مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۱۷، شماره‌ی ۱، بهار و تابستان ۱۳۹۹، صص ۱–۱۸

J. Statist. Res. Iran **17** (2020): 1–18
DOI: 10.52547/jsri.17.1.1
DOR: 20.1001.1.25385771.2020.17.1.1.2

Balanced Acceptance Sampling^{+m}: A Balance Between Entropy and Spatially Balance

Hossein Veisipour[†], Mohammad Moradi[†,*] and Jennifer Brown[‡]

[†] Razi University
[‡] University of Canterbury

Received: 2022/19/01 Approved: 2022/18/04

Abstract. Balanced acceptance sampling is a relatively new sampling scheme that has potential to improve the efficiency of spatial studies. There are two drawbacks of the design, it can have low entropy and some of the unbiased estimates can not be calculated. In this paper, such shortcomings have been addressed by integrating simple random sampling with balanced acceptance sampling. In a simulation study on two datasets, the entropy and spatially balance of the introduced sampling design are calculated and are compared with the same results from balanced acceptance sampling and simple random sampling. Simulation results show that the introduced sampling design has the flexibility to balance the entropy and spatially balance.

Keywords. Entropy of sampling design, Halton sequence, inclusion probability, spatially balanced sampling.

MSC 2010: 62D05.

1 Introduction

In two-dimensional populations where the variable of interest has a spatial trend, using sampling designs which are spatially balanced can improve

* Corresponding author

the efficiency of estimators (e.g., Stevens and Olsen (2004); Theobald et al. (2007); Grafström (2012); Grafström et al. (2012); Grafström and Lundström (2013)). In spatially balanced sampling (SBS) designs, sample units are selected that are well-spread over the area, with few nearby units. Some of the interesting properties of SBS in natural resource applications, discussed by Theobald et al. (2007), include high efficiency, good spatial distribution, easy to calculate estimators and high cost-effective.
Spatially balanced sampling designs have been applied in soil studies, environmental protection, and mapping. Systematic sampling with a regular grid, has been studied by Bickford et al. (1963) , Messer et al. (1986), Hazard (1989), and is an example of the early SBS designs.

Robertson et al. (2013) introduced another SBS design, balanced acceptance sampling (BAS), which can give better spatial balance than some of the other SBS designs.

One limitation in the calculation of the statistical estimators with BAS, and some other SBS designs, is that the second order inclusion probabilities are approximately zero for neighborhood units. In such situations, the variance estimator, introduced by Horvitz and Thompson (1952) and Yates and Grundy (1953), cannot be calculated. Also, as with systematic sampling, BAS has low entropy. Sampling designs with high entropy generate highly randomized and more robust samples (Tillé and Haziza (2010)). Grafström (2010) mentioned that sampling designs that yield low entropy in general should be avoided. Another advantage of high entropy, for unequal probability sampling, is that an approximation of the second order inclusion probabilities can be derived from the first order inclusion probabilities (Brewer and Donadio (2003); Matei and Tillé (2005); Rao et al. (2008)).

Poisson sampling and Bernoulli sampling designs have high entropy. Tillé and Haziza (2010) show that the simple random sampling without replacement (SRSWOR) and Bernoulli sampling have approximately the same entropy, when the population size is enough large.

We introduce a modification to BAS, BAS^{+m}, where the sample is selected in two stages, a BAS sample is selected in the first stage and in the second stage, the remained sample size is selected by SRSWOR. BAS^{+m} can be considered to be a flexible sampling design between the two extremes of BAS and SRSWOR. The SRSWOR yields a good entropy sample and the BAS yields a good spatially balanced sample. Also, as shown in Section 2, the second order inclusion probabilities in BAS^{+m} are non-zero.

In section 2, BAS, BAS^{+m}, and the first and the second order inclusion

probabilities of BAS^{+m} are introduced. When the researcher is free to define sampling units a closed form is given for inclusion probabilities. In section 3, a simulation study is undertaken with two datasets and an artificial population. Finally, concluded remarks are given in section 4.

2 Inclusion Probabilities for BAS$^{+\mathrm{m}}$

Let a population with units $u_1, u_2, ..., u_N$ and the corresponding study variables $y_1, y_2, ...y_N$ be located in a d-dimensional area. In the BAS design, random points are determined sequentially from the random start Halton sequence (Wang and Hickernell (2000)). The d-dimensional Halton sequence (Halton (1960)) $\{X_k\}_{k=0}^{\infty} = \{(x_k^{(1)}, x_k^{(2)}, ..., x_k^{(d)})\}_{k=0}^{\infty}$ is a deterministic sequence in $[0,1)^d$, where the ith component of the kth point, $x_k^{(i)}$, is defined, using van der Corput sequence, as follows:
Any natural number, k, has a base b_i representation of the form

$$k = \lambda_M \lambda_{M-1} ... \lambda_1 \lambda_0 = \sum_{j=0}^{M} \lambda_j b_i^j,$$

where $\lambda_j \in \{0, 1, ..., b_i - 1\}$ is the jth digit of k in its base b_i expansion and M is a positive integer. Van der Corput sequence is given by defining the radical inverse function $x_k^{(i)}$ as follows:

$$x_k^{(i)} = 0.\lambda_0 \lambda_1 ... \lambda_{M-1} \lambda_M = \sum_{j=0}^{M} \frac{\lambda_j}{b_i^{j+1}}.$$

The sequence is uniformly distributed, if all bases b_i for $i = 1, \ldots, d$, are pairwise co-prime (Faure et al. (2015)). If the first term is randomly selected from the sequence, then the given sequence is known as a random start Halton sequence.

First, the N points are replaced with N non-overlapping equally sized boxes with positive Lebesgue measure, where each box contains exactly one point. Then, a random-start Halton sequence is defined over a minimal bounding box containing all N boxes. If the random selected value from Halton sequence, say x_1, is within a unit's box, that unit is included in the sample. Otherwise, no unit is selected. The next point in the sequence, x_2, is then considered and the method repeats (Robertson et al. (2013)).

The exact first order inclusion probability for unit i is simply the fraction of sequences that contain unit i, and is given by

$$\pi_i = \frac{1}{R^d} \sum_{j=1}^{R^d} I_i(\{X_k^{(j)}\}_{k=1}^{v}),$$

where ν is the least integer such that n distinct sample units are observed, $I_i(\{X_k^{(j)}\}_{k=1}^{v}) = 1$ if unit i is selected by the jth random-start Halton sequence, $X^{(j)}$, and zero otherwise. The exact second order inclusion probability for units i and j is given similarly as the fraction of sequences that contain both units i and j, and is given as follows:

$$\pi_{ij} = \frac{1}{R^d} \sum_{j'=1}^{R^d} I_{ij}(\{X_k^{(j')}\}_{k=1}^{v}),$$

where $I_{ij}(\{X_k^{(j')}\}_{k=1}^{v}) = 1$ if both units i and j are selected by the j'th random-start Halton sequence, $X^{(j')}$, and zero otherwise (Robertson et al. (2013)).

For inclusion probabilities, closed forms can be found, if sample units are set based on a feature of the Halton series, explained by Price and Price (2012); Halton (1960); Robertson et al. (2017). Price and Price (2012), in Proposition 2, show a feature of Halton sequence, where in two-dimensional case, for $(b_1, b_2) = (2, 3)$, it can be written as follows:

"Let the unit hypercube be subdivided into boxes of equal size and shape, where each box is of the form

$$\left[m_1 b_1^{-J_1}, (m_1+1) b_1^{-J_1}\right) * \left[m_2 b_2^{-J_2}, (m_2+1) b_2^{-J_2}\right) \tag{1}$$

for integers m_1 and m_2 satisfying $0 \le m_1 < b_1^{J_1}$, $0 \le m_2 < b_2^{J_2}$ and J_1, J_2 are positive integers.

Then any consecutive $N = b_1^{J_1} b_2^{J_2} = 2^{J_1} 3^{J_2}$ points in the scrambled prime recycling Halton sequence have exactly one point in each box."

Boxes are usually indexed by $\{0, 1, ..., B-1\}$ and to adopt sampling literature, we have indexed it as $\{1, ..., N\}$. Here, sample units are set as the given boxes and the initial sample unit, say u_r, is selected randomly from the set $\{u_1, ..., u_N\}$. The remaining sample units are selected consecutively

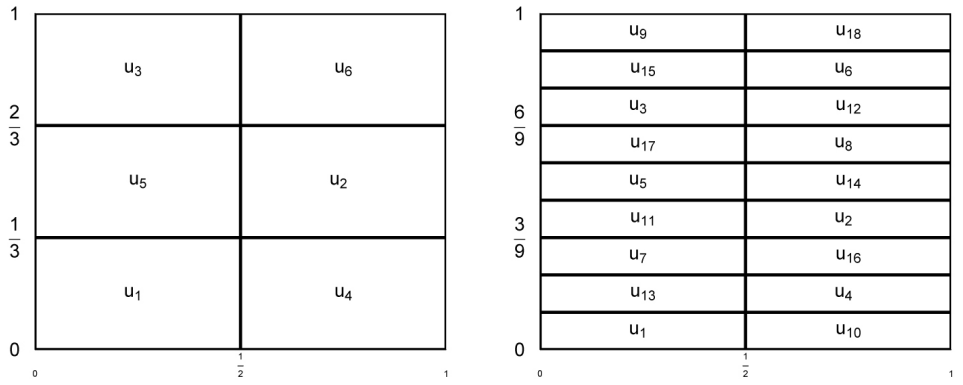

Figure 1. Halton boxes with $b_1 = 2$ and $b_2 = 3$ for different J_i values. Left: $J = (1, 1)$ and $N = 2 \times 3 = 6$; Right: $J = (1, 2)$ and $N = 2 \times 3^2 = 18$.

from the set $\{u_1, ..., u_N\}$ similar to circular systematic sampling design. More detailed are explained in Example 1.

Example 1. In a population of size $N = b_1^{J_1} b_2^{J_2} = 2 \times 3 = 6$, all of the sample units (Halton boxes (1)) and BAS samples of size $n = 3$ are shown in Table (1).
Since, $0 \leq m_1 < 2$ and $0 \leq m_2 < 3$ the intervals $\left[m_1 b_1^{-J_1}, (m_1 + 1) b_1^{-J_1}\right)$ are as $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ and intervals $\left[m_2 b_2^{-J_2}, (m_2 + 1) b_2^{-J_2}\right)$ are as $[0, \frac{1}{3})$, $[\frac{1}{3}, \frac{2}{3})$ and $[\frac{2}{3}, 1)$.

For example, if the random value is $r = 5$, then $5 \; mod \; 2 = 1$, the first interval is $[0, \frac{1}{2})$, and $5 \; mod \; 3 = 2$, the second interval is $[\frac{1}{3}, \frac{2}{3})$, and the first sample unit of the BAS sample is then $u_5 = [0, \frac{1}{2}) * [\frac{1}{3}, \frac{2}{3})$ and the BAS sample is given by $\{u_5, u_6, u_1\}$. Partitioning the area into 6 sample units is shown in the left panel of Figure 1 and for $N = 2 * 3^2 = 18$ sample units is shown in the right panel.

Here, the first and the second order inclusion probabilities are given in Lemma 1.

***Lemma* 1.** *In a BAS sample of size n, where sample units are set as the given boxes by Halton series, the inclusion probabilities for units (boxes)* u_i

Table 1. All BAS samples of size $n = 3$, in a population of size $N = 6$, $b_1 = 2$ and $b_2 = 3$.

r	1	2	3	4	5	6
(m_1, m_2)	$(0,0)$	$(1,1)$	$(0,2)$	$(1,0)$	$(0,1)$	$(1,2)$
$\lceil m_1/2, (m_1+1)/2 \rceil$	$[0, \frac{1}{2})$	$[\frac{1}{2}, 1)$	$[0, \frac{1}{2})$	$[\frac{1}{2}, 1)$	$[0, \frac{1}{2})$	$[\frac{1}{2}, 1)$
$\lceil m_2/3, (m_2+1)/3 \rceil$	$[0, \frac{1}{3})$	$[\frac{1}{3}, \frac{2}{3})$	$[\frac{2}{3}, 1)$	$[0, \frac{1}{3})$	$[\frac{1}{3}, \frac{2}{3})$	$[\frac{2}{3}, 1)$
The first sample unit	$u_1 = [0, \frac{1}{2})[0, \frac{1}{3})$	$u_2 = [\frac{1}{2}, 1)[\frac{1}{3}, \frac{2}{3})$	$u_3 = [0, \frac{1}{2})[\frac{2}{3}, 1)$	$u_4 = [\frac{1}{2}, 1)[0, \frac{1}{3})$	$u_5 = [0, \frac{1}{2})[\frac{1}{3}, \frac{2}{3})$	$u_6 = [\frac{1}{2}, 1)[\frac{2}{3}, 1)$
BAS sample	$s_1 = \{u_1, u_2, u_3\}$	$s_2 = \{u_2, u_3, u_4\}$	$s_3 = \{u_3, u_4, u_5\}$	$s_4 = \{u_4, u_5, u_6\}$	$s_5 = \{u_5, u_6, u_1\}$	$s_6 = \{u_6, u_1, u_2\}$

and u_j are given as follows:

$$\pi_{ij} = \left(\frac{n - |i-j|}{N}\right) I_{\{0,1,2,\ldots,n-1\}}(|i-j|) + \left(\frac{n - N + |i-j|}{N}\right) I_{\{N-n+1, N-n+2, \ldots, N-1\}}(|i-j|), \qquad (2)$$

obviously, for $i = 1, 2, ..., N$; where $\pi_{ii} = \pi_i$ and $\pi_i = n/N$.

Proof. *Proof in Appendix.* □

To check the accuracy of the inclusion probabilities, some general properties must be satisfied. For example, $\sum_{i=1}^{N} \pi_i = n$, is satisfied with the introduced the first order inclusion probabilities, where it can be proven easily. Also, the property of the second order inclusion probability, that summation of joint inclusion probabilities is equal to $n(n-1)$, is shown in Appendix (؟؟).

2.1 The first order inclusion probabilities in BAS^{+m}

As mentioned in Section (1), in the first stage of BAS^{+m}, a BAS of size $n-m$ is selected and in the next stage, by SRSWOR a sample of size m is selected from the $N-n+m$ remaining sample units, where m can be varied between 0 and n.

Let π_i be the first order inclusion probability for unit i in a BAS sample of size $n-m$, then in BAS^{+m}, the first order inclusion probability for unit i, shown by π_i^{+m}, are given as follows;

$$\begin{aligned}\pi_i^{+m} &= Pr(u_i \in s_{_{BAS^{+m}}}) = Pr(A_i \cup A_i'B_i)\\ &= Pr(A_i) + Pr(A_i'B_i)\\ &= Pr(A_i) + Pr(A_i')Pr(B_i|A_i')\\ &= \pi_i + (1-\pi_i)\frac{m}{N-n+m},\end{aligned}$$

where A_i and B_i are the events that unit u_i is selected by BAS and SRSWOR, respectively. Obviously, for $\pi_i = (n-m)/N$ it can be proven that $\pi_i^{+m} = n/N$.

2.2 The Second Order Inclusion Probabilities in BAS^{+m}

Let π_{ij} be the second order inclusion probability for units i and j in a BAS sample of size $n-m$, then in BAS^{+m}, the second order inclusion probability for such units, shown by π_{ij}^{+m}, are given as follows:

$$\begin{aligned}\pi_{ij}^{+m} &= Pr(\{u_i, u_j\} \subset s_{_{BAS^{+m}}})\\ &= Pr\left(A_iA_j \bigcup A_iA_j'B_j \bigcup A_i'A_jB_i \bigcup A_i'A_j'B_iB_j\right)\\ &= Pr(A_iA_j) + Pr(A_iA_j')Pr\big(B_j|A_iA_j'\big)+\\ &\quad Pr(A_i'A_j)Pr\big(B_i|A_i'A_j\big) + Pr(A_i'A_j')Pr(B_iB_j|A_i'A_j')\\ &= \pi_{ij} + (\pi_i - \pi_{ij})\frac{m}{N-n+m} + (\pi_j - \pi_{ij})\frac{m}{N-n+m}+\\ &\quad (1-\pi_i-\pi_j+\pi_{ij})\frac{m}{N-n+m} \times \frac{m-1}{N-n+m-1},\end{aligned}$$

It can be easily shown that when $m > 0$, the second order inclusion probability for each pair is non-zero.

2.3 Entropy

The entropy of a sampling design $p(.)$, denoted by $I(p)$, is defined as

$$I(p) = -\sum_{s\in Q} Pr(s)\log(Pr(s))$$

where $Pr(s)$ is the probability of selecting the sample s, $Q = \{s|Pr(s) > 0\}$ is the support of the sampling design $p(.)$ (Tillé and Haziza (2010)).

In a population of size N, the entropy of SRSWOR is given by

$$I(p_{SRS}) = \log N! - \log n! - \log(N-n)!,$$

In BAS^{+m}, if sample units are defined as Halton boxes, the probability of selecting a sample s is given by

$$Pr_{BAS^{+m}}(s) = N^{-1} \left(\begin{array}{c} N-n+m \\ m \end{array} \right)^{-1}$$

Therefore, the entropy is given by

$$I(p_{_{BAS^{+m}}}) = \log N + \log(N-n+m)! - \log m! - \log(N-n)!$$

Generally, when sample units do not correspond to Halton boxes, in a two-dimensional population with $d = 2$, the probability of selecting the sample s and the entropy is given by

$$\begin{aligned} p_{_{BAS^{+m}}}(s) =& R_{dist}^{-d} \left(\begin{array}{c} N-n+m \\ m \end{array} \right)^{-1} = R_{dist}^{-2} \left(\begin{array}{c} N-n+m \\ m \end{array} \right)^{-1} \\ I(p_{_{BAS^{+m}}}) =& d \log R_{dist} + \log(N-n+m)! - \log m! - \log(N-n)! \\ =& 2 \log R_{dist} + \log(N-n+m)! - \log m! - \log(N-n)!, \end{aligned}$$

where R_{dist} is the greatest integer that yields distinct samples. It should be noted that if the population area is continuous and sample units are assumed to be the selected points from the Halton sequence, then all the selected points will be distinct and R_{dist} can be large as possible. If sample units are assumed to be a sub-area like a rectangular, many points from the Halton sequence will be located inside each sample unit, therefore each sample unit would be selected many times by Halton sequence. For example, in the Volcano population explained in Section (3), each BAS sample of size n=30 would be repeated after the next 13436928 points of the Halton sequence. In such a cases, to calculate the exact entropy the R_{dist} should be assumed to be the greatest integer that yields distinct samples. Clearly, by set $m = 0$, entropy of BAS can be derived as follows:

$$I(p_{BAS}) = 2 \log R_{dist}$$

2.4 Spatial Balance

The spatial balance (SB) can be evaluated with different methods. In this paper, following Robertson et al. (2018), the method of Voronoi polygons suggested by Stevens and Olsen (2004) is used. For a given sample, $s = \{u_1, u_2, \ldots, u_n\}$, it is calculated as follows:

$$SB = \frac{1}{n} \sum_{u_i \in s} (v_i - 1)^2.$$

where $v_i = \sum_{j:u_j \in \omega_i} \pi_j$ and ω_i is the Voronoi polygon generated by unit u_i. The SB value is non-negative, the minimum value $SB = 0$ is given from a sample that has the maximum spatially balance and greater values of SB show weak spatially balances. For each sampling design, this criteria is given by a Monte Carlo simulation, so that SB is calculated for a large number of samples, say $B = 200000$, then they have been averaged as follow:

$$\overline{SB} = \frac{\sum_{b=1}^{B} SB_b}{B}.$$

In Figure (2), two samples of size $n = 30$ are selected from a population of size $N = 10 \times 10 = 100$ by BAS and SRSWOR, their spatial balances are as $SB_{BAS} = 0.064$ and $SB_{SRSWOR} = 0.361$. In this example, the sample from BAS has more spatial balance than the one from SRSWOR.

3 Simulation Study

In this section, the entropy and spatial balance of BAS^{+m} are calculated for two real populations. Also, in an artificial population, for the case that sample units are defined as Halton boxes, the entropy and spatial balance of BAS^{+m} are calculated. In a simulation study, by 200000 selected BAS^{+m} samples, the average SB, $\overline{SB}$, of BAS^{+m} is calculated for all combinations of sample sizes $n = \{30, 90\}$ and $m = \{0, \frac{n}{6}, \frac{n}{3}, \frac{n}{2}, n\}$.

The case study datasets are from a study of volcanoes in New Zealand and a study of marine crabs in Qatar. The datasets are just two examples

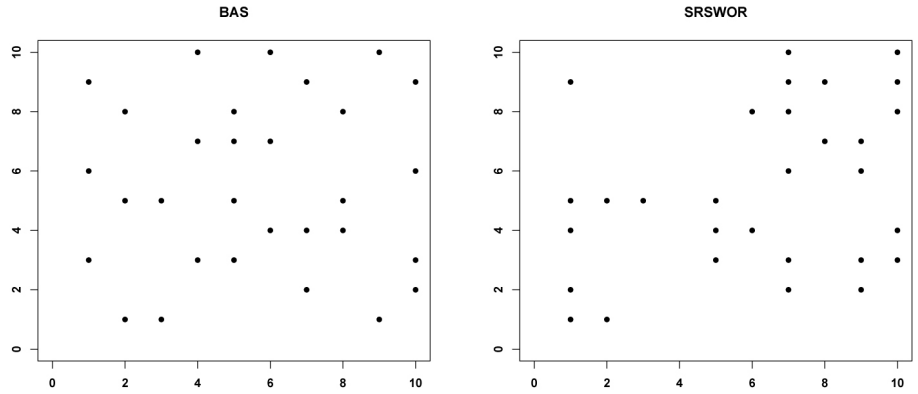

Figure 2. Two generated samples of size $n = 30$ by BAS and SRSWOR from a population of size $N = 10 \times 10 = 100$. The given spatial balances are as $SB_{BAS} = 0.064$ and $SB_{SRSWOR} = 0.361$.

of populations where spatial information is available. The populations are quite different in size so that the effect of population size can be evaluated. The volcano dataset is available in R (Team, R (2017)) and is data from Maunga Whau, one of the volcanoes in the Auckland volcanic field. The dataset has topographic information for *Auckland's Maunga Whau Volcano* on a 10m by 10m grid presented in a matrix with 87 rows and 61 columns. The artificial population, the edited Volcano, was created to use Halton boxes as sample units in the Volcano population. Then we assume the Volcano area is partitioned into 5184 sample units with $81 = 3^4$ rows and $64 = 2^6$ columns.

The other dataset is of the count of crab burrows for *a Nasima dotilliformis* poulation in the intertidal mudflat of Al Khor from $400 \times 400 = 160000$ equal quadrats. Salehi et al. (2015) investigated the effect of quadrat size and constructed different quadrat sizes by partitioning the area into 50×50 equal quadrats.

The results are shown in Table 2, spatial balance and entropy values are shown in two columns. Since in the case $m = 0$, BAS^{+m} is transformed to a BAS of size n and in the case $m = n$, BAS^{+m} is transformed to a SRSWOR of size n, we have compared both entropy and spatial balance of BAS^{+m} rather than BAS and SRSWOR with equal sample size.

The results show that when m is increased from 0 to n, the $\overline{SB}$ is decreased and entropy is increased. It can be concluded that by determining an

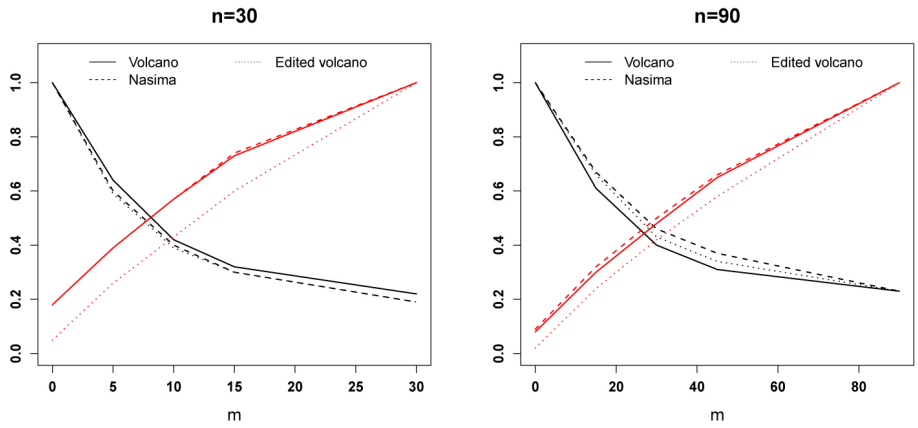

Figure 3. The red curves show the relative entropy of BAS^{+m} compared with $SRSWOR$, $\frac{I(p_{BAS^{+m}})}{I(p_{SRS})}$, for different m values. The black curves show the relative spatial balance of BAS compared with BAS^{+m}, $\frac{\overline{SB}_{BAS}}{\overline{SB}_{BAS^{+m}}}$, for different m values.

appropriate value of m, a balance can be made between spatial balance and entropy. For example, in Volcano population, if 25 units be selected by BAS and 5 units by SRSWOR, the spatially balance is equal to 64% of the case that 30 units are selected by BAS and the entropy is 38% of the case that 30 units are selected by SRSWOR. Also, comparing the entropy of BAS with equal sample sizes in the Volcano and the edited Volcano datasets, Tables 2 and 3, shows that in the edited Volcano entropy is smaller than for the Volcano dataset. When sample units are defined as Halton boxes (refer to Formula(1) and Example 1), simple formulae are given for inclusion probabilities, although we note that the entropy is decreased compared with when the sample units are defined arbitrary. Determining the value of m depends on the degree of interest in both entropy and SB. Figure 3 shows that SB decreases rapidly with increasing m. We recommend that, if SB is more of interest, m should not be more than about 20% of the sample size. As Figure 3 shows, the relative entropy in both populations are approximately equal but their relative SB are slightly different, especially for larger sample size n=90. When the sample units are defined as the generated Halton boxes as in the edited Volcano population, the relative entropy is decreased compared with the Volcano population.

Table 2. The spatial balance and the entropy of BAS^{+m}, in Volcano population $R_{dist}(n = 30) = 13436928$ and $R_{dist}(n = 90) = 107495424$, in Nasima population $R_{dist}(n = 30) = 1353024$ and $R_{dist}(n = 90) = 21695488$.

n-m	m	$\frac{\overline{SB}_{SRS}}{\overline{SB}_{BAS+m}}$	$\frac{\overline{SB}_{BAS}}{\overline{SB}_{BAS+m}}$	$\overline{SB}_{BAS+m}$	$\frac{I(p_{BAS+m})}{I(p_{BAS})}$	$\frac{I(p_{BAS+m})}{I(p_{SRS})}$	$I(p_{BAS+m})$
Volcano population							
30	0	4.54	1	0.07	1.00	0.18	32.83
25	5	2.91	0.64	0.11	2.16	0.39	70.89
20	10	2.00	0.42	0.16	3.15	0.57	103.44
15	15	1.52	0.32	0.21	4.07	0.73	133.52
0	30	1.00	0.22	0.32	5.56	1.00	182.56
90	0	4.43	1.00	0.07	1.00	0.08	36.99
75	15	2.82	0.61	0.11	3.72	0.30	137.51
60	30	1.94	0.40	0.16	5.93	0.48	219.21
45	45	1.55	0.31	0.20	7.93	0.65	293.25
0	90	1.00	0.23	0.37	12.50	1.00	453.00
Nasima population							
30	0	5.33	1.00	0.06	1.00	0.18	28.24
25	5	2.91	0.60	0.11	2.21	0.39	62.51
20	10	2.00	0.40	0.16	3.23	0.57	91.27
15	15	1.52	0.30	0.21	4.16	0.74	117.56
0	30	1.00	0.19	0.32	5.56	1.00	159.89
90	0	4.29	1	0.07	1.00	0.09	33.79
75	15	2.73	0.67	0.11	3.63	0.32	122.75
60	30	1.88	0.46	0.16	5.71	0.50	192.94
45	45	1.50	0.37	0.20	7.56	0.66	255.52
0	90	1.00	0.23	0.30	11.11	1.00	384.39

Table 3. The spatial balance and the entropy of BAS^{+m} in a population of size $N = 3^4 \times 2^5 = 5184$, where sample units are defined as Halton boxes.

		The edited Volcano					
n-m	m	$\frac{\overline{SB}_{SRS}}{\overline{SB}_{BAS+m}}$	$\frac{\overline{SB}_{BAS}}{\overline{SB}_{BAS+m}}$	$\overline{SB}_{BAS+m}$	$\frac{I(p_{BAS+m})}{I(p_{BAS})}$	$\frac{I(p_{BAS+m})}{I(p_{SRS})}$	$I(p_{BAS+m})$
30	0	5.33	1	0.06	1.00	0.05	8.55
25	5	2.91	0.59	0.11	5.44	0.26	46.51
20	10	2.00	0.39	0.16	9.23	0.43	78.93
15	15	1.52	0.30	0.21	12.74	0.60	108.89
0	30	1.00	0.19	0.32	20.00	1.00	181.86
90	0	4.43	1.00	0.07	1.00	0.02	8.55
75	15	3.10	0.66	0.10	12.71	0.24	108.71
60	30	1.94	0.43	0.16	22.23	0.42	190.06
45	45	1.55	0.34	0.20	30.85	0.58	263.74
0	90	1.00	0.23	0.31	50.0	1.00	450.87

4 Discussion

We have suggested the BAS^{+m} sampling design, to overcome the problem in calculating unbiased estimators for some parameters like variance, covariance, correlation coefficients and regression coefficient, and to improve the entropy of the design.

We also introduced a close form for calculating the first and the second order inclusion probabilities, when sample units in a unit square population are set as generated Halton boxes. Since box sides are set as $[0, 1/2^{J_1}] \times [0, 1/3^{J_2}]$ and values J_1, J_2 can be determined arbitrary, setting sample units as Halton boxes is not a cumbersome requirement. Defining sample units, as arbitrary equal rectangular shapes, is used conventionally in many fields of research such as agriculture, environment, mining, oil exploration, etc. In future works by taking into account efficiency, entropy and spatial balance, it can be seek to find the optimal m.

Acknowledgment

The authors are very grateful to the associate editor and two anonymous reviewers for their many useful comments.

References

Bickford, C.A., Mayer, C., and Ware, K. (1963). An efficient sampling design for forest inventory: The northeastern forest survey. *Journal of Forestry*, **61**, 826-833.

Brewer, K.R., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, **29**, 189-196.

Faure, H., Kritzer, P., and Pillichshammer, F. (2015). From van der Corput to modern constructions of sequences for quasi-Monte Carlo rules. *Indagationes Mathematicae*, **26**, 760-822.

Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statistical Methodology*, **7**, 84-97.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, **142**, 139-147.

Grafström, A., Lundström, N.L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, **68**, 514-520.

Grafström, A., and Lundström, N.L. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, **3**, 36-41.

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

Hazard, J.W. (1989). *Forest survey methods used in the USDA Forest Service.* Environmental Research Laboratory, Office of Research and Development, US Environmental Protection Agency.

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.

Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics,* **21**, 543-570.

Messer, J.J. (1987). *National Surface Water Survey: National Stream Survey, Phase I–pilot Survey.* US Environmental Protection Agency, Environmental Monitoring Systems Laboratory.

Price, C.J., and Price, C.P. (2012). Recycling primes in Halton sequences: an optimization perspective. *Adv Model Optim*, **14**, 17-29.

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

Robertson, B.L., Brown, J.A., McDonald, T., and Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, **69**, 776-784.

Robertson, B.L., McDonald, T., Price, C.J., and Brown, J.A. (2017). A modification of balanced acceptance sampling. *Statistics & Probability Letters*, **129**, 107-112.

Robertson, B., McDonald, T., Price, C., and Brown, J. (2018). Halton iterative partitioning: spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, **25**, 305-323.

Salehi, M., Moradi, M., Al Khayat, J.A., Brown, J., and Yousif, A.E.M. (2015). Inverse adaptive cluster sampling with unequal selection probabilities: case studies on crab holes and arsenic pollution. *Australian & New Zealand Journal of Statistics*, **57**, 189-201.

Stevens Jr, D.L., and Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the american Statistical association*, **99**, 262-278.

Team, R.C. (2013). R: A language and environment for statistical computing.

Theobald, D.M., Stevens, D.L., White, D., Urquhart, N.S., Olsen, A.R., and Norman, J.B. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management*, **40**, 134-146.

Tillé, Y., and Haziza, D. (2010). An interesting property of the entropy of some sampling designs. *Survey Methodology*, **36**, 229-231.

Wang, X., and Hickernell, F.J. (2000). Randomized halton sequences. *Mathematical and Computer Modelling*, **32**, 887-899.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, **15**, 253-261.

Appendix

If in a BAS sample of size n, sample units are set as the given Halton boxes, the inclusion probabilities for units (boxes) u_i and u_j are given as follows:

$$\pi_{ij} = (\frac{n-|i-j|}{N})I_{\{0,1,2,...,n-1\}}(|i-j|) + (\frac{n-N+|i-j|}{N})I_{\{N-n+1,N-n+2,...,N-1\}}(|i-j|),$$

Proof. The quasi-periodic property of the Halton sequence is that for any point X_k, the points $X_{k+N}, X_{k+2N}, ...$ will be located in the same Halton box (defined in 1) including X_k, where N is the number of boxes. The property means consecutive points from the sequence cyclically visit the boxes in a specific order determined by the mod N values (Price and Price (2012); Halton (1960)) .

Consider a finite population of size N and let $U = \{u_1, u_2, \ldots, u_N\}$ be the set of units , where u_i is corresponding to ith Halton box. Since the first box is selected randomly with equal chance and the next boxes are selected consecutively, then the number of BAS samples with size n in this case is equal to N, that can be shown as follows:

$$\left|\begin{array}{l} s_1 = \{\ u_1, u_2, ..., u_n\ \} \\ s_2 = \{\ u_2, u_3, ..., u_{n-1}\ \} \\ \vdots \\ s_{N-n+1} = \{u_{N-n+1},\ u_{N-n+2}, \ldots,\ u_N\} \\ s_{N-n+2} = \{u_{N-n+2},\ u_{N-n+3}, \ldots,\ u_N,\ u_1\} \\ \vdots \\ s_N = \{u_N, u_1, ..., u_{n-2}\} \end{array}\right|$$

therefor the second order inclusion probability, π_{ij}, is given as follows:

$$\pi_{ij} = (\frac{n-|i-j|}{N})I_{\{0,1,2,...,n-1\}}(|i-j|) + (\frac{n-N+|i-j|}{N})I_{\{N-n+1,N-n+2,...,N-1\}}(|i-j|)$$

□

We have proven that $\sum_{i\neq j=1}^{N} \pi_{ij} = n(n-1)$ is satisfied for π_{ij} from the formula (2).

Proof. Clearly, it can be written as:

$$\begin{aligned}
\sum_{i<j=1}^{N} \pi_{ij} &= \sum_{i<j=1}^{N} \Big[(\frac{n-|i-j|}{N}) I_{\{0,1,2,\ldots,n-1\}}(|i-j|) + (\frac{n-N+|i-j|}{N}) I_{\{N-n+1,\ldots,N-1\}}(|i-j|) \Big] \\
&= \sum_{n-(j-i)>0}^{N} \frac{n-(j-i)}{N} + \sum_{(n-N)+(j-i)>0}^{N} \frac{(n-N)+(j-i)}{N} \\
&= \sum_{i=1}^{N-(n-1)} \sum_{j=i+1}^{i+(n-1)} \frac{n-(j-i)}{N} + \sum_{i=N-(n-2)}^{N-1} \sum_{j=i+1}^{N} \frac{n-(j-i)}{N} \\
&+ \sum_{(n-N)+(j-i)>0}^{N} \frac{(n-N)+(j-i)}{N} \qquad *
\end{aligned}$$

The first term of equation * is simplified as follows:

$$\begin{aligned}
\sum_{i=1}^{N-(n-1)} \sum_{j=i+1}^{i+(n-1)} \frac{n-(j-i)}{N} &= \sum_{i=1}^{N-(n-1)} \frac{(n-1)+(n-2)+\ldots+(n-(n-1))}{N} \\
&= \sum_{i=1}^{N-(n-1)} \frac{1+2+\ldots+(n-1))}{N} = (\frac{N-(n-1)}{N})(\frac{n(n-1)}{2}) \\
&= \frac{n(n-1)}{2} - \frac{n(n-1)^2}{2N} \qquad \texttt{A}
\end{aligned}$$

The second term of equation * is simplified as follows:

$$\begin{aligned}
&\sum_{i=N-(n-2)}^{N-1} \sum_{j=i+1}^{N} \frac{n-(j-i)}{N} = \sum_{i=N-(n-2)}^{N-1)} \frac{(n-1)+(n-2)+\ldots+(n-(N-i))}{N} \\
&= \sum_{i=N-(n-2)}^{N-1)} \frac{(N-i)n-(1+2+\ldots+N-i)}{N} \\
&= \sum_{i=N-(n-2)}^{N-1)} \frac{(N-i)n}{N} - \sum_{i=N-(n-2)}^{N-1)} \frac{(N-i)(N-i+1)}{N} \\
&= \frac{n}{N}\left[1+2+\ldots+(n-2)\right] - \frac{1}{2N}\left[1\times 2+2\times 3+\ldots+(n-2)\times(n-1)\right] \\
&= \frac{n}{N}\left[\frac{(n-1)(n-2)}{2}]\right] - \frac{1}{2N}\sum_{i=1}^{n-2} i(i-1) \\
&= \frac{n}{N}\left[\frac{(n-1)(n-2)}{2}]\right] - \frac{1}{2N}\left[\frac{(n-2)(n-1)(2(n-2)+1)}{6} + \frac{(n-2)(n-1)}{2}\right] \\
&= \frac{n(n-1)(n-2)}{3N} \qquad \texttt{B}
\end{aligned}$$

The last term of equation * is simplified as follows:

$$\begin{aligned}\sum_{(n-N)+(j-i)>0}^{N} \frac{(n-N)+(j-i)}{N} &= \sum_{i=1}^{N-1} \frac{(1+2+\ldots+(n-i))}{N} \\ &= \frac{1}{N}\sum_{i=1}^{n-1} \frac{(n-i)(n-i+1)}{2} \\ &= \frac{1}{2N}\sum_{i=1}^{n-1} i(i+1) = \frac{1}{2N}\left[\frac{n(n-1)(2(n-1)+1)}{6} + \frac{n(n-1)}{2}\right] \\ &= \frac{n(n-1)(n+1)}{6N} \qquad \text{c}\end{aligned}$$

By substituting the obtained A, B and C in equation $*$, the $\sum_{i\neq j=1}^{N} \pi_{ij}$ can be written as:

$$\begin{aligned}\sum_{i<j=1}^{N} \pi_{ij} &= \frac{n(n-1)}{2} - \frac{n(n-1)^2}{2N} + \frac{n(n-1)(n-2)}{3N} + \frac{n(n-1)(n+1)}{6N} \\ &= \frac{n(n-1)}{2} \\ &\Rightarrow \sum_{i\neq j=1}^{N} \pi_{ij} = 2\sum_{i<j=1}^{N} \pi_{ij} = n(n-1)\end{aligned}$$

□

Hossein Veisipour
Department of Statistics,
Razi University,
Kermanshah, Iran.
email: *h.veisipour@razi.ac.ir*

Mohammad Moradi
Department of Statistics,
Razi University,
Kermanshah, Iran.
email: *moradi_ m@razi.ac.ir*

Jennifer Brown
School of Mathematics and Statistics,
University of Canterbury,
Christchurch, New Zealand.
email: *jennifer.brown@canterbury.ac.ir*