.

Chipperfield, J., and Bell, P. (2010). Embedded experiments in repeated and overlapping surveys. *Journal of the Royal Statistical Society,* Series A, 173, 51-66.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. and Lensvelt-Mulders, G. (2007). The influence of advance letters in response in telephone surveys. *Public Opinion Quarterly*, 71, 413-443.

Fienberg, S.E., and Tanur, J.M. (1987). Experimental and Sampling Structures: Parallels Diverging and Meeting. *International Statistical Review*, 55, 75-96.

Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16, 135-151.

Fienberg, S.E., and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. *Science*, 243, 1017-1022.

Fienberg, S.E., and Tanur, J.M. (1996). Reconsidering the Fundemental Contributions of Fisher and Neyman on Experimentation and Sampling. *International Statistical Review*, 64, 237-253.

Groves, R.M., Cialdini R.B. and Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56, 475-495.

Groves, R.M., and Couper, M.P. (1998). *Nonresponse in household interview surveys*, New York: John Wiley.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by D. Basu, in *Foundations of Statistical Inference* (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart, and Winston.

Hidiroglou, M.A., and Lavallée, P. (2005). Indirect two-phase sampling: Applying it to questionnaire field-testing. *Proceedings of Statistics Canada Symposium 2005: Methodological challenges for future information needs*.

Hinkelmann, K., and Kempthorne, O. (1994). *Design and Analysis of Experiments, Volume 1: Introduction to experimental design*, New York: John Wiley.

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.

Jäckle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78, 3-20.

Kempthorne, O. (1955). The Randomization Theory of Experimental Inference. *Journal of the American Statistical Association*, 50, 946-967.

Luiten, A., Campanelli, P., Klaasen, D. and Beukenhorst, D. (2008). Advance letters and the language and behaviour profile, paper presented at the 19th International Workshop on Household Survey Nonresponse.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-370.

Montgomery, D.C. (2001). *Design and Analysis of Experiments*, New York: John Wiley.

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

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

Scheffé, H. (1959). *The Analysis of Variance*, New York: John Wiley.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, Chichester: John Wiley.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society*, Series A, 171, 581-613.

van den Brakel, J.A. (2010). Design-based analysis of factorial designs embedded in probability samples. Discussion paper 201014, Statistics Netherlands, Heerlen.

van den Brakel, J.A., and Binder, D. (2000). Variance estimation for experiments embedded in complex sampling schemes. *Proceedings of the section on Survey Research Methods*, American Statistical Association, 805-810.

van den Brakel, J.A., and Van Berkel, C.A.M. (2002). A Design-based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys. An Application in the Dutch Labor Force Survey. *Journal of Official Statistics*, 18, 217-231.

van den Brakel, J.A., and Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14, 277-295.

van den Brakel, J.A., and Renssen, R.H. (2005). Analysis of experiments embedded in complex sampling designs. *Survey Methodology*, 31, 23-40.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. American Mathematical Society*, 54, 426-482.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 351-368
Statistics Canada, Catalogue No. 12-001-X

Estimation and replicate variance estimation of deciles for complex survey data from positively skewed populations

Stephen J. Kaputa and Katherine Jenny Thompson[1]

Abstract

Thompson and Sigman (2000) introduced an estimation procedure for estimating medians from highly positively skewed population data. Their procedure uses interpolation over data-dependent intervals (bins). The earlier paper demonstrated that this procedure has good statistical properties for medians computed from a highly skewed sample. This research extends the previous work to decile estimation methods for a positively skewed population using complex survey data. We present three different interpolation methods along with the traditional decile estimation method (no bins) and evaluate each method empirically, using residential housing data from the Survey of Construction and via a simulation study. We found that a variant of the current procedure using the 95th percentile as a scaling factor produces decile estimates with the best statistical properties.

Key Words: Median; Modified half-sample replication; Interpolation; Deciles.

1 Introduction

Developing viable decile estimates for positively skewed populations from complex survey data poses interesting challenges. The literature supports two different approaches to percentile estimation with complex survey data. The first method (the "traditional" method) obtains decile estimates from empirical cumulative-distribution functions, selecting the item value that corresponds to the sample percentile computed by summing associated survey weights. This approach yields decile estimates that are "close to unbiased" but unstable. An alternative approach is to group the continuous data into disjoint intervals (bins), then use linear interpolation over the bin containing the decile. With appropriately defined bins, this approach also produces nearly unbiased decile estimates while improving their stability – at least for the percentiles that are far from the tail of the distribution. For the upper percentiles, often the binned data contain very few observations, with little or no uniformity. Hence, the reliability of the large decile estimates (*e.g.*, 90th percentile or greater) is rarely comparable to that of the other deciles.

Although the usage of interpolation is advantageous for developing stable estimates, developing an optimal set of bins for a given characteristic is not always an easy task. Often, the distributions change over time, and the bin widths/locations in the sample should reflect this change in scale. For example, the average sales price for single-family homes in a geographic area may increase over time due to inflation, but the population of single-family homes in that area is still characterized by a skewed distribution, with a few expensive homes located in the tail. Many economic data programs share this trait. Consequently, developing a fixed set of bins for interpolation with an ongoing survey is unwise. To address this, Thompson and Sigman (2000) introduced an estimation procedure for estimating *medians* from highly positively skewed population data. Their procedure uses interpolation over data-dependent intervals (bins), after scaling by the 75th percentile. The earlier paper examined the estimation and variance

1. Stephen J. Kaputa and Katherine Jenny Thompson, Office of Statistical Methods and Research for Economic Programs, US Census Bureau, 4600 Silver Hill RD, Washington, DC 20233. E-mail: Stephen.kaputa@census.gov.

estimation properties of the considered methods, using modified half sample (MHS) replication for variance estimation (Fay 1989; Judkins 1990).

This research extends the previous work to decile estimation methods using complex survey data sampled from a positively skewed population. We present three different interpolation methods along with the traditional decile estimation method (no bins) and evaluate each method empirically, using residential housing data from the Survey of Construction (SOC) conducted by the U.S. Census Bureau and via a simulation study. Our research was motivated by a recent request from the SOC data users to estimate and publish complete sets of decile estimates for several housing characteristics. Thus, our research was conducted under the constraints of maintaining comparably reliable median estimates as those currently published and using MHS replication for variance estimation.

Section 2 presents the candidate decile estimation methods and gives an overview of modified half-sample replication. Section 3 evaluates these procedures, using empirical and simulated data from the Survey of Construction (SOC). Finally, we conclude with recommendations in Section 4.

2 Methodology

2.1 Decile estimation

We consider two approaches to decile estimation for continuous data: the sample decile (SD) method and interpolation. The SD method uses ordered sample weights to locate the estimate (Rao and Shao 1996). For this, the characteristics values are sorted in ascending order, and the sample weights are accumulated until they exceed the desired decile's percent of the total weight.

Interpolation methods group the continuous data in bins and interpolate over the bin containing the decile. To obtain the decile estimate $\left(\xi^{d}\right)$, we use the Woodruff formula (Woodruff 1952) for interpolation provided below:

$$\xi^{d} = F^{-1}\left(d\hat{N}\right) \approx ll + \left(\frac{d\hat{N} - cf}{f_i}\right) * (i) \tag{2.1}$$

where

F = cumulative frequency of the characteristic using sample weights,

ll = lower limit of the bin containing the decile,

$\hat{N}$ = estimated total number of elements in the population,

cf = cumulative frequency in all intervals preceding the bin containing the sample decile,

f_i = decile class frequency (estimated total number of elements in the population of the interval containing the sample decile),

i = width of the bin containing the sample decile,

d = desired decile (0.1, 0.2, 0.3, …, 0.9).

Notice that this formula does not require that each bin to be of equal length. However, it does require that the data within each bin be uniformly distributed. This later requirement poses the true challenge with a highly skewed population, especially in the upper tail.

Figure 2.1 below illustrates how to use the Woodruff method to estimate the 80th decile. The sample data have been grouped into twelve separate bins. The empirical CDF is produced from the complete set of weighted sample data (as one referee noted, the empirical CDF is extremely smooth for sample survey data; in practice, the curve would include discrete steps. The Woodruff method procedure would be the same, however). The decile estimate is located at the intersection of the empirical CDF curve and red asymptote at $Y = 0.80$. The 80th decile is $F^{-1}(0.80)$, contained in the 5th bin; the interpolated estimate of the 80th percentile would therefore be obtained by using (2.1) over the fifth bin.

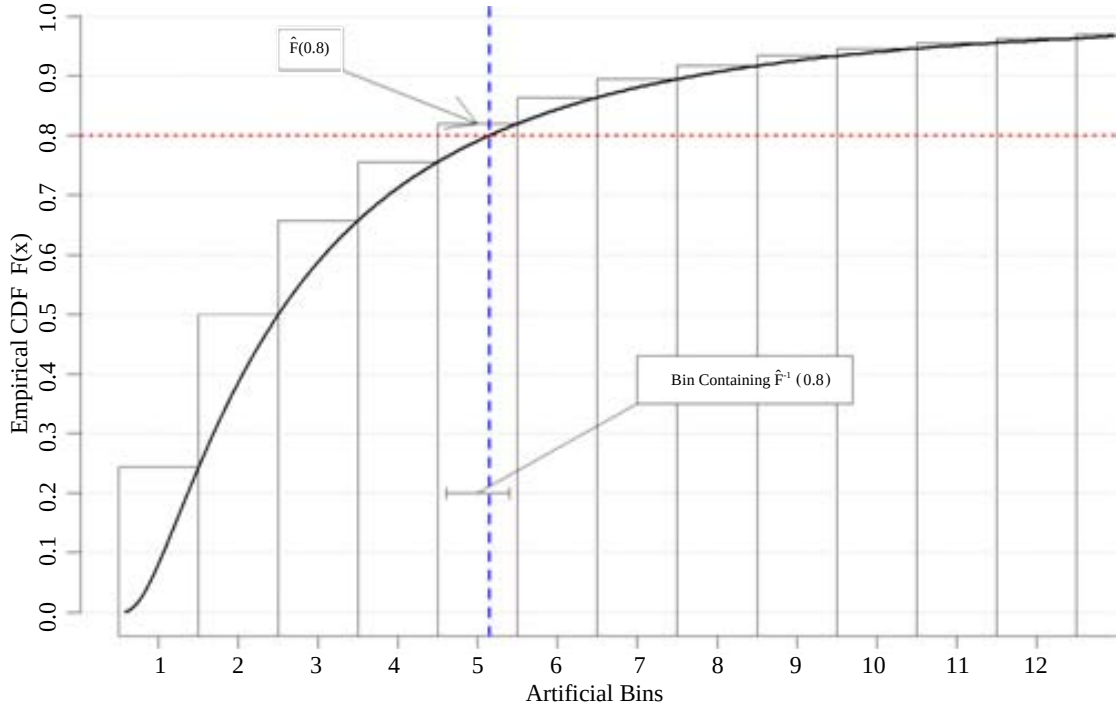


Figure 2.1 Illustration of the Woodruff method

Determining the optimal bin size for both estimation and variance estimation can be difficult. As the bins narrow (approaching width 1), then the variance estimates become more unstable. Smoothing the estimates via the interpolation reduces the instability of variance, but increases the bias in the estimate. The bias component increases as the bin widths increase.

Economic data generally have a positive skewed distribution. Moreover, the subdomains' characteristic distributions will vary, and their respective moments change over time as the economy changes. Consequently, developing a standard set of fixed bins for interpolation that work consistently over time is nearly impossible. Instead, Thompson and Sigman (2000) developed a "data-dependent" binning procedure, where the width of each bin is determined separately by the estimation cell. Their

recommended method linearly transforms each characteristic to a standard scale and then uses a standard set of bins for every characteristic. The authors use the following linear transformation

$$X_i' = X_i \times \frac{1{,}000}{Q_{75}}$$

where Q_{75} is the 75th percentile (3rd quartile) of the sample distribution, obtained using the SD method. The interpolated-median estimate of the X' is multiplied by $\left(Q_{75}/1{,}000\right)$ to obtain a value on the original scale. This procedure is equivalent to simply dividing the original sample in each estimation cell from 0 to Q_{75} into Z bins of equal width and placing the remainder of the sample into one bin, which, by design, is much larger than the others. With the highly positive skewed housing data, this transformation works well for estimating the median because it is far from the Q_{75} scaling parameter. However, it does not permit estimation of either the 80th or 90th deciles. Thus, if we wanted to continue using an interpolation method, we needed to consider alternative transformations.

The simplest approach is to use the original data-dependent bin method with a higher scaling parameter, *i.e.*, use any percentile value larger than 90%. We use the 95th percentile as the scaling factor and hereafter refer to this method as the "P95 method".

The P95 method does create uniform distributions within the majority of the bins but is still problematic at the upper end of the distribution for two reasons. First, the final bin contains only five percent of the sample distribution, and the values within this bin are generally very different. Second, the data-dependent binning procedure requires that each decile be "far from" the large final bin; if not, then the decile estimates exhibit the same instability as those obtained using the "SD method." Unfortunately, the bin containing the 90th percentile is often close to the final bin when using a scaling parameter of 95%.

To address the second issue, we considered another data-binning approach, denoted as the "P75 method." For this, we create two sets of bins per estimation cell, each with different widths above and below the cell's Q_{75} value. This requires two separate linear transformations per estimation cell, given by

$$X_i' = X_i \times \frac{1{,}000}{X_{75}} \text{ when } X_i < X_{75}$$

$$X_i'' = \left(X_i - X_{75}\right) \times \frac{1{,}000}{\left(X_{100} - X_{75}\right)}$$

$$\text{when } X_i \geq X_p \text{ and } X_{100} = \text{ maximun value in sample.}$$

The X_i' is then placed into Z equal length bins, and X_i'' into K equal length bins, where $Z \neq K$. The interpolation is performed independently for each decile, with the appropriate inverse transformation being applied to each interpolated decile. This procedure ensures that median estimates exactly match those obtained with the current procedure.

Our third considered interpolation approach makes parametric assumptions about the characteristics. Often, economic data are approximately log-normally distributed (*e.g.*, Steel and Fay 1995). The Normal Binning method (denoted "NB") uses the properties of the normal distribution applied to the log-transformed data to obtain data-dependent bins. The binning technique ensures that areas of high

probability have smaller bin widths to limit the amount of observations per bin and areas of low probability have larger bin width to increase the amount of observations per bin.

The NB method centers the log-transformed data around the weighted sample median, then scales the centered data by an estimate of the population standard deviation. We use the sample median because it is more outlier resistant than the sample mean. Of course, the mean and median are equivalent with normally distributed data. Given a standard normal distribution where $\mu = 0$ and $\sigma = 1$, then

$$IQR = Q_3 - Q_1$$
$$IQR = \left(0.67449*\sigma\right) - \left(-0.67449*\sigma\right) = \sigma\left(0.67449 + 0.67449\right) = \sigma * 1.34898.$$

We estimated the standard deviation (sigma) as the ratio $\sigma \approx IQR/1.34898$, where the IQR is obtained from the empirical CDF in the estimation cell. To normalize the data, we applied the following transformation

$$Y_i = \text{Log}\left(X_i\right) \quad Y_i' = \frac{Y_i - Y_{\text{med}}}{\sigma_y} = \frac{Y_i - Y_{\text{med}}}{IQR_y/1.34898}$$

where

Y_{med} = log-transformed sample median over domain i,

IQR_y = log-transformed sample interquartile range over domain i.

Again, the sample deciles and interquartile ranges are obtained via the SD method. If the data are log-normally distributed, Y_i' should have a standard normal distribution, so that roughly 68.3% of the data are within one standard deviation of the mean and 95.4% of the data are within two standard deviations of the mean. Using those properties, we split the transformed Y_i' into the five different zones and created the 45 bins shown in Table 2.1.

Table 2.1
Bins for the log-normal transformation (Normal method)

Zone	1	2	3	4	5
Range	[Low, -2)	[-2, -1)	[-1, 1)	[1, 2)	[2, High]
Percent in Zone	2.3	13.6	68.2	13.6	2.3
Bins	1	6	31	6	1
Average Percent of Sample Units per Bin	2.3	2.3	2.2	2.3	2.3

There are four different bin widths with roughly the same average percentage of sampled units per bin. Woodruff's method is applied to the transformed data to obtain the deciles and we exponentiate these decile estimates to obtain values on the original scale. Unlike the linear rescaling methods presented above, there is an additional induced estimation bias caused by the power transformation. It may have

been possible to make a bias adjustment for the transformation via a Taylor expansion, as suggested by a referee, but we did not consider this approach.

2.2 Variance estimation

The MHS replication method (aka "Fay's method") is a "compromise" between the stratified jackknife and the BRR method (Fay 1989). Rao and Shao (1999) demonstrate that the MHS variance estimator is asymptotically consistent for both smooth statistics such as ratio estimators and for non-smooth statistics such as sample quantiles estimated using the SD method outlined in 2.1. Their paper does not extend this property to interpolated decile estimates, although it does follow that these variance estimates should be consistent as the bin width approaches width 1. Like BRR, MHS replication uses a Hadamard matrix to form replicates, but uses replicate weights of 1.5 and 0.5 instead of the values of 2 and 0 used in BRR. The MHS formula for standard error estimation of any estimate $\hat{\theta}$ is

$$\hat{S}\left(\hat{\theta}\right) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}_0\right)^2} \tag{2.2}$$

where $\hat{\theta}_r$ is the r^{th} replicate estimate $\left(r = 1, 2, \ldots, R\right)$ and $\hat{\theta}_0$ is the full sample estimate. The sum of squared error term is adjusted by a factor of $4 = 1\Big/\left(1 - 0.5\right)^2$ to prevent negative bias in the variance estimate (Judkins 1990).

3 Empirical analysis

3.1 SOC sample design

As mentioned in the introduction, our research was motivated by a request from data users of the Survey of Construction (SOC). The SOC is a national survey that collects information on characteristics of new residential housing in the United States. The SOC data are used to produce three principle economic indicators published each month by the U.S. Census Bureau: housing starts, housing completions, and housing sales (single-family homes only). In addition, SOC publishes monthly, quarterly, and annual estimates on a variety of housing characteristics, such as sales price and average sales price per square foot of sold houses, length of time from permit authorization to housing start, and length of time from start to completion of housing construction. This paper examines two key housing characteristics that are published annually: sales price and price per square foot of sold homes. Both characteristics are collected monthly as they become available from the homebuilders. Currently, average and median estimates for both characteristics are included in the annual reports; average and median sales price of sold homes are also published monthly.

The SOC universe comprises two sub-populations: areas that require building permits and areas that do not. Areas that require building permits are covered in the Survey of the Use of Permits (SUP) and non permit issuing areas are covered by the Nonpermit Survey (NP). The vast majority of the sample comes from SUP. Both populations are sampled from the same PSUs, but are independent samples at subsequent

stages. Since the majority of the SOC sample consists of sampled permits, we focus entirely on the SUP portion of the SOC in our research.

The SUP is selected in three stages. The first stage selects a probability proportional to size (PPS) subsample of Primary Sampling Units (PSUs) from the 2000 Current Population Survey design (CPS) and is performed once every ten years. The CPS PSUs are land areas such as counties or townships. The second stage of the SUP sample is a stratified systematic sample of permit-issuing places within sampled PSUs, also performed once every ten years. The third stage of sampling is performed monthly in each of the sampled permit-issuing places. Each month, the Field Representatives develop complete lists of new building permits from the permit offices in the sampled places and select a systematic sample of building permits. Sampling rates are assigned to permit offices to obtain an overall sampling rate of one-in-fifty for one to four unit structures. Larger buildings of five or more units are included with certainty (*i.e.*, are self-representing).

The SOC uses the MHS replication method to estimate variances with a 200×200 Hadamard matrix, assigning a total of 198 rows to replicate groups. Since SOC does not have a two-PSU per stratum design, SOC uses a collapsed stratum approach for creating replicates: see Thompson (1998) for details.

3.2 Empirical data analysis of SOC data

Our empirical analyses uses SOC data collected from 2006 through 2009. SOC uses the data-dependent binning method described in Section 2.1 with 41 bins to produce median estimates. We use 51 bins for the P95 method, with 95% of the sample spread over 50 equally sized bins; the 51st bin contains any data greater than the 95th percentile. The P75 method uses 40 equal-sized bins for all values below the 75th percentile and 10 equal-sized bins for all values between the 75th percentile and the maximum value of the sample distribution. Finally, the NB method uses a total number of 45 bins.

For sales price and price per square foot, all of the decile estimates obtained via the P75, P95, and SD methods are quite comparable to each other for the 10th through 70th deciles; the NB deciles were generally slightly larger than their other counterparts. However, the P75 method decile estimates for the 80th and 90th deciles were consistently larger than the other three methods' estimates. The explanation is straightforward: with the P75 method, both the 80th and 90th deciles are almost always located in the same bin. Both characteristic's distributions are quite skewed. Consequently, the majority of the upper 25 percent of the sample is contained in the bin closest to the 75 percentile.

The patterns displayed by the decile estimates for each method were extremely consistent at both the national and regional levels. Unlike the estimate comparisons, there are fewer clear patterns with the variance estimates. The variance estimates for 80th decile obtained using the P75 method were considerably larger than those obtained by the three other methods and the 90th was likewise considerably smaller. With respect to the other three methods, the NB variances tend to be smaller than the corresponding P95 method and SD variances; these differences are more pronounced with the sales price per square foot decile estimates.

Three of the four considered methods yielded comparable sets of decile estimates. The P75 method proved intractable given the highly skewed distributions considered; we simply could not find an adequate "bin width" for the upper quartile of data. Thus, the empirical data evaluation reduced our candidate set of estimation methods to three. However, although the corresponding estimates were quite similar, the

variance estimates were clearly different. Consequently, we decided to conduct a simulation study to evaluate the statistical properties of the alternative estimators over repeated samples.

3.3 Simulation study

3.3.1 Modeling and sample selection procedure

For our simulation, we developed a population that mimics the qualities of the majority of the SUP population. That is, we developed stratified populations of permit offices (PSUs), from which we selected permit samples (SSUs). There were several advantages to developing such a complex simulation set up. From a practical immediate perspective, it was beneficial for interpreting the empirical results presented in Section 3.2. More important, the earlier research conducted by Thompson and Sigman (2000) obtained nearly perfect results for the data-dependent interpolated medians on a simulated population that did not include clustering; the distinctions between the statistical properties of each method only became apparent when clustering was incorporated into the design.

We used a "bottom-up" approach to develop viable simulated population data. First, we modeled multivariate populations of permit data within each region. Next, we combined the modeled permit data to form "clusters" representing the permit offices (the primary sampling units). To guard against model misspecification, we independently developed two artificial populations of permit data with each permit record containing sales price and price per square foot, modeling one population as log-normally distributed within region using the algorithm outlined in Lienhard (2004) and modeling the other using a nonparametric SIMDATA algorithm (Thompson 2000) in each region.

In general, the modeled permit data in the nonparametric population is a better representation of the corresponding levels at each decile of the training data. However, the distributions of permits in the log-normal population are quite smooth, whereas the nonparametric population distributions are "choppy," with large breaks (steps) between adjacent point estimates. Table 3.1 presents key percentiles and means from the simulated populations for both modeled characteristics, comparing them to the empirical values from the SOC data (denoted by the "Training Data" column).

Our simulation procedure developed populations of permits (the SSUs), then created the artificial first stage clusters (permit offices) and stratified them. The two-step cluster creation and stratification process described below assumes that the permits within population strata are heterogeneous with respect to sales price and price per square foot and that permits issued from the same permit office have similar housing characteristics. These criteria were obtained from the subject matter experts, who believe that modeling variation between permit offices was more realistic than modeling variation within permit offices. The multivariate log-normal population lends itself better to this than the nonparametric population; the assigned elements within each cluster tend to be homogeneous because of the smoother distribution.

We used discriminant analysis to group the simulated permits into disjoint strata. After applying the same discriminant function to each simulated permit data population (log-normal and nonparametric), we clustered the permits within strata to form approximately 14,000 active permit offices per population. The cluster analysis application created permit offices of variable size with homogenous characteristics within office.

Table 3.1
Simulated populations' statistics and empirical data statistics

			Simulated Population		Training Data (Weighted)
			Log-normal	Nonparametric	
Sales Price	Decile	1	74,747.74	94,323.27	95,000.00
		5	105,009.64	120,073.33	120,000.00
		10	126,492.43	136,009.85	140,000.00
		20	158,517.61	158,931.57	160,000.00
		30	186,927.58	181,941.12	180,000.00
		40	215,346.44	205,003.46	210,000.00
		50	245,502.91	230,188.69	230,000.00
		60	280,064.08	260,524.68	260,000.00
		70	322,790.68	301,374.90	300,000.00
		80	381,501.24	359,138.64	360,000.00
		90	482,209.62	488,517.45	490,000.00
		95	586,359.10	622,185.68	630,000.00
		99	855,983.47	1,167,704.85	1,300,000.00
		Mean	283,085.94	287,134.16	290,000.00
			Simulated Population		Training Data (Weighted)
			Log-normal	Nonparametric	
Sales Price per Square Foot	Decile	1	32.76	32.68	35.00
		5	43.11	47.27	47.00
		10	49.86	54.70	55.00
		20	59.26	63.74	64.00
		30	67.22	70.71	72.00
		40	74.91	77.14	78.00
		50	83.11	83.60	84.00
		60	92.42	90.60	91.00
		70	103.75	98.70	99.00
		80	119.61	109.57	110.00
		90	147.62	130.02	130.00
		95	178.94	155.08	160.00
		99	265.10	262.68	270.00
		Mean	93.58	94.90	96.00

We selected 5,000 repeated samples from our simulated population using a much simplified version of the SOC design described above. The first stage of sampling selects permit offices. The largest 250 offices at the US level were selected with a probability of one (certainty), so that each repeated sample contains the same self-representing offices. Then, we selected a probability proportionate to size sample of two non self-representing permit offices in each stratum, with each office receiving its own permit office weight.

At the second sampling stage, we selected permit records from each sampled permit office. We selected a simple random sample (SRS) of permits from each office, with an office sampling rate obtained by dividing the permit office's weight by 50, thus obtaining an overall one-in-fifty sample of permits (if the permit office weight is greater than fifty, all permits within the office are sampled). Final weights for each record were calculated by multiplying the permit office weight and the permit weight. The permits selected from the certainty offices vary in each repeated sample due to the independent sampling unless the office contained more than 50 permits.

Finally, in each sample, we assigned permits or permit offices to replicates. We did not mimic the SOC partially balanced half sample application described in Section 3.1. Collapsing strata induces bias in the variance estimates. To eliminate this bias component from our simulation, we used a two-PSU per stratum design and 572 replicates (*i.e.*, a 572×572 Hadamard matrix), so that each of the 250 self-representing offices and each of the 321 non self-representing strata (pairs of sampled permit offices) received its own Hadamard matrix row. Mimicking the SOC production method, each self-representing office is treated as a "pseudo stratum," and replicate panels are obtained by randomly splitting the permits within each office.

Within each sample, a set of estimates at the US and regional level were calculated for the three considered decile estimation methods in each replicate and the MHS replicate variance estimates for each decile were computed using (2.2) with $R = 572$.

3.3.2 Evaluation methodology

The simulation study examines the statistical properties of each decile estimation method and associated variance estimates over repeated samples. Let ζ_m^d represent the decile d estimate calculated by using method $m\,(m = \text{SD, P95, NB})$.

To assess *estimation* properties of method m for decile d over repeated samples, we computed the relative bias and the empirical mean squared error. The relative bias of each decile estimate for each estimation procedure is given by

$$\hat{B}\left(\xi_m^d\right) = 100 \times \left[\overline{\xi}_m^d \Big/ \xi_p^d\right] - 1$$

where ξ_{ms}^d is the estimated decile d estimate from method m in sample s, $\overline{\xi}_m^d$ is the average over the 5,000 samples, and ξ_p^d is the population decile (evaluation measures for estimates and variance estimates for domain i (Northeast, Midwest, South, and West in the simulation study presented in Section 4) are available upon request to the authors, but are omitted for brevity).

The empirical mean squared error (MSE) of each decile estimate for each estimation method is given by

$$\text{MSE}\left(\xi_m^d\right) = \frac{1}{5{,}000}\sum_{s=1}^{5{,}000}\left(\xi_{ms}^d - \xi_p^d\right)^2 = \frac{1}{5{,}000}\sum_{s=1}^{5{,}000}\left(\xi_{ms}^d - \overline{\xi}_m^d\right)^2 + \left(\overline{\xi}_m^d - \xi_p^d\right)^2$$

$$\text{MSE}\left(\xi_m^d\right) = \hat{\sigma}\left(\xi_m^d\right) + \hat{B}^2\left(\xi_m^d\right).$$

To assess the *variance estimation* properties of the estimation method for decile d over repeated samples, we computed the following statistics:

Relative bias of the variance $100 \times \left[\hat{\bar{v}}\left(\xi_m^d\right) \Big/ \text{MSE}\left(\xi_m^d\right)\right] - 1$ where $\hat{\bar{v}}\left(\xi_m^d\right)$ is the average *variance* estimate of decile d from method m over 5,000 samples *i.e.*, $\hat{\bar{v}}\left(\xi_m^d\right) = \hat{v}\left(\xi_{ms}^d\right) \Big/ 5{,}000$. In our case

study, the variance estimates in each sample $s, \hat{v}\left(\xi_{ms}^{d}\right)$, are modified half-sample replicate variance estimates described in Section 3.3.1.

Stability of the variance estimate $\sqrt{\sum_{s=1}^{5,000}\left(\hat{v}\left(\xi_{ms}^{d}\right) - \mathrm{MSE}\left(\xi_{m}^{d}\right)\right)^{2} \Big/ 5,000} \Big/ \mathrm{MSE}\left(\xi_{m}^{d}\right)$.

Coverage rates (CR) = the proportion of 90% confidence intervals for a given method that contain the true population decile ξ_{p}^{d}.

The stability of the variance is a measure of the variance of the variance estimates. Ideally, both the relative bias and stability measures should be near zero. Coverage rates demonstrate the combined effect of the estimate and variance estimate on inference.

3.3.3 Simulation study results

The following sections summarize our simulation study results, presenting in illustrative graphs (tables available upon request to the authors).

3.3.3.1 Estimation properties of each method

Figure 3.1 plots the relative biases of the national level decile estimates by estimation method in the log-normal population. Recall that unbiased estimates will have a relative bias of zero indicated by the grey horizontal asymptote on each figure. Caution is advised in visual comparisons of bias levels, as the two characterics' graphs may not be on the same scale.

The SD method produces the least biased decile estimates for both sales price and price per square foot. That said, the biases of the decile estimates for both characteristics obtained using the P95 and NB methods are trivial. The largest biases can be found at the 10th percentile and the 90th percentile, that is, near the tails of the distribution where the sample is expected to be less stable. Although the SD decile estimates are less biased than their P95 and NB counterparts, they are less precise. In general, the P95 deciles have the minimum MSE among the three competing methods, although in many cases, the differences between the P95 and NB MSEs are negligible.

Figure 3.2 plots the relative biases of the national level decile estimates by estimation method in the nonparametric population. The bias patterns for price per square foot follow the same patterns as above, as do the MSEs. However, the pattern of the bias and MSE of sales price is different. Here, the SD estimates are the least biased, but the largest bias occurs at the median (0.005). This is also true for the two interpolation methods, with the P95 and NB medians each having a relative positive bias of seven tenths of a percent. For the 50th and 60th deciles, the MSE of the P95 estimates is somewhat larger than the other corresponding estimates, reflecting the impact of this estimator.

Some of the biases from the nonparametric population are large enough to warrant concern, especially for the median estimate. That said, the log-normal population does appear to more closely mimic the true SOC data, so the nonparmetric results are not necessarily reflective of SOC's "reality." These results do reflect the impact of the constant bias term in the decile estimates caused by interpolation.

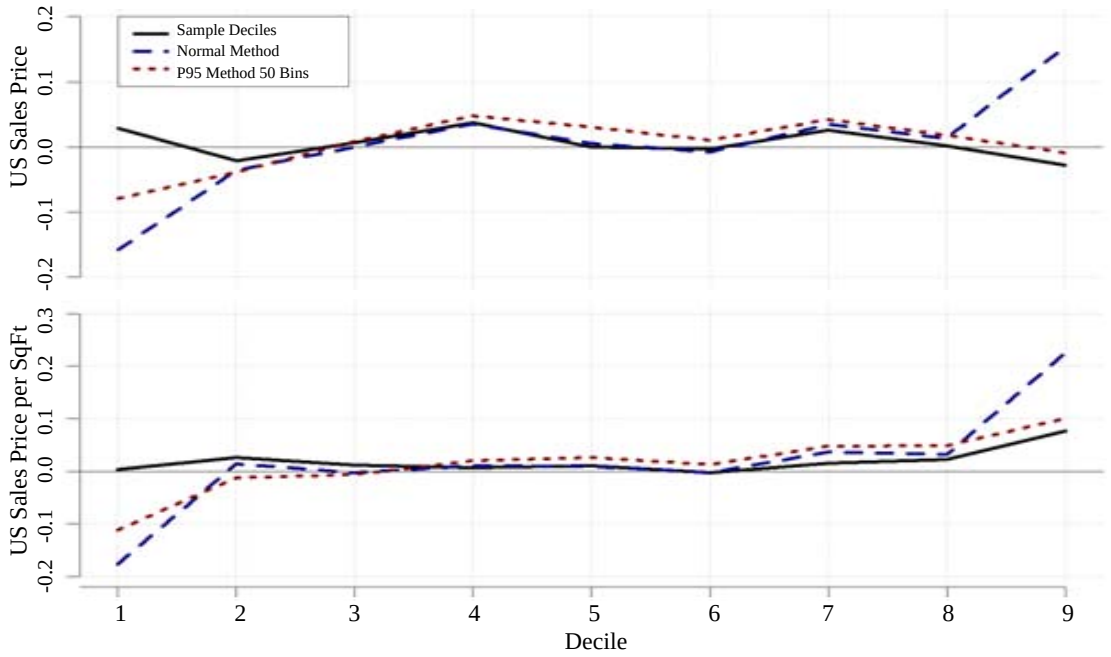


Figure 3.1 Relative bias of sales price and sales price per square foot estimates from the log-normal population (Expressed in percentages)

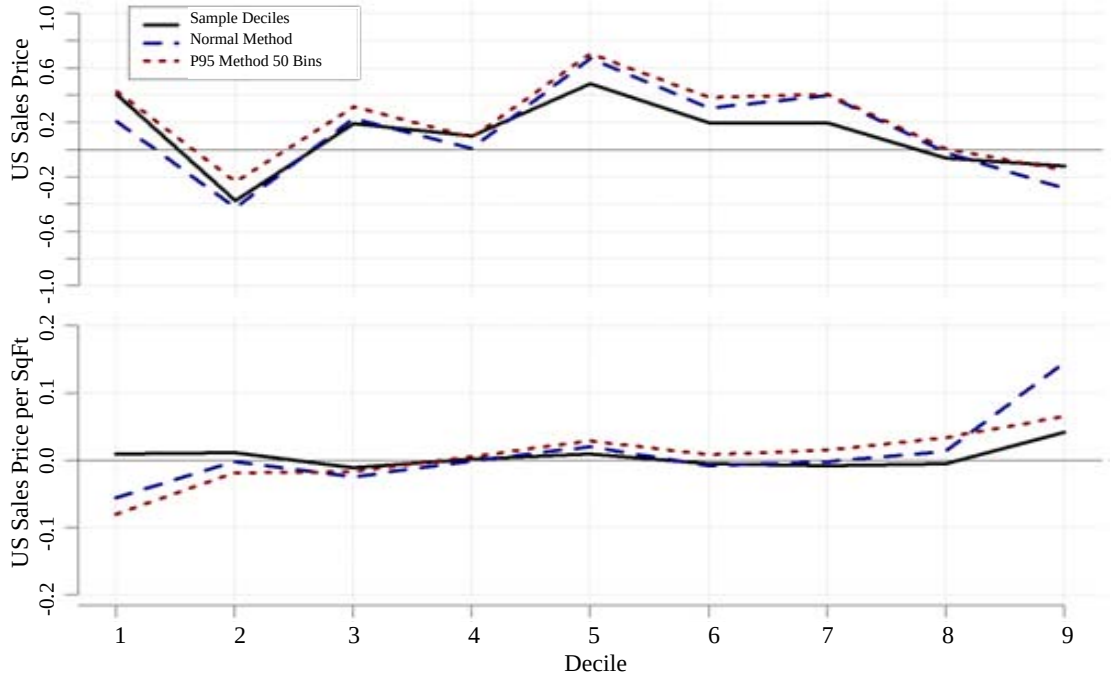


Figure 3.2 Relative bias of the sales price and sales price per square foot estimates from the nonparametric population (Expressed in percentages)

Overall, the MSEs follow similar patterns for both populations and both characteristics. Minimal MSEs can be found around the center deciles. The lower tail deciles have slightly higher MSEs and the upper tail decile MSEs increase at a rapid rate.

3.3.3.2 Variance estimation properties of each method (Given MHS replication)

As demonstrated in Figure 3.3, *all* relative variance biases of the MHS variance estimates in the log-normal population are positive regardless of estimator, and the SD variance estimates are the most biased. The P95 and NB methods have similar relative biases for all characteristics, with all being less biased than those obtained via the SD method. Overall, the P95 variance estimates are the least biased. Notice that *all* of the variance estimates for sales price are positively biased with all estimators, to save space, the y-axis begins at 5%. The same cautions about visual comparisons stated in Section 3.3.3.1 apply to the figures in this section.

The SD variance estimates for both characteristics are by far the least stable. This result is expected, since the interpolation variance estimates benefit from smoothing. Of the two interpolation methods, the P95 method yields more stable variance estimates for all deciles except for a handful of the upper tail deciles. The more stable variances in the NB upper deciles are likely a result of using properties of a normal distribution to obtain equal percentages of the sample in each bin.

None of the three considered methods yielded 90% coverage rates for sales price or price per square foot (Figure 3.4). Most of the coverage rates are slightly anti-conservative (below the 90% horizontal asymptote), and no decile estimation method appears to exhibit superior coverage properties over the others.

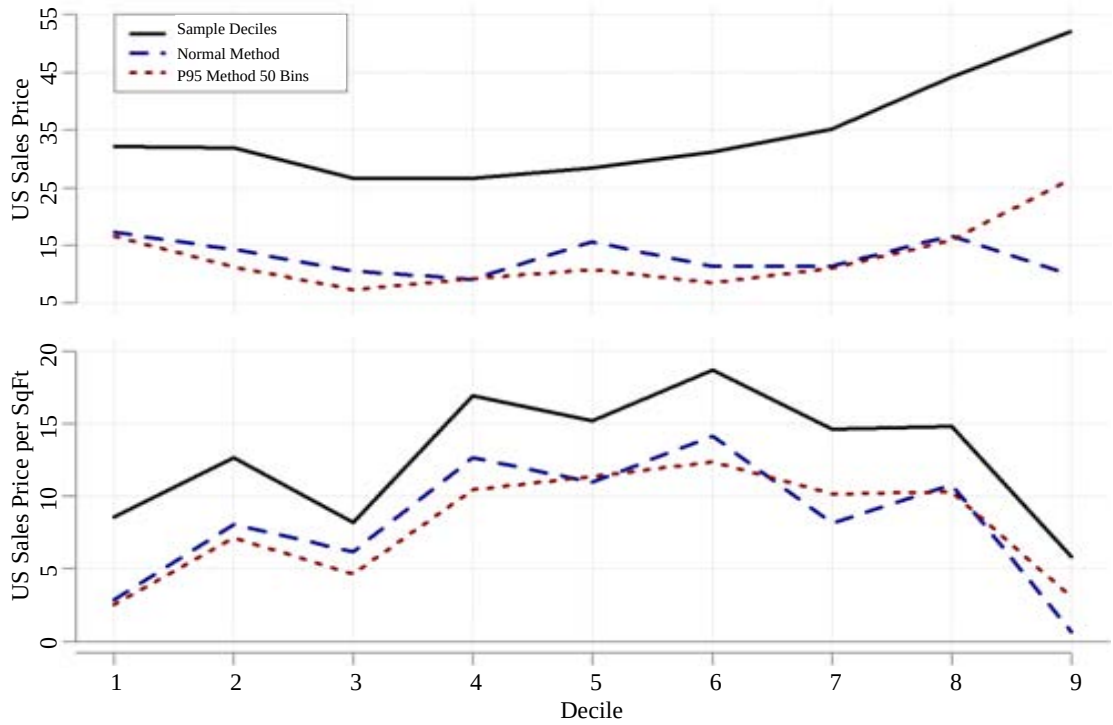


Figure 3.3 Relative bias of the variance for sales price and sales price per square foot from the log-normal population (Expressed in percentages)

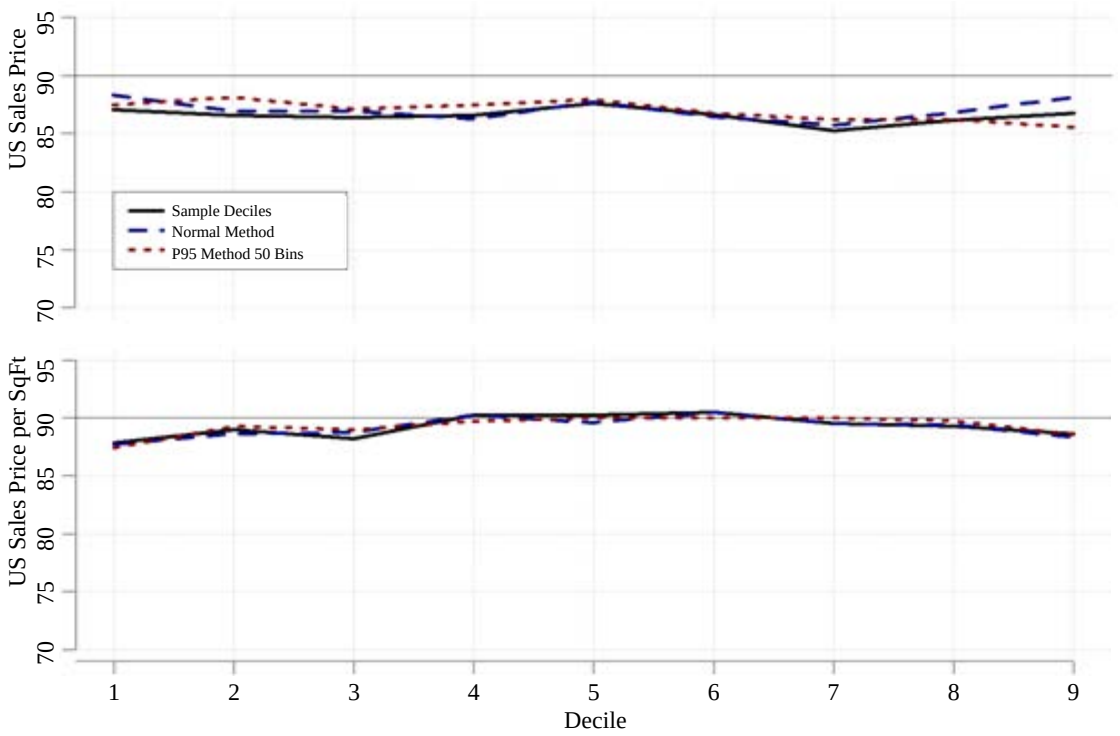


Figure 3.4 Coverage rates for sales price and sales price per square foot (Log-normal population)

Figure 3.5 illustrates the relative biases of the MHS variances obtained from the nonparametric population where both the P95 and NB interpolation methods are generally much smaller than their SD counterparts, and the P95 methods tend to produce less biased variance estimates for both characteristics. The relative bias of sales price for the nonparametric population does not follow the same pattern as the log-normal characteristics. The SD's relative biases are always positive and higher than the two interpolation methods. The two interpolation methods produced different results across populations of sales price. The nonparametric population contains many negative relative biases as opposed to all positive biases. The relative biases for sales price per square foot does follow the same pattern as the log-normal population, with large positive biases for the SD method, and lower similar positive biases for the two interpolation methods.

The stability of the nonparametric variance estimates matched up well with those obtained from the log-normal population, except for a few differences with sales price decile estimates. Sales price stability estimates for the SD method are still always larger than the other two interpolation methods, but follow a more erratic pattern. The NB method has a large stability estimate for the 40^{th} decile, which does not follow the expected trend.

The coverage rates for sales price per square foot follow the same pattern as in the log-normal population (Figure 3.6). However, the pattern of the sales price rates is more variable.

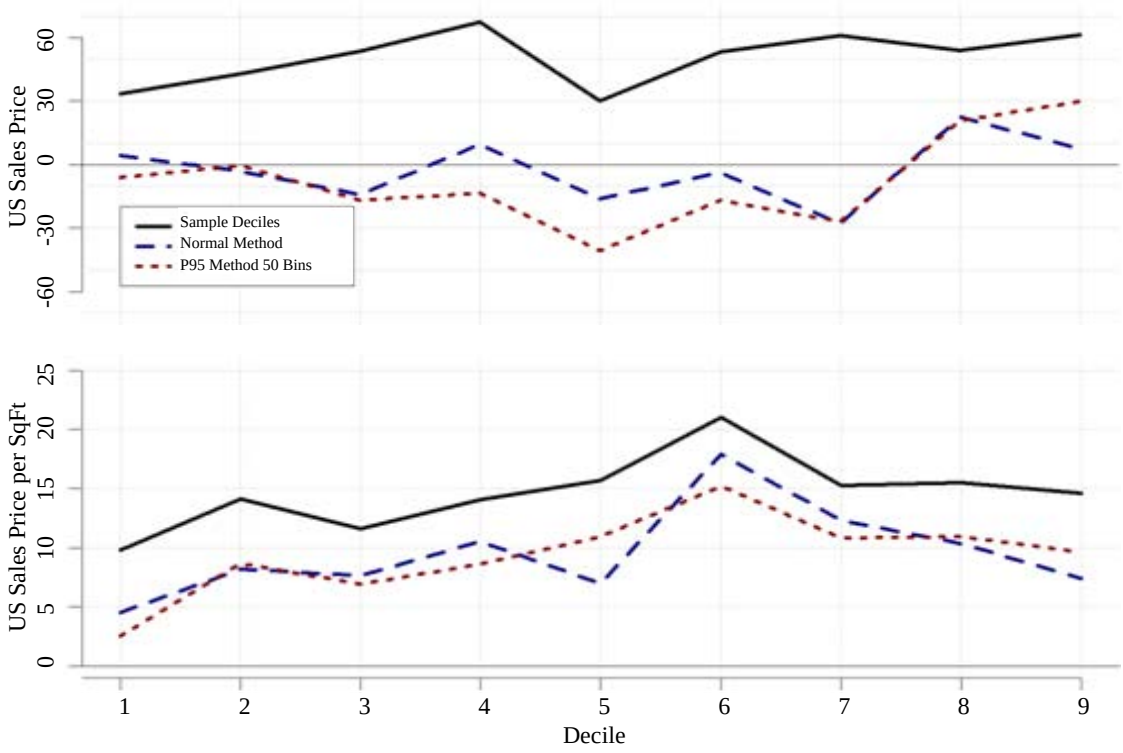


Figure 3.5 Relative bias of the variance of sales price and sales price per square foot (Nonparametric population)

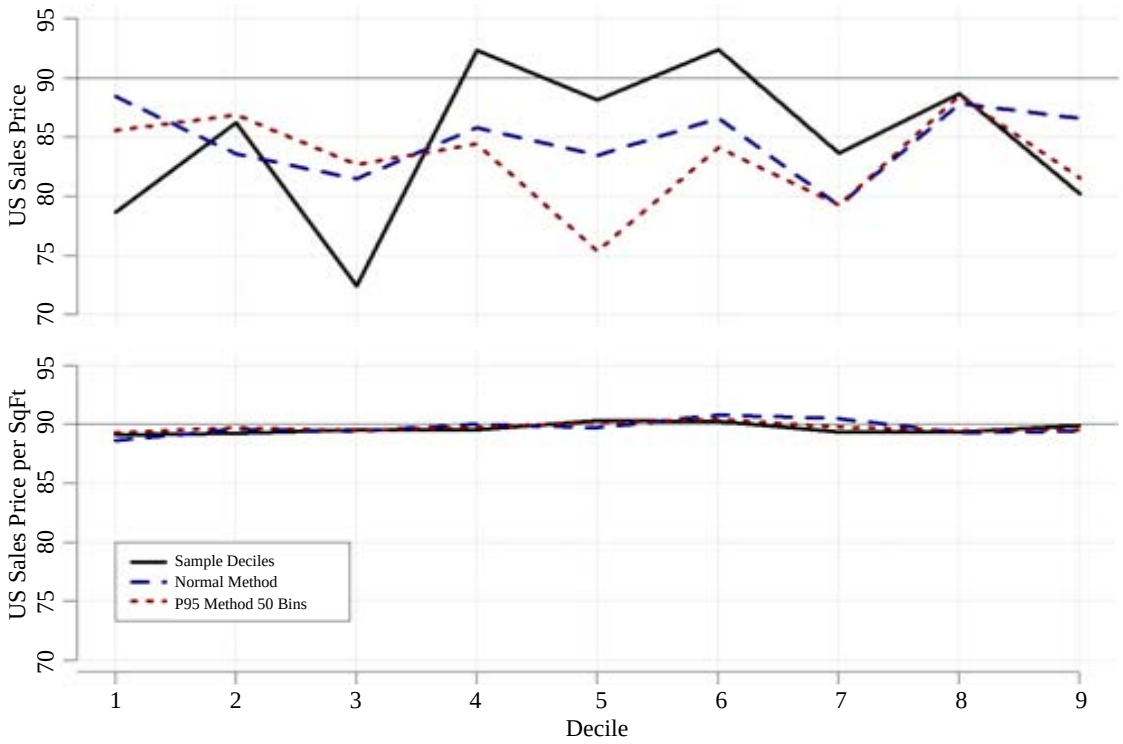


Figure 3.6 Coverage rates of sales price and sales price per square foot (Nonparametric population)

3.3.3.3 Additional simulations to assess bin size effects

Overall, the statistical properties of the P95 estimates and variance estimates obtained from the log-normal population (both characteristics) and from the nonparametric population for sales price per square foot are quite promising. However, none of the considered methods have nearly as solid properties for sales price in the nonparametric population. This is troubling, despite the previous caveats about the nonparametric population modeling.

In the nonparametric population, the sales price bins near the median contained more observations than bins in the distribution's tail [Note: this is true for both the P95 and NB methods.] These large bins can over-smooth the distribution, resulting in very stable estimates. The over-smoothing manifests itself in the replication variance method as underestimation due to lack of variability between replicate estimates for the "middle" deciles. Conversely, as expected, the SD method produces unstable estimates throughout and consequently overestimates the variance (positive bias).

The purpose of transforming the data before binning is to obtain uniform distributions within bins. For sales price, neither transformation achieves a uniform distribution within the bins, resulting in non-negligible interpolation bias, in turn affecting the MSE estimates.

To better understand how the estimation procedure affects the variance estimation procedure, recall that our variance estimates are evaluated with respect to the mean squared error (MSE) obtained under the estimation procedure. The SD procedure yields essentially unbiased estimates, but the trade-off is a large and unstable variance. Using interpolation reduces the sampling variance and improves its stability, but can substantively increase the bias squared term of the MSE.

Ultimately, the P95 method had the most promising results for most characteristics. However, there were still several concerns about the bias of the median estimate for nonparametric sales price. To address these concerns, we conducted additional simulations on both populations and characteristics, using the P95 method with 50 bins, 75 bins, and 100 bins.

In most cases, using 75 bins with the P95 method generally reduces the estimate bias and MSE without detrimentally affecting the variance estimate properties. This is definitely a balancing act. As the number of bins increases, the corresponding evaluation statistics begin to mimic those obtained with the SD method. This improves the interpolated estimates in the cases where the SD deciles had better statistical properties. However, increasing the number of bins has a detrimental effect on the statistical properties of the decile estimates and variance estimates when the P95 method with 50 bins yielded less biased estimates or more stable variance estimates.

4 Conclusion

The fundamental finding from Thompson and Sigman (2000) was that interpolation methods can be used to produce stable *median* estimates for samples from positively skewed populations, but the effectiveness of the interpolation was highly dependent on both the width of the bins and their location in the sample. Their primary contribution was to develop a data-dependent binning approach that used each individual estimation cell distribution.

Our approach to determining a decile estimation method for complex samples from a positively skewed population builds on these earlier findings, recognizing both that data-dependent binning is a necessity and that the binning method selected must account for a positively skewed distribution to facilitate the complete set of decile estimates. We considered three interpolation methods, each of which took a different approach to resolving the sparse data problem at the 90^{th} decile posed by the skewed distributions. Our empirical analysis showed that all of the studied approaches yielded complete sets of decile estimates with reasonable statistical properties, at least for the Survey of Construction. However, the properties of the corresponding MHS variance estimates were not as good and exhibited different patterns. At the U.S. level, our simulation results demonstrate consistently good statistical properties for decile estimation and variance estimation using the P95 transformation and 75 bins in one simulated population in terms of estimate bias, MSE, bias of the variance estimates, and stability, while rarely achieving a 90% coverage rate.

Of course, it is much more challenging to estimate a complete set of deciles than a single median, especially from positively skewed distributions. However, our recommended method appears to work quite well for most decile estimates and could certainly be modified to produce viable quartile estimates if the production decile estimates prove too unstable. In the meantime, the SOC program has decided to implement the P95 interpolation method and produce complete sets of decile estimates for selected annual characteristics in future reports.

Although we believe that our findings can be extended to other survey designs, we recognize that our research is conducted under extremely restrictive conditions, namely multi-stage cluster sampling from a highly skewed population, with a two-PSU per stratum design at the first stage. In other applications, interpolation with data-dependent bins could be combined with the variance estimator proposed in the 1952 Woodruff paper, as suggested by J.N.K. Rao. For surveys that are not well suited to BRR or MHS replication that publish decile estimates, our data-dependent binning and interpolation approach could be used in conjunction with a bootstrap replication method such as the Rao-Wu bootstrap (Rao and Wu 1988).

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors acknowledge Erica Filipek, Bonnie Kegan, Amy Newman-Smith for their valuable contributions to this research project. In addition, we thank Wan-Ying Chang, Laura Bechtel, Xijian Liu, J.N.K. Rao, the Associate Editor, and two anonymous referees for their helpful comments of earlier drafts of this manuscript.

References

Fay, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Lienhard, S. (2004). Multivariate Lognormal Simulation with Correlation. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=6426&objectType=File.

Rao, J.N.K., and Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

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

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

Steel, P., and Fay, R.W. (1995). Variance estimation for finite populations with imputed data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Thompson, J.R. (2000). Simulation: A Modeler's Approach. New York: John Wiley & Sons, Inc., 87-110.

Thompson, K.J. (1998). Evaluation of Modified Half Sample Replication for Estimating Variances for the Survey of Construction (SOC). Technical Report #ESM9801, available upon request to the Office of Statistical Methods and Research for Economic Programs from the U.S. Census Bureau.

Thompson, K.J., and Sigman, R.S. (2000). Estimation and replicate variance estimation of median sales prices of sold houses. *Survey Methodology*, 26, 2, 153-162.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 369-393
Statistics Canada, Catalogue No. 12-001-X

Joint determination of optimal stratification and sample allocation using genetic algorithm

Marco Ballin and Giulio Barcaroli[1]

Abstract

This paper offers a solution to the problem of finding the optimal stratification of the available population frame, so as to ensure the minimization of the cost of the sample required to satisfy precision constraints on a set of different target estimates. The solution is searched by exploring the universe of all possible stratifications obtainable by cross-classifying the categorical auxiliary variables available in the frame (continuous auxiliary variables can be transformed into categorical ones by means of suitable methods). Therefore, the followed approach is multivariate with respect to both target and auxiliary variables. The proposed algorithm is based on a non deterministic evolutionary approach, making use of the genetic algorithm paradigm. The key feature of the algorithm is in considering each possible stratification as an individual subject to evolution, whose fitness is given by the cost of the associated sample required to satisfy a set of precision constraints, the cost being calculated by applying the Bethel algorithm for multivariate allocation. This optimal stratification algorithm, implemented in an R package (SamplingStrata), has been so far applied to a number of current surveys in the Italian National Institute of Statistics: the obtained results always show significant improvements in the efficiency of the samples obtained, with respect to previously adopted stratifications.

Key Words: Genetic algorithm; Optimal stratification; Sample design; Sample allocation; R package.

1 Introduction

The optimality of a sample can be defined in terms of costs (associated to fieldwork, especially in terms of the number of units to be interviewed) and accuracy (related to the sampling variance of target estimates). Stratified sampling is a widely adopted design that may ensure savings in costs and gains in accuracy of estimates, when stratification variables are available in the sampling frame.

Many studies have been published on the problem of the optimization of stratified sample design. We can classify them accordingly to the object of the optimization:

1. the allocation has to be optimized, while stratification is considered as given;
2. stratification has to be optimized, while the allocation issue is postponed to a later stage;
3. stratification and allocation are optimized in a single step.

In the first group we can include Cochran (1977), Bethel (1985, 1989), Chromy (1987), Huddleston, Claypool and Hocking (1970), Kish (1976), Stokes and Plummer (2004), Day (2006, 2010), Díaz-García and Cortez (2008), Kozak, Zieliński and Singh (2008), Khan, Maiti and Ahsan (2010), Kozak and Wang (2010). Bethel (1985, 1989) and Chromy (1987) propose similar algorithms for the extension of the Neyman allocation to the multivariate case by using Convex Programming methods. Stokes and Plummer (2004) show how to make use of the Non Linear Programming tool available in Excel spreadsheets as a solver for the same problem. In Day (2006, 2010) the evolutionary algorithm approach is proposed to solve the multivariate allocation problem under the same setting indicated by Bethel and Chromy. In

1. Marco Ballin and Giulio Barcaroli, Istituto Nazionale di Statistica, via C.Balbo 16 - 00184 Roma (Italy). E-mail: ballin@istat.it, barcarol@istat.it.

Díaz-García and Cortez (2008), the optimum multivariate allocation problem is solved as a problem of multi-objective optimization of integers. Kozak *et al.* (2008) investigate the case of stratified two-stage sampling.

In the second group, we can consider Dalenius and Hodges (1959), Singh (1971), Hidiroglou (1986), Lavallée and Hidiroglou (1988), Gunning and Horgan (2004), Khan, Nand and Ahmad (2008). In general, the problem dealt with is related to the optimization of the stratification obtainable by one or more continuous variables, correlated to one or more target variables.

A number of papers deal with both problems (stratification and allocation) jointly. Kozak, Verma and Zieliński (2007) propose a method to obtain multivariate stratification while minimising the overall sample size. The method is defined only on a theoretical base, and the claim is that in the univariate case the optimization is not difficult, while in the multivariate case more research is needed. Keskintürk and Er (2007) make use of the genetic algorithm to solve jointly the allocation and strata boundaries problems, in the case of only one continuous stratification variable, and considering as given both the number of strata and the total sample size. The proposal of Benedetti, Espa and Lafratta (2008) is based on the use of a tree-based approach: their procedure defines a path from the null stratification towards the so called atomic stratification (characterised by the maximum number of strata, obtained by using all auxiliary variables, with the most detailed classifications), generally without reaching it, given that a number of stopping rules are used. Baillargeon and Rivest (2009, 2011) propose a method that can jointly optimise stratum boundaries and sample size, using an iterative algorithm: stratum boundaries (related to only one stratification variable) are obtained by minimising the anticipated sample size required for estimating the population total of only one survey variable (so this approach is univariate with respect to both stratification and target variables). In conclusion, most contributions in this group are dedicated to solving the problem of finding best strata boundaries for only one, continuous, auxiliary variable: only Benedetti *et al.* (2008) deal with the multivariate stratification case.

In case of categorical stratification variables, we could consider the stratification given by their Cartesian product; but when the number of produced strata is high, this could determine a huge sample size, far beyond the one affordable, or the one necessary to ensure the required precision levels. So, a crucial task is to choose the "best" auxiliary variables cross product, *i.e.*, the best partition of the frame, that takes the maximum advantage of the auxiliary information, but at the same time does not lead to an explosion of the number of the strata.

This paper proposes a solution to the problem of jointly determining the optimal stratification of a sampling frame together with the optimal sample size and allocation, in the full multivariate case (*i.e.*, with regard to both stratification and target variables). The only restriction is on the nature of the stratification variables that must be categorical (but we give indications on a suitable way to transform continuous ones into categorical ones). The proposed solution is based on the use of the genetic algorithm. The general procedure has been implemented in an R package, named `SamplingStrata`, available on the CRAN (Barcaroli, Pagliuca and Willighagen 2013a). This package makes use of a modified version of some functions of another R package, `genalg` (Willighagen 2012).

The paper is structured as follows: Section 2 contains a formalization of the optimization problem. Section 3 details how the genetic algorithm is employed in order to optimally solve the problem of finding the best stratification that allows the minimal cost of the required sample. To better illustrate this, Section 4 contains an example based on a well known dataset (the "iris flowers" data). Section 5 reports and

analyses the results of the application of the algorithm to a real survey, the *Italian Farm Structure Survey,* and these results are compared to the practical solution adopted by survey statisticians. A further application, to the *Monthly Survey on milk and milk products,* is reported in Section 6. Final conclusions are contained in Section 7.

2 Formalization of the optimization problem

Universe of alternative stratifications

We define as *sampling frame* F a set of N records containing information (organised in variables) related to N individuals of the reference population. Some variables are useful for the identification of units, while some other can be used in order to define the sampling strategy. The values of the latter (from now on: *auxiliary variables*) can be observed by means of a census, or from other sources as administrative registers.

We assume that in the frame a set of M auxiliary variables $X_m \left(m = 1, \ldots, M\right)$ are available. This set may contain different typologies of variables (nominal, ordinal, or continuous). We assume also that continuous auxiliary variables are split into classes by applying suitable transformation algorithms.

All such variables can potentially be used to stratify the units in the frame.

Under these assumptions, we can associate to each auxiliary variable a vector $d_m = \left\{x_1, \ldots, x_{k_m}\right\}$ of contiguous integer values, each of them representing an original value in the domain set.

Then, the most detailed stratification of F can be considered as the result of the Cartesian product $CP = X_1 \times X_2 \times \ldots \times X_M$.

The maximum number of strata will be $K = \prod_{m=1}^{M} k_m - I^*$, where I^* is the number of impossible or absent combinations of values in the frame. So, the most detailed stratification of the frame is such that it contains K strata, corresponding to all possible combinations of values in the M auxiliary variables. We call *atomic strata* the strata belonging to this particular stratification. Each atomic stratum is characterised by a *unique* combination of values of the M auxiliary variables. We can assign a label $l_k \left(k = 1, \ldots, K\right)$ to each atomic stratum.

If we consider the labelled set of atomic strata $L = \left\{l_1, l_2, \ldots, l_K\right\}$, we can define the set of all its possible partitions $P_1, P_2, \ldots, P_B$, where B can be calculated by using the Bell formula:

$$B_K = \sum_{i=0}^{K-1} \binom{K-1}{i} \cdot B_i \qquad \left(B_0 = 1\right)$$

We define the set $\left\{P_1, P_2, \ldots, P_B\right\}$ of partitions of L as the *universe (or space) of stratifications.*

Assessment of a given stratification

Given a partition P_i of L, characterized by H strata, let N_h and $S^2_{h,g}, h = 1, \ldots, H$, $g = 1, \ldots, G$ be respectively the number of units and variances in stratum h of the G different survey target variables

$Y_1, \ldots, Y_G$. Assuming a simple random sampling of n_h units without replacement in each stratum, the variance of the Horvitz-Thompson estimator of the total of the g^{th} target variable $\left(\hat{T}_g\right)$ is

$$\text{Var}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \quad g = 1, \ldots, G \tag{2.1}$$

Consider the following cost function

$$C\left(n_1, \ldots, n_H\right) = C_0 + \sum_{h=1}^{H} C_h n_h \tag{2.2}$$

where C_0 indicates a fixed cost (not dependent on the sample size) and C_h represents the average cost of observing a unit in stratum h.

Given $V_g \left(g = 1, \ldots, G\right)$, the upper bounds for the expected sampling variance for $\hat{T}_1, \ldots, \hat{T}_G$, the classical optimal multivariate allocation problem (Bethel 1985) can be defined as the search for the solution of the minimum (with respect to n_h) of the linear function C under the convex constraints $\text{Var}\left(\hat{T}_g\right) \leq V_g \quad g = 1, \ldots, G$:

$$\begin{cases} \min C\left(n_1, \ldots, n_H\right) = C_0 + \sum_{h=1}^{H} C_h n_h \\ \text{Var}(\hat{T}_g) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \leq V_g \qquad g = 1, \ldots, G \end{cases} \tag{2.3}$$

Bethel (1989) suggested that the problem can be more easily solved by considering the following function of n_h :

$$x_h = \begin{cases} 1/n_h & \text{if } n_h \geq 1 \\ \infty & \text{otherwise} \end{cases} \tag{2.4}$$

Using x_h the cost function can be written as

$$C\left(x_1, \ldots, x_H\right) = C_0 + \sum_{h=1}^{H} \frac{C_h}{x_h} \tag{2.5}$$

and the variances as

$$\text{Var}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{1}{x_h N_h}\right) S_{h,g}^2 x_h = \sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \quad g = 1, \ldots, G \tag{2.6}$$

Consequently, the multivariate allocation problem can be defined as the search for the minimum (with respect to x_h) of the convex function (2.5) under a set of linear constraints

$$\sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \leq V_g \quad g = 1, \ldots, G \tag{2.7}$$

An algorithm, that is proved to converge to the solution (if it exists), was provided by Bethel by applying the Lagrangian multipliers method to this problem (an easier algorithm was previously proposed

by Chromy (1987); as Bethel pointed out, the Chromy algorithm works in most of the practical cases but there is no proof that it converges if a solution exists).

The optimization approach here illustrated yields a continuous solution, which must be rounded to provide integer stratum sample sizes. The implementation we made of the Bethel algorithm provides the n_h values as the values $1/x_h$ rounded up to the upper integer.

It should be noted that the same approach can be used to deal with the multidomain problem. Let us consider the usual transformation for the domain estimation problem:

$$Y_i^d = \begin{cases} Y_i & \text{if the unit } i \text{ belongs to domain } d \\ 0 & \text{otherwise} \end{cases}$$

If the quantities previously defined to describe the Bethel approach are computed using the variables $Y^d \ (d = 1, \ldots, D)$, then the multivariate allocation solution is the solution for the multidomain case.

Selection of the best stratification on the basis of a complete enumeration

In order to choose the best stratification of a given frame, *i.e.*, the one that ensures the minimum cost $C(n_1, \ldots, n_H)$ associated to a sample whose total size and allocation are compliant to precision constraints, it is possible to proceed as follows:

- generate the most detailed stratification associated with F, that is the set L of atomic strata;
- enumerate all partitions P_i of L;
- for each partition P_i, solve the corresponding allocation problem, that is equivalent to determine the vector $(n_1, \ldots, n_H)$, and calculate the value $C_i(n_1, \ldots, n_H)$ associated to P_i;
- choose the partition P_i for which $C_i(n_1, \ldots, n_H)$ is minimized.

By so doing, the optimization of the solution is obtained by considering the whole universe of stratifications.

Unfortunately, this procedure is applicable only in situations where the dimension K of L is low: in fact, the number of partitions (given by the Bell formula) grows very rapidly (for example, $B_4 = 15$, $B_{10} = 115{,}975$ and $B_{100} \approx 4.76 \times 10^{115}$). Therefore, in most cases, the complete enumeration of the space of the solutions is not feasible. The present proposal, based on the genetic algorithm, allows to explore the universe of stratifications and to identify the one that is expected not to be far from the optimal.

The genetic algorithm

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are a particular class of evolutionary algorithms that make use of techniques inspired by evolutionary biology, such as *inheritance, mutation, selection* and *crossover* (also called *recombination*) (Vose 1999) (Schmitt 2001 and 2004).

A *GA* is implemented as an iterative computer simulation, in which an initial set of *individuals,* each one being a potential solution to the current problem (represented by a vector called *genome*), evolves by

inheritance, mutation, selection and *crossover,* increasing the average *fitness* of next *generations*. Here, the *fitness* corresponds to the objective function defined in the optimization problem so that the evolution results into the maximization (or minimization) of the objective function.

The set of individuals treated in each iteration of the GA is called *generation*. The *evolution* is the set of changes that occurs in producing consecutive *generations* by iterating the process.

At each iteration of the GA, after having evaluated the fitness of every individual in the generation, a set of individuals are stochastically selected (privileging those with higher fitness), and modified (recombined and sometimes randomly mutated) to form a new generation. This new generation is then evaluated in the next iteration of the algorithm. As individuals with the best fitness are more likely to be selected for generating individuals for the next generation, the GA produces an increase of average fitness in the course of the evolution.

The parameter *mutation rate* is expressed as the rate of *chromosomes* (the *genome* elements) that can be mutated for each individual at the moment of the generation of *children* for the next generation. A high value guarantees large differences between successive generations. It should be noted that a high mutation rate makes the GA more likely to avoid stagnating at local optima, at the price of a slower convergence to the optimal solution; whilst a low value accelerates the convergence speed, increasing the risk of local optima.

Usually, the algorithm terminates when either a maximum number of iterations has been reached, or the current solution is not improved by continuing the iteration. In both cases, the optimal solution may or may not have been reached.

3 Application of the genetic algorithm to the optimal stratification problem

On the basis of the GA setting, the stratification allocation problem can be represented as follows:

- a given stratification is considered as an *individual;*
- the *genome* of an individual is a vector whose dimension is given by the number K of atomic strata;
- each position $i\,(i = 1, \ldots, K)$ in the vector is associated to a given atomic stratum, and contains an integer value $v_i\,(1 < v_i < U)$ with $U \leq K$, where U is defined as the maximum number of strata in the final solution: if some elements of the vector have the same value, it means that the corresponding atomic strata collapse into a new stratum identified by this value;
- in this way, a stratification $P(\nu)$ can be identified by a vector $\boldsymbol{v} = [v_1, \ldots, v_K]$, where each value v_i is positionally associated to the atomic stratum identified by the label l_i and can assume an integer value internal to an interval $[1, U]$. The *space of all potential stratifications (or partitions)* $P(\nu)$ (space of solutions) is given by all possible vectors $\boldsymbol{v}$;

- the *fitness* function of an individual $P(\nu)$ is the value of the cost function $C\left(n_1,\ldots,n_{H_{P(\nu)}}\right) = C_0 + \sum_{h=1}^{H_{P(\nu)}} C_h n_h$, where the terms C_0 and C_h are given constants, and the $n_1,\ldots,n_{H_{P(\nu)}}$ are calculated by applying the Bethel algorithm to the stratification, under precision constraints set on the target variables.

It is worth while noting that, if we set $C_0 = 0$, and $C_h = 1$ for all the atomic strata, then the value of the cost function simply coincides with the sample size required to satisfy precision constraints.

Having defined a suitable representation of the domain of all possible solutions, and the fitness function to be calculated for each solution, in the following it is reported how *GA* operates.

Step 0: Creation of the initial generation of individuals

The first step consists in forming an initial set of different stratifications (the initial generation of individuals): on the basis of the value of the parameter *size of the generations,* p different individuals are generated. This means that, for the j^{th} individual, K integer values (one for each element of the vector representing the genome) are randomly generated from a uniform distribution in the interval $[1,U]$. Fixing $U \leq K$ we can set an upper limit to the maximum number of distinct aggregate strata.

Step 1: Evaluation of fitness for each individual in the population

For each individual in the population (that is for each one of p stratifications), its related fitness is evaluated by calculating the total cost required to satisfy precision constraints on the G different $\hat{T}_g$ estimates (in order to remove the dependence on the scale (or range) of the values associated with the G target variables, instead of considering the constraints expressed in the (2.7) as an upper limit to the variance of the target variables, we set constraints on their coefficient of variation $\text{CV} = \sqrt{\text{var}(\hat{T}_G)}\big/\hat{T}_G$). The evaluation is carried out by applying the Bethel algorithm, that requires as input, for each stratum of the current solution:

- means and standard deviations of target variables;
- cost of interviewing per unit;
- population (number of units).

Each one of the above items is computed on the basis of corresponding values in the atomic strata.

Let us consider a particular partition $P(\nu)$ of L determined by a given solution $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_K]$. Let $D_i\left(i = 1,2,\ldots,Q_{P(\nu)}\right)$ be one stratum in this partition. There are two possibilities:

1. D_i coincides with an atomic stratum l_k;
2. $D_i = \left\{l_j^i,\ldots,l_l^i\right\}$ is the result of the aggregation of a subset $\left\{l_j^i,\ldots,l_l^i\right\}$ of the atomic strata.

In the first case, means and variances of target variables in the stratum are known. In the second case, means and variances in D_i may be calculated by using the following formulas:

$$\bar{Y}_{g,D_i} = \frac{\sum_{l_k \in D_i} \bar{Y}_{g,l_k} N_{l_k}}{\sum_{l_k \in D_i} N_{l_k}} \tag{3.1}$$

$$S^2_{g,D_i} = \left(\sum_{l_k \in D_i} N_{l_k} - 1 \right)^{-1} \left\{ \sum_{l_k \in D_i} \left(N_{l_k} - 1 \right) S^2_{g,l_k} + \sum_{l_k \in D_i} N_{l_k} \left(\bar{Y}_{g,l_k} - \bar{Y}_{g,D_i} \right)^2 \right\} \tag{3.2}$$

where:

$\bar{Y}_{g,D_i}$ and $\bar{Y}_{g,l_k}$ are the mean values in aggregated stratum D_i and atomic strata l_k;

N_{l_k} is the number of units in atomic stratum l_k;

S^2_{g,D_i} and S^2_{g,l_k} are the variances in aggregated stratum D_i and atomic strata l_k.

The expected cost of observing a unit in a given aggregate stratum is calculated by averaging the costs in each contributing atomic stratum, weighted by their population:

$$C_{D_i} = \frac{\sum_{l_k \in D_i} C_{l_k} N_{l_k}}{\sum_{l_k \in D_i} N_{l_k}} \tag{3.3}$$

Finally, we can compute the population in any aggregate stratum as the sum of the units in the contributing atomic strata:

$$N_{D_i} = \sum_{l_k \in D_i} N_{l_k} \tag{3.4}$$

So, in correspondence of each potential solution, we are able to calculate dynamically all the information required to apply the optimal allocation algorithm, that produces the total cost

$$C\left(n_1, \ldots, n_{H_{P(v)}}\right) = C_0 + \sum_{h=1}^{H_{P(v)}} C_h n_h$$

that is the fitness of the individual.

Step 2: Breeding a new generation

Once the fitness of each individual is evaluated, a proportion of them are selected to breed a new generation. Individuals are selected through this fitness-based process, where fitter individuals are more likely to be selected, while only a small proportion of less fit individuals are selected. The presence of this second component helps to keep the diversity of the generation large enough, preventing premature convergence on poor solutions. There is also the option of indicating the number of the best individuals (expressed as a percentage of the p size of the generation) that in any case must be present also in the next generation (parameter *elitism*).

The next generation will thus be composed by a number of individuals from the previous generation (the best ones), plus a number of "children", obtained by selecting and crossing "parents" from the current

generation. In the *GA* approach, the *genome* of a "child" individual is formed using the *crossover* and *mutation* operators:

- *crossover:* many crossover techniques exist for *GA*, which use different data structures and different criteria of chromosomes selection, but the general approach is to exchange a subset of chromosomes between two parents. In our implementation, once two parents have been selected with probability proportional to their fitness, a *crossover-point* is generated, still on a random basis. This crossover-point is an integer belonging to the interval $[1, K]$. Let c be this generated crossover-point: then, the child individual will be formed by inheriting the first c chromosomes from the first parent, and the remaining $(K - c)$ chromosomes from the second parent;
- *mutation:* given the probability that an arbitrary value in a genetic sequence will be changed from its original state *(mutation chance)*, *GA* proceeds to draw, for each chromosome in the genome, a random value to decide if the value will be changed or not.

By applying the above methods of crossover and mutation, a new individual is created which typically shares many of the characteristics of its "parents". New parents are selected to produce new children, and the process continues until a new generation of individuals (stratifications) of appropriate size is generated.

Step 3: Iteration and stopping criteria

Usually, the average fitness is increased moving from one generation to the next. Steps 1 and 2 are repeated until a termination condition has been reached. Common terminating conditions are:

1. the maximum number of iterations has been reached;
2. a "plateau" has been reached, such that successive iterations no longer produce better results;
3. a combination of the above.

In our case, the terminating condition can be considered as a combination of the above. Actually, the used rule is the maximum number of iterations, but this number is determined by analysing previous runs, in order to detect the "plateau" and be sure that additional iterations are not likely to improve the final solution.

Critical parameters of the optimal stratification algorithm

Here a distinction is made between the parameters that are common to genetic algorithm, and the ones that are peculiar to the particular problem to which it is applied, *i.e.*, the optimal stratification of a population frame (the names of the parameters are those used in the R package `SamplingStrata`).

Among the first we list:

- size of generation of individuals *(pop)*;
- number of iterations *(iterations)*;
- mutation chance *(mut_chance)*;
- elitism (elitism_rate).

Instead, the context parameters are:

- minimum number of units per stratum *(minnumstrat)* (the Bethel algorithm is forced to allocate in each stratum at least the number of units indicated by this parameter);
- initial number of strata *(initialStrata)*;
- possibility to increase the maximum number of strata *(addStrataFactor).*

As for the first group, there are no strict rules to assign values to these parameters. Given a particular problem, it is suggested to carry out a number of trials in order to assess the sensitivity of the solutions to the values of the parameters.

It is important to take into account that parameters as *size of generation* and *elitism* are in general influent on the rapidity of convergence, and not so much on the final solution, given that a "reasonable" number of iterations is given.

The reasonability of the parameter *number of iterations* can be assessed by analysing the behaviour of the fitness function: if the values of this function are no longer decreasing after a certain number of iterations, it is reasonable to expect that to increase the number of iterations will not produce better results.

On the contrary, the value of *mutation chance* has effects on both rapidity of convergence and the goodness of the final solution: a high mutation chance allows to avoid local minima, at the cost of a slower convergence.

Conversely, parameters of the second group should be given on the basis of practical considerations, related to the characteristics and requirements of the survey that is under design.

As for the parameter *minimum number of units per stratum,* if an adequate number of observations in all strata is to be ensured (in order to take into account the expected non response, the need of calculating sampling variance, fieldwork reasons, *etc*.), a value can be set higher than the default one (which is set to 2).

The parameter *initial number of strata* is very important. First of all, its value, if associated with a value of the parameter *addStrataFactor* equal to zero, determines the maximum acceptable number of strata in the final solution. This possibility may be useful not only for fieldwork reasons (if, for example, for organizational considerations the number of strata is to be limited), but especially because the final solution is very sensitive to the value of this parameter. We have experimented that if the algorithm with different values of *initialStrata* is run, from low values up to the maximum given by the number of atomic strata, solutions can be very different. It is possible to let the algorithm to choose for us, in this way: we set *initialStrata* by assigning a low value to it, together with a high value of parameter *addStrataFactor* (the parameter *addStrataFactor* is used to increase dynamically the value set by parameter *initialStrata:* each time a mutation takes place, a random number between 0 and 1 is generated, and if it is greater than the quantity (1-*addStrataFactor*), the maximum number of strata is increased of one unit) (by default, it is equal to 0). Manoeuvring these two parameters, there are different possibilities:

1. for any given value of *initialStrata,* if *addStrataFactor* is set equal to 0, then the algorithm has to consider that value as a fixed limit, and all solutions to be explored will be characterised by that maximum number of strata;

2. otherwise, if *addStrataFactor* is set to a value greater then 0, then the algorithm may explore solutions varying the number of strata, from an initial value given by *initialStrata,* up to a maximum number given by the number of atomic strata.

4 An example based on the *Iris flowers* dataset

To show how to apply the algorithm for finding the optimal stratification, the well known *Iris flowers* dataset can be considered. This dataset consists of a total of 150 observations, equally distributed by the three species of Iris flowers (*setosa, virginica* and *versicolor*). Four features are measured for each observation (*i.e.*, the length and the width of sepal and petal, in centimetres).

We will consider this dataset as a possible sampling frame from which to draw a sample, under a stratified design, in order to estimate two target variables:

- Y_1 : Petal.Length;
- Y_2 : Petal.Width.

For sake of simplicity, we suppose there are only two auxiliary variables available in the frame:

- X_1 : Sepal.Length;
- X_2 : Species.

While the second auxiliary variable is categorical, the first one is continuous, and needs to be transformed into a categorical ordered variable. To this aim, we make use of the k*-means univariate clustering method* (Hartigan and Wong 1979), obtaining the following ranges: [4.3; 5.5], (5.5; 6.5], (6.5; 7.9].

The Cartesian product of the two auxiliary variables should produce $3 \times 3 = 9$ different strata. Actually, one of these contains no units, the one related to Species = "setosa" and Sepal.Length $\in$ (6.5; 7.9]. So the one reported in table 4.1 will be considered as the initial atomic stratification.

Table 4.1
Information concerning atomic strata

stratum	X_1 = Sepal.Length	X_2 = Species	N	Y_1 = Petal.Length		Y_2 = Petal.Width		cost
				Mean	Standard deviation	Mean	Standard deviation	
1	[4.3; 5.5] (1)	Setosa (1)	45	1,47	0,17	0,24	0,11	1
2	[4.3; 5.5] (1)	Versicora (2)	6	3,58	0,49	1,17	0,21	1
3	[4.3; 5.5] (1)	Virginica (3)	1	4,50	0,00	1,70	0,00	1
4	[5.5; 6.5] (2)	Setosa (1)	5	1,42	0,17	0,26	0,08	1
5	[5.5; 6.5] (2)	Versicora (2)	35	4,27	0,37	1,32	0,19	1
6	[5.5; 6.5] (2)	Virginica (3)	23	5,23	0,32	1,95	0,29	1
7	[6.5; 7.9] (3)	Versicora (2)	9	4,68	0,19	1,46	0,11	1
8	[6.5; 7.9] (3)	Virginica (3)	26	5,88	0,49	2,11	0,23	1

For sake of simplicity, we assume that the fixed cost C_0 is null, and all C_h are set equal to 1: by so doing, the cost of a solution coincides with the sum of sampling units allocated in the strata, *i.e.*, with the total sample size $\left(C = n = \sum_{h=1}^{H} n_h\right)$.

We set as precision constraints to the estimates of both target variables an upper limit of 0.05 (5%) to their expected coefficient of variation.

Finally, we set a minimum number of units to be selected in each stratum equal to 2 (the minimum required in order to calculate sampling variance).

Under these assumptions, and using the atomic stratification, the Bethel algorithm solves the optimal allocation problem by defining a minimum sample size of 17 units, with an allocation vector $\mathbf{a} = (2, 2, 1, 2, 3, 3, 2, 2)$.

If we proceed to partition the set of atomic strata, the resulting number of all possible stratifications (given by the Bell formula) is $B_8 = 4{,}140$. This number is such that we can afford to enumerate all partitions of atomic strata, and for each of them we are able to calculate the minimum sample size by applying the Bethel algorithm (to enumerate all the partitions in this example, we made use of the function `setparts()`, contained in the R package `partitions` (Hankin 2011)).

The range of sample sizes steps from a minimum of 11 to a maximum of 78 (this latter corresponds to the "*no stratification* solution") (see figure 4.1).

We notice that the minimum value $(n = 11)$ that has been found is considerably lower than the one calculated in correspondence with the atomic stratification $(n = 17)$. This minimum value characterizes only 8 partitions out of 4,140.

Now, the genetic algorithm is applied in order to evaluate its capability to find the optimal solution (or at least one that is not far from it), without being obliged to explore all solutions, but only a strict subset of them.

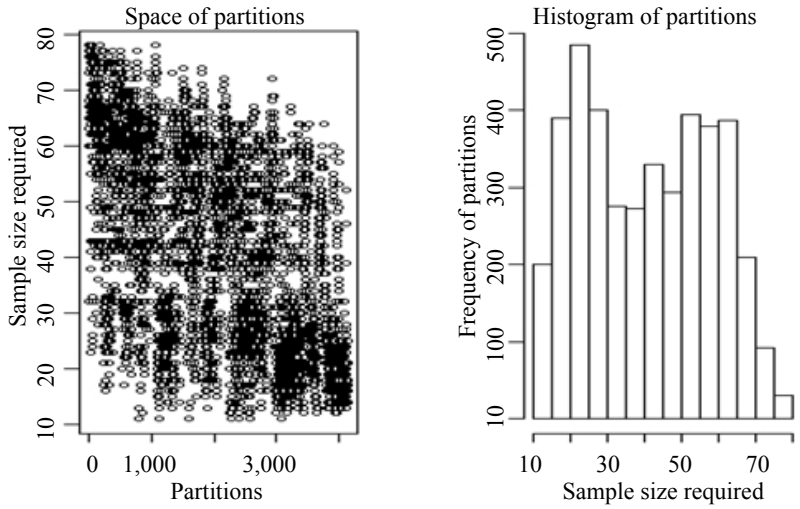


Figure 4.1 Space of partitions

Step 0: Creation of the initial generation

First, we set $U = 8$ (we can accept a number of final strata that is equal to the number of atomic strata, so $U = K$). The *generation size* parameter *pop* is set equal to 10. So, an initial set containing 10 different individuals (stratifications) is generated. Each of them is represented by a vector of 8 elements, *i.e.*, the number of different atomic strata. An individual $\boldsymbol{v} = (1, 2, 3, 4, 5, 6, 7, 8)$ or, equivalently, $\boldsymbol{v} = (3, 6, 4, 2, 1, 8, 7, 5)$ corresponds to the most detailed stratification (as all strata are labelled with different labels), while $\boldsymbol{v} = (1, 1, 1, 1, 1, 1, 1, 1)$ or equivalently $\boldsymbol{v} = (4, 4, 4, 4, 4, 4, 4, 4)$ corresponds to "null stratification" (as atomic strata are labelled with identical labels).

Step 1: Evaluation of fitness for each individual in the generation

To each one of the 10 individuals in the current generation, the Bethel algorithm is applied in order to find the cost of the sample required to comply with fixed precision constraints.

To do this, first of all related strata and information are calculated for each individual. For example, for a generated individual $\boldsymbol{v} = (4, 1, 1, 4, 8, 7, 8, 1)$ the information is derived by the one available from atomic strata, by applying (3.1) and (3.2) (see table 4.2).

Table 4.2
Information concerning generated aggregated strata

Aggregated stratum	Original atomic strata	(X_1, X_2)	N	Y_1		Y_2	
				Mean	Standard deviation	Mean	Standard deviation
1	2,3,8	(1,2) or (1,3) or (3,3)	33	5.41	1.01	1.92	0.44
2	1,4	(1,1) or (2,1)	50	1.46	0.17	0.25	0.10
3	6	(2,3)	23	5.23	0.31	1.95	0.28
4	5,7	(2,2) or (3,2)	44	4.35	0.37	1.35	0.18

The fitness of this individual is measured by the corresponding required sample size, that results to be 14, with an allocation vector $\mathbf{a} = (6, 2, 3, 3)$.

All individuals are sorted accordingly with their performance: the individual in the first position is the one supporting the minimum sample size, the 10^{th} individual is the one requiring the maximum sample size.

Step 2: Breeding a new generation

By setting the *elitism* parameter to 20% (a common default value) we always take the best 2 individuals in the current generation and directly move them to the next generation, without any change of their genome.

Then, we proceed in generating new individuals in the following way:

1. we select couples of individuals of the current generation with probability proportional to their fitness: for instance, assume to select $\mathrm{v}_k = (1,1,3,4,3,2,2,2)$ and $\mathrm{v}_j = (2,2,2,2,2,1,1,1)$;
2. a crossover point is randomly generated, *i.e.*, an integer internal to the interval $[1,8]$: suppose to set it equal to 3;
3. the crossover is performed by assigning to the child the first three elements of parent v_k and the last five elements of parent v_j, obtaining in this way $\mathrm{v}_{\text{new}} = (1,1,3,2,2,1,1,1)$;
4. having set a *mutation rate* parameter equal to 0.05, for each element of the child a random number is generated in the interval $[0,1]$: if it is less than 0.05, the value of the element is changed (by generating a new value comprised between 1 and 9), otherwise it is not changed.

Step 3: Iteration and stopping criteria

The number of iterations has been set equal to 25. So, steps 1 and 2 are repeated 25 times. The individual with the best fitness alongside all the generations is retained as the best solution.

The graph in figure 4.2, obtained during the execution of the program, shows the convergence of the algorithm. In the graph, two different curves are reported: the lower one is related to the best solution found until the k^{th} iteration (as the best solution is memorised, it can only decrease as the algorithm proceeds); the upper one reports the mean of the 10 solutions evaluated in each iteration.

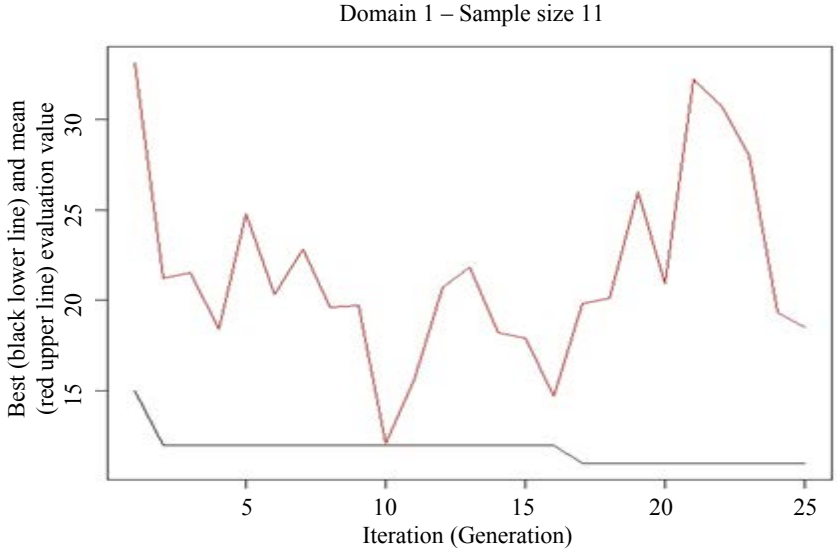


Figure 4.2 Best and mean evaluation values during GA execution

The resulting best solution is $\boldsymbol{v} = (4,1,3,4,1,3,3,2)$. It corresponds to the stratification reported in table 4.3, with an allocation vector $\mathbf{a} = (3,2,4,2)$.

Table 4.3
Information concerning final strata

Aggregated stratum	Original atomic strata	(X_1, X_2)	N	Y_1		Y_2	
				Mean	Standard deviation	Mean	Standard deviation
1	2,5	(1,2) or (2,2)	41	4.16	0.45	1.30	0.19
2	8	(3,3)	26	5.88	0.49	2.10	0.22
3	3,6,7	(1,3) or (2,3) or (3,2)	33	5.06	0.38	1.80	0.33
4	1,4	(1,1) or (2,1)	50	1.46	0.17	0.25	0.10

In conclusion, by applying the genetic algorithm, we succeeded in finding the optimal solution by exploring only $25 \times 10 = 250$ alternative stratifications instead of the 4,140 belonging to the universe of partitions.

In order to verify that this result is not due to a "lucky strike", we perform different executions of the algorithm: each execution iterates 10 times the application of the genetic algorithm, varying the values of the parameter "number of iterations". Results are reported in table 4.4.

Table 4.4
Capability of GA to find the optimal solution

Execution of the GA (10 times each)	Value of parameter "number of iterations" in the GA	Solutions with n = 11 (optimal)	Solutions with n = 12	Solutions with n = 14
(a)	25	5	4	1
(b)	50	7	3	-
(c)	100	9	1	-
(d)	200	10	-	-

In execution (a), we discover that, with only 25 iterations, to succeed in finding the optimal solution is actually a "lucky strike", as in half of the trials the found solution is higher than the optimal. But increasing the number of the iterations up to 200 (execution (d)), the genetic algorithm proves to be reliable with respect to its capability to reach optimality, as in all the trials the optimal solution is found.

As for the number of the strata corresponding to the found optimal solutions, on average it is 4, with a range of $[3, 5]$.

Finally, we also want to verify that the found solutions are compliant with the precision constraints (maximum CV equal to 5% for both target variables). So, in execution (d) (iterations = 200), for each one of the 10 produced solutions we proceed to draw 1,000 samples from the frame and to calculate the related CV's. Corresponding results are shown in figure 4.3: the average of CV's for the first target variable (Petal.Lenght) is around 3%, while for the second one is around 5%. So, we can say that, on average, precision constraints have not been violated.

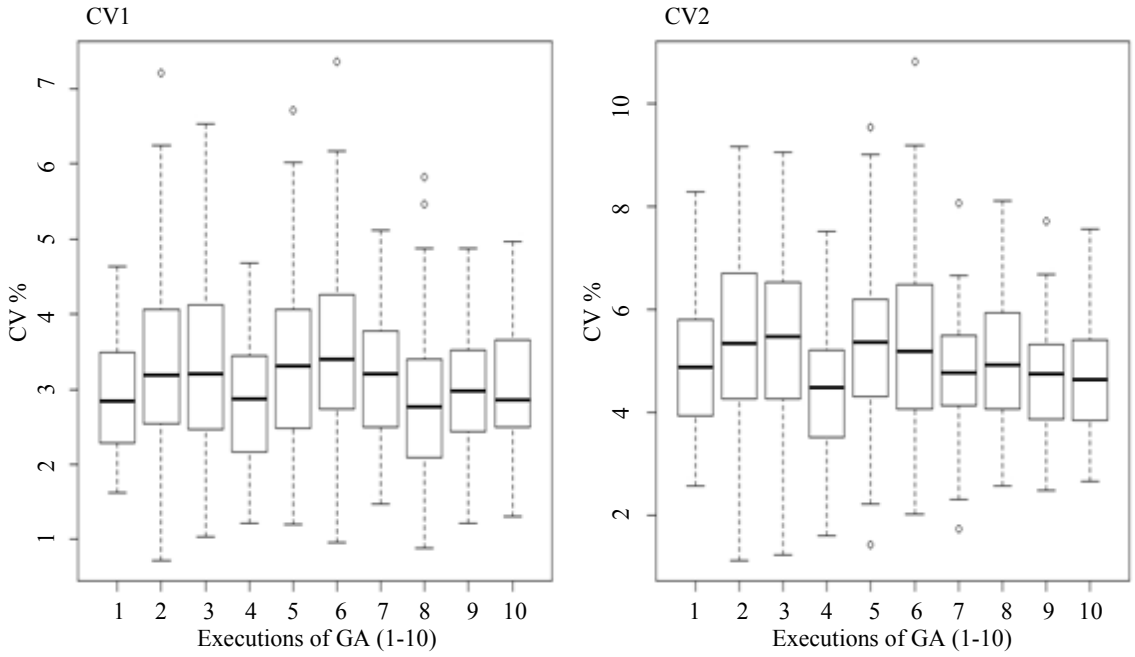


Figure 4.3 Distributions of CV's for target variables in the simulation

A more complete example involving the use of all the functions in the package `SamplingStrata` is reported in Barcaroli (2013b).

5 An application: the *Italian Farm Structure Survey* (FSS)

The sampling frame used for the selection of *2003 Italian Farm Structure Survey* (FSS) sample contains 2,153,710 farms. For the purposes of FSS sample design, the auxiliary variables considered are the following:

1. regions (21 different values);
2. provinces (103 different values);
3. legal status (2 classes);
4. sector of economic activity (9 classes);
5. economic size unit (3 classes);
6. agricultural area utilized (3 classes);
7. livestock unit (3 classes);
8. altimetry of the headquarter of the holding (5 classes).

Fourteen different target variables have been considered as the main target of FSS, on which the required precision levels (in terms of maximum coefficient of variations) have been fixed at regional level (domains of interest). The list of variables and related precision constraints are reported in table 5.1.

Both the 8 auxiliary and the 14 target variables have been observed during the previous 2000 Agricultural Census, so their values are available for each unit in the frame. This gives the possibility to calculate means and standard deviations related to whichever defined stratum.

Firstly, the current "manual" procedure followed in 2003 to choose the most suitable stratification for sample selection is described.

2003 manual configuration of strata to select the FSS sample

In the first step, a take-all stratum was defined in each region on the basis of local characteristics. The thresholds for defining the take-all strata were determined using the Hidiroglou method (1986).

In the second step, a choice between a stratification based on provinces or on the region as a whole, was chosen region by region, on the basis of local organizational considerations.

In the third step, the other six variables were alternatively used in each region or provinces (depending on the result obtained in the second step) as stratification variables. For each of such alternative stratifications, the optimal sample size was computed (the minimum sample size in each stratum had been fixed to 50) (in the cost function, fixed cost has been set to zero, and variable costs were set equal to 1 in each atomic stratum: so the cost function coincides with the total sample size). The stratification supporting the overall minimum sample size in each region (usually defined on different variables) was considered as the output of this step.

In the fourth step, the remaining five variables were used separately to refine the stratification previously obtained. For each of these refined stratifications the optimal sample size was computed considering the same constraints used in the third step.

This stepwise procedure was repeated on a regional basis, by refining the best stratification obtained in each step, using the remaining available variables until the obtained stratification revealed to be less efficient than the stratification in the previous step.

By so doing, the total amount of planned sample size was fixed to 42,465 units (actually, the sample size used for 2003 FSS was increased to 52,713 in order to obtain better estimates at national level. Here we consider the number of 42,465 to correctly compare the results obtained with the genetic algorithm).

Use of the genetic algorithm to identify optimal strata and best allocation

The most detailed available stratification of the frame, obtained as a Cartesian product of all the auxiliary variables, consists of 24,454 different strata, 1,787 of which have been defined as take-all strata. So, the atomic strata are given by the 22,667 sampling strata obtained by subtracting the 1,787 take-all strata. The latter are collapsed in only one stratum, whose 6,971 units will always be selected for whatever sample.

Actually, one of the auxiliary variables, *region,* is considered as the domain variable. So, our task consists in optimising the frame stratification and the sample allocation distinctly for each one of the different 21 Italian regions. For instance, the first region (Piemonte) is characterised by 105,074 units in 1,646 sampling strata, and 597 units in 129 take-all strata.

Precision constraints (once again expressed in terms of upper limits on coefficients of variation) have been set, for each one of the 14 different target variables, at the same values chosen on the occasion of

manual configuration of strata carried out for the 2003 survey: they are 5%, 6% or 10% for the most important variables in each region. Table 5.1 reports the complete set of the coefficient of variations used in planning the 2003 FSS.

Table 5.2 reports the results of the two solutions in terms of required sample size: the one planned in 2003 by the expert sample designer of the FSS (column 6), and the one obtained by applying the genetic algorithm (column 7).

As the determination of the best stratification has been carried out separately for each region, 21 independent results can certify the great convenience of the algorithm in most domains. A dramatic decrease of the required overall sample size can be observed, as shown by a 38.17 % saving on the previous total. This result is differentiated region by region, with a maximum decrease for Sardegna (-57.85%) and a minimum for Sicilia (-20.61%). Also in terms of strata, from the initial number of atomic strata (22,667), a huge reduction occurs to the final stratification, characterised by only 213 different strata (ranging from a minimum of 6 strata in region Friuli, up to 22 strata in Sicilia).

Table 5.1
Maximum expected coefficients of variation (%) used in the 2003 FSS

Region	Cereals	Industrial crops	Fresh vegetables	Flowers	Vineyards	Olives	Citrus fruit	Fruits	Bovines	Pigs	Sheep	Economic size units	Utilized agricultural surface	Livestock unit
Piemonte	5.0	10.0			5.0				5.0			5.0	6.0	6.0
Val d'Aosta									5.0			5.0	6.0	6.0
Lombardia	5.0	10.0							5.0	5.0		5.0	6.0	6.0
Bolzano								5.0				5.0	6.0	6.0
Trento								5.0				5.0	6.0	6.0
Veneto	5.0	10.0			5.0					5.0		5.0	6.0	6.0
Friuli V.G.	5.0	10.0										5.0	6.0	6.0
Liguria				5.0								5.0	6.0	6.0
Emilia R.	5.0	10.0			5.0			5.0	5.0	5.0		5.0	6.0	6.0
Toscana	5.0	10.0			5.0							5.0	6.0	6.0
Umbria						5.0						5.0	6.0	6.0
Marche												5.0	6.0	6.0
Lazio	5.0		5.0		5.0	5.0						5.0	6.0	6.0
Abruzzi						5.0						5.0	6.0	6.0
Molise						5.0						5.0	6.0	6.0
Campania	5.0	10.0	5.0			5.0		5.0				5.0	6.0	6.0
Puglia	5.0		5.0		5.0	5.0						5.0	6.0	6.0
Basilicata	5.0											5.0	6.0	6.0
Calabria	5.0					5.0	5.0					5.0	6.0	6.0
Sicilia	5.0		5.0		5.0	5.0	5.0				5.0	5.0	6.0	6.0
Sardegna	5.0										5.0	5.0	6.0	6.0

As for the setting of the parameters used to obtain the above result, the most important revealed to be the following:

1. number of iterations (or generations);
2. generation size (number of individuals, or solutions, evaluated at each iteration);
3. mutation chance;
4. initial number of strata;
5. factor for increasing the initial number of strata.

Table 5.2
2003 FSS sample size determination: Comparison of results

(1) Domain (region)	(2) Total number of units in the frame	(3) Number of atomic sampling strata in the frame	(4) Number of units in the sampling strata	(5) Number of units in take-all strata	(6) Sample size by 2003 stratification	(7) Sample size by Genetic Algorithm solution	(8) Number of strata in GA solution	(9) % relative difference (7) *vs* (6)
Piemonte	105,671	1,646	105,074	597	2,687	1,497	9	-44.29
Valle d'Aosta	6,125	65	6,074	51	408	317	7	-22.30
Lombardia	71,257	1,902	69,495	1,762	3,428	2,151	7	-37.25
Bolzano	23,362	127	23,202	160	692	430	7	-37.86
Trento	30,021	124	29,908	113	676	523	7	-22.63
Veneto	176,999	1,450	176,064	935	3,531	1,868	11	-47.10
Friuli	32,981	638	32,805	176	807	498	6	-38.29
Liguria	29,992	584	29,967	25	766	485	7	-36.68
Emilia R.	103,702	2,157	102,922	780	2,584	2,022	11	-21.75
Toscana	107,288	1,959	106,964	324	2,099	1,337	16	-36.30
Umbria	46,074	435	45,897	177	1,354	751	7	-44.53
Marche	60,439	1,005	60,271	168	918	488	8	-46.84
Lazio	162,109	1,304	161,801	308	3,233	2,216	14	-31.46
Abruzzi	67,117	888	66,941	176	1,035	743	10	-28.21
Molise	28,890	375	28,834	56	1,190	630	6	-47.06
Campania	212,145	1,271	211,833	312	2,559	1,883	13	-26.42
Puglia	288,087	1,026	287,877	210	4,712	2,009	14	-57.36
Basilicata	68,470	504	68,355	115	703	493	7	-29.87
Calabria	145,812	1,624	145,654	158	2,798	1,792	17	-35.95
Sicilia	295,637	2,345	295,472	165	3,955	3,140	22	-20.61
Sardegna	91,532	1,238	91,329	203	2,330	982	7	-57.85
Italia	**2,153,710**	**22,667**	**2,146,739**	**6,971**	**42,465**	**26,255**	**213**	**-38.17**

Their final values have been determined, after numerous trials, on the basis of the analysis of the runs for each region.

In particular, by inspecting the convergence plot, it is possible to understand if the number of iterations is sufficient to ensure that the final solution is definitely the best obtainable, or if otherwise a higher number of iterations is needed. This can be done by analysing the behaviour of the two curves in the plot: the lower one reports the *best* evaluation value, while the upper one refers to the *mean* evaluation value. When the mean evaluation value is still decreasing, together with the best evaluation value, it is worthwhile to go on iterating. When the best value line becomes stably constant (and typically the mean value line begins to fluctuate up and down), no further gain can be expected by new iterations. This is the case, for instance, of the convergence plot for Trento region, shown in figure 5.1.

A convenient value for *iterations* parameter was found to be 5,000. As for the mutation chance, a suitable value was found to be 0.001: this means that, for any chromosome in the genome (any value in vector $\boldsymbol{v}$), a mutation occurred on average only once out of a thousand. A critical point is in fixing the initial number of strata. Since the final solution is very sensitive on the number of strata, we decided to let the algorithm to choose it. This can be done, as already said in section 4, by assigning a low value to *initialStrata*, and by giving a value greater than zero to *addStrataFactor:* this enables the algorithm to explore solutions characterized by a wide range of number of strata. In our experiment, we set the initial number of strata to the value 5, while assigning a value 0.01 to the factor for increasing the initial number of strata (this means that, any time a mutation occurs, there is a probability of 1% to increase by 1 the current number of strata).

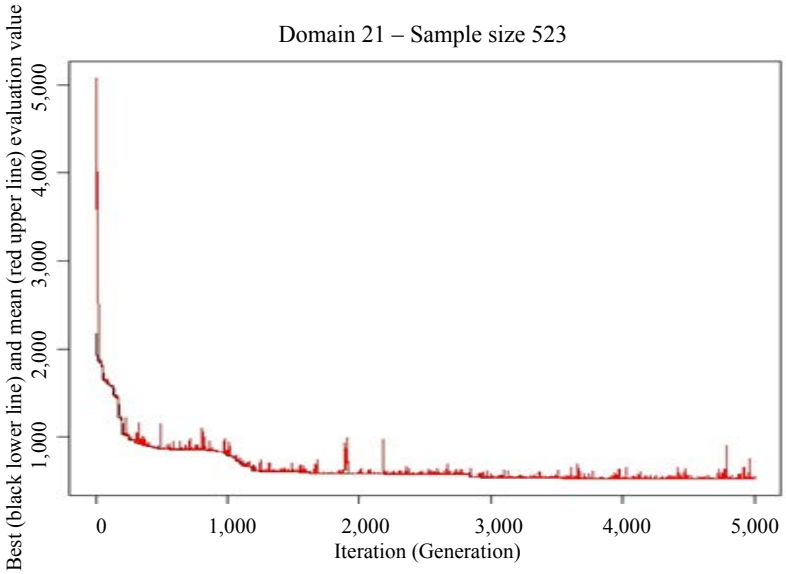


Figure 5.1 Best and mean evaluation value for the Trento region

From a computational point of view, the overall task required an elapsed time of 641,820 seconds (more than 178 hours, nearly one week) (the job was run on a desktop AMD Athlon 64×2 (2.90 Ghz, 3 GB RAM)).

6 A further application: the *Monthly survey on milk and milk products*

A further application of our algorithm concerned the *2010 Monthly survey on milk and milk products*. This is a sample survey that depends strictly on the "Annual survey on milk and milk products", which is a census of all Italian farms producing milk and milk products. Both surveys collect the same information: the amount of milk collected at the national level and its use (in processing dairy products: milk, cheese, butter, *etc.*); the purpose of the monthly sample survey is to obtain timely information before the results of the annual survey (carried out in the year before) become available. The sample for 2010 has been planned in this way:

1. the information collected on the 2,250 respondent units in the 2008 round of the Annual survey were organised as a frame: in particular, four of the target variables of the Annual survey, which are continuous, were transformed into categorical variables (ordered factors) by using the k-means clustering method, and considered as auxiliary information in the frame;
2. the cross-product of the obtained categorical variables, produced a stratification of the frame consisting in 152 (atomic) strata;
3. the information related to means and standard deviations of the four target variables of the monthly survey were calculated for each one of the atomic strata by using Annual survey data.

Constraints on the coefficients of variation of the estimates of the totals are reported in table 6.1.

Table 6.1
Coefficients of variation (%) used in planning the 2010 Monthly Survey on Milk

Variable	Maximum acceptable CV on total estimates (%)
Collected milk	1
Milk	15
Butter	3.8
Cow's milk cheeses	3

After this, the Bethel algorithm was applied in order to verify the sample size required with the initial (atomic) stratification available for the frame (also in this application, the function cost coincides with the total sample size, as the fixed cost was set to zero, and variable costs were set equal to 1 in each atomic stratum): it resulted in 290 units to be interviewed, allocated in the 152 different strata. The usual procedure terminates here: at this point, the 290 units would be selected from the frame represented by the Annual Survey, and the Monthly Survey would start.

Instead, the application of the genetic algorithm suggested a collapsing of the 152 initial atomic strata into 88 aggregate strata, requiring a sample size of only 247 to satisfy the same constraints, with a consequent decrease of about 15%.

After a considerable amount of attempts, the following values were given to the most important parameters:

1. generation size was set equal to 50;
2. the number of iterations was set equal to 4,000;
3. a minimum of 2 units per stratum was required;
4. the initial number of strata (coinciding with the maximum number of them, as parameter *addStrataFactor* was set to 0) was set equal to the number of atomic strata (152);
5. the mutation chance was set to 0.0005.

The combination of parameter "generation size" and "number of iterations" determined the evaluation of 200,000 $(50 \times 4{,}000)$ solutions. The convergence plot reported in figure 6.1 shows that after 2,700/2,800 there has been no further improvement of the identified best solution.

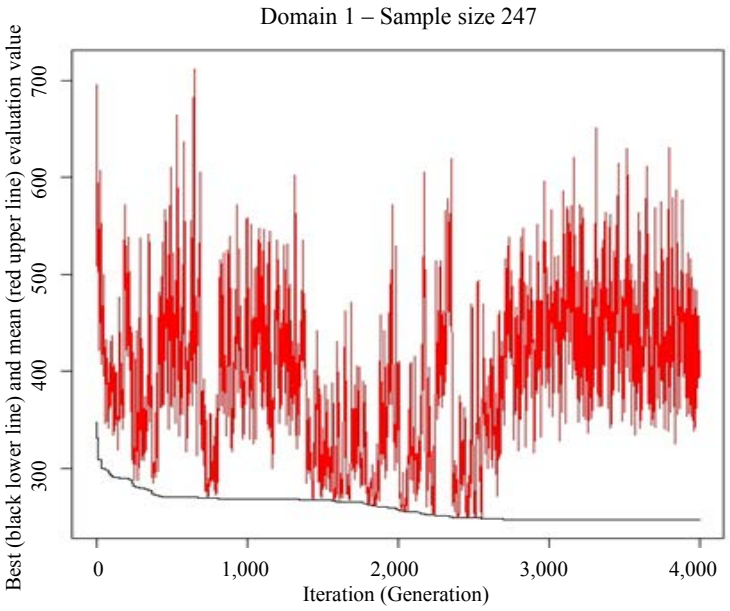


Figure 6.1 Best and mean evaluation value in the optimization of Monthly Survey on Milk

7 Conclusions and future work

For any given multipurpose and multidomain sample survey, the optimal stratification of the sampling frame can be determined together with the optimal sample size and allocation of units among strata, by means of a combined use of the Bethel algorithm (or, more generally, of a NLP solver) for the determination of the minimum sample size required to satisfy precision constraints, and of the genetic algorithm for the exploration of the universe of potential stratifications, rigorously generated accordingly to the theory of partitions. The information required is nearly the same as the one required by the allocation problem: desired precision on estimates of total (or means) of target variables, and information

regarding the distributions of each target variable in population strata. Initial stratification should be considered at the most detailed level (atomic stratification), *i.e*. the one determined by the Cartesian product of values of all available stratification variables.

The complete exploration of the set of all possible stratifications is in practical cases computationally prohibitive. The use of the genetic algorithm permits to explore the space of solutions in a very efficient manner. By carefully tuning the execution parameters, it is possible to determine the optimal solution, or at least a solution likely to be not far from the optimal one.

The application of this algorithm to two different surveys (the 2003 *Italian Farm Structure Survey* and the 2010 *Monthly milk and milk products*) shows that the obtained solutions are much better, in terms of sample efficiency, than the ones manually produced by expert methodologists (in Istat, the algorithm has been applied to three more surveys: *"Economic outcomes of agricultural holdings", "Structure and production of main wooden cultivations", "Survey on forecasting of some herbal crops sowing"*).

In all the cases reported, it has been possible to calculate the values required as input to our algorithm (in particular: means and standard deviations of the target variables in the different atomic strata), because of the availability of related values for each unit in the frames. In more realistic situations, this kind of information is not directly available. Instead, we could use estimates produced by alternative sources: administrative data, other surveys, or previous rounds of the same survey, or even hypothesis (usually conservative) on the variability of target variables within the strata. Accordingly to Rivest (2002) it is also possible to model target variables assuming auxiliary variables X's as explanatory variables, in order to estimate means and standard deviations on the basis of predicted values of Y's. Of course, the less "direct" is the information on the target variables, the less robust is the proposed method, because of the uncertainty caused by the use of proxy information, or model-based predictions.

Another limit affecting this approach still lies in the handling of continuous auxiliary variables. In our approach, we simply suggest to transform them into categorical ones, in order to be considered in the determination of the universe of all possible stratifications of the sampling frame. A first element for future work is in giving indications on how to transform these variables in order to get the best from them. A second one is in the fact that some of the strata contained in the optimal solution may be characterized by non contiguous values of the transformed continuous variables, or of the categorical ordinal variables, which is something odd that should not be allowed: this could be prevented by imposing constraints on the generation of candidate solutions.

References

Baillargeon, S., and Rivest, L.-P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77, 3, 331-344.

Baillargeon, S., and Rivest, L.-P. (2011). The construction of stratified designs in R with the package *stratification*. *Survey Methodology,* 37, 1, 53-65.

Barcaroli, G., Pagliuca, D. and Willighagen, E. (2013a). SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys. R package version 1.0-1. http://cran.r-project.org/web/packages/SamplingStrata/index.html.

Barcaroli, G. (2013b). Optimization of sampling strata with the SamplingStrata package. http://cran.r-project.org/web/packages/SamplingStrata/vignettes/SamplingStrataVignette.pdf.

Benedetti, R., Espa, G. and Lafratta, G. (2008). A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology,* 34, 2, 195-203.

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. *American Statistical Proceedings of the Survey Research Methods Section,* 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology,* 15, 1, 47-57.

Chromy, J.B. (1987). Design optimization with multiple objectives. Proceedings of the American Statistical Association Section on Survey Research Methods 1987, 194-199.

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

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of American Statistical Association,* 54, 88-101.

Day, C.D. (2006). Application of an evolutionary algorithm to multivariate optimal allocation in stratified sampling designs. *Proceedings of the American Statistical Association Section on Survey Research Methods 2006* [CD-ROM].

Day, C.D. (2010). A multi-objective evolutionary algorithm for multivariate optimal allocation. *Section on Survey Research Methods - JSM 2010,* 3351-3358.

Díaz-García, J.A., and Cortez, L.U. (2008). Multi-objective optimisation for optimum allocation in multivariate stratified sampling. *Survey Methodology,* 34, 2, 215-222.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology,* 30, 2, 159-166.

Hankin, R.K.S., and West, L.J. (2007). Set Partitions in R. *Journal Of Statistical Software,* Code Snippet 2. December 2007, 23, http://www.jstatsoft.org/.

Hankin, R.K.S. (2011). Partitions: Additive partitions of integers. R package version 1.9-19. http://cran.r-project.org/web/packages/partitions/index.html.

Hartigan, J.A., and Wong, M.A. (1979). A k-means clustering algorithm. *Applied Statistics,* 28, 100-108.

Hidiroglou, M.A. (1986). The costruction of self-representing stratum of large units in survey design. *The American Statistician,* 40, 27-31.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Applied Statistics,* 19, 273-278.

Keskintürk, T., and Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics and Data Analysis,* 15 September 2007, 52, 1, 53-67.

Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology,* 34, 2, 205-214.

Khan, M.G.M., Maiti, T. and Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics,* 26, 4, 695-708.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society,* Series A, 159, 80-95.

Kozak, M., Verma, M.R. and Zieliński, A. (2007). Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition,* 8(2), 223-250.

Kozak, M., Zieliński, A. and Singh, S. (2008). Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages*. Statistics & Probability Letter,* June 2008, 78, 8, 970-974.

Kozak, M., and Wang, H.Y. (2010). On stochastic optimization in sample allocation among strata. *Metron - International Journal of Statistics,* LXVIII, 1, 95-103.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology,* 14, 1, 33-43.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology,* 28, 2, 191-198.

Schmitt, L.M. (2001). Theory of genetic algorithms. *Theoretical Computer Science,* 259, 1-61.

Schmitt, L.M. (2004). Theory of genetic algorithms II: Models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling. *Theoretical Computer Science,* 310, 181-231.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variables. *Journal of the American Statistical Association,* 66, 829-833.

Stokes, L., and Plummer, J. (2004). Using spreadsheet solvers in sample design. *Computational Statistics & Data Analysis,* 44, 527-546.

Vose, M.D. (1999). The Simple Genetic Algorithm: Foundations and Theory, MIT Press, Cambridge, MA.

Willighagen, E. (2012). Genalg: R Based Genetic Algorithm. R package version 0.1.1. http://cran.r-project.org/web/packages/genalg/index.html.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 395-417
Statistics Canada, Catalogue No. 12-001-X

An appraisal-based generalized regression estimator of house price change

Jan de Haan and Rens Hendriks[1]

Abstract

The house price index compiled by Statistics Netherlands relies on the Sale Price Appraisal Ratio (SPAR) method. The SPAR method combines selling prices with prior government assessments of properties. This paper outlines an alternative approach where the appraisals serve as auxiliary information in a generalized regression (GREG) framework. An application on Dutch data demonstrates that, although the GREG index is much smoother than the ratio of sample means, it is very similar to the SPAR series. To explain this result we show that the SPAR index is an estimator of our more general GREG index and in practice almost as efficient.

Key Words: Generalized regression estimation; House price index; Property assessments; Sampling.

1 Introduction

When attempting to construct constant-quality house price indexes, statistical agencies face a number of problems. First, exact matching of properties over time is problematic as their quality will likely have changed; houses depreciate and they may also have had major repairs, additions or remodelling done to them. In other words, every property in each period can be viewed as a unique good. Second, the turnover of houses is generally low compared to the housing stock and the mix of properties sold changes over time, so a quality mix problem arises. Third, there is often a lack of data on characteristics. Data availability issues have implications for the choice of measurement method.

Three main types of house price indexes can be found in the literature: median or mean indexes, repeat sales indexes and hedonic indexes. A median (mean) index tracks the change in the price of the median (mean) house traded from one period to the next. This method is problematic in that the characteristics of, *e.g.*, the median house changes over time. The problem is often tackled by stratifying the samples according to region, type of dwelling, *etc*., a procedure which is also known as *mix adjustment*. Stratification obviously requires additional data.

Repeat sales methods address the quality mix problem by restricting the data set to houses that have been sold twice or more during the sample period. This ensures that 'like is compared with like', assuming that the quality of the individual houses remains unchanged. Repeat sales methods are based on regressions where the repeat sales data pertaining to different periods are pooled. A potential drawback is revision; when new data is added to the sample, previously computed index numbers will change. The repeat sales method is originally due to Bailey, Muth and Nourse (1963). Case and Shiller (1987, 1989) argue that changes in house prices include components whose variances increase with the interval of sales and propose a Weighted Least Squares approach to adjust for this type of heteroskedasticity. An alternative weighted method has been suggested by Calhoun (1996). Jansen, de Vries, Coolen, Lamain

1. Jan de Haan, OTB Research Institute for the Built Environment, Delft University of Technology and Division of Process Development, IT and Methodology, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands. E-mail: j.dehaan@cbs.nl; Rens Hendriks, Division of Economic and Business Statistics and National Accounts, Statistics Netherlands. E-mail: r.hendriks@cbs.nl.

and Boelhouwer (2008), using Dutch data, compare the unweighted repeat sales method with various weighted methods and conclude that the unweighted method performs satisfactorily.

Unlike repeat sales methods, *hedonic regression methods* can in principle adjust for quality changes of individual properties (in addition to quality mix changes). These methods utilize information on housing characteristics, such as number of bedrooms, lot size and location, to estimate quality adjusted price indexes using regression techniques. Today, hedonic house price indexes are computed in many countries. For example, the French statistical agency (INSEE), jointly with Conseil Supérieur du Notariat, compiles a hedonic index (Gouriéroux and Laferrère 2009) as does Statistics Finland (Saarnio 2006). The UK has three hedonic house price indexes, compiled by different institutes. RPData-Rismark computes hedonic indexes for the capital cities in Australia (Hardman 2011). Hedonic indexes come in two main varieties. The time dummy method models the log of price as a function of property characteristics and a set of dummy variables indicating the time periods. Since the data of all periods are pooled, this method suffers from revision as well. Hedonic imputation methods, which estimate the 'missing prices', do not have this drawback. Hill and Melser (2008) discuss numerous hedonic imputation methods in the housing context. Diewert, Heravi and Silver (2009) and de Haan (2010) provide a comparison between time dummy and hedonic imputation price indexes.

A fourth approach to estimating house price indexes is the *use of assessment or appraisal data*. One option is to augment a repeat sales dataset by using assessment data as estimates for past or current values of properties that have not been resold during the sample period. Some of the data on which the repeat sales index is based would then be pseudo rather than genuine repeat data. For more on the use of assessment information in a repeat sales price index and the removal of appraisal bias, see *e.g.*, Geltner (1996), Edelstein and Quan (2006), and Leventis (2006). Another option, which also controls for quality-mix changes, is to combine current selling prices with appraisals from an earlier period to compute price relatives in a standard matched-model framework. An advantage over the repeat sales approach is that index numbers will not be revised. This so-called Sale Price Appraisal Ratio (SPAR) method has been applied in New Zealand for a long time now and is currently also being used in the Netherlands and a few other European countries. Bourassa, Hoesli and Sun (2006) describe the New Zealand SPAR index which is compiled by Quotable Value, a state-owned property valuation company. Other studies into the SPAR method include Rossini and Kershaw (2006), van der Wal, ter Steege and Kroese (2006), de Vries, de Haan, van der Wal and Mariën (2009), de Haan, van der Wal and de Vries (2009), Shi, Young and Hargreaves (2009), and Grimes and Young (2010).

In this paper we outline an alternative appraisal-based method to measure house price change. The appraisals serve as auxiliary information in a *generalized regression* (GREG) estimation framework. GREG is a model-assisted technique that can be used to increase efficiency as compared to simpler estimators such as sample means (Särndal, Swensson and Wretman 1992), provided that population information is known for one or more variables that exhibit a strong linear correlation with the variable under study. In our case we regress selling prices in each time period on appraisals. Appraised values are available in the Netherlands for all properties in stock in some reference period, and we expect them to be highly collinear with selling prices. Although the method is based on regression, the resulting price index is not a hedonic index as the regression model is descriptive rather than explanatory.

The paper is organized as follows. To set the stage, in Section 2 we describe the SPAR method and its relation to the sample means of sale prices and appraisals. Due to compositional change and the relatively

low number of transactions, the Dutch SPAR series exhibits strong volatility, especially for small market segments. In Section 3 we outline a simple GREG estimator of house price change and two alternatives. The first alternative is a stratified version of the original index whereas the second one uses an alternative model specification. Section 4 contains empirical evidence using Dutch data. The GREG index numbers turn out to be very similar to the SPAR index numbers and are equally volatile. In Section 5 we explain this result by showing that the SPAR index is in fact an estimator of the GREG index and almost as efficient. Section 6 concludes and suggests a topic for further research in this field.

2 Horvitz-Thompson estimators and the SPAR index

The typical aim of survey sampling is to estimate the total or (arithmetic) mean of some variable for a finite population. In a housing context we may want to estimate the total value of the housing stock in, say, period 0. Let U^0 denote the housing stock of size N^0 and p_n^0 the value of house $n\left(n = 1,\ldots, N^0\right)$. The target to be estimated is

$$V^0 = \sum_{n\in U^0} p_n^0. \tag{2.1}$$

Suppose we have a sample S^0 consisting of n^0 houses sold in the base period. If the houses were selected by simple random sampling from the housing stock U^0, where each house had the same inclusion probability, then the Horvitz-Thompson estimator

$$\hat{V}^0 = \left(N^0/n^0\right)\sum_{n=1}^{n^0} p_n^0 \tag{2.2}$$

is an unbiased estimator of (2.1); see *e.g.*, Cochran (1977).

A natural target – though not the only possibility – for a house price index would be the value change of a fixed housing stock. Conditioning on the *base period stock* has two implications: additions to the stock (mostly newly-built houses) should be excluded and the price changes of existing properties should be adjusted for quality changes, *i.e.*, for the impact of depreciation, renovations and extensions. For convenience we assume that such quality changes are negligible. In that case the target price index going from the base period 0 to the comparison period $t\left(> 0\right)$ is defined as

$$P^{0t} = \frac{\sum_{n\in U^0} p_n^t}{\sum_{n\in U^0} p_n^0}, \tag{2.3}$$

with obvious notation. Suppose that we also have a sample S^t, consisting of n^t houses sold in period t and assume that it is an independent random draw from the base period stock. The ratio of the Horvitz-Thompson estimators (the sample means) in both periods

$$\hat{P}^{0t} = \frac{\left(N^0/n^t\right)\sum_{n\in S^t} p_n^t}{\left(N^0/n^0\right)\sum_{n\in S^0} p_n^0} = \frac{\sum_{n\in S^t} p_n^t/n^t}{\sum_{n\in S^0} p_n^0/n^0} \tag{2.4}$$

might seem a natural estimator of our target index (2.3). However, if the samples S^0 and S^t are independently drawn, the variance of estimator (2.4) can be substantial. Moreover, an estimated ratio such as (2.4) has a bias that depends on the variance of the numerator and the covariance of the numerator and the denominator (Cochran 1977). From an index number perspective the issue at stake is that the mix of properties traded in period t differs from that in period 0. That is, we are not comparing like with like.

The standard approach to estimating price indexes relies on the matched model methodology where prices p_n^0 and p_n^t are observed for a fixed panel of items. The use of panel data ensures that like is compared with like and will reduce the variance of the ratio estimator because p_n^0 and p_n^t are typically positively correlated. However, unless the samples S^0 and S^t are extraordinary large, there will only be few matched houses, if any. Hence, while prices p_n^t are observed for the houses belonging to S^t, for most of those houses the base period prices p_n^0 are 'missing'. What may be available instead are government assessments a_n^0. We could use these as base period values and construct the following (pseudo) matched-model estimator of house price change:

$$\tilde{P}^{0t} = \frac{\sum_{n\in S^t} p_n^t/n^t}{\sum_{n\in S^t} a_n^0/n^t}. \tag{2.5}$$

A problem associated with estimator (2.5) is that the base period index number will differ from 1 because the appraisals a_n^0 differ from the selling prices p_n^0. Rescaling (2.5) by dividing it by its base period value is an obvious solution, yielding

$$\hat{P}^{0t}_{\text{SPAR}} = \frac{\sum_{n\in S^t} p_n^t/n^t}{\sum_{n\in S^t} a_n^0/n^t}\left[\frac{\sum_{n\in S^0} p_n^0/n^0}{\sum_{n\in S^0} a_n^0/n^0}\right]^{-1} = \frac{\sum_{n\in S^t} p_n^t/n^t}{\sum_{n\in S^0} p_n^0/n^0}\left[\frac{\sum_{n\in S^0} a_n^0/n^0}{\sum_{n\in S^t} a_n^0/n^t}\right]. \tag{2.6}$$

Note that the rescaling factor is stochastic, as it is a ratio of sample means for the base period, and will increase the variance of (2.6) as compared to the estimator given by (2.5), depending on the correlations between the appraisals and the selling prices. Details can be found in de Haan (2007). But we cannot circumvent rescaling since a price index that does not start at the value 1 would be meaningless.

Expression (2.6) is called a Sale Price Appraisal Ratio (SPAR) index. The SPAR method has been applied in the Netherlands since January 2008 to measure the price change of owner-occupied dwellings. As mentioned earlier, we assume that the SPAR index aims at tracking the price change of the *housing stock*, which is a measure of the change in wealth. In the context of the Harmonized Index of Consumer Prices on the other hand, the house price index should measure the price change of the *houses sold* during the base period (Makaronidis and Hayes 2006; Eurostat 2010). Under the latter concept there would be no sampling involved if all transactions are recorded and used in the compilation of the index, as is the case in the Netherlands.

The second expression on the right-hand side of (2.6) writes the SPAR index as the product of two factors, the ratio of sample means and a factor between brackets. As the SPAR index is essentially based on the matched model methodology (using base period appraisals instead of sale prices), this factor adjusts the ratio of sample means for changes in the quality mix of the samples that occur between period 0 and period t. A potential problem is that the SPAR index is *not a panel-type estimator*. A SPAR time series, say for periods $t = 0,\ldots,T$, might therefore suffer from short-term volatility due to mix changes, especially when the number of sales is low.

3 Generalized regression estimation

3.1 A simple GREG method

In this section we will outline an alternative approach to measuring house price change that makes use of appraisal data. The appraisals now serve as auxiliary information in a generalized regression (GREG) framework. Consider the following simple two-variable linear regression model:

$$p_n^0 = \alpha^0 + \beta^0 a_n^0 + \varepsilon_n^0, \tag{3.1}$$

where ε_n^0 is the error term. Unlike hedonic regression models, which postulate a causal relation between the selling price p_n^0 and a set of characteristics relating to the structure and the location of the housing units, this model does not say anything about how house prices are generated; equation (3.1) is merely a descriptive model.

Estimating model (3.1) by least squares regression on the data of sample S^0 yields predicted prices

$$\hat{p}_n^0 = \hat{\alpha}^0 + \hat{\beta}^0 a_n^0. \tag{3.2}$$

The regression residuals for $n \in S^0$ are $e_n^0 = p_n^0 - \hat{p}_n^0$. Assuming random sampling, as before, we can write the Horvitz-Thompson estimator $\sum_{n\in S^0} p_n^0 / n^0$ of the mean value $\sum_{n\in U^0} p_n^0 / N^0$ as

$$\sum_{n\in S^0} p_n^0 / n^0 = \sum_{n\in S^0} \hat{p}_n^0 / n^0 + \sum_{n\in S^0} e_n^0 / n^0 = \hat{\alpha}^0 + \hat{\beta}^0 \sum_{n\in S^0} a_n^0 / n^0 + \sum_{n\in S^0} e_n^0 / n^0. \tag{3.3}$$

Replacing the sample average of appraisals, $\sum_{n\in S^0} a_n^0 / n^0$, by its population counterpart $\sum_{n\in U^0} a_n^0 / N^0$ yields the generalized regression (GREG) estimator:

$$\hat{\bar{P}}_{\text{GREG}}^0 = \hat{\alpha}^0 + \hat{\beta}^0 \sum_{n\in U^0} a_n^0 / N^0 + \sum_{n\in S^0} e_n^0 / n^0 = \sum_{n\in U^0} \hat{p}_n^0 / N^0 + \sum_{n\in S^0} e_n^0 / n^0. \tag{3.4}$$

Model-assisted sampling theory shows that GREG estimators are *asymptotically design unbiased* (Särndal, *et al.* 1992), irrespective of the choice of regressors. Unless the sample would be small, the bias can be neglected. It is obvious that the GREG estimator (3.4) will be more efficient – in the sense that it has a lower variance – than the Horvitz-Thompson estimator (3.3). As a result, the GREG estimator will usually outperform the Horvitz-Thompson estimator in terms of the mean square error (the sum of the variance and the squared bias).

The same procedure can be applied to the comparison period t. After estimating the model

$$p_n^t = \alpha^t + \beta^t a_n^0 + \varepsilon_n^t \tag{3.5}$$

through least squares regression on the data of the current period sample S^t, we obtain predicted prices

$$\hat{p}_n^t = \hat{\alpha}^t + \hat{\beta}^t a_n^0, \tag{3.6}$$

which lead to the GREG estimator of the mean value of the housing stock in period t :

$$\hat{\bar{P}}_{\text{GREG}}^t = \hat{\alpha}^t + \hat{\beta}^t \sum_{n \in U^t} a_n^0 / N^t + \sum_{n \in S^t} e_n^t / n^t = \sum_{n \in U^t} \hat{p}_n^t / N^t + \sum_{n \in S^t} e_n^t / n^t, \tag{3.7}$$

where $e_n^t = p_n^t - \hat{p}_n^t$ denote the period t regression residuals. For a fixed housing stock we have $U^t = U^0$, hence $\sum_{n \in U^t} a_n^0 / N^t = \sum_{n \in U^0} a_n^0 / N^0$, and it follows that

$$\hat{\bar{P}}_{\text{GREG}}^t = \hat{\alpha}^t + \hat{\beta}^t \sum_{n \in U^0} a_n^0 / N^0 + \sum_{n \in S^t} e_n^t / n^t = \sum_{n \in U^0} \hat{p}_n^t / N^0 + \sum_{n \in S^t} e_n^t / n^t. \tag{3.8}$$

The GREG estimator of house price change results simply from taking the ratio of equations (3.8) and (3.4):

$$\hat{P}_{\text{GREG}}^{0t} = \frac{\hat{\bar{P}}_{\text{GREG}}^t}{\hat{\bar{P}}_{\text{GREG}}^0} = \frac{\hat{\alpha}^t + \hat{\beta}^t \bar{a}^0 + \sum_{n \in S^t} e_n^t / n^t}{\hat{\alpha}^0 + \hat{\beta}^0 \bar{a}^0 + \sum_{n \in S^0} e_n^0 / n^0} = \frac{\sum_{n \in U^0} \hat{p}_n^t / N^0 + \sum_{n \in S^t} e_n^t / n^t}{\sum_{n \in U^0} \hat{p}_n^0 / N^0 + \sum_{n \in S^0} e_n^0 / n^0}, \tag{3.9}$$

where $\bar{a}^0 = \sum_{n \in U^0} a_n^0 / N^0$. Some additional small sample bias will be introduced due to the non-linear (ratio) structure. When using Ordinary Least Squares (OLS) regression to estimate the models (3.1) and (3.5), the unweighted sample means of regression residuals in (3.9), $\sum_{n \in S^0} e_n^0 / n^0$ and $\sum_{n \in S^t} e_n^t / n^t$, will be equal to 0 and the GREG index reduces to

$$\hat{P}_{\text{GREG,OLS}}^{0t} = \frac{\sum_{n \in U^0} \hat{p}_n^t / N^0}{\sum_{n \in U^0} \hat{p}_n^0 / N^0} = \frac{\hat{\alpha}^t + \hat{\beta}^t \bar{a}^0}{\hat{\alpha}^0 + \hat{\beta}^0 \bar{a}^0} = \frac{\hat{\alpha}^t / \bar{a}^0 + \hat{\beta}^t}{\hat{\alpha}^0 / \bar{a}^0 + \hat{\beta}^0}. \tag{3.10}$$

As the first expression on the right-hand side of (3.10) indicates, the (OLS) GREG approach essentially imputes prices pertaining to the base period and the current period using equations (3.2) and (3.6). The difference with the hedonic *double imputation* method is twofold: a descriptive model, not a hedonic one, is used to estimate predicted prices – so that we cannot speak of unbiased predicted prices – and prices are imputed for all houses of the housing stock instead of the sub-set of sampled houses.

3.2 Properties of the GREG index

The (OLS) GREG index has several properties worth mentioning. First, the computation of the GREG index is very simple. Once the population mean of appraisals $\bar{a}^0$ and the base period regression coefficients $\hat{\alpha}^0$ and $\hat{\beta}^0$ have been calculated, all that is needed is running a regression each month of

selling prices against appraisals and plugging the coefficients $\hat{\alpha}^t$ and $\hat{\beta}^t$ into (3.10). Note that the GREG index can be written as a pseudo chain index:

$$\hat{P}^{0t}_{\text{GREG,OLS}} = \frac{\hat{\alpha}^t/\bar{a}^0 + \hat{\beta}^t}{\hat{\alpha}^0/\bar{a}^0 + \hat{\beta}^0} = \prod_{\tau=1}^{t} \frac{\hat{\alpha}^\tau/\bar{a}^0 + \hat{\beta}^\tau}{\hat{\alpha}^{\tau-1}/\bar{a}^0 + \hat{\beta}^{\tau-1}}. \tag{3.11}$$

This can be helpful in practice, particularly when new appraisal data becomes available. New appraisal data often becomes available to the statistical agency with a considerable time lag, up to more than a year. There are two reasons for using the latest appraisal information. The quality of the appraisals may improve over time, which seems to have been the case in the Netherlands (de Vries *et al*. 2009). Also, the assumption of a fixed housing stock can be relaxed so that newly-built properties can be incorporated through chaining; the resulting chained GREG index takes the dynamics of the housing stock into account. The same advantages of chaining apply to the SPAR method. Suppose new appraisals, relating to period $T\,(0 < T \leq t)$, are available in period $t+1$. The time series can then be updated through chain-linking, *i.e.*, by multiplying $\hat{P}^{0t}_{\text{GREG,OLS}}$ by the month-to-month change $\left(\tilde{\alpha}^{t+1}/\bar{a}^T + \tilde{\beta}^{t+1}\right)\Big/\left(\tilde{\alpha}^t/\bar{a}^T + \tilde{\beta}^t\right)$, where the coefficients now pertain to a regression of selling prices on the period T appraisals.

Second, *standard errors* of the GREG index can be estimated rather easily using the variance-covariance matrix of the regression coefficients, which is standard output of most statistical packages. An expression for the approximate standard error is derived in the Appendix. The standard error of the GREG index depends on the goodness of fit (R^2) of the regression model. It is most likely that R^2 for the base period regression is higher than that for the current period regressions. This is because we expect to find a strong linear relation between appraisals and sale prices in the appraisal reference period while in later periods this relation will probably be weaker due to differing price trends across different types of houses or regions. The derivation of approximate standard errors for the SPAR index is a bit more complex because there is an additional source of sampling error, namely the sampling variability of the mean appraisals; see de Haan (2007).

The latter point brings us to the third property of the GREG index, namely its dependence on the *quality of the appraisal data*. For two reasons at least the appraisals may not exactly represent the transaction prices during the base period so that the model fit is not perfect $(R^2 < 1)$. The assessment authorities may not have (real time) access to the actual sale prices and therefore have to make their own judgements based on other information. But even if they knew the selling prices, the authorities may still decide to make adjustments when determining the property values. It can be argued that selling prices do not always properly measure the unknown market values – which can be seen as a latent variable – and tend to be more volatile. In this respect, Francke (2010) and others have used the term transaction noise.

The way in which the appraisals have been determined will affect the standard error of the GREG index. As long as the quality of the appraisal data is the same for all houses in stock, no bias arises since the appraisals only serve as an auxiliary variable in regressions run on the samples S^0 and S^t of properties sold in periods 0 and $t\,(t = 1, \ldots, T)$. However, in general we expect the quality of the appraisals to be higher for properties belonging to the appraisal reference (base) period sample S^0, although this will most likely differ across valuation methods. In the Netherlands the properties are assessed for tax purposes, both for income tax and local taxes. The municipalities are responsible for the valuations. Several municipalities value the houses which are sold during the reference period (January)

by the selling price. Houses which were not sold are sometimes valued by comparing them to similar traded houses. Some municipalities apparently use a form of hedonic regression to value the houses, but the methodology is unfortunately not made publicly available. For more information on the Dutch appraisal system, see de Vries *et al*. (2009).

So far we have assumed that the quality of the individual houses stays the same over time. This is a strong assumption. Thus, the fourth property – and most important drawback – of the GREG method is that the resulting price index suffers from *quality change bias* since explicit quality adjustments are not carried out. The same drawback holds true for the SPAR method and for the standard repeat sales method. In principle, hedonic regression methods can deal with the quality change problem, although it may prove difficult to control for all relevant price determining characteristics, in particular micro location. The SPAR method automatically controls for micro location, provided of course that the appraisals sufficiently account for this, as it is based on the matched-model methodology where the matching is done at the address level.

3.3 Alternative GREG estimators

Statistics Netherlands not only computes house price indexes for the whole country but also for segments of the housing market, according to type of house (family dwellings and apartments) and region (provinces and large cities), mainly because of user needs. Another motivation behind stratifying the sample can be to mitigate the effect of *sample selection bias*. This type of bias may arise if the set of houses sold in a particular period is not a random selection from the housing stock. The nationwide index should then be indirectly computed as a weighted average of the stratum indexes instead of directly from all observations.

Suppose the total housing stock U^0 is sub-divided into K non-overlapping strata U_k^0 of size $N_k^0 \left(\sum_{k=1}^K N_k^0 = N^0\right)$. The target price index (2.3) can now be rewritten as

$$P^{0t} = \frac{\sum_{n\in U^0} p_n^t}{\sum_{n\in U^0} p_n^0} = \frac{\sum_{k=1}^K \sum_{n\in U_k^0} p_n^t}{\sum_{k=1}^K \sum_{n\in U_k^0} p_n^0} = \sum_{k=1}^K s_k^0 P_k^{0t}, \tag{3.12}$$

where $P_k^{0t} = \sum_{n\in U_k^0} p_n^t \Big/ \sum_{n\in U_k^0} p_n^0$ is the target price index for stratum $U_k^0 \left(k = 1,\ldots,K\right)$. The base period stock value shares $s_k^0 = \sum_{n\in U_k^0} p_n^0 \Big/ \sum_{n\in U^0} p_n^0$, which serve as weights for the stratum indexes, are unknown and have to be estimated. Assuming the variables that define the strata are known for all $n \in U^0$, a natural choice for the weights would be the appraisal shares $\hat{s}_k^0 = \sum_{n\in U_k^0} a_n^0 \Big/ \sum_{n\in U^0} a_n^0 = \left(N_k^0 / N^0\right)\left(\overline{a}_k^0 / \overline{a}^0\right)$. Obviously, the stratum-defining housing variables should be included in the appraisal data set. In the Netherlands address and type of dwelling are included. This allows a sub-division of the population into cross classifications of location and type of dwelling. Appraisals may not always be accurate estimates of the 'true' market values of the individual properties

but at the stratum level we expect the accuracy of the average appraisals to be sufficient for the computation of the weights.

Statistical techniques such as GREG estimation are typically applied to estimate totals or means for small domains for which the number of observations is so small that the standard errors using traditional (Horvitz-Thompson) estimators – in our case the ratio of sample means – would become unacceptably high. It should be mentioned that, even with the GREG method, the stratification scheme should not be too detailed since that might unduly raise the variance of the stratum indexes and hence of the aggregate index. More importantly perhaps, small sample bias will increase and may become non-negligible with very small samples.

OLS regressions of selling prices on appraisals should now be run in every time period for each stratum in order to compute the aggregate GREG index. The stratified (OLS) GREG index is

$$\hat{P}^{0t}_{\text{StrGREG}} = \sum_{k=1}^{K} \hat{s}_k^0 \hat{P}^{0t}_{k,\text{GREG,OLS}} = \sum_{k=1}^{K} \hat{s}_k^0 \left(\frac{\hat{\alpha}_k^t / \overline{a}_k^0 + \hat{\beta}_k^t}{\hat{\alpha}_k^0 / \overline{a}_k^0 + \hat{\beta}_k^0} \right); \tag{3.13}$$

Differences in the slope coefficients $\hat{\beta}_k^s \left(s = 0, t\right)$ across the strata could be the result of sampling error or reflect a real phenomenon. The latter can be of particular importance for periods t which are very distant from period 0 as different housing market segments tend to show varying price trends. Whether any differences in the slope coefficients reflect a real phenomenon could be tested.

An alternative model, to be estimated on the entire data set, is one with a single intercept term, but where the β's are allowed to differ across the strata. Let $D_{n,k}$ be a dummy variable that has the value 1 if property n belongs to stratum k and 0 otherwise. In period $s\left(s = 0, t\right)$ the model

$$p_n^s = \alpha^s + \sum_{k=1}^{K} \beta_k^s D_{n,k} a_n^0 + \varepsilon_n^s \tag{3.14}$$

is estimated by OLS regression on the data of the sample S^s, yielding predicted prices $\tilde{p}_n^s = \tilde{\alpha}^s + \tilde{\beta}_k^s a_n^0$ for $n \in U_k^0$. The residuals again sum to zero and the new (unstratified) OLS GREG index becomes

$$\tilde{P}^{0t}_{\text{GREG,OLS}} = \frac{\sum_{n \in U^0} \tilde{p}_n^t / N^0}{\sum_{n \in U^0} \tilde{p}_n^0 / N^0} = \frac{\sum_{k=1}^{K} \sum_{n \in U_k^0} \tilde{p}_n^t / N^0}{\sum_{k=1}^{K} \sum_{n \in U_k^0} \tilde{p}_n^0 / N^0} = \frac{\tilde{\alpha}^t + \sum_{k=1}^{K} \left(\frac{N_k^0}{N^0} \right) \tilde{\beta}_k^t \overline{a}_k^0}{\tilde{\alpha}^0 + \sum_{k=1}^{K} \left(\frac{N_k^0}{N^0} \right) \tilde{\beta}_k^0 \overline{a}_k^0}. \tag{3.15}$$

Model (3.14) is more flexible than the original model given by equations (3.1) and (3.5), and could be useful if the proportionality between sale prices and appraisals fails. Estimator (3.15) reduces to the original GREG index (3.10) if the $\tilde{\beta}_k^s$'s are all equal. In practice this will not happen, and (3.15) and (3.10) will give different answers. A common justification for the use of GREG estimators is that, being asymptotically unbiased, they are relatively *robust to model choice*. So we would expect the impact of the alternative model specification (3.15) to be moderate. On the other hand, it is well recognized in the literature that model dependence can be an issue under specific circumstances, notably when dealing with highly variable and outlier-prone populations. For example, Hedlin, Falvey, Chambers and Kokic (2001) stress the importance of a careful model specification search while Beaumont and Alavi (2004) focus on

the treatment of outliers. It would therefore be worthwhile examining the effect of this alternative model specification.

4 Empirical illustration

For the empirical study we used two data sets from different sources. The first data set contains the sale prices of nearly all transactions of existing houses (excluding newly-built houses) in the Netherlands between January 2003 and March 2009 as registered by the Dutch land registry office. The total number of observations amounts to 1,126,242 or approximately 15 thousand per month. The sales were recorded at the time the final agreement was made at the notary's office, on average six weeks after the preliminary sale was agreed on. The second data set contains the government appraisals, relating to January 2003, for all owner-occupied dwellings in the housing stock. Because addresses are available in both data sets, we know the sale price and the appraisal value for each transaction. Because the type of dwelling is also available, we were able to stratify by dwelling type and location.

The first thing we did was run unstratified OLS regressions of selling prices on appraisals, using model (3.8), for all 75 months. A selection of the results is listed in Table 4.1; detailed empirical material is available from the authors upon request. Not surprisingly, the coefficients $\hat{\beta}^t$ are different from zero at very low significance levels. In most cases the intercepts $\hat{\alpha}^t$ differ significantly from zero at the 5% level. Roughly 80 to 90% of the variation in selling prices is 'explained' by the variation in appraisals, as shown by the R^2 values. In other words, the correlation coefficient between selling prices and base period appraisals ranges from 0.89 to 0.95. Figure 4.1 shows that R^2 diminishes slightly over time. As mentioned earlier, one of the reasons could be that different segments of the market exhibit different price changes. We were a bit surprised to find though that R^2 is not the highest in January 2003, being the appraisal reference period.

Based on the above regression results, we computed GREG price index numbers according to equation (3.10). From January 2003 until mid 2008 house prices increased by some 25% in the Netherlands but then started to fall, probably due to the financial and economic crisis. Importantly, the GREG index turns out to be a lot smoother than the simple ratio of sample means as Figure 4.2 makes clear, which is precisely what the index has been designed for.

Table 4.1
Regression results

Month	Alpha	t	Beta	t	R squared
January 2003	1,900.49	2.26	0.98	275.19	0.87
January 2004	5,039.16	5.96	1.01	269.26	0.88
January 2005	-2,555.12	2.43	1.08	237.54	0.84
January 2006	1,282.14	1.41	1.11	286.39	0.87
January 2007	-7,567.99	6.36	1.19	243.72	0.83
January 2008	11,007.39	8.48	1.26	231.93	0.83
January 2009	16,677.31	9.83	1.30	184.24	0.81

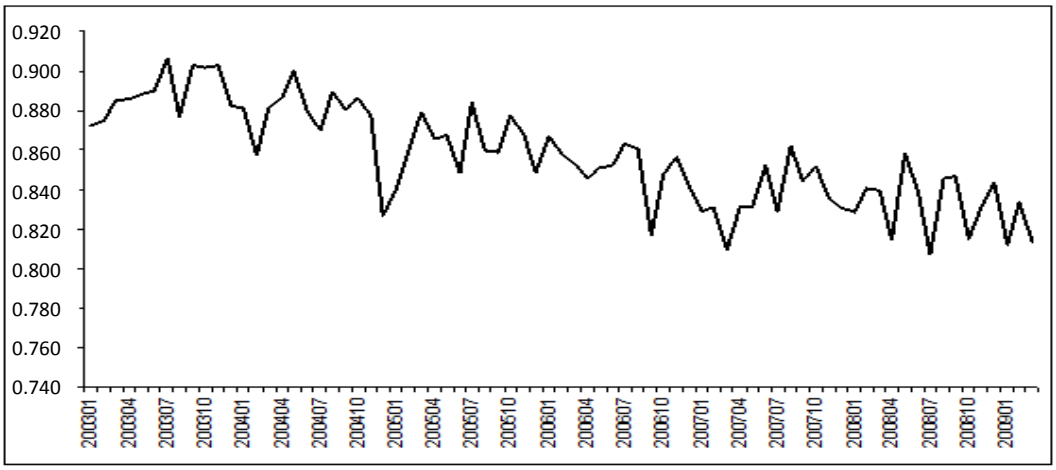


Figure 4.1 R squared values

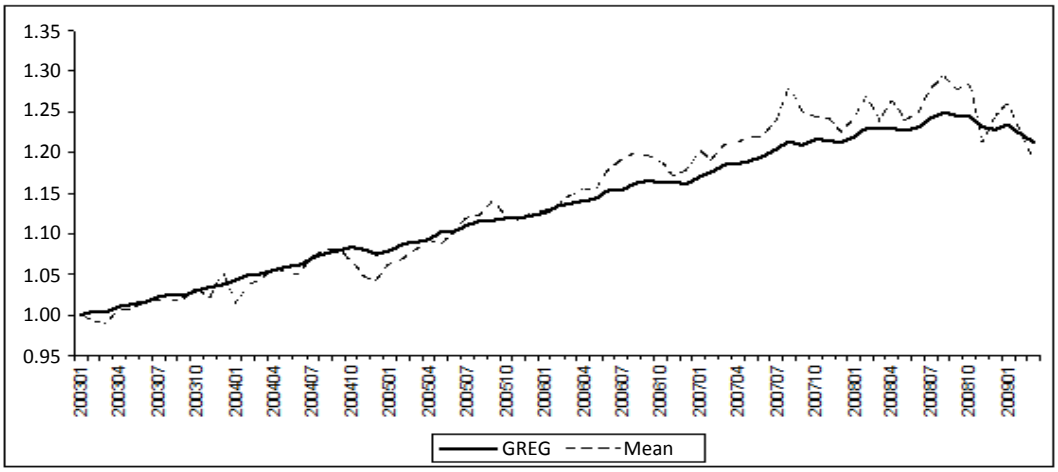


Figure 4.2 GREG index and ratio of sample means

Figure 4.3 compares the GREG index with the SPAR index. In general the trend of both indexes is very similar, although there appears to be a small difference by the end of the period. Figure 4.4 shows that the month-to-month changes in the GREG and SPAR indexes do not differ much either, the GREG index being just a little bit less volatile. So we can conclude that, at the nationwide level, both methods generate more or less equal results. Note that the SPAR index in Figures 4.3 and 4.4 is not the official SPAR index published by Statistics Netherlands. We computed a fixed base index using appraisals for January 2003 only whereas the official index is a chained index, based on appraisals for various reference periods; see also Section 5.3.

Next we stratified the data by thirteen provinces and five types of dwellings, ran OLS regressions per month for the resulting 65 strata and calculated GREG indexes as well as sample means ratios. Figure 4.5 displays the results for one stratum, apartments in the province of Friesland. Due to the relatively low number of observations there are some dramatic spikes, for instance in September 2009 when the ratio of sample means increases by 50%. Again, the GREG index is smoother than the ratio of sample means (but still very volatile) and strikingly similar to the SPAR. The same picture emerges for the other strata, so we do not present those results.

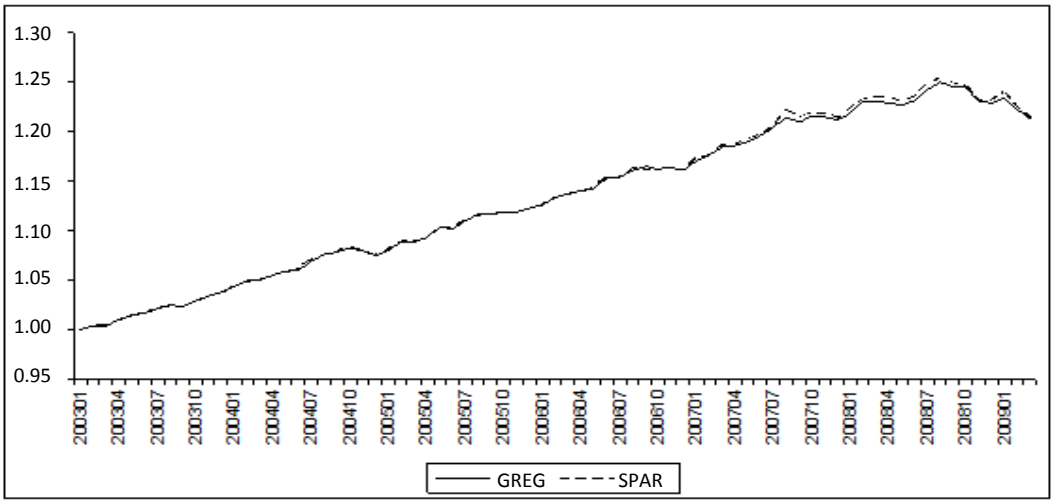


Figure 4.3 GREG and SPAR indexes

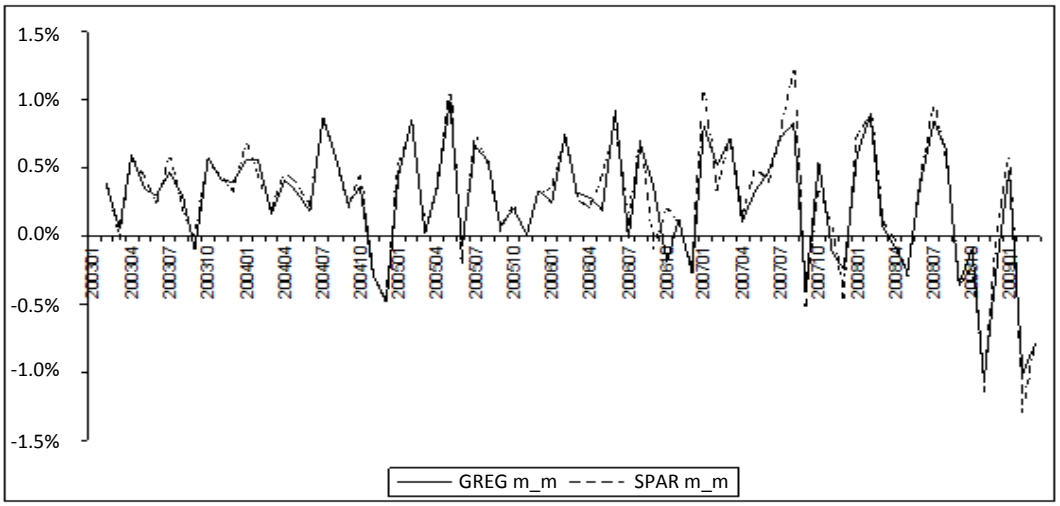


Figure 4.4 GREG and SPAR: month-to-month percentage changes

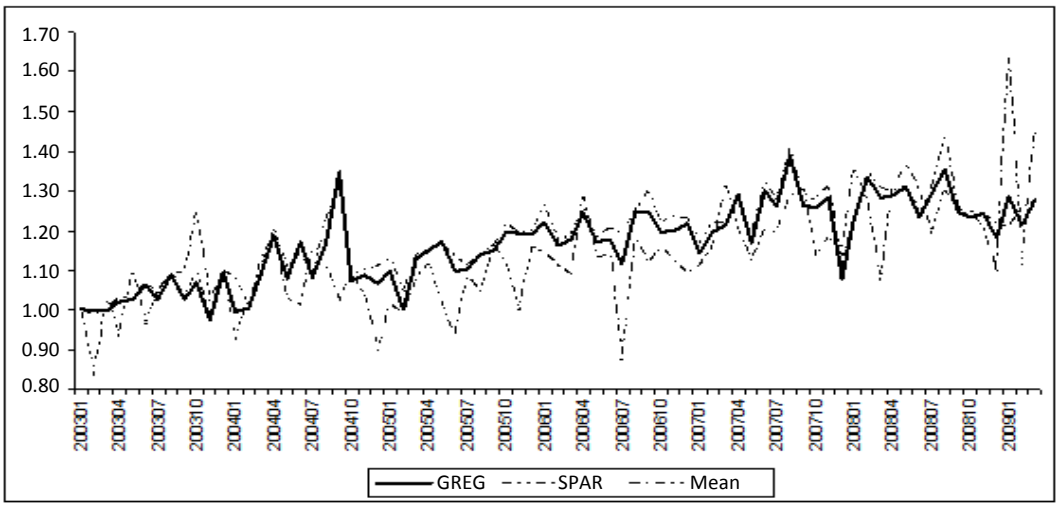


Figure 4.5 GREG and SPAR indexes and ratio of sample means; apartments in the province of Friesland

Finally, using the stratum results, we computed stratified GREG indexes for the whole country according to equation (3.13), where the base period appraisal shares serve as stock value weights. As can be seen from Figure 4.6, there are hardly any differences between the stratified and unstratified GREG indexes, suggesting that sample selection bias is not a major issue. Figure 4.6 also shows a second alternative GREG price index, computed according to equation (3.15), which is based on OLS regressions of the dummy variable model (3.14). And again, the differences with the original GREG index appear to be small.

It should be noted that even within strata some houses are still more likely to sell than others, in particular during the crisis after 2008, so that some sample selection bias in the GREG and SPAR indexes will remain. The direction and magnitude of this bias can only be predicted if data on property characteristics was available to estimate the likelihood of houses to sell. Also, as was mentioned earlier, a too detailed stratification will increase both the sampling variance and sample bias if the number of houses sold is extremely low, and may raise rather than reduce the mean square error of the estimators.

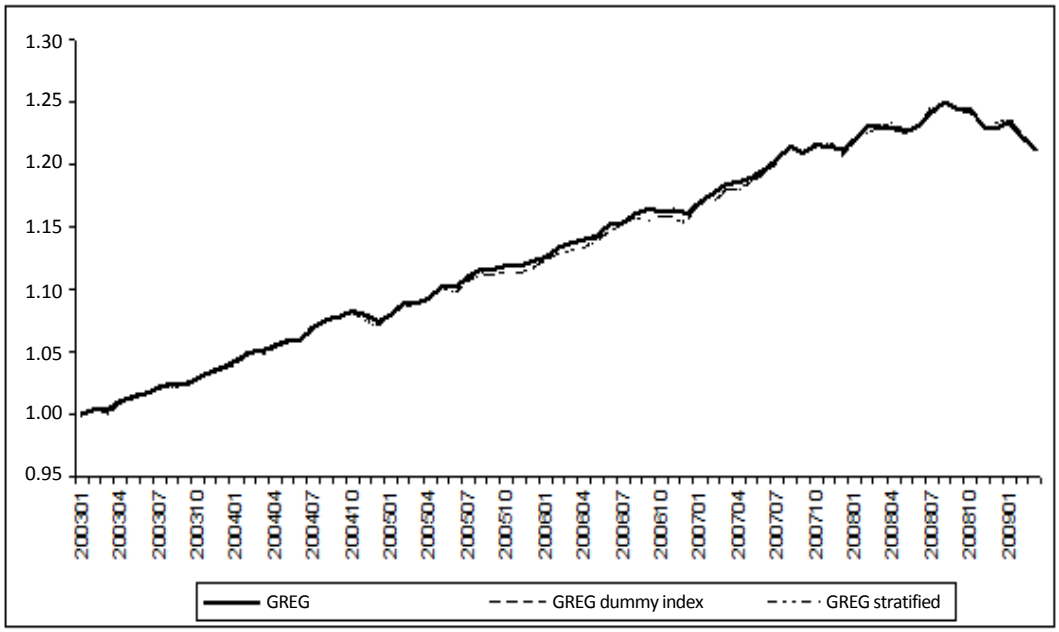


Figure 4.6 GREG, stratified GREG and dummy variable GREG indexes

5 Discussion

5.1 Comparing GREG to SPAR

The most interesting question arising from Section 4 is: why are the GREG and SPAR index numbers so similar in spite of their very different construction methods? It is not remarkable that the trends are similar: although the GREG index does not rely on the matched-model methodology, this index does aim at the same target as the SPAR index. If the sample sizes n^0 and n^t would approach the population size

N^0 – which in reality will of course never happen – then both price indexes approach the value change of the fixed housing stock. Put differently, the two methods are both asymptotically unbiased or 'consistent'.

What may come as a surprise is that the GREG index exhibits roughly the same amount of volatility over time as the SPAR index. To understand the reason why, recall that, with OLS, the regression residuals sum to zero in every time period. This implies $\sum_{n\in S^0} p_n^0/n^0 = \sum_{n\in S^0} \hat{p}_n^0/n^0$ and $\sum_{n\in S^t} p_n^t/n^t = \sum_{n\in S^t} \hat{p}_n^t/n^t$. For the basic regression models (3.1) and (3.5), the SPAR index can thus alternatively be written as

$$\hat{P}_{\text{SPAR}}^{0t} = \frac{\sum_{n\in S^t} \hat{p}_n^t/n^t}{\sum_{n\in S^0} \hat{p}_n^0/n^0}\left[\frac{\sum_{n\in S^0} a_n^0/n^0}{\sum_{n\in S^t} a_n^0/n^t}\right] = \frac{\left(\hat{\alpha}^t + \hat{\beta}^t\bar{a}^{0(t)}\right)/\bar{a}^{0(t)}}{\left(\hat{\alpha}^0 + \hat{\beta}^0\bar{a}^{0(0)}\right)/\bar{a}^{0(0)}} = \frac{\hat{\alpha}^t/\bar{a}^{0(t)} + \hat{\beta}^t}{\hat{\alpha}^0/\bar{a}^{0(0)} + \hat{\beta}^0}, \tag{5.1}$$

using (3.2) and (3.6) for $n \in S^0$ and $n \in S^t$, respectively, where $\bar{a}^{0(0)} = \sum_{n\in S^0} a_n^0/n^0$ and $\bar{a}^{0(t)} = \sum_{n\in S^t} a_n^0/n^t$ for short. There is a striking similarity between the last expression on the right-hand sides of (5.1) and (3.10). The only difference is that the SPAR index (5.1) divides the coefficients $\hat{\alpha}^0$ and $\hat{\alpha}^t$ by the sample means of appraisals, $\bar{a}^{0(0)}$ and $\bar{a}^{0(t)}$, whereas the GREG index (3.10) divides them both by the fixed, non-stochastic population mean $\bar{a}^0$. Essentially, the SPAR index is a fully sample-based estimator of the GREG index.

Compared with the SPAR method, the GREG approach eliminates one source of sampling error, *i.e.*, the sampling variability of the mean appraisals. In accordance with generalized regression theory, we would intuitively expect the GREG method to reduce the sampling error of the price index and produce a less volatile time series (under the reasonable assumption that $\bar{a}^{0(t)}$ and $\hat{\alpha}^t$ are uncorrelated across periods $t = 0,\ldots,T$). Put differently, while the GREG method has been designed as an improvement over the ratio of sample means, we might have expected it to work as a smoothing procedure for the SPAR index also. But, as was shown in Section 4, in practice this is hardly the case. This result can be explained as follows.

The variance reduction of the GREG index relative to the SPAR depends on the value of the intercept terms from the regressions in periods 0 and t. If the regression lines passed exactly through the origin $\left(\hat{\alpha}^t = \hat{\alpha}^0 = 0\right)$, then the GREG index and SPAR index would both be equal to the ratio of the slope coefficients $\hat{\beta}^t/\hat{\beta}^0$ and no reduction in variance would be achieved. In the less extreme case, when $\hat{\alpha}^t$ and $\hat{\alpha}^0$ are close to 0 and the ratios $\hat{\alpha}^t/\bar{a}^0$, $\hat{\alpha}^t/\bar{a}^{0(t)}$, $\hat{\alpha}^0/\bar{a}^0$ and $\hat{\alpha}^0/\bar{a}^{0(0)}$ in (3.10) and (5.2) are very small compared to $\hat{\beta}^t$ and $\hat{\beta}^0$, the GREG and SPAR indexes will differ only slightly and the variance reduction will be marginal; see also the Appendix.

The latter is indeed what happens in practice, as can be seen from Figures 5.1 and 5.2 where the values of $\hat{\alpha}^t/\bar{a}^0$ and $\hat{\alpha}^t/\bar{a}^{0(t)}$ and those of $\hat{\beta}^t$ are plotted over time. The ratios $\hat{\alpha}^t/\bar{a}^0$ and $\hat{\alpha}^t/\bar{a}^{0(t)}$ are remarkably similar and small as compared to the $\hat{\beta}^t$'s. Although we cannot ignore those ratios, it is the change in $\hat{\beta}^t$ that mainly drives the GREG and SPAR indexes. The SPAR index is not only a fully sample-based estimator of the GREG index, as mentioned above, it appears to be almost as efficient.

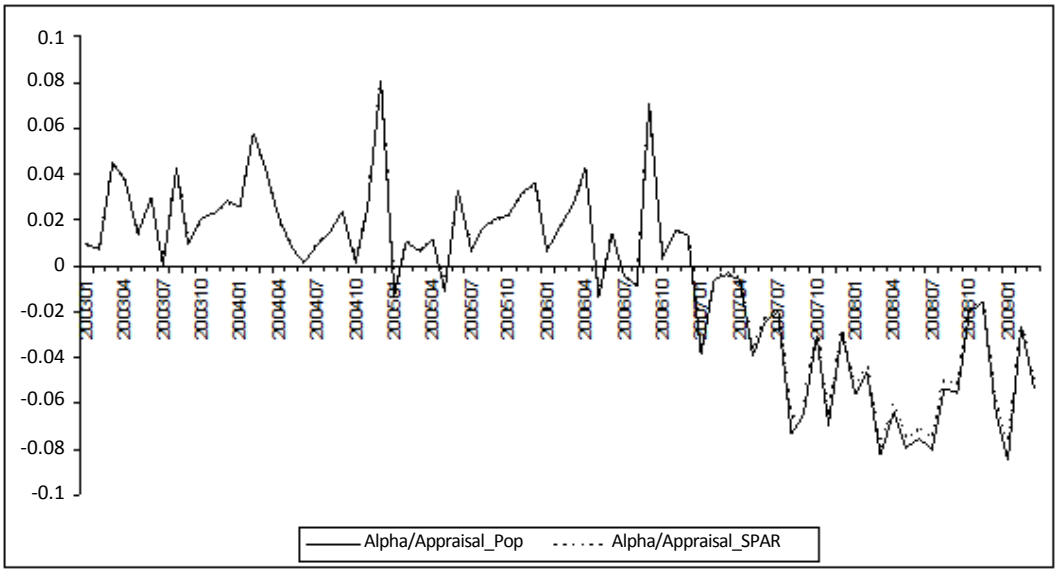


Figure 5.1 Intercepts divided by appraisal means

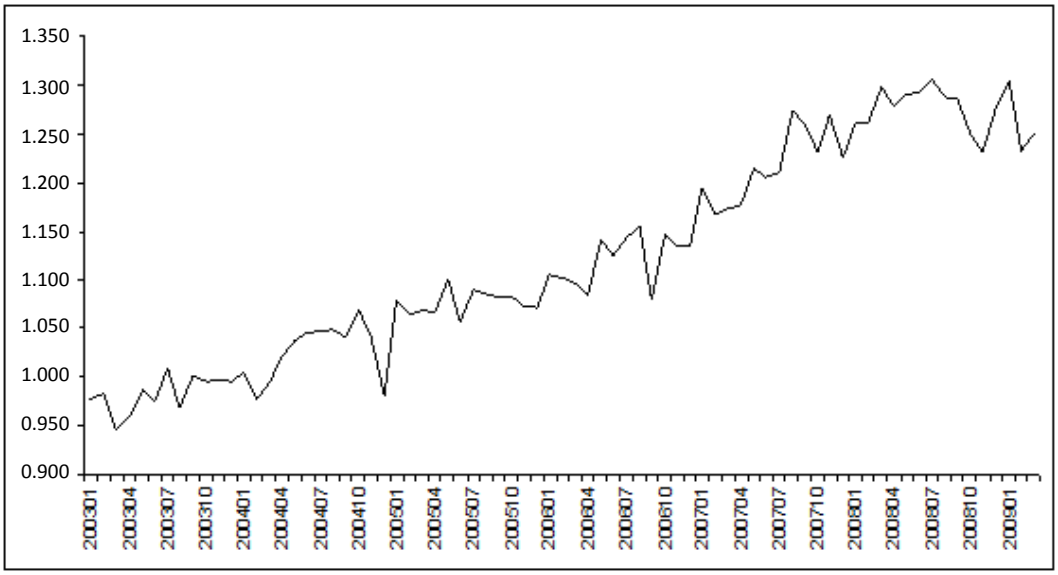


Figure 5.2 Slope coefficients

5.2 The volatility of the slope coefficient

Several factors may have contributed to the volatility of the slope coefficients $\hat{\beta}^t$ in our regressions of selling prices on appraisals and hence of the GREG and SPAR indexes. We will briefly discuss three of these factors: sample mix change, heteroskedasticity and outliers.

A sample of houses can be viewed as a sample of locations, or addresses, since houses are attached to the land they are built on. A change in the sample mix is nothing else than a change in the observed mix of locations at the lowest level. A *location mix change* affects the sample composition in terms of the average quality characteristics of the properties, such as the number of rooms, surface area, *etc*. In our

simple framework, where we observe only one (non-physical) characteristic, namely the appraised value, a location mix change boils down to a change in the sample distribution of the appraisals. This, together with any varying price changes across market segments, induces a change in the sample distribution of the ratios p_n^t / a_n^0, which in turn leads to a change in $\hat{\beta}^t$ in the two-variable regression model (3.5).

Other than by stratification there is little we can do about the effect of changes in the sample mix of locations (but stratifying by province and type of dwelling did not help much), so the volatility of $\hat{\beta}^t$ and therefore of the GREG and SPAR indexes, will be difficult to reduce. Controlling for location at the address level is also impossible in hedonic imputation methods. Here, the effect of (location) mix change is mitigated by controlling for region plus a range of physical characteristics. However, this does not necessarily mean that hedonic imputation will produce more stable index series than the GREG or SPAR methods. Most standard hedonic models fit the cross sectional data less well than our model does, and the characteristics' coefficients typically exhibit a great deal of variability over time. So maybe it is not surprising that Bourassa, *et al*. (2006) find that "the SPAR index [….] reliably tracks house price changes, but exhibits less volatility than index methods that require more parameter estimates."

We can alternatively look at the variability of the slope coefficient from a purely statistical perspective. It is well known that in a two-variable model the OLS estimator $\hat{\beta}^t$ can be written as

$$\hat{\beta}^t = r\left(p^t, a^0\right)\frac{s\left(p^t\right)}{s\left(a^0\right)}, \tag{5.2}$$

where $r\left(p^t, a^0\right)$ denotes the sample correlation coefficient in period t between selling prices and appraisals, which is equal to the square root of R^2; $s\left(p^t\right)$ and $s\left(a^0\right)$ are the corresponding sample standard deviations. A comparison of Figures 4.1 and 5.2 suggests that sudden changes in R^2 are largely responsible for the volatility of $\hat{\beta}^t$. In December 2004 for example, a substantial drop in R^2 coincides with a significant decrease of $\hat{\beta}^t$ (and with a decrease in the GREG and SPAR indexes, as shown by Figure 4.4).

Least squares regression can either be weighted or unweighted. In the absence of *heteroskedasticity, i.e.*, when the variance of the errors is constant, OLS should be used. Weighted Least Squares (WLS) is preferred if there is evidence of heteroskedasticity; using appropriate weights, WLS will lead to more stable coefficients than OLS. In this case the unweighted sample sum of the residuals differs from zero so the estimator (3.9) has to be applied. To facilitate the interpretation of the GREG index and the comparison with the SPAR index, in Section 3 we assumed away the problem of heteroskedasticity and restricted ourselves to OLS. Note that the (OLS) GREG estimator (3.10) remains asymptotically design unbiased if heteroskedasticity is present.

The most interesting form of (classical) heteroskedasticity – and, given our data set, the only form we would have been able to reduce – would arise if the variance of the errors of our regression model (3.5) depended on the appraisal value, being the only regressor. However, the residuals from our OLS regressions do not point to substantial heteroskedasticity of this type. This is illustrated in Figure 5.3 for three months, including the base period (January 2003), where the sale prices are plotted against the appraisals; the regression lines are also given. To be sure, we also performed the White (1980) test. This test did not point towards the presence of this form of heteroskedasticity either.

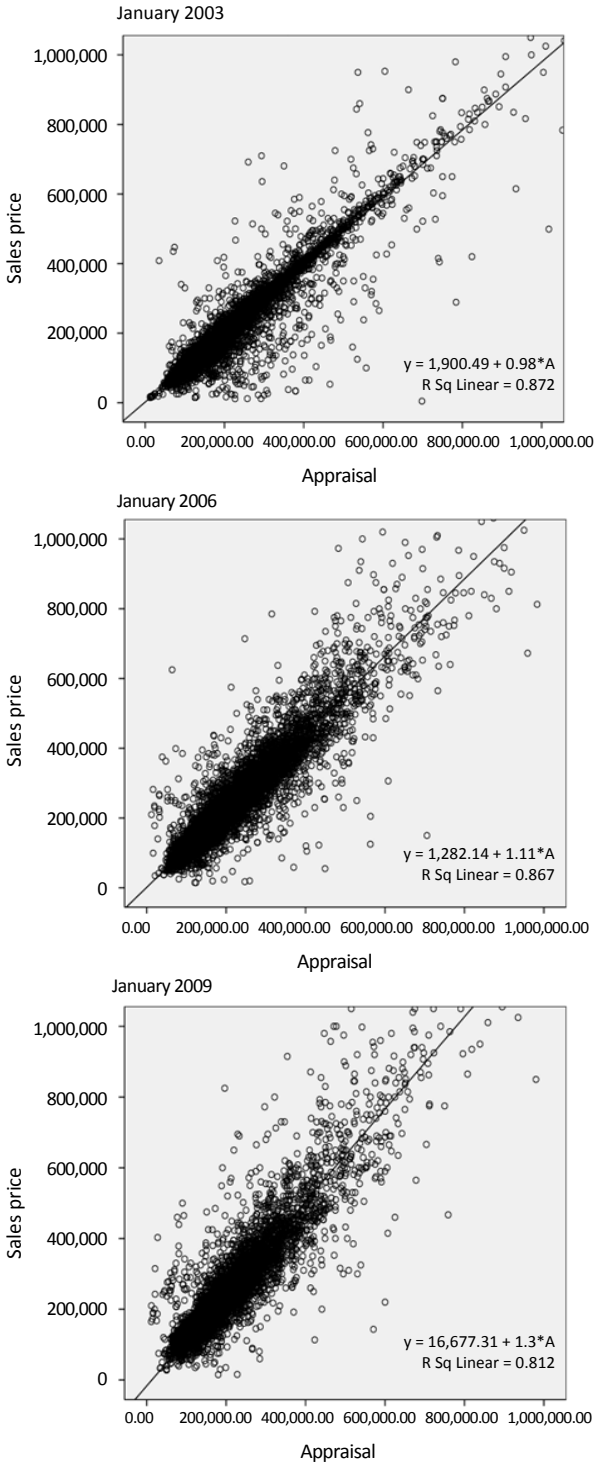


Figure 5.3 Scatter plots and regression lines

Our initial data set of sale prices and appraisals included some obvious *outliers*. To estimate the GREG index we therefore made use of a cleaned data set that has been prepared to compute the official Dutch house price index. Statistics Netherlands applies several data cleaning procedures. Houses that were sold

more than once in a month are excluded from the data set. To delete entry errors and outliers that may unduly affect the results, properties with sale prices or appraisals below €10,000 or above €5,000,000 and properties with 'unrealistic' sale price-appraisal ratios are also removed. The removal of 'unrealistic' observations is done by looking at the distribution of the logarithm of the sale price-appraisal ratios; all observations are deleted for which the log ratio differs more than 5 standard deviations from the mean. For more information, see Statistics Netherlands (2008).

These procedures are rather arbitrary. For regression-based estimators such as the GREG it is more appropriate to delete observations with high leverage, *i.e.*, to delete those sample units that have a big impact on the regression coefficients when they are excluded from the sample. A well-known measure in this context is the DFBETA of a sample unit (Cook and Weisberg 1982). Since the SPAR can be written as a regression-based index, this measure could be used here as well to detect and delete outliers. The scatter plots in Figure 5.3 show that the cleaned data set still contains some big outliers. Whether these have high leverage, and whether removing them will reduce the volatility of the $\hat{\beta}$'s and the GREG and SPAR indexes, remains to be seen.

5.3 Some further points

The GREG method is based on the premise of a fixed housing stock. That is, we have assumed that there are no entries (*e.g.*, newly-built houses) or exits (discarded houses) and that housing quality remains fixed over time. Our approach is non-symmetric in that we condition on the *base period* stock. From an index number point of view we estimate a Laspeyres price index for the housing stock where the quantities are all equal to 1 because every house is treated as a unique property. An equally justifiable approach would be to measure the price change of the current period stock, which includes additions to the stock in each period, using a Paasche index. Taking the geometric mean of both indexes would lead to the Fisher index. The Fisher index is a preferred measure of price change due to its symmetric form. The construction of a Fisher-type GREG index is, however, infeasible since the Paasche component requires real time assessed values for houses that are new to the stock, which are obviously not available.

The assumption of a fixed (base period) housing stock can be relaxed through annual chaining, provided that the housing stock is re-assessed annually. This is the current state of affairs in the Netherlands; in the past, assessments were undertaken once every three or four years. Annual updating of the appraisals might also adjust for quality changes of the properties, to some extent at least, because the updated appraisals likely account for major repairs, remodelling and depreciation.

One final remark is in order. For some purposes it is desirable to decompose the overall house price index into two components: a component that measures the change in the price of the structure and a component that measures the change in the price of the land. Neither our GREG method nor SPAR and repeat sales methods are fit for that purpose. Hedonic imputation methods might work, notwithstanding practical problems like multicollinearity; see Diewert, de Haan and Hendriks (2012) for a first attempt. If data on structure size, plot size and other price-determining attributes became available for all properties in the housing stock, then we would be able to estimate a "hedonic imputation GREG index", including the land-structure split. The chances of getting such data in the Netherlands are unfortunately negligible.

6 Conclusion

The simple GREG method outlined in this paper, which is based on OLS regressions of selling prices on appraisals, substantially reduces the volatility of a house price index as compared to the ratio of sample means. The SPAR index can be viewed as an estimator of the OLS GREG index (which itself is an estimator, of course) where the base period population mean of appraisals is replaced by the sample means in the base period and the comparison period. Our empirical results for the Netherlands indicate that the SPAR index is almost as efficient as the GREG index, even for small sub-populations. We have checked this by drawing a random sample of 50 observations each month from the total number of monthly sales (15,000 on average). The month-to-month changes of the SPAR index were only slightly bigger than those of the GREG.

Due to compositional change of the properties sold, the GREG (and SPAR) time series exhibit strong short-term volatility. An increase in a particular month is typically followed by a decrease in the next month. Put differently, the month-to-month changes do not tell us much about the true price change of the housing stock which, except under unusual circumstances, should behave smoothly. An improved outlier detection method might help reduce the index volatility, but the effect will probably be limited. Applying a smoothing procedure would seem to be an option. However, that will typically lead to revisions of previously published price index numbers, and the lack of revisions is one of the strenghts of the GREG and SPAR approaches. Another option would be to reduce the frequency of observation, for example to quarters, but that may be undesirable as well.

From a purely statistical point of view, in our two-variable model the variability of R^2 seems to be responsible for a large part of the volatility of the slope coefficient and therefore of the volatility of the price index series. Future research could focus on the relation between compositional changes in terms of the property characteristics and changes in R^2. As many housing characteristics are unavailable, we cannot investigate this issue with our data. Fortunately, Statistics Netherlands has access to a data set from the largest Dutch association of real estate agents that might be useful for this purpose. This data set covers around 70% of all housing sales in the Netherlands during 1999-2008, includes many property characteristics and has been enriched with appraisal data. In the past we already used the data set to compare the SPAR index with various types of hedonic indexes.

Acknowledgements

The authors would like to thank the participants at the Economic Measurement Group Workshop, 1-3 December 2010, University of New South Wales, Sydney, Australia and the participants at an Applied Economics Seminar, 22 November 2011, University of Queensland, Brisbane, Australia for their helpful comments on preliminary versions of the paper. Comments and suggestions made by the editor and two anonymous referees also helped to improve the paper. The assistance of Erna van der Wal, who provided us with the data, is gratefully acknowledged. The views expressed in this paper are those of the authors and do not necessarily reflect the views of Statistics Netherlands.

Appendix

Approximate Standard Errors of the GREG Index

The GREG index defined by equation (3.10) in the main text is a ratio of two estimators, $\hat{\bar{p}}^{t}_{\text{GREG}}$ and $\hat{\bar{p}}^{0}_{\text{GREG}}$; for brevity we delete "OLS". Using a first-order Taylor expansion, the variance of the index can be approximated by (see *e.g.*, Kendall and Stuart 1976)

$$\text{var}\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \left[\frac{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}\right]^2 \left[\frac{\text{var}\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{\left\{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)\right\}^2} + \frac{\text{var}\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}{\left\{E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)\right\}^2} + \frac{\text{cov}\left(\hat{\bar{p}}^{t}_{\text{GREG}}, \hat{\bar{p}}^{0}_{\text{GREG}}\right)}{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}\right], \qquad \text{(A.1)}$$

where $E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)$ and $E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)$ denote expected values.

The covariance term in (A.1) is equal to 0 since, by assumption, the samples in periods 0 and t are independently drawn. Replacing the expected values in (A.1) by the estimators and subsequently taking the square root leads to the following expression for the standard error of $\hat{P}^{0t}_{\text{GREG}}$:

$$se\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \hat{P}^{0t}_{\text{GREG}} \left[\frac{\text{var}\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)^2} + \frac{\text{var}\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)^2}\right]^{1/2}. \qquad \text{(A.2)}$$

Equation (A.2) can be estimated in practice using $\hat{\bar{p}}^{s}_{\text{GREG}} = \hat{\alpha}^{s} + \hat{\beta}^{s}\bar{a}^{0}\left(s = 0, t\right)$, hence $\text{var}\left(\hat{\bar{p}}^{s}_{\text{GREG}}\right) = \text{var}\left(\hat{\alpha}^{s}\right) + \left(\bar{a}^{0}\right)^2 \text{var}\left(\hat{\beta}^{s}\right) + 2\bar{a}^{0}\,\text{cov}\left(\hat{\alpha}^{s}, \hat{\beta}^{s}\right)$. Estimates of the (co)variances are readily available in most statistical packages from the variance-covariance matrix.

Dividing (A.2) by $\hat{P}^{0t}_{\text{GREG}}$ yields an expression for the relative standard error or coefficient of variation, $CV\left(\hat{P}^{0t}_{\text{GREG}}\right) = se\left(\hat{P}^{0t}_{\text{GREG}}\right)\Big/\hat{P}^{0t}_{\text{GREG}}$, of the GREG index:

$$CV\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \left[\frac{\text{var}\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)^2} + \frac{\text{var}\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)^2}\right]^{1/2} = \left[\left\{CV\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)\right\}^2 + \left\{CV\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)\right\}^2\right]^{1/2}. \qquad \text{(A.3)}$$

Of more importance is the relative standard error of the *percentage change* of the index, *i.e.*, $CV\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) = se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right)\Big/\left(\hat{P}^{0t}_{\text{GREG}} - 1\right)$. This is generally greater than $CV\left(\hat{P}^{0t}_{\text{GREG}}\right)$, given that $se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) = se\left(\hat{P}^{0t}_{\text{GREG}}\right)$ and $\hat{P}^{0t}_{\text{GREG}} - 1 < \hat{P}^{0t}_{\text{GREG}}$.

If both regression lines almost pass through the origin, hence $\hat{\alpha}^{s} \cong 0\left(s = 0, t\right)$, we have $\hat{P}^{0t}_{\text{GREG}} \cong \hat{\beta}^{t}\big/\hat{\beta}^{0}$ and (A.2) simplifies to

$$se\left(\hat{P}^{0t}_{\text{GREG}}\right) = se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) \cong \hat{P}^{0t}_{\text{GREG}} \left[\frac{\text{var}\left(\hat{\beta}^{t}\right)}{\left(\hat{\beta}^{t}\right)^2} + \frac{\text{var}\left(\hat{\beta}^{0}\right)}{\left(\hat{\beta}^{0}\right)^2}\right]^{1/2}. \qquad \text{(A.4)}$$

In this particular case the GREG and SPAR indexes nearly coincide, so (A.4) also holds for the SPAR index (using $\hat{P}^{0t}_{\text{SPAR}}$ rather than $\hat{P}^{0t}_{\text{GREG}}$).

References

Bailey, M.J., Muth, R.F. and Nourse, H.O. (1963). A regression method for real estate price construction. *Journal of the American Statistical Association*, 58, 933-942.

Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 2, 195-208.

Bourassa, S.C., Hoesli, M. and Sun, J. (2006). A simple alternative house price index method. *Journal of Housing Economics*, 15, 80-97.

Calhoun, C.A. (1996). OFHEO House Price Indexes: HPI Technical Description. Office of Federal Housing Enterprise Oversight, Washington, DC.

Case, K.E., and Shiller, R.J. (1987). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review*, September-October, 45-56.

Case, K.E., and Shiller, R.J. (1989). The efficiency of the market for single family homes. *The American Economic Review*, 79, 125-137.

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

Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall.

de Haan, J. (2007). Formulae for the Variance of (Changes in) the SPAR Index. Unpublished manuscript, Statistics Netherlands, Voorburg (Dutch only; available from the author upon request).

de Haan, J. (2010). Hedonic price indexes: A comparison of imputation, time dummy and 'Re-Pricing' methods, *Journal of Economics and Statistics* (Jahrbucher fur Nationalokonomie und Statistik), 230, 772-791.

de Haan, J., van der Wal, E. and de Vries, P. (2009). The measurement of house prices: A review of the sale price appraisal method. *Journal of Economic and Social Measurement*, 34, 51-86.

de Vries, P., de Haan, J., van der Wal, E. and Mariën,G. (2009). A house price index based on the SPAR method. *Journal of Housing Economics*, 18, 214-223.

Diewert, W.E., de Haan, J. and Hendriks, R. (2012). The decomposition of a house price index into land and structures components: A hedonic regression approach. *Econometric Reviews* (forthcoming).

Diewert, W.E., Heravi, S. and Silver, M. (2009). Hedonic imputation versus time dummy hedonic indexes. In *Price Index Concepts and Measurement*, (Eds., W.E. Diewert, J. Greenlees and C. Hulten), NBER Studies in Income and Wealth, Chicago: Chicago University Press, 70, 161-196.

Edelstein, R.H., and Quan, D.C. (2006). How does appraisal smoothing bias real estate returns measurement? *Journal of Real Estate Finance and Economics*, 32, 41-60.

Eurostat (2010). *Technical Manual on Owner-Occupied Housing for Harmonised Index of Consumer Prices*, Version 1.9. Available at www.epp.eurostat.ec.europa.eu/portal/page/portal/hicp/documents/Tab/Tab/03_METH-OOH-TECHMANUAL_V1-9.pdf.

Francke, M.K. (2010). Repeat sales index for thin markets: A structural time series approach. *Journal of Real Estate Finance and Economics*, 41, 24-52.

Geltner, D. (1996). The repeated-measures regression-based index: A better way to construct appraisal-based indexes of commercial property value. *Real Estate Finance*, 12, 29-35.

Gouriéroux, C., and Laferrère, A. (2009). Managing hedonic house price indexes: The french experience. *Journal of Housing Economics*, 18, 206-213.

Grimes, A., and Young, C. (2010). A Simple Repeat Sales House Price Index: Comparative Properties Under Alternative Data Generation Processes. Motu Working Paper 10-10, Motu Economic and Public Policy Research, New Zealand.

Hardman, M. (2011). Calculating High Frequency Australian Residential Property Price Indices. Rismark Technical Paper, available at www.rpdata.com/images/stories/content/PDFs/technical_method_paper.pdf.

Hedlin, D., Falvey, H., Chambers, R. and Kokic, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17, 527-544.

Hill, R.J., and Melser, D. (2008). Hedonic imputation and the price index problem: An application to housing. *Economic Inquiry*, 46, 593-609.

Jansen, S.J.T., de Vries, P., Coolen, H.C.C.H., Lamain, C.J.M. and Boelhouwer, P. (2008). Developing a house price index for the Netherlands: A practical application of weighted repeat sales. *Journal of Real Estate Finance and Economics*, 37, 163-186.

Kendall, M., and Stuart, A. (1976). *The Advanced Theory of Statistics – Volume 1: Distribution Theory*, 4th Edition, London: Charles Griffin & Company.

Leventis, A. (2006). Removing Appraisal Bias from a Repeat Transactions House Price Index: A Basic Approach. Paper presented at the OECD-IMF Workshop on Real Estate Price Indexes, Paris, 6-7 November 2006.

Makaronidis, A., and Hayes, K. (2006). Owner Occupied Housing for the HICP. Paper presented at the OECD-IMF Workshop on Real Estate Price Indexes, Paris, 6-7 November 2006.

Rossini, P., and Kershaw, P. (2006). Developing a Weekly Residential Price Index Using the Sales Price Appraisal Ratio. Paper presented at the twelfth Annual Pacific Rim Real Estate Society Conference, Auckland, 22-25 January 2006.

Saarnio, M. (2006). Housing Price Statistics at Statistics Finland. Paper presented at the OECD-IMF Workshop on Real Estate Price Indexes, Paris, 6-7 November 2006.

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

Shi, S., Young, M. and Hargreaves, B. (2009). Issues in measuring a monthly house price index in New Zealand. *Journal of Housing Economics*, 18, 336-350.

Statistics Netherlands (2008). Price Index Owner-occupied Existing Dwellings; Method Description. Statistics Netherlands, The Hague, available at www.cbs.nl/NR/rdonlyres/A49D8542-26EC-40FD-9093-82A519247F4B/0/MethodebeschrijvingPrijsindexBestaandeKoopwoningene.pdf.

van der Wal, E., ter Steege, D. and Kroese B. (2006). Two Ways to Construct a House Price Index for the Netherlands: The Repeat Sale and Sale Price Appraisal Ratio. Paper presented at the OECD-IMF Workshop on Real Estate Price Indexes, Paris, 6-7 November 2006.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 419-434
Statistics Canada, Catalogue No. 12-001-X

Does the first impression count? Examining the effect of the welcome screen design on the response rate

Roos Haer and Nadine Meidert[1]

Abstract

Web surveys are generally connected with low response rates. Common suggestions in textbooks on Web survey research highlight the importance of the welcome screen in encouraging respondents to take part. The importance of this screen has been empirically proven in research, showing that most respondents breakoff at the welcome screen. However, there has been little research on the effect of the design of this screen on the level of the breakoff rate. In a study conducted at the University of Konstanz, three experimental treatments were added to a survey of the first-year student population (2,629 students) to assess the impact of different design features of this screen on the breakoff rates. The methodological experiments included varying the background color of the welcome screen, varying the promised task duration on this first screen, and varying the length of the information provided on the welcome screen explaining the privacy rights of the respondents. The analyses show that the longer stated length and the more attention given to explaining privacy rights on the welcome screen, the fewer respondents started and completed the survey. However, the use of a different background color does not result in the expected significant difference.

Key Words: Web surveys; Welcome screens; Breakoffs; Design.

1 Introduction

With the growing number of internet users and the increasing popularity of broader bandwidth, the use of Web surveys to collect data is proliferating at a rapid pace (Vicente and Reis 2010). The advantages of this survey mode have been well documented; they save a significant amount of time and money. Along with the positive aspects of Web surveys, there are methodological concerns that cannot be ignored if survey quality is to be guaranteed (Vicente and Reis 2010). These concerns are mainly about nonresponse and coverage. Although coverage is of less concern for surveys of specifically named persons, such as students, nonresponse remains a major concern in Web survey research (Crawford, Couper and Lamias 2001).

Web surveys are connected with relatively low response rates compared to other modes of survey research (Lozar Manfreda, Bosnjak, Berzelak, Haas and Vehovar 2008). It affects all types of Web surveys, from list-based samples to pre-recruited probability-based panels and opt-in or volunteer panels (Couper and Miller 2008, page 833). The nonresponse rate is the sum of those respondents that did not participate in the Web survey, although they were invited, together with those respondents that broke off and dropped out prematurely. In other words, nonresponders are those respondents who do not view all questions and answer all questions (Bosjnak and Tuten 2001). It is important to note that we use the terms ‘dropout’ and ‘breakoff’; as synonymous throughout this study. Nonresponse is of particular importance to researchers because the unknown characteristics and attitudes of non-respondents may cause inaccuracies in the results of the study in question (Bosnjak and Tuten 2001). This problem poses a challenge for any survey mode, but in particular to Web surveys (Galesic 2006). In general, dropout rates

1. Roos Haer and Nadine Meidert, Department of Political Science and Public Administration, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany. E-mail: Roos.vanderHaer@uni-konstanz.de, Nadine.Meidert@uni-konstanz.de.

in Web surveys may be as high as 80 percent, with an average of about 30 percent. For individually targeted Web surveys, these rates are lower, but still average at about 15 percent (Galesic 2006, page 313; Peytchev 2009).

By far, the largest number of respondents who drop out do so on the initial page – the welcome screen (Couper 2008). On this splash screen, the invitee is reassured that they have arrived at the right place, is informed about the content of the survey, (given the country context) privacy rights, and encouraged to proceed to the survey itself. Consequently, this particular page plays an important role in sealing the deal; in turning the invitee into a respondent (Couper 2008, page 330). However, despite its importance, the influence of the welcome screen on breakoff rates has received little to no research attention (Couper 2008, page 330).

This is even more surprising, considering the rich multimedia capabilities of Web surveys that allow text to be supplemented with a variety of visual elements, such as color, graphics, typography, and animation. Experimental research has shown that the content of the text as well as these auxiliary features are potentially powerful tools for maintaining respondents' interest in the survey and for encouraging completion of the instrument (Couper, Traugott and Lamias 2001). Although respondents are exposed to these design features from the very first screen that they see, most of the conducted experimental survey research has been limited to the influence of the design of the actual complete survey on breakoff rates and not to the influence of specific lay-out and design features of the welcome screen.

With this in mind, the goal of this study is to systematically explore some of the factors connected to the welcome screen that may affect the decision to break off the survey prematurely. As such, this research falls into the category of research focused on how to increase response rates (Bosjnak and Tuten 2001). In identifying these factors, special attention is devoted to the use of color, the announced length of the survey, and to variations in describing the privacy rights of the respondent. These three factors are standard elements of the welcome screen often discussed in textbooks, but neglected in empirical research.

The remainder of the article is structured as follows. Firstly, empirical evidence is used from Web surveys to develop our hypotheses about the influence of particular design features of the welcome screen on the breakoff rates. Next, we describe the experimental study we conducted aiming to test these hypotheses. Finally, we conclude this article with a discussion and implications for Web survey design.

2 Theoretical background

One of the most prevalent threats to Web surveys inference is breakoffs or so-called dropouts. These are respondents that quit prior to completing the survey (Bosjnak and Tuten 2001). Breakoff, as a part of the nonresponse rate, can harm the quality of survey statistics; the larger this rate, the larger the risk of nonresponse error. As a result, much effort of survey researchers has been focused on reducing this rate (Groves *et al.* 2004). Within the Web survey methodology literature, most research on how to improve data quality by reducing this rate is focused on the use of follow-ups, incentives, length, wording, and presentation of the questionnaire (Deutskens, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004, page 22). Most of these response-enhancing features focus on making changes in the content and its lay-out. To our knowledge, limited attention has so far been given to guidelines concerning the lay-out and wording of the welcome screen. For example, Dillman's (2007, page 377) recommendation on how to

construct an effective welcome screen only highlights that this particular page has to be motivational, emphasizing the ease of responding, and should instruct the respondents about how to proceed to the next page. Detailed and practical instructions on how to design an effective welcome screen are lacking. This is surprising because most respondents drop out after the first screen (*i.e.*, the so-called unit nonresponders) (Couper 2008; Bosjnak and Tuten 2001). Moreover, this particular splash screen provides the potential respondent with a first impression of the survey: it evokes emotions towards the questionnaire that might induce respondents to not only start with the Web survey, but also to provide answers faster, to overlook imperfections of the design, and perhaps even answer more honestly (Dillman, Gertseva and Mahon-Haft 2005). Furthermore, it is a question of aesthetics as visual traits determine an individual's feelings and emotional reaction. In Web surveys, the welcome screen is the first visual contact to the respondent and thus its lay-out impacts the respondents' feelings toward the whole survey. It can be even assumed that an appealing survey design can distract from the bad quality of the questionnaire itself (Mahon-Haft and Dillman 2010).

To fill this niche in empirical research, we test three potential factors embedded on the welcome screen on the breakoff rates of a Web survey. These factors are; the use of background color, the number of words devoted to explaining the privacy rights and the data security to the potential respondent, and the announced length of the Web survey. These three elements of the welcome screen are chosen because they are not only essential elements of welcome screens, they also influence the first impression respondents might get of the survey.

2.1 Background color

Unlike paper surveys, the use of the Web opens a batch of visual possibilities. This visual potential is of importance since respondents attend to many features of survey questions not just the words that convey the question or the literal meaning of those words (Tourangeau, Couper and Conrad 2007). These nonverbal features compromise numeric, symbolic, and graphic language (Redline and Dillman 2002; Dillman 2007). Because numbers or symbols are hardly ever used in designing effective welcome screens, the graphical non-verbal element (*i.e.*, brightness, size, shape, spatial arrangement, contrast, figure/ground, and even color) might play an important role in increasing response rates.

With the flexibility of the Web, it is, for example, simple for the designer of the survey to create text and background combinations of a variety of differing colors (Hall and Hanna 2004). As a result, myriad of different color combinations proliferate Web surveys. The choice of a particular color relates to the visual contrast of the verbal information presented on the colored background. This is partly determined by their wavelengths. For example, saturated colors have different wavelengths that need to be focused at different depths behind the lens of the eye, which lead to visual fatigue (Couper 2008, page 164). In addition, research has shown that respondents tend to find short wavelength colors (blues and greens) more pleasant than long wavelength colors (reds and yellows) (Hall and Hanna 2004). For example, Pope and Baker (2005) varied the background color of a survey of college students, using blue or pink background for a survey on alcohol-related issues. The survey with the blue background took less time to complete (although the differences were not statistically significant).

Besides wavelength, colors also may direct communication in other ways. Color has meaning, whether through cultural conventions, learned associations, or the actions associated with color in the instrument

itself (Couper 2008, page 168). That is to say, color can affect respondents emotionally. The color red, for example, is often associated with danger or hotness, especially when linked with blue for cold (see for example, Gorn, Chattopadhyay, Yi and Dahl 1997). A few studies have focused on user emotions when filling out colored Web surveys. Weller and Livingston (1988) for example, found out that the color of the questionnaire did indeed affect the received responses. Specifically, the color pink produced less emotional response than the color blue.

Some studies have been conducted on the influence of color on response rates. For example, Etter, Cucherat and Perneger (2002) concluded in their meta-analysis of 10 experimental studies, that printing questionnaires on colored paper does not substantially influence the speed of response or the proportion of missing items. More importantly, when all colors (blue, green, or yellow) were pooled, no study in the meta-analysis found a statistically significant effect of colored paper (versus white paper) on response rate. The only color that had some minor effect (in comparison to white) was pink. Also the studies that have been conducted to examine influence of color in self-administered Web surveys on response rates show that background color may have some effect, although not in all cases and not always a very large effect (Couper 2008). Dillman, Conradt and Bowker (1998) and Hall and Hanna (2004) for example, show in their studies that a design of black letters on a white background is the most effective design concerning response rates. This is also confirmed with a recent meta-analysis for mail surveys conducted by Edwards, Roberts, Clarke, Diguiseppi, Wentz, Kwan, Cooper, Felix and Pratap (2009). They found that the odds of response were increased by a third using a white background. Extending this argument to the usage of color on welcome screens, we expect that the group of respondents receiving a welcome screen consisting of colors with long wavelengths that boast negative emotional response, will have a higher breakoff rate than the group of respondents receiving a simple welcome screen without many colors.

2.2 Privacy rights

One of the possible explanations for why Web surveys are troubled with low response rates compared to other modes of survey research, may relate to confidentiality concerns with respect to electronic mail and to the Web in general (Couper 2000). Although self-administered Web surveys have the ability to collect sensitive information with less desirability bias, concerns about the security of the Web may negate this benefit, potentially producing higher nonresponse rates (or less honest reporting).

It is therefore not a surprise that most Web survey researchers, depending on the legal regulations of their country, provide their potential respondents with information on what will be done with the information they give. In addition, they emphasize the voluntary aspect of the survey and often assure that they will never match respondent's names with the results in any way. These rights to privacy of respondents are not only mentioned in the invitation e-mail but also on the welcome screen. The welcome screen also plays an important part in reassuring respondents and motivating them to start the Web survey.

A few studies have examined how assurances of privacy and confidentiality have an effect on the response rate. Most of the early studies concerned the U.S. household decennial census and were based on the assumption that privacy assurances were a 'good' thing - it increases response rates by overcoming respondent's concerns (Singer, Hippler, and Schwarz 1992, 258; Singer, Von Thurn, Miller 1995, 66-67). However, these early studies show the contrary; these assurances reduced the willingness to participate (Fay, Bates, and Moore 1991; Singer, Mathiowetz, and Couper 1993; Singer, Van Hoewyk, and

Neugebauer 2003; Hillygus, Nie, Prewitt and Pals 2006). For example, Singer *et al.* (1992) demonstrated that mentioning privacy rights negatively affects response, whether measured as item nonresponse, unit nonresponse, response rate, or response quality. These studies uncovered unanticipated consequences. Assurances of confidentiality and privacy protection rights might actually increase participate concerns about the survey content. These assurances seem to change respondents' perception of the threat of the survey: they suggest that it might contain unpleasant, difficult or even embarrassing questions. In other words, these guarantees result in a priming response effect, *i.e.*, it activates the concept of confidentiality and privacy rights in the respondent's memory, which is then given increased weight in the subsequent decision to participate or not. Extending this to the issue of mentioning the privacy rights on the welcome screen, we expect that the more words researchers use to explain these rights, the more likely potential respondents become aware of possible problems with the issue, and the less likely they will start the Web survey. However, we have no clear expectations concerning the influence of stressing the privacy rights on the breakoff rates during the Web survey.

2.3 Announced length

The decision to fill out a Web survey and to carry it out till the end is to a great extent influenced by the effort required of the respondent (Vicente and Reis 2010). This is partly determined by the perceived length of the survey (Bradburn 1978). Common sense tells us that longer surveys increase the perceived costs of participation and make it more likely that people will break off the survey prematurely.

Several studies have examined the effect of questionnaire length on response rates in Web surveys with mixed result. The meta-analysis conducted by Cook, Heath and Thompson (2000) for example, found no significant correlation between questionnaire length and response rates in Web surveys. However, questionnaire length was found to affect response rates in subsequent studies (Vicente and Reis 2010, page 256). For example, Deutskens *et al.* (2004) and Ganassali (2008) affirmed that the breakoff rate was higher in the long version of their Web survey than in the short version. Also Marcus, Bosnjak, Linder, Pilishenko and Schütz (2007) tested the relationship between the length of the Web survey and response rates in a field experiment. They found a significant effect: 30.8% responded to the short survey but only 18.6% to the longer one. This strong effect was significant throughout several other models in which they control for alternative explanations, such as the salience of the survey topic or the use of incentives.

A related issue is the announcement a priori of questionnaire length. The relationship between this announcement and the response rate has, however, more to do with the perception of the length than with the actual length of the survey. The announced length is also an indicator of the respondent's perceived burden and influences the decision to participate and to continue to participate. A few studies have experimented with this announcement. For example, Crawford *et al.* (2001) conducted an experiment to evaluate whether the previous announcement of questionnaire length would affect the percentage of people who begin the survey and whether breakoffs would be higher when the survey took longer than the promised completion time. As hypothesized, the authors found that respondents who were informed that the survey would take only eight to ten minutes to complete had a lower overall nonresponse rate than those who were told it would take 20 minutes. However, the 20-minutes group had a lower rate of breakoffs once they started the survey. These results are also found in other studies, such as those of Hogg and Mill (2003), Baker-Prewitt (2003), and of Galesic (2006).

The literature on the effect on the announced length on the response rates is closely related to the discussion on the advantages and disadvantages of using a progress indicator (see for example, Galesic and Bosnjak 2009; Heerwegh 2004). Yan, Conrad, Tourangeau and Couper (2010) for example, found that the effect of the progress indicators depend on respondents' expectations and the degree to which they were realized; the presence of a progress indicator led to fewer breakoffs when respondents expected a short task based on the invitation and when the questionnaire was indeed shorter when they expected the task to be longer.

In accordance with Crawford *et al.* (2001) and unlike the other two possible design factors, we expect that the announced length of the survey on the welcome screen influences not only the initial nonresponse but also the breakoff rate later in the survey. To be more precise, we expect that fewer respondents will start a Web survey when the announced length on the welcome screen is longer. In addition, these respondents are less likely to drop out once they have started since the real length of the survey will hardly exceed the perceived duration.

3 Research design and implementation

The experiments we describe here were included in a survey of University of Konstanz first year Bachelor and Master students conducted by the quality management unit of the university. This unit was interested in why students choose to study in Konstanz. We designed the questionnaire in close cooperation with the quality management unit, whereas the various designs of the welcome screen were conceptualized solely by us.

The different designs of the welcome screen were added after the content of the study was determined. We tested the three features in a 2 x 2 x 2 experimental design. Table 3.1 gives an overview of the control and treatment groups. See Appendix A for an example of one of the six possible welcome screens.

Table 3.1
Research design

		Privacy Right			
		Available via link		**On the welcome screen**	
		Background color			
		white	**red**	**white**	**red**
Announced survey duration	short (8min)	short white link	short red link	short white screen	short red screen
	long (20min)	long white link	long red link	long white screen	long red screen

To test the influence of background color on the likelihood of breakoff, we selected two lay-outs: one with black text on a white background and another one with black text on a red background. We are aware of the fact that the red is not a very realistic background color. Nevertheless, due to the mixed-results of

previous research, we have chosen this color as a most likely case of breakoff directly after the welcome screen. Red is a saturated color with a long wavelength. Additionally, this color might cause a negative effect on the emotional response of the respondent since it is usually used as a warning sign. However, we are acquainted with the fact that our research design cannot clearly determine which of the discussed mechanisms (*i.e.*, wavelength, saturation, or emotional response) has a possible impact. Nevertheless, we can give first insights if the color on the welcome screen is relevant at all. Note that we have verified that the display of the background colors was the same across different browsers.

To examine the effect of privacy rights on the breakoff rate, we again came up with two designs: a version in which the privacy rights were described in detail directly on the welcome screen, and another version in which the privacy rights were only briefly mentioned and respondents could use a Web link that opened a new window where their privacy rights were made clear in the same way it was done in the first version.

To test the effect of the perception of duration *i.e.*, the announced length of the Web survey on the welcome screen, we announced two different time durations needed to fill in the questionnaire. We used the result of the pretest as guidance to estimate the duration. The result of the pretest indicated that, completing the survey took on average around 12 minutes. Consequently, we decided to inform the sampled persons in one version of the survey that it would only take about 8 minutes to complete, which corresponds to the expected minimum to complete the questionnaire, whereas another group of respondents were told it would take about 20 minutes to complete. The duration to complete the survey depended to a great deal on how many answers respondents would give in the survey, since the questionnaire contained many filters. The real mean time to complete the questionnaire was 17.81 minutes (with a standard deviation of 9.01), which is considerable higher than the pretest indicated.

The invitation of the Web survey was sent to all 2,629 first year students' university e-mail accounts (See Appendix B). We focus on this particular student population since we assume that they had not been frequently exposed to Web surveys from the university and they were therefore more inclined to fill out such a survey without 'satisficing' (Toepoel, Das and Van Soest 2008). Once the students clicked on the survey link (n = 1,419), they were randomly assigned to one of the eight groups, with a minimum of 151 students and a maximum of 185 students in each treatment group. On average, there were around 177 students per treatment group with a standard deviation of 8.5 (see Table 4.1 for the exact number of respondents per treatment group).

Since the information provided in the invitation e-mail often overlaps with that provided on the welcome screen, we limited the instruction in the e-mail as much as possible to isolate the possible effect of the welcome screen. Furthermore, to keep the e-mail's layout as simple as possible no HTML format was used. With the invitation e-mail, the students received a unique URL in which their personal password was integrated, which prevented multiple completions of the survey. These e-mail messages inviting students to participate in the survey were sent on November 8, 2011 after having conducted a qualitative pretest with 15 individuals, who were working in the administration of the university, who were students, or who had experience with survey research. This pretest was focused on technical aspects of the survey and on the wording of the text and the questions.

Five days after the initial e-mail had been sent, a reminder was sent to those who had not participated in the survey yet. A final reminder was sent on November 18, 2011 not only to those who had not participated yet but also to those who had started but not completed the survey. The survey closed on

December 5, 2011. From the 2,629 students, 1,419 started the survey, from which 1,118 completed the entire survey. Completion of the survey means that the participant arrived at the last page. Since all important questions were implemented as forced-choice, item non-response is not relevant. These figures result in response rates of 43 percent using the *American Association of Public Opinion Research* (AAPOR) RR1 and 54 percent using AAPOR RR2 which takes partial response into account (AAPOR 2011).

The Web survey was implemented using the Unipark program, which is an online survey software allowing users to create Web surveys with minimal effort. The program allows for straightforward programming and is in comparison to other providers low priced for scientific use. Unipark is rather flexible in the different aspects. For example, participants can interrupt filling in the questionnaire and continue later. Furthermore, the system records information when participants break off or the time they need to complete the questionnaire and individual screens. In addition, it allows for integrating sophisticated and non standardized tools to the questionnaire design. Despite the increasing use of mobile devices, we did not implement a mobile webpage, which has proven to be a good decision as only 65 participants used mobile devices to fill in the questionnaire. However, these participants, although more likely to access the survey, were also more likely to break off (21 percent in the group without mobile devices and 35 percent with mobile devices, $p<0,01$).

4 Results

Before presenting the statistical results, we first compared the shares of our sample (*i.e.*, those students that participated in the survey) with that of the population (*i.e.*, all students that received an invitation). The invitation was sent to 2,629 students, from which just over 47 percent were male. From those that participated (1,419), just over 44 percent were male. There does not seem to be a different response behavior between men and women. However, some differences are observable when comparing the different faculty departments. Whereas the Science faculty seems to be represented adequately in the sample (both the sample and the population around 30 percent), the faculty of Humanities seems to be overrepresented (43 percent in the sample, 29 percent in the population), and the faculty of Politics, Law and Economics underrepresented (25 percent in the sample, 40 percent in the population).

In Table 4.1, the breakoff rates and the absolute number of those students who participated in the Web survey are displayed per treatment group. Notice that 83 students broke off after seeing the first screen, while 218 others broke off on other pages of the Web survey (in total the overal breakoff rate is 300). This descriptive table also shows that the breakoff rates (whether directly after the welcome screen or the breakoff rate on all other pages) are generally lower for those respondents who received a welcome screen in which privacy rights were not emphasized and in which the announced length of the survey is underestimated. However, the influence of color on the welcome screen on the breakoff rates seems to be mixed.

To test whether the patterns observed in Table 4.1 are robust and statistically significant, we conducted Logit regressions with three different dependent variables. First, a dichotomous variable comprising whether the respondent broke off directly on the welcome screen (coded as 1 or 0 otherwise). Second, a dichotomous variable taking the value of 1 if the respondent broke off on any other page than the welcome

screen (otherwise coded as 0). Third, a dichotomous variable measuring whether the respondent broke off on any page of the survey (both the welcome screen and any other page, coded as 1 or 0 otherwise). However, in case of the latter measures we cannot clearly prove if the observed effect is primarily due to the welcome screen or to an interaction between the welcome screen and the whole Web survey (or particular pages).

Table 4.1
Breakoff in the different experimental groups

	Breakoff on welcome screen		Overall breakoff rate		Total respondent n per treatment group
	n	**%**	***n***	**%**	
White, short, link	2	1.07	35	18.72	187
White, short, screen	4	2.30	31	17.83	174
White, long, link	15	7.89	49	25.79	190
White, long, screen	18	10.40	50	28.90	173
Red, short, link	3	1.79	18	10.71	168
Red, short, screen	12	6.78	35	19.77	177
Red, long, link	13	7.10	37	20.22	183
Red, long, screen	16	9.58	46	27.54	167
Total n	83		301		1,419
Mean of n (Stand. Dev.)	10.4 (6.4)		37.6 (10.7)		177.4 (8.5)

The results of the Logit regressions are presented in Table 4.2. The second column of this table presents the effects that the different treatments have on the likelihood that respondents break off on the welcome screen. The third column shows the impact of the different treatments on the breakoff likelihood during the survey excluding those respondents that broke off on the welcome screen. The fourth column looks at the effect of the design features on the overall breakoff likelihood. We have also estimated the models including all possible interaction effects between the experimental variables. However, the results did not clearly show that a combination of various treatments consequently increases the effects. Furthermore, we also included interactions effects between the experimental variables and subgroup variables such as gender or faculty. Since unambiguous subgroup differences in effects could not be identified or the variations within the subgroups were too small to estimate the model, the results of these interactions are also not presented and discussed in the following results section where we present only the parsimonious models.

In line with previous research, we expected that those respondents who received a red welcome screen are more likely to break off than those who received a white screen. Although the positive Logit coefficient in the second column indicates that there is indeed a positive relation between the red background and having a higher level of breakoffs on the welcome screen, this relationship is not statistically significant. However, there is a statistically significant negative effect of the red welcome screen on the breakoff rate on any screen except the welcome screen, *i.e.*, the combination of the red welcome screen and the other white screens of the questionnaire seems to encourage the participants to continue. When looking at the effect of the red welcome screen on the overall breakoff rate, the coefficient presented in the fourth column of the table suggests that the color of the welcome screen has no significant effect. This observation indicates that the welcome screen, although important, is just one screen of the

Web survey. As one of our pre-testers suggested, it might be the case that the color red boasts such a negative feeling that respondents click immediately further without looking at the screen. This idea was tested with an Ordinary Least Squares (OLS) regression in which the color and the other treatments as control were regressed on the amount of time spent on the welcome screen. However, the results (available upon request) did not prove any statistical significant effect. Note that the Pseudo R square values reported for the model (and across all models) is quite low. However, this is a common result for Logit regressions analyzing experimental outcomes. For example, Marcus *et al.* (2007) report a Nagelkerke's R squared of 0.041 and Bandilla, Couper and Kaczmirekt (2012) a Pseudo R square of 0.05.

Table 4.2
Logit regression

	(1) Breakoff on the welcome screen	(2) Breakoff on any page except the welcome screen	(3) Breakoff at any time of the survey
Background color: red	0.17	-0.33**	-0.20
	(0.23)	(0.15)	(0.13)
Announced duration: 20minutes	1.15***	0.23	0.53***
	(0.26)	(0.15)	(0.13)
Data security information: available via	-0.52**	-0.14	-0.28**
link	(0.23)	(0.15)	(0.13)
Constant	-3.34***	-1.61***	-1.37***
	(0.27)	(0.15)	(0.13)
N	1,419	1,419	1,419
Pseudo R-squared	0.04	0.01	0.02
Prob > chi2	0.00	0.05	0.00

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We also suspected that respondents receiving a welcome screen with the announcement that the survey only takes 20 minutes were less likely to start than those who received a welcome screen on which it was stated that the Web survey would only take eight minutes. This theoretical expectation is statistically supported by a positive and statistically significant Logit coefficient of 1.15. Additionally, we assumed that those respondents that started the 'long' survey were less inclined to break off during the Web survey. However, we did not find any support for this hypothesis. The non-significant coefficient of 0.23 means that there is no significant difference of the breakoff rates on any screen except the first page between those that received an announced 'eight minute' survey and those that received the announcement that the survey would take 20 minutes to complete. In sum, the positive and significant coefficient of 0.53 in the fourth column indicates that those respondents that received a welcome screen on which it was stated that the Web survey would take 20 minutes are more likely to breakoff the survey than those respondents who received the announcement that it would only take eight minutes. Overall, the coefficients of the announced duration are the most important ones across the different models, *i.e.*, the most important factor that explains the breakoff rates is the announced length on the welcome screen. This result is in line with the study conducted by Galesic and Bosnjak (2009) who found out that the longer stated length, the fewer respondents started and completed the questionnaire.

The last design feature that we varied on the welcome screen was the amount of emphasis that was placed on the privacy rights of respondents. We expected that the more these rights were emphasized, the more respondents would become aware of possible problems with these rights, and the less likely they were willing to start the Web survey in the first place. The results of the Logit models support this idea. The negative coefficient of -0.52 indicate that when the privacy rights are explained via a link on the Web survey (only six respondents actually opened this hyperlink), *i.e.*, few words are spend on the explanation of these rights on the screen itself, the breakoff rates on the welcome screen decreases. In other words, the priming of the privacy rights on the welcome screen increases nonresponse. In addition, explaining the privacy rights more in depth on the welcome screen has also influence on the breakoff rates during the complete survey. However, we are not sure whether the decline in response rate is due to the amount of emphasis on privacy rights or because of the length of the welcome screen (explaining the privacy rights on the welcome screen resulted in a longer screen). Further research should try to distinguish these two related processes.

5 Conclusion and discussion

One of the biggest quests of Web survey designers is to get respondents to sign on to the survey site and to keep respondents motivated to complete the survey once they have begun. However, most designers have looked into the lay-out and the design of the Web survey itself, and how this affects breakoff rates. In their quest, they have paid less attention to the important role that the welcome screen plays. This initial screen of a Web survey turns the invitee into a respondent and influences their first impression of the survey. In addition, empirical research has determined that most respondents breakoff after this initial page.

The purpose of this study was to ascertain some of the factors related to the design of the welcome screen for Web surveys affecting response rates in the electronic environment. To examine this influence, we embedded a 2x2x2 design a Web survey. The findings suggest that the lay-out of the welcome screen plays an important role in communicating to the respondent. The longer the expected announced length on the welcome screen and the more emphasis is placed on the explanation of the respondent's privacy rights, the fewer respondents started and completed the Web survey. However, background color did not have a statistical significant influence on the level of breakoff rates on the welcome screen. Only an impact during the Web survey itself was observed but as there is no significant impact on the overall breakoff rate this design feature is not regarded as relevant. Overall, based on these results we can state some more practical implications, which may help to improve the Web survey practice: (1) Keep the Web survey as short as possible. (2) Use elaborate pretests to determine reliable information concerning the time necessary to complete the survey. (3) Providing privacy rights is an important element of the welcome screen, but most respondents prefer a short description of these rights and do not want to spend too much time reading them. An appropriate way of fulfilling these wishes is providing respondents a link to a more detailed description of these rights.

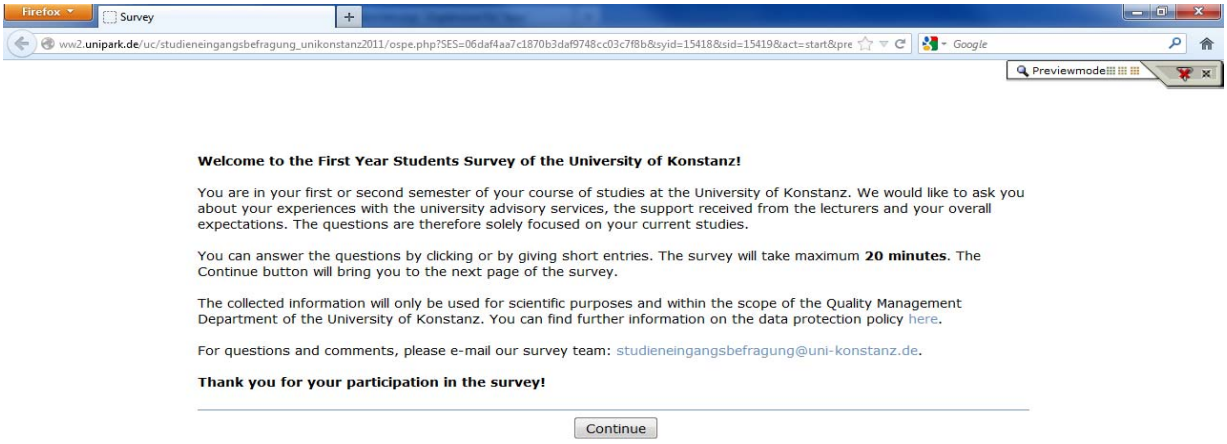
One important limitation of this study is that it was conducted on a sample of a very specific population – first-year university students. It is highly likely that the topic of the Web survey, *i.e.*, the choice of study, is highly salient among students compared to other survey topics and this might increase

the general level of response. Web survey research on populations other than students should then also determine whether the presented results are robust. In addition, it would be interesting to determine the precise effect of the emphasis on privacy rights on response rates: is it determined by the amount of words devoted to this topic, or is the emphasis given on possible problems relating to privacy rights. The specific mechanisms of the observed effect remain unclear and further research is, therefore, necessary.

Acknowledgements

We thank Christine Abele, Valentin Gold, Katharina Holzinger, and Elena Sewelies for their helpful suggestions, comments, and their (data gathering) support, and collaboration.

Appendix A



Appendix B

Dear Mrs Example,

You are now in your first or second semester of your course of studies at the University of Konstanz. In these first few weeks you have gotten to know your department, the university, and the city. We would like to know about your experience with the university advisory services, the support received from the lecturers and your overall expectations. Therefore, we would like to invite you to participate in our First Year Students Survey. Filling in this survey will help us to improve the study conditions of the University of Konstanz.

Please click on the following link to the survey (English version):

http://personalizedlink

If you cannot enter the survey via the link, please copy and paste the address in your Web browser.

The participation in this survey is voluntary. The survey is subjected to the data protection regulations and the information you provide will only be used by the Quality Management Department of the University of Konstanz and for scientific purposes. More information on data protection can be found on the welcome screen of the survey.

For questions and comments, please e-mail our survey team: studieneingangsbefragung@uni-konstanz.de.

Thank you for your participation and we wish you all the best with your course of studies!

Kind regards

References

American Association for Public Opinion Research (AAPOR) (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved February, 27, 2013, from http://aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/StandardDefinitions/StandardDefinitions2011.pdf.

Baker-Prewitt, J. (2003). *All Web Surveys are not Created Equal: Your Design Choices Can Impact Results.* Paper presented at the SumIT03 Global Market Research Symposium, Montreal.

Bandilla, W., Couper, M.P. and Kaczmirekt, L. (2012). The mode of invitation for web surveys. *Survey Practice*, 5.

Bosnjak, M., and Tute, T.L. (2001). Classifying response behavior in web surveys. *Journal of Computer Medicated Communication*, 6. Retrieved November, 21, 2012, from http://jcmc.indiana.edu/vol6/issue3/boznjak.html.

Bradburn, N.M. (1978). Respondent Burden. Proceedings of the Section of Survey Research Methods, American Statistical Association.

Brewer, P.B., Graf, J. and Willnat, L. (2003). Priming or Framing. Media Influence on Attitudes toward Foreign Countries. *Gazette: The International Journal for Communication Studies*, 65, 493-508.

Cook, C., Heath, F. and Thompson, R.L. (2000). A meta-analysis of response rates in web- or Internet-based surveys. *Educational and Psychological Measurement*, 60, 821-836.

Couper, M.P. (2000). Web surveys. A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.

Couper, M.P., Traugott, M.W. and Lamais, M.J. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65, 230-253.

Couper, M.P., and Miller, P.V. (2008). Web survey methods. Introduction. *Public Opinion Quarterly*, 72, 831-835.

Crawford, S.D., Couper, M.P. and Lamias, M.J. (2001). Web surveys. Percepsintions of Burden. *Social Science Computer Review*, 19, 146-62.

Dillman, D.A., Conradt, J. and Bowker, D. (1998). *Influence of Plain vs. Fancy Design on Response Rates for Web Surveys.* Paper presented at annual meeting of the American Statistical Association, Dallas, TX.

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

Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons, Inc.

Deutskens, E., De Ruyter, K., Wetzels, M. and Oosterveld, P. (2004). Response rate and response quality of Internet-based surveys: An experimental study. *Marketing Letters*, 15, 21-36.

Edwards, P.J., Roberts, I., Clarke, M.J., Diguiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L.M. and Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, 8, 3. Retrieved November, 21, 2012, from http://www.ncbi.nlm.nih.gov/pubmed/19588449.

Etter, J., Cucherat, M. and Perneger, T.V. (2002). Questionnaire color and response rates to mailed surveys: A randomized trial and a meta-analysis. *Evaluation & The Health Professions*, 25, 185-99.

Faubert, J. (1994). Seeing depth in colour: More than just what meets the eyes. *Vision Research*, 34, 1165-1186.

Fay, R.E., Bates, N. and Moore, J. (1991). Lower mail response in the 1990 Census: A preliminary interpretation. In *Proceedings of the Annual Research Conference of the U.S. Census Bureau*, 3-32. Washington DC: Census Bureau. Retrieved November, 27, 2012 from https://www.census.gov/srd/papers/pdf/rsm2010-13.pdf.

Galesic, M. (2006). Dropouts on the web: Effects of interest and Burden experienced during an online survey. *Journal of Official Statistics*, 22, 313-328.

Galesic, M., and Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 349-360.

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods*, 2, 21-32.

Gorn, G.J., Chattopadhyay, A., Yi, T. and Dahl, D.W. (1997). Effects of color as an executional cue in advertising: They're in the shade. *Management Science*, 43, 1387-1400.

Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.

Hall, R.H., and Hanna, P. (2004). The impact of web page text-background color combinations on readability, retention, aesthetics, and behavioral intention. *Behaviour & Information Technology*, 23, 183-195.

Heerwegh, D. (2004). Using Progress Indicators in Web Surveys. Paper prepared for the 59th AAPOR conference (Phoenix, Arizona May 13-16 2004). Retrieved November, 27, 2012, from https://perswww.kuleuven.be/~u0034437/public/Files/Heerwegh%20Using%20Progress%20Indicators.pdf.

Hillygus, S., Nie, N., Prewitt, K., and Pals, G. (2006). Civic Mobilization and Privacy Concerns in the 2000 Census. New York: Russell Sage Foundation.

Hogg, A., and Miller, J. (2003). Watch out for Dropouts. Retrieved April, 18, 2011, from http://www.quirks.com.

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

Mahon-Haft, T.A., and Dillman, D.A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods*, 4, 43-59.

Marcus, B, Bosnjak, M., Linder, S., Pilischenko, S. and Schütz, A. (2007). Compensating for low topic interest and long surveys. A field experiment on nonresponse in web surveys. *Social Science Computer Review*, 25, 372-383.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, 73, 74-97.

Pope, D., and Baker, R.P. (2005). Experiments in Color for Web-Based Surveys. Paper presented at the FedCASIC Workshops, Washington, D.C.

Redline, C., and Dillman, D.A. (2002). The influence of alternative visual designs of respondent's performance with branching instructions in self-administered questionnaire. In *Survey Response*, (Eds., R. Groves, D.A. Dillman, E. Eltinge and R. Little), 179-196. New York: John Wiley & Sons, Inc.

Singer, E., Hippler, H.J. and Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 256-268.

Singer, E., Mathiowetz, N. and Couper, M.P. (1993). The role of privacy and confidentiality as factors in response to the 1990 census. *Public Opinion Quarterly*, 57, 465-82.

Singer, E., Von Thurn, D.R. and Miller, E.R. (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly*, 59, 66-77.

Singer, E., Van Hoewyk, J. and Neugebauer, R.J. (2003). Attitudes and Behavior - the impact of privacy and confidentiality concerns on participation in the 2000 census. *Public Opinion Quarterly*, 67, 368-384.

Terhanian, G. (2005). How to Produce Credible, Trustworthy Information through Internet-Based Survey Research. Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR.

Tourangeau, R., Couper, M.P. and Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71, 91-112.

Toepoel, V., Das, M. and Van Soest, A. (2008). Effects of design in web surveys. Comparing trained and fresh respondents. *Public Opinion Quarterly*, 72, 985-1007.

Vicente, P., and Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, 28, 251-267.

Weller, L., and Livingston, R. (1988). Effect of color of questionnaire on emotional responses. *The Journal of General Psychology*, 115, 433-440.

Yan, T., Conrad, F.G., Tourangeau, R. and Couper, M.P. (2010). Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, 23, 131-147.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2013.

S.R. Amer, *RTI International, US*
R. Andridge, *Ohio State University*
N. Bates, *U.S. Census Bureau*
R. Bautista, *NORC at the University of Chicago*
J.-F. Beaumont, *Statistics Canada*
W. Bell, *U.S. Census Bureau*
D. Bellhouse, *University of Western Ontario*
E. Benhin, *Statistics Canada*
Y.G. Berger, *University of Southampton*
P. Biemer, *RTI*
H.J. Boonstra, *Statistics Netherlands*
J. Breidt, *Colorado State University*
B. Buelens, *Statistics Netherlands*
M. Callegaro, *Google Ltd, London*
P. Cantwell, *U.S. Census Bureau*
I.A. Carrillo, *RTI*
R. Chambers, *NIASRA*
R. Chambers, *University of Wollongong, Australia*
G. Chauvet, *ENSAI, France*
J. Chipperfield, *Australian Bureau of Statistics*
G. Datta, *University of Georgia*
T. DeMaio, *U.S. Census Bureau*
S. Dipko, *Westat*
G.B. Durrant, *University of Southampton*
S. Er, *Istanbul University School of Business*
V. Estevao, *Statistics Canada*
R. Fecso, *Ernst & Young LLP*
O. Fischer, *U.S. Census Bureau*
T.I. Garner, *U.S. Bureau of Labour Statistics*
C. Girard, *Statistics Canada*
Y. He, *National Center for Health Statistics*
R. Janicki, *U.S. Census Bureau*
J. Jiang, *U. of California Davis*
D. Judkins, *Abt Associates*
M.G.M. Khan, *University of the South Pacific, Fiji*
J.-K. Kim, *Iowa State University*
B. Klemens, *U.S. Census Bureau*
P. Kokic, *CSIRO, Australia*
S. Kolenikov, *Abt SRBI*
P.S. Kott, *RTI*
P. Lavallée, *Statistics Canada*
H. Lee, *Westat*
L. Lee, *NORC at the University of Chicago*
D. Liao, *RTI*
M. Link, *Neilson*
Y. Lu, *University of New Mexico*
P. Lynn, *University of Essex*
H. Mantel, *Statistics Canada*
D. Marker, *Westat*
A. Matei, *Université de Neuchâtel*
T. Merkouris, *Statistics Canada*
A. Natei, *University of Neuchatel*
J. Opsomer, *Colorado State University*
N. Prasad, *University of Alberta*
G. Ranalli, *University of Perugia, Italy*
A.P. Rao, *Statistical Consultant*
J.N.K. Rao, *Carleton University*
J. Ridenhour, *RTI*
L.-P. Rivest, *Université Laval*
L. Rizzo, *Westat*
E. Robison, *U.S. Bureau of Labor Statistics*
A. Scott, *University of Aukland*
H.-C. Shin, *National Center for Health Statistics*
N. Shlomo, *U of Manchester*
Y. Si, *Columbia University*
J. Siddique, *Northwestern University Feinberg School of Medicine*
R. Sigman, *Westat*
E.V. Slud, *U.S. Census Bureau*
P. Smith, *Office for National Statistics*
M. Sverchkov, *U.S. Bureau of Labor Statistics*
A. Théberge, *Statistics Canada*
Y. Tillé, *University of Neuchatel*
R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
M. de Toledo Vieira, *Federal University of Juiz de For a*
R. Varriale, *ISTAT*
M. Williams, *National Agricultural Statistics Service, USDA*
J. Wood, *Office for National Statistics*
C. Yu, *Iowa State University*
G. Zhang, *National Center for Health Statistics*
L.-C. Zhang, *University of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2013 issues: Céline Ethier of Statistical Research and Innovation Division, Joana Bérubé of Business Survey Methods Division, the team from Dissemination Division, in particular: Daniel Piché, Jasvinder Jassal, Martin Lachance, Jacqueline Luffman, Kathy Charbonneau and Lucie Gauthier as well as Julie Dion of Administration and Dissemination Systems Division.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

ANNOUNCEMENTS

Nominations Sought for the 2015 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2015 Waksberg Invited Address at the Statistics Canada Symposium to be held in the autumn of 2015. The paper will be published in a future issue of *Survey Methodology* (targeted for December 2015).

The author of the 2015 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2014 to the chair of the committee, Cynthia Clark (cynthia_clark@nass.usda.gov).

Previous Waksberg Award honorees and their invited papers are:

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.

2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.

2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.

2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.

2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.

2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.

2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.

2012 Lars **Lyberg**, "Survey Quality". *Survey Methodology*, vol. 38, 2, 107-130.

2013 Ken **Brewer**, "Three controversies in the history of survey sampling". *Survey Methodology*, vol. 39, 2, 249-262.

2014 Connie **Citro**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2013-2014)

Cynthia Clark, *USDA* (Chair)
Louis-Paul Rivest, *Université de Laval*
Tommy Wright, *U.S. Bureau of the Census*
J.N.K. Rao, *Carleton University*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)

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 29, No. 2, 2013

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

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

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

Contents
Volume 29, No. 3, 2013

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

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

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 2, June/juin 2013

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 3, September/septembre 2013