Journal of Statistical Planning and Inference 149 (2014) 98–115

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Maximizing the conditional overlap in business surveys

Ioana Schiopu-Kratina [a], Jean-Marc Fillion [b], Lenka Mach [c,*], Philip T. Reiss [d,e]

[a] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward, Ottawa, Ontario, Canada K1N 6N5
[b] Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6
[c] Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6
[d] Department of Child & Adolescent Psychiatry and Department of Population Health, New York University School of Medicine, 1 Park Avenue, 7th Floor, New York, NY 10016, USA
[e] Nathan S. Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA

ARTICLE INFO

Article history:
Received 10 April 2012
Received in revised form
3 February 2014
Accepted 4 February 2014
Available online 19 February 2014

Keywords:
Sample coordination
Stratified SRSWOR
Linear programming
Expected sample overlap
Row error variance

ABSTRACT

This article presents novel sequential methods of sample coordination appropriate for a repeated survey, with a stratified design and simple random sampling without replacement (SRSWOR) selection within each stratum, when the composition or definition of strata changes. Such changes could be the result of updating the frame for births, deaths, or the modification of the industry classification system. Given that a sample has already been selected according to a first (before the frame updates) SRSWOR design, our general aim is to select a minimum number of new units for the second (after the updates) survey while preserving the first-order inclusion probabilities of units in the second SRSWOR design. Sequential methods presently in use can attain a large expected overlap, but do not control the overlap on each pair of selected samples. In this article we present a set of new methods for maximizing the expected overlap, which can handle realistic situations when strata and the associated sample sizes are large. These methods include one that not only maximizes the expected overlap but, for any initially selected sample, maximizes its overlap with the second sample; its superior performance is illustrated with numerical examples.

1. Introduction

1.1. Motivation for sample coordination

Statistical agencies often need to control the overlap of samples drawn from overlapping populations. For instance, maximizing the overlap (*positive* sample coordination) can increase the precision of estimates of change between two occasions of a repeated survey and reduce the costs of first contacts. Minimizing the overlap (*negative* sample coordination) aims at minimizing the number of surveys for which the same unit is selected, to avoid overburdening of individual respondents.

Maximizing the expected overlap of samples has been a topic of interest for many years. Raj (1956) considered maximizing the expected overlap of sites (villages) to be visited by interviewers to collect information for two surveys. In this case, the overlap is controlled at the level of primary sampling units (PSUs) and the resulting design minimizes the

* Corresponding author. Tel.: +1 613 951 4754.
E-mail address: Lenka.Mach@statcan.gc.ca (L. Mach).

http://dx.doi.org/10.1016/j.jspi.2014.02.002
0378-3758

cost of interviewers' travel. On the other hand, maximizing the expected overlap at the level of ultimate sampling units reduces the cost of initiation interviews. Methods for selecting a sample on two different occasions of a repeated survey are called *sequential*, whereas *simultaneous* methods select samples for several surveys taken at the same time (Ernst, 1999).

In this article, we propose several sequential methods appropriate for a repeated survey with a stratified design and simple random sampling without replacement (SRSWOR) selection within each stratum, when the composition or definition of strata changes. Such changes could result, for example, from updating the frame for births, deaths, or the modification of the industry classification system. Sequential methods presently in use can attain a large expected overlap, but do not maximize the overlap for each initially selected sample. Our methods, by contrast, have good "conditional" properties: in addition to maximizing the expected overlap, they maximize the overlap of the initial sample with the one selected on the second occasion.

1.2. Some previous related work

Keyfitz (1951) introduced a sequential method for positive coordination when, in each stratum, one unit is selected with probability proportional to size. Kish and Scott (1971) considered the more complex problem of stratified surveys, in which units can change strata. Their procedure does not attain the maximum expected overlap of samples. Brewer et al. (1972) introduced *permanent random numbers* (PRN). Their procedure, based on Poisson sampling, attains the maximum expected overlap and applies to surveys with large sample sizes, but does not guarantee a fixed sample size.

Raj (1956) was the first to solve sample coordination problems by linear programming (LP). For small samples (e.g., when selecting PSUs in a multi-stage design), LP methods can be used to optimize the expected overlap, subject to design constraints. Such is the case in Causey et al. (1985), where sample coordination is set up as a transportation problem (TP), a type of LP problem. LP formulations have the advantage that the objective can be expressed mathematically, and attained. However, their use has been limited by the very large number of variables required in practical situations, necessitating methods to reduce the size of the problem. Earlier sample coordination methods, and methods for reducing the number of variables, are discussed by Ernst (1999) and Mach et al. (2006). More recently, the TP formulation was used by Ernst and Paben (2002), Reiss et al. (2003), Mach et al. (2007), Matei and Tillé (2006), and Matei and Skinner (2009). For stratified SRSWOR designs, Reiss et al. (2003) and Mach et al. (2006, 2007) obtain an optimal solution by grouping samples and solving first a smaller problem at the level of groups of samples, followed by a simple optimization within each group. Matei and Tillé (2006) derive conditions under which the absolute upper or lower bound of the expected overlap can be reached, using the Iterative Proportional Fitting procedure. Matei and Skinner (2009) solve a more weakly constrained problem using controlled sampling (Rao and Nigam, 1990).

Due to their flexibility and ease of implementation, PRN methods have been used by many statistical agencies for coordinating business surveys (Ohlsson, 1995; Ernst et al., 2000; Nedyalkova et al., 2008). These methods can accommodate changes in the composition of the strata and can control the overlap between samples of a repeated survey as well as between different surveys. However, some do not guarantee a fixed sample size. Furthermore, many PRN methods are not optimal in the sense that, even though the overlap of samples is controlled, the optimal expected overlap is not attained (Mach et al., 2006; Nedyalkova et al., 2008).

1.3. Objective and contribution of the present paper

Given that a sample has already been selected according to a first (before the frame updates) SRSWOR design, our general objective is to select a minimum number of new units for the second (after the updates) survey, while attaining the first-order inclusion probabilities of units, as required by the second SRSWOR design. This general problem can be solved separately in each stratum of the second survey (see Proposition A1 in Appendix A).

To attain our objective, we formulate two problems, for which we provide theoretical and numerical solutions. In Problem 1, we require the maximization of the expected overlap of samples, under the constraint that the first-order inclusion probabilities of units are attained on the second occasion. In Problem 2, we minimize the variance of row errors, which, under the same constraints, minimizes the expected value of row errors and attains the maximum overlap for each possible pair of samples.

When our solutions are applied, the overlap with the newly selected sample is large for any initial sample, and varies little from one initial sample to the next. This property is very important, since in practice an updated sample is drawn only once. It would be hard to justify to the user a procedure that discards many old units and selects many new ones, on the premise that a maximum expected overlap will have been attained, had all possible initial selections been made. The "conditional optimality" that one of our solutions attains does not seem to have been addressed elsewhere.

Problems 1 and 2 entail a large initial number of variables, so we solve each problem in two steps. We seek an optimal solution at each step and then put them together to obtain an overall optimal solution. We use LP techniques to solve the problem at the first step. At the second step, we propose several solutions, including LP formulations.

As in Mach et al. (2006), we group initial samples into *configurations* and first solve a reduced problem at the level of configurations. The technical innovation of this article is the use of *expected configurations* as variables at the first step, which exploits the weaker design constraints of this problem and considerably reduces its size. Furthermore, while in Mach et al. (2006), maximizing the overlap within blocks of configurations is trivial, in the current set-up, we must characterize