STATISTICA, anno LXXIX, n. 3, 2019

QUASI RANDOM RESAMPLING DESIGNS FOR MULTIPLE FRAME SURVEYS

Cherif Ahmat Tidiane Aidara [1]
Department of Mathematics, University of The Gambia, Brikama Campus, The Gambia

1. INTRODUCTION

Multiple frame surveys have recently received a great deal of attention from researchers (see Skinner and Rao, 1996; Lohr and Rao, 2000, 2006; Mecatti, 2007; Lohr, 2007). Although the initial use of multiple frame surveys was motivated by a reduction of survey sampling costs (see Hartley, 1962), its current use is more geared towards addressing frame undercoverage shortcomings such as capturing special, rare, and difficult-to-sample populations (see Kalton and Anderson, 1986).

In multiple frame surveys, it is assumed that the population of interest is covered adequately by a combination of sampling frames, each covering partially the population. Estimation is usually carried out by sampling independently from each of the sampling frames and creating estimators that account for units appearing in different sampling frames.

As in classical survey sampling, the standard error plays a crucial role in multiple frame survey sampling as it is used to measure the quality of survey estimators. Techniques to generate standard errors of survey estimators have been designed and can be put into two main categories namely the analytic techniques and the replication techniques. The analytic techniques include the Taylor linearization which involves differentiation of some functions of interest; whereas the replication techniques include the Jackknife and the bootstrap which rely solely on computation. The reader is invited to look into (see Lohr and Rao, 2000; Lohr, 2007).

Among all these standard error estimation techniques, the bootstrap is the most advantageous one in that it works for both smooth and non smooth statistics; in addition, it allows the user to choose the number of replication runs needed. Statistics Canada has gradually settled into using the bootstrap technique; and $R = 500$ to $R = 1000$ simulation runs are used in practice in the estimation of the standard error through the bootstrap technique (see Girard, 2009).

[1] Corresponding Author. E-mail: cataidara@utg.edu.gm

It is worth noting that the bootstrap method is based on Monte Carlo simulation which uses pseudo random numbers and has a convergence rate of $1/\sqrt{R}$ where R is the number of simulation runs. However, there exists a more efficient simulation technique (called quasi Monte Carlo simulation) which has a faster convergence rate $1/R$ and relies on quasi random numbers that are more uniformly distributed than the pseudo random numbers in the unit hypercube.

The main contribution of the present paper is that it proposes a couple of techniques that use quasi Monte Carlo simulation to generate efficient resamples for the bootstrap standard error estimation in multiple frame surveys. These techniques are inspired from Aidara (2013) and Teytaud *et al.* (2006).

The rest of the article is organized as follows. Section 2 gives a review of the standard Sobol sequence and at the same time presents the shuffled Sobol sequence. Section 3 presents a brief review of the multiplicity estimator. Section 4 proposes applications of quasi Monte Carlo simulation in bootstrap variance estimation of multiple frame survey estimators. Section 5 presents empirical studies that investigate the performance of the proposed techniques to the bootstrap variance estimation of the multiplicity estimator. Section 6 gives some concluding remarks.

2. THE SHUFFLED SOBOL SEQUENCE

As the most popular and used quasi random sequence, the Sobol sequence has been studied extensively in the applied mathematics, computer science, as well as other fields of knowledge. As a consequence we only consider a brief review of the construction of the Sobol sequence and follow the presentation in Joe and Kuo (2008).

The shuffled Sobol sequence is derived from the standard Sobol sequence which is a D-dimensional sequence (where $D \geq 2$) that uses exclusively base two in all its dimensions. The construction of the shuffled Sobol sequence follows a five-step procedure. The first four steps form the standard Sobol sequence and the fifth step performs the randomization. Since these steps are identical for each of the dimensions of the Sobol sequence, the procedure for one dimension is illustrated below.

Step 1: choose an arbitrary primitive polynomial of degree α as follows

$$P(x) = x^{\alpha} + a_1 x^{\alpha-1} + \cdots + a_{\alpha-1} x + 1 \tag{1}$$

with coefficients $a_i \in \{0, 1\}$. For instance the polynomials $x+1$ and x^2+x+1 are primitive polynomials of degree one and two respectively.

Step 2: choose arbitrarily the set of the first α initialization numbers $\{m_1, m_2, \cdots, m_\alpha\}$ such that each m_i is odd and less than 2^i for $i = 1, \cdots, \alpha$; then generate the rest of the initialization numbers through the recurrence relation

$$m_j = 2a_1 m_{j-1} \oplus 2^2 a_2 m_{j-2} \oplus \cdots \oplus 2^{\alpha-1} a_{\alpha-1} m_{j-\alpha+1} \oplus 2^{\alpha} m_{j-\alpha} \oplus m_{j-\alpha} \tag{2}$$

where a_i are the coefficients of the primitive polynomial, $j > \alpha$, and $\oplus$ is the xor operation defined as $1 \oplus 0 = 0 \oplus 1 = 1$ and $0 \oplus 0 = 1 \oplus 1 = 0$.

Step 3: define the direction numbers by

$$v_j = \frac{m_j}{2^j}.$$

This is equivalent to expressing m_j in binary representation and then shifting the position of the fractional point by j places to the left.

Step 4: use the more efficient recursive algorithm in Antonov and Saleev (1979) to calculate the $(k+1)^{th}$ Sobol number as

$$\psi_{k+1} = \psi_k \oplus v_\ell \tag{3}$$

where ℓ is the index of the first 0 digit from the right in the binary representation of k, v_ℓ is the ℓ^{th} direction number, and $\psi_0 = 0$ is assumed to be the starting point of the sequence.

Step 5: reshuffle randomly the generated standard Sobol numbers to obtain the shuffled Sobol sequence.

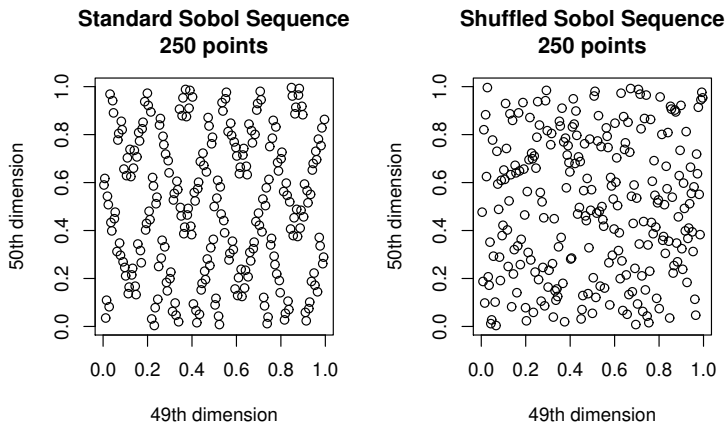

Figure 1 – 250 Sobol points.

It is worth noting that a distinct primitive polynomial is chosen for each dimension of the Sobol sequence.

The quality of the Sobol sequence depends on the choice of the initialization numbers. Any improper choice of these numbers leads to high correlations between different

dimensions of the Sobol sequence (see left hand side plot of Figure 1). Fortunately, tables of good initialization numbers together with primitive polynomials are available in the literature (for more details, see Joe and Kuo, 2008).

The shuffled Sobol sequence is contrasted with the standard Sobol sequence in Figure 1. It is obvious that the shuffled Sobol sequence displays a more uniform distribution than the standard Sobol sequence in high dimension (see right hand side plot of Figure 1).

3. Review of the multiplicity estimator

Suppose our population of interest U which consists of N units is adequately covered by $Q(\geq 2)$ overlapping sampling frames each of size $N^{(q)}$ and denoted by $A^{(q)}$ where $q = 1, \cdots, Q$. The frames divide naturally the population of interest into $2^Q - 1$ non overlapping domains. The majority of estimators in multiple frame surveys are based on these non overlapping domains (see, e.g., Halton, 1960; Skinner and Rao, 1996). However Mecatti (2007) proposed a new approach that does not make use of these non overlapping domains. This approach makes use of a partial membership information, termed *multiplicity* which is *the number of frames* a unit belongs. It is worth noting that unit multiplicity can be collected easily during the data collection process.

To illustrate the details of the multiplicity approach, suppose Q independent probability samples each of size $n^{(q)}$ and denoted by $S^{(q)}$ is selected from $A^{(q)}$ for $q = 1, \cdots, Q$. The probability sampling design for $A^{(q)}$ generates for unit k a known inclusion probability, $\pi_k^{(q)} > 0$, and a corresponding sampling design weight $d_k^{(q)} = 1/\pi_k^{(q)}$. The values (y_{qk}, m_{qk}) of the study variable y and the multiplicity m are recorded for all $k \in S^{(q)}$. At this juncture it is worth observing that every population unit $i \in U$ belongs to a finite number of sampling frames m_i and therefore any frame-specific unit (qk) corresponds to a unique $i \in U$. If the objective is to estimate a population total

$$T_y = \sum_{i=1}^{N} y_i,$$

then it is easy to verify that

$$T_y = \sum_{q=1}^{Q} \sum_{k \in A^{(q)}} y_{qk} m_{qk}^{-1}.$$

As a result, the multiplicity estimator of the population total T_y is defined as

$$t_y = \sum_{q=1}^{Q} \sum_{k \in S^{(q)}} d_k^{(q)} y_{qk} m_{qk}^{-1}.$$

The estimator t_y is indeed a function of weights and can thus be expressed as

$$t_y = f\left(\boldsymbol{d}^{(1)}, \cdots, \boldsymbol{d}^{(q)}, \cdots, \boldsymbol{d}^{(Q)}\right),$$

where $\boldsymbol{d}^{(q)}$ is the vector of weights from frame q.

4. THE PROPOSED ALGORITHMS

This section presents two algorithms that generate efficiently independent bootstrap samples for multiple frame surveys while using quasi random sequences. These algorithms are applicable to all multiple frame survey estimators. A notation similar to the one used in Lohr and Rao (2006) is adopted throughout the rest of the paper.

Let $A^{(1)},\cdots,A^{(Q)}$ be the Q overlapping sampling frames that cover adequately the population of interest. Let $A^{(q)}$, $q=1,\cdots,Q$ be partitioned into $H^{(q)}$ nonoverlapping strata $A^{(q)}=\left\{A_1^{(q)},\cdots,A_{H^{(q)}}^{(q)}\right\}$, where $A_h^{(q)}$ is comprised of $N_h^{(q)}$ primary sampling units (psu). Let $S^{(1)},\cdots,S^{(Q)}$ be Q independent samples drawn respectively from the Q frames. A probability sample $S_h^{(q)}$ is selected from the h-th stratum of the q-th sampling frame $A_h^{(q)}$. The probability sampling design generates for psu i within the h-th stratum of the q-th sampling frame a known inclusion probability, $\pi_{hi}^{(q)}>0$, and a corresponding sampling design $d_{hi}^{(q)}=1/\pi_{hi}^{(q)}$. It is worth noting that $S_h^{(q)}=\{1,2,\cdots,n_h^{(q)}\}$, $S^{(q)}=\bigcup_{h=1}^{H^{(q)}}S_h^{(q)}$, and $n^{(q)}=\sum_{h=1}^{H^{(q)}}n_h^{(q)}$ is the size of $S^{(q)}$.

The first proposed algorithm works generally in the following manner. First, a Sobol sequence of $D=\sum_{q=1}^{Q}H^{(q)}$ dimensions is generated. Then, the first $H^{(1)}$ components of the Sobol sequence are mapped to the $H^{(1)}$ strata of the sample $S^{(1)}$ on a one-to-one basis, the next $H^{(2)}$ components of the Sobol sequence are mapped to the $H^{(2)}$ strata of the sample $S^{(2)}$ on a one-to-one basis, and so on. It is worth noting that the Sobol sequence elements are mapped through the quantile function of the binomial distribution.

For a detailed illustration of the proposed algorithm, suppose a bootstrap sample $S_h^{*(q)}$ is drawn from $S_h^{(q)}$ by sampling with replacement for $h=1,2,\cdots,H^{(q)}$ independently. Sampling with replacement the units of $S_h^{(q)}$ ensures a selection probability of $1/n_h^{(q)}$ for each unit. If $x_{hi}^{(q)}$ denotes the number of times unit i of stratum h appears in the bootstrap sample $S_h^{*(q)}$ and $m_h^{(q)}(\leq n_h^{(q)})$ the number of draws, then $S_h^{*(q)}$ can be represented as $\left\{x_{h1}^{(q)},\cdots,x_{hn_h^{(q)}}^{(q)}\right\}$ which has a multinomial distribution with size $m_h^{(q)}=\sum_{i=1}^{n_h^{(q)}}x_{hi}^{(q)}$ and cell probabilities $p_{hi}^{(q)}=1/n_h^{(q)}$. It is worth noting that each $x_{hi}^{(q)}$ is binomially distributed with size $m_h^{(q)}$ and probability $p_{hi}^{(q)}$. In addition, if $i-1$ units have already been observed, then the i-th unit $x_{hi}^{(q)}$ has a binomial distribution with size $m_h^{(q)}-\sum_{j=1}^{i-1}x_{hj}^{(q)}$ and probability $p_{hi}^{(q)}/\left(1-\sum_{j=1}^{i-1}p_{hj}^{(q)}\right)$.

The bootstrap sample $S_h^{*(q)}$ is generated by mapping the elements of the h-th component of the $H^{(q)}$ components of the Sobol sequence to the units of $S_h^{(q)}$ as shown in

the following procedure.

1. Generate $\psi_{h1}^{(q)}$ from the h-th dimension of the Sobol sequence.
2. Define $x_{h1}^{(q)}$ minimal such that $\mathfrak{D}\left(\psi_{h1}^{(q)}\right) = \inf\left\{x_{h1}^{(q)} : \psi_{h1}^{(q)} \leq \text{Prob}\left(X_{h1}^{(q)} \leq x_{h1}^{(q)}\right)\right\}$ where $X_{h1}^{(q)}$ is binomially distributed with size $m_h^{(q)}$ and probability $p_{h1}^{(q)}$.
3. For $i = 2$ to $n_h^{(q)}$
   1. Generate $\psi_{hi}^{(q)}$ from the h-th dimension of the Sobol sequence.
   2. Define $x_{hi}^{(q)}$ minimal such that $\mathfrak{D}\left(\psi_{hi}^{(q)}\right) = \inf\left\{x_{hi}^{(q)} : \psi_{hi}^{(q)} \leq \text{Prob}\left(X_{hi}^{(q)} \leq x_{hi}^{(q)}\right)\right\}$ where $X_{hi}^{(q)}$ is binomially distributed with size $m_h^{(q)} - \sum_{j=1}^{i-1} x_{hj}^{(q)}$ and probability $p_{hi}^{(q)} / \left(1 - \sum_{j=1}^{i-1} p_{hj}^{(q)}\right)$.
4. Repeat steps 1 to 3 for $h = 1, \cdots, H^{(q)}$ to obtain the bootstrap sample from the stratified sample $S^{*(q)} = \bigcup_{h=1}^{H^{(q)}} S_h^{*(q)}$
5. Repeat steps 1 to 4 for $q = 1, \cdots, Q$ to obtain the bootstrap sample for the multiple frame survey.

The second proposed algorithm uses a Sobol sequence whose dimension is equal to the number of sampling frames i.e. $D = Q$. In this method, the first component of the Sobol sequence is mapped to the elements of the sample drawn from the first sampling frame, the second component of the Sobol sequence is mapped to the elements of the sample drawn from the second sampling frame, and so on. The details of the mapping are presented in the following algorithm:

1. Generate $\psi_{h1}^{(q)}$ from the q-th dimension of the Sobol sequence.
2. Define $x_{h1}^{(q)}$ minimal such that $\mathfrak{D}\left(\psi_{h1}^{(q)}\right) = \inf\left\{x_{h1}^{(q)} : \psi_{h1}^{(q)} \leq \text{Prob}\left(X_{h1}^{(q)} \leq x_{h1}^{(q)}\right)\right\}$ where $X_{h1}^{(q)}$ is binomially distributed with size $m_h^{(q)}$ and probability $p_{h1}^{(q)}$.
3. For $i = 2$ to $n_h^{(q)}$
   1. Generate $\psi_{hi}^{(q)}$ from the q-th dimension of the Sobol sequence.
   2. Define $x_{hi}^{(q)}$ minimal such that $\mathfrak{D}\left(\psi_{hi}^{(q)}\right) = \inf\left\{x_{hi}^{(q)} : \psi_{hi}^{(q)} \leq \text{Prob}\left(X_{hi}^{(q)} \leq x_{hi}^{(q)}\right)\right\}$ where $X_{hi}^{(q)}$ is binomially distributed with size $m_h^{(q)} - \sum_{j=1}^{i-1} x_{hj}^{(q)}$ and probability $p_{hi}^{(q)} / \left(1 - \sum_{j=1}^{i-1} p_{hj}^{(q)}\right)$.

4. Repeat steps 1 to 3 for $h = 1, \cdots, H^{(q)}$ to obtain the bootstrap sample from the stratified sample $S^{*(q)} = \bigcup_{h=1}^{H^{(q)}} S_h^{*(q)}$. Notice that the first $n_1^{(q)}$ elements of the Sobol sequence are used for the first stratum of $S^{*(q)}$ that is $S_1^{*(q)}$, the second $n_2^{(q)}$ elements of the Sobol sequence are used for the second stratum of $S^{*(q)}$ that is $S_2^{*(q)}$, and so on.

5. Repeat steps 1 to 4 for $q = 1, \cdots, Q$ to obtain the bootstrap sample for the multiple frame survey.

At this juncture, it is befitting to present both a flowchart and an example that illustrate the use of these algorithms. Since the algorithms are similar, only a flowchart for the first algorithm from steps 1 to 3 is presented and captured in Figure 2.

As for the example, consider the population in Table 1 where the column names are uid, h, i, y represent respectively the unique identity number of the element, the stratum, the relative number of the stratum element, some parameter of interest. Without loss of generality, $Q(=2)$ overlapping frames (as in Tables 2, 3) with $H^{(1)} = H^{(2)}(=2)$ strata are selected from the population independently. Notice that a further column denoted m is added in these tables to specify whether or not the element is only captured in one frame or both. Consider further that two stratified samples $S^{(1)}$ and $S^{(2)}$ are selected independently, using the sampling package (Tillé and Matei, 2016), from the frames such that $S^{(1)} = S_1^{(1)} \cup S_2^{(1)}$ and $S^{(2)} = S_1^{(2)} \cup S_2^{(2)}$. Suppose we selected five elements in $S_1^{(1)}$ and four elements in $S_2^{(1)}$; three elements in $S_2^{(2)}$ and four elements in $S_2^{(2)}$. Using the uid, the selected elements are as follows $S_1^{(1)} = \{1, 3, 5, 10, 11\}$ and $S_2^{(1)} = \{16, 32, 34, 35\}$ from the first and second stratum of the first frame; and $S_1^{(2)} = \{8, 9, 10\}$ and $S_2^{(2)} = \{19, 20, 24, 27\}$ from the first and second stratum of the second frame. From the first proposed algorithm, we have to generate a reshuffled Sobol sequence of 4 dimensions and use the first 2 dimensions for the resample $S^{*(1)}$ and the remaining 2 dimensions for the resample $S^{*(2)}$. Upon applying this algorithm, we have obtained the following stratum resamples $S_1^{*(1)} = \{1, 0, 1, 0, 2\}, S_2^{*(1)} = \{1, 1, 0, 1\}$, and $S_1^{*(2)} = \{0, 1, 1\}, S_2^{*(2)} = \{0, 1, 1, 1\}$. The numbers in the stratum resample should be understood as the number of times the corresponding elements are selected.

From the second proposed algorithm, we have to generate a reshuffled Sobol sequence of 2 dimensions and use the elements in the first dimension for the resample $S^{*(1)}$ and the elements in the second dimension for the resample $S^{*(2)}$. In details, we first generate a sequence of length $n^{(1)} = n_1^{(1)} + n_2^{(1)}$ in the first dimension. Then we use the first $n_1^{(1)}$ points for the first stratum and the next $n_2^{(1)}$ points for the second stratum. The same approach is used for the resample $S^{*(2)}$. Upon applying this algorithm, we have obtained the following stratum resamples $S_1^{*(1)} = \{1, 0, 1, 0, 2\}, S_2^{*(1)} = \{1, 1, 1, 0\}$ and $S_1^{*(2)} = \{1, 1, 0\}, S_2^{*(2)} = \{2, 0, 0, 1\}$.

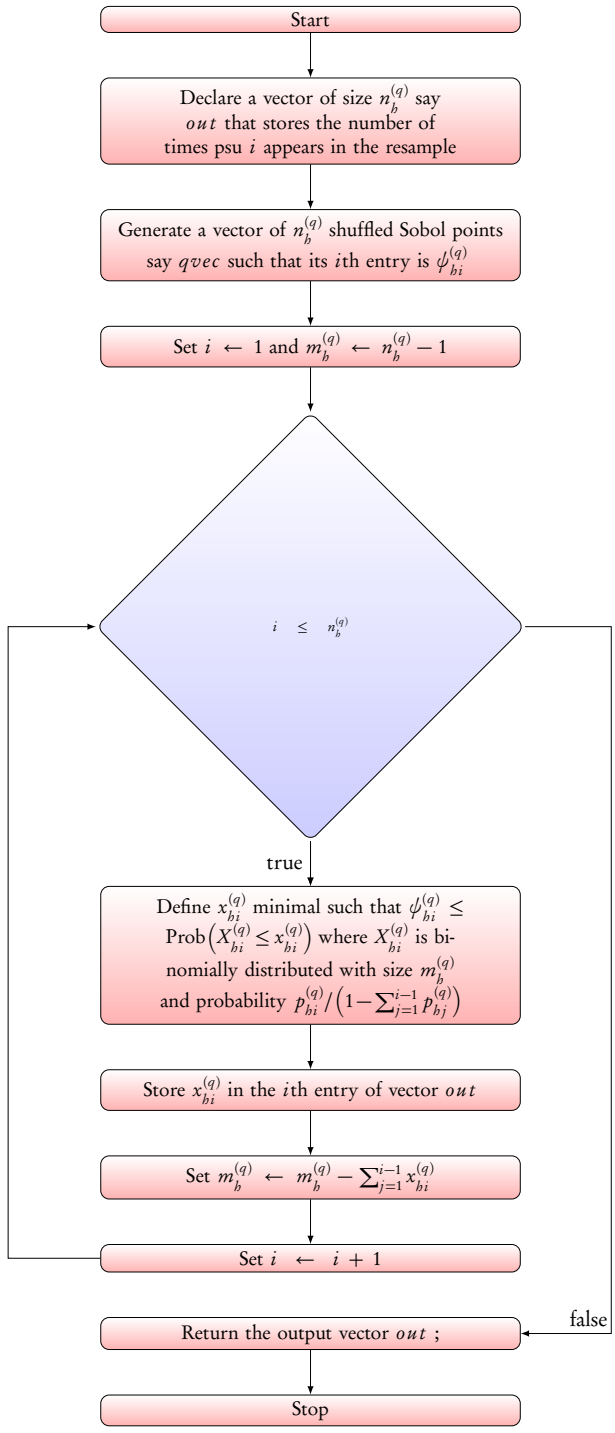

Figure 2 – Flowchart for the selection of the resample $S_h^{*(q)}$ through the first algorithm steps 1-3.

TABLE 1
A toy population from which two overlapping frames are extracted.

uid	h	i	y
1	1	1	1.693
2	1	2	1.980
3	1	3	0.904
4	1	4	0.536
5	1	5	0.990
6	1	6	2.147
7	1	7	1.014
8	1	8	0.895
9	1	9	1.918
10	1	10	1.302
11	1	11	3.667
12	1	12	0.150
13	1	13	1.059
14	1	14	3.100
15	1	15	3.503
16	2	1	3.187
17	2	2	0.456
18	2	3	1.389
19	2	4	2.466
20	2	5	2.923
21	2	6	1.546
22	2	7	6.133
23	2	8	0.848
24	2	9	1.003
25	2	10	1.677
26	2	11	3.040
27	2	12	3.444
28	2	13	1.781
29	2	14	5.052
30	2	15	0.530
31	2	16	1.690
32	2	17	0.970
33	2	18	1.660
34	2	19	0.755
35	2	20	1.817

TABLE 2
First frame extracted from the toy population.

uid	h	i	y	m
1	1	1	1.693	1
2	1	2	1.980	2
3	1	3	0.904	1
4	1	4	0.536	2
5	1	5	0.990	2
7	1	6	1.014	1
10	1	7	1.302	2
11	1	8	3.667	1
13	1	9	1.059	1
14	1	10	3.100	1
16	2	1	3.187	2
17	2	2	0.456	1
20	2	3	2.923	2
25	2	4	1.677	2
26	2	5	3.040	1
28	2	6	1.781	2
29	2	7	5.052	2
31	2	8	1.690	1
32	2	9	0.970	1
34	2	10	0.755	1
35	2	11	1.817	2

TABLE 3
Second frame extracted from the toy population.

uid	h	i	y	m
2	1	1	1.980	2
4	1	2	0.536	2
5	1	3	0.990	2
6	1	4	2.147	1
8	1	5	0.895	1
9	1	6	1.918	1
10	1	7	1.302	2
12	1	8	0.150	1
15	1	9	3.503	1
16	2	1	3.187	2
18	2	2	1.389	1
19	2	3	2.466	1
20	2	4	2.923	2
21	2	5	1.546	1
22	2	6	6.133	1
23	2	7	0.848	1
24	2	8	1.003	1
25	2	9	1.677	2
27	2	10	3.444	1
28	2	11	1.781	2
29	2	12	5.052	2
30	2	13	0.530	1
33	2	14	1.660	1
35	2	15	1.817	2

It is worth noting that the second algorithm presents an advantage over the first one in that the dimension of the Sobol sequence is reduced significantly. Knowing that quasi random sequences present a better uniform distribution of their elements in low dimension, the results from this algorithm are expected to be more accurate.

Then the extended Rao-Wu bootstrap weight for the frames proposed in Lohr (2007) was used to compute the variance. Suppose $d_{hi}^{(q)}$ is the weight attached to unit i of stratum h for the q-th sampling frame. Then the corresponding bootstrap weight for the b-th simulation run is denoted by $d_{hi}^{(q)}[b]$ and defined as

$$d_{hi}^{(q)}[b] = d_{hi}^{(q)} \frac{n_h^{(q)}}{m_h^{(q)}} x_{hi}^{(q)}[b],$$

where $x_{hi}^{(q)}[b]$ is the number of times unit i of stratum h is selected in the b-th simulation run.

To estimate the bootstrap variance of an estimator say t, the separated and combined bootstrap approaches proposed in Lohr (2007) are used. For the separated bootstrap approach, B_q bootstrap samples are created from the sample $S^{(q)}$ of the q-th frame A_q using the above algorithm. For each of the $H^{(q)}$ dimensions associated with the sample $S^{(q)}$, $B_q n_h^{(q)}$ elements are generated and the first $n_h^{(q)}$ elements are used for the creation of the h-th stratum of the first bootstrap sample $S_h^{*(q)}[1]$, the next $n_h^{(q)}$ elements are used for the creation of the h-th stratum of the second bootstrap sample $S_h^{*(q)}[2]$, and so on. Therefore, the b-th bootstrap sample is defined as $S^{*(q)}[b] = \bigcup_{h=1}^{H^{(q)}} S_h^{*(q)}[b]$. The separated bootstrap method is estimated by

$$v_s = \sum_{q=1}^{Q} \frac{1}{B_q} \sum_{b=1}^{B_q} (t_y^q[b] - t_y)^2. \tag{4}$$

Note that $t_y^q[b] = f(\boldsymbol{d}^{(1)}, \cdots, \boldsymbol{d}^{(q)}[b], \cdots, \boldsymbol{d}^{(Q)})$ which means that the original weights in $S^{(q)}$ are replaced by the bootstrap weights for just the frame q in the b-th simulation run

For the combined bootstrap approach, B bootstrap samples are created from the sample $S^{(q)}$ of the q-th frame A_q using the above algorithm. For each of the $H^{(q)}$ dimensions associated with the sample $S^{(q)}$, $Bn_h^{(q)}$ elements are generated and the first $n_h^{(q)}$ elements are used for the creation of the h-th stratum of the first bootstrap sample $S_h^{*(q)}[1]$, the next $n_h^{(q)}$ elements are used for the creation of the h-th stratum of the second bootstrap sample $S_h^{*(q)}[2]$, and so on. Therefore, the b-th bootstrap sample is

defined as $S^{*(q)}[b] = \bigcup_{h=1}^{H^{(q)}} S_h^{*(q)}[b]$. The combined bootstrap method is given by

$$v_c = \frac{1}{B}\sum_{b=1}^{B}(t_y[b] - t_y)^2. \tag{5}$$

Note that $t_y[b] = f(\boldsymbol{d}^{(1)}[b], \cdots, \boldsymbol{d}^{(q)}[b], \cdots, \boldsymbol{d}^{(Q)}[b])$ which means that in the b-th simulation run, the original weights in the Q frames are simultaneously replaced by the bootstrap weights from the Q independent samples.

For the second algorithm, the method works as follows. For the q-th dimension, $B_q n^{(q)}$ elements are generated and the first $n^{(q)}$ elements are used for the creation of the first bootstrap sample. The next $n^{(q)}$ elements are used for the creation of the second bootstrap sample and so on. It is worth noting that the size of $S^{(q)}$ is $n^{(q)} = \sum_{h=1}^{H^{(q)}} n_h^{(q)}$ and that the b-th bootsrap sample is obtained by using the first $n_1^{(q)}$ elements of the sequence for the sampling of the units of the first stratum of $S^{(q)}$, the next $n_2^{(q)}$ elements of the sequence for the sampling of the units of the second stratum of $S^{(q)}$, and so on.

5. SIMULATION STUDY

A limited simulation study was carried out to investigate the performance of the proposed methods in a three-frame setup. Each of the six different stratified finite populations described in Chen and Sitter (1999) was used in some preliminary simulations. For each population, we created three overlapping frames and tested the proposed methods; and there were no significant differences in the performance of the proposed methods. Hence we decided to report the results of our simulation study using only one of these six populations namely population 2 of Chen and Sitter (1999). As a reminder, population 2 had $H = 4$ strata with stratum sizes $N_h = 8000 - 300h$ for $h = 1, 2, 3, 4$. For the ith unit within the hth stratum, the characteristics x_{hi} were generated by adding $h/2$ to χ^2_{2h} variate and the y_{hi} were generated using the model

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \gamma_h x_{hi}^2 + \xi_h x_{hi}^a \epsilon_{hi} \tag{6}$$

for specific values of α_h,β_h,γ_h,a and ξ_h, where ϵ_{hi} are random variables, independent and identically distributed over i, from a chi-square distribution with b_h degrees of freedom, χ^2_{hi}.

For each parameter combination, N_h pairs of characteristic variables (x_{hi}, y_{hi}) were generated using (6). The six parameter combinations used to generate the stratified finite population 2 are given in Table 4.

After constructing population 2, we formed $Q = 3$ overlapping sampling frames namely A_1, A_2 and A_3 in the following manner. First every pair (y_j, x_j) where j is a unique element of population 2 that corresponds to stratum specific unit hi was randomly assigned to the $Q = 3$ sampling frames according to 3 independent Bernoulli

TABLE 4
Parameter settings for generated finite Population 2.

h	α_h	β_h	γ_h	ξ_h	a	ϵ_h
1	2	0.5	0	0.2	-0.5	χ^2_3
2	6	1.0	0	0.2	-0.5	χ^2_4
3	10	-0.5	0	0.2	-0.5	χ^2_5
4	14	-1.0	0	0.2	-0.5	χ^2_6

trials with probability $\alpha_q = N^{(q)}/N$ for $q = 1,2,3$. Then we made sure that none of the sampling frames was empty and that when combined, they covered adequately the population of interest.

Three stratified random samples $S^{(1)}$, $S^{(2)}$ and $S^{(3)}$ of sizes $n^{(1)}$, $n^{(2)}$ and $n^{(3)}$ units were then selected from frame A_1, A_2 and A_3 respectively. The parameters considered here are the population total of the study variable y denoted by T_y, the population size N, and the ratio T_y/T_x respectively.

For the purpose of comparison, the following variance estimators were considered: the separated bootstrap (LSEP) as described in Lohr (2007), the combined bootstrap (LCOM) as described in Lohr (2007), the separated bootstrap using the first proposed algorithm (QSEP1), the combined bootstrap using the first proposed algorithm (QCOM1), the separated bootstrap using the second proposed algorithm (QSEP2), and the combined bootstrap using the second proposed algorithm (QCOM2).

A total of $B = 1000$ simulation runs were performed. For each simulation run, three independent samples $S^{(1)}$, $S^{(2)}$ and $S^{(3)}$ were selected from each frame independently and variance estimates were created using the above four variance estimators. The sizes for the bootstrap samples were $m^{(1)} = n^{(1)} - 1$, $m^{(2)} = n^{(2)} - 1$ and $m^{(3)} = n^{(3)} - 1$ for $S^{(1)}$, $S^{(2)}$ and $S^{(3)}$ respectively. For all these methods, the number of replications was $R = 100$. The true MSEs are approximated by $10,000$ simulation runs.

All computations were performed in R (R Core Team, 2018). The R package randtoolbox (Christophe and Petr, 2015) was used to generate the Sobol numbers for the proposed quasi Monte Carlo methods. The Rcpp package (Eddelbuettel and Balamuta, 2017) was used to implement the ratio and total estimators.

The performance of the above variance estimators was measured and compared in terms of the simulated relative percentage bias (RB %), coefficient of variation (CV), and empirical coverage probabilities of 95 % confidence intervals (Coverage). The simulated values of RB and CV for a particular variance estimator v were computed as

$$RB = 100 \times \frac{1}{B} \sum_{b=1}^{B} \frac{v_b - MSE}{MSE} \tag{7}$$

and

$$CV = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(v_b - MSE)^2}/MSE, \tag{8}$$

where v_b is the variance estimate of v for the b-th simulated sample.

TABLE 5
Comparison of variance estimators for t_y.

Method	RB %	CV	Coverage(95%)
LSEP	1.35	0.1871	94.4
LCOM	1.03	0.2179	94.1
QSEP1	1.26	0.1874	94.9
QCOM1	1.98	0.2320	94.5
QSEP2	0.97	0.1839	94.7
QCOM2	1.21	0.2248	94.4

TABLE 6
Comparison of variance estimators for $\widehat{N}$.

Method	RB %	CV	Coverage(95%)
LSEP	1.66	0.1290	94.6
LCOM	1.98	0.1806	94.0
QSEP1	1.35	0.1295	94.7
QCOM1	1.13	0.1749	94.4
QSEP2	1.53	0.1276	94.7
QCOM2	1.59	0.1723	95.0

From Tables 5 and 6, we clearly see that the relative bias of the population total t_y and population size $\widehat{N}$ is small and positive. It is also clear that the highest RB for these estimators is 1.98 whereas the smallest RB is 0.97. In terms of probabilities of coverage, the two estimators perform comparably and well in the sense that both estimators produce a coverage very close to 95 %. From Table 7, it is clear that the RB for the ratio estimator t_y/t_x is smaller and negative. In terms of absolute values the smallest RB for the ratio estimator is 0.08 and the highest RB is 0.86. The coverage probabilities of Table 7 are within an acceptable range though they are a little less than those for the estimators t_y and $\widehat{N}$.

From Tables 5, 6, 7, it is clear that the coefficients of variation for t_y and t_y/t_x are very similar. It is worth noting that the CV for the separate bootstrap methods is a little smaller than that of the combined bootstrap methods for all estimated population quantities. From Table 6, the CV for the separate bootstrap methods of the estimator $\hat{N}$ is the smallest and varies around 0.13.

TABLE 7
Comparison of variance estimators for t_y/t_x.

Method	RB %	CV	Coverage(95%)
LSEP	-0.45	0.1917	93.6
LCOM	0.86	0.2279	93.4
QSEP1	-0.21	0.1899	93.5
QCOM1	-0.27	0.2351	93.4
QSEP2	-0.15	0.1879	93.4
QCOM2	0.08	0.2238	93.8

6. Conclusion

The bootstrap variance estimation technique is very useful for assessing the quality of estimators in complex surveys, particularly when non linear estimators are involved. This paper has proposed two new algorithms that generate efficiently resampling designs using the reshuffled Sobol sequences in multiple frame surveys. The methods perform well and comparably with already established bootstrap methods in Lohr (2007).

Future work will involve the use of quasi random numbers to generate without replacement subsamples for multiple frame surveys as well as establishing formally their asymptotic properties.

References

C. A. T. Aidara (2013). *Bootstrap variance estimation for complex survey data: A quasi Monte Carlo approach*. Sankhya B, 75, no. 1, pp. 29–41.

I. Antonov, V. Saleev (1979). *An economic method of computing LP-sequences*. USSR Computational Mathematics and Mathematical Physics, 19, pp. 252–256.

J. Chen, R. Sitter (1999). *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys*. Statistica Sinica, 9, pp. 385–406.

D. Christophe, S. Petr (2015). *Randtoolbox: Generating and Testing Random Numbers*. R package version 1.17.

D. Eddelbuettel, J. J. Balamuta (2017). *Extending R with C++: A Brief Introduction to Rcpp*. PeerJ Preprints, 5:e3188v1. URL `https://doi.org/10.7287/peerj.preprints.3188v1`.

C. Girard (2009). *The Rao-Wu rescaling bootstrap: From theory to practice*. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. Federal Committee on Statistical Methodology, Washington, DC, pp. 2–4.

J. HALTON (1960). *On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals*. Numerische Mathematik, 2, pp. 84–90.

H. HARTLEY (1962). *Multiple frame surveys*. In *Proceedings of the Social Statistics Section*.

S. JOE, F. Y. KUO (2008). *Notes on generating Sobol sequences*. URL `https://web.maths.unsw.edu.au/ fkuo/sobol/joe-kuo-notes.pdf`. Available online.

G. KALTON, D. ANDERSON (1986). *Sampling rare populations*. Journal of the Royal Statistical Society, Series A, 149, no. 1, pp. 65–82.

S. LOHR (2007). *Recent developments in multiple frame surveys*. Journal of the American Statistical Association, pp. 3257–3264.

S. LOHR, J. RAO (2000). *Inference from dual frame surveys*. Journal of the American Statistical Association, 95, no. 449, pp. 271–280.

S. LOHR, J. RAO (2006). *Estimation in multiple- frame surveys*. Journal of the American Statistical Association, 101, no. 475, pp. 1019–1030.

F. MECATTI (2007). *A single frame multiplicity estimator for multiple frame surveys*. Component of Statistics Canada, Catalogue n0.12-001-X, Business Survey Methods Division.

R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

C. SKINNER, J. RAO (1996). *Estimation in dual frame surveys with complex design*. American Statistical Association, 91, no. 433, pp. 349–356.

O. TEYTAUD, S. GELLY, S. LALLICH, E. PRUDHOMME (2006). *Quasi-random resamplings, with applications to rule extraction, cross-validation and (su-)bagging*. Dans International Workshop on Intelligent Information Access III A 2006.

Y. TILLÉ, A. MATEI (2016). *Sampling: Survey Sampling*. URL `https://CRAN.R-project.org/package=sampling`. R package version 2.8.

SUMMARY

In this paper, we present two new algorithms that use the shuffled Sobol sequence to generate the bootstrap resampling designs in multiple frame surveys. We investigate the performance of the proposed algorithms in a simulation study using a three-overlapping frame setup design. The samples were selected independently from the frames using a stratified simple random sampling design. The performance of the proposed methods is comparable with the already established ones such as the Lohr-Rao bootstrap methods for multiple frame surveys in terms of relative percentage bias, coefficient of variation, and empirical coverage probabilities of 95 percent confidence interval.

Keywords: Multiple frame surveys; Bootstrap variance method; Shuffled Sobol sequence.