A modification of Balanced Acceptance Sampling

B. L. Robertson[a,*], T. McDonald[b], C. J. Price[a], J. A. Brown[a]

[a] *School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand*
[b] *Western EcoSystems Technology, Inc, Cheyenne, Wyoming, USA*

Abstract

This article presents a modification of balanced acceptance sampling (BAS) that causes inclusion probabilities to better approximate targeted inclusion probabilities. A new sample frame constructor for BAS is also introduced from which equi-probable spatially balanced samples are drawn.

Keywords: Environmental sampling; Halton sequence; spatial balance.

1. Introduction

A spatially balanced sampling design selects sample locations that are well spread over the study area, a sample with few 'clumps' or 'voids'. Natural resources are often spatially autocorrelated because nearby locations interact with one another and are influenced by the same factors (Stevens & Olsen, 2004). Hence, spreading the sample over of the study area is known to be efficient, and many variations of spatially balanced designs have been proposed (Stevens & Olsen, 2004; Grafström et al., 2012; Robertson et al., 2013). This article considers balanced acceptance sampling (BAS) (Robertson et al., 2013).

BAS uses a quasi-random number sequence to select spatially balanced samples from either continuous or point resources in multidimensional space. In particular, BAS uses a random-start Halton sequence (Wang & Hickernell, 2000), which maps the natural numbers to vectors $\{\mathbf{x}_k\}_{k=1}^{\infty}$ in $[0,1)^d$ (Robertson et al.,

*Corresponding author
Email address: blair.robertson@canterbury.ac.nz (B. L. Robertson)

2013). The ith coordinate of each vector in the sequence has an associated base, b_i, and all bases $\{b_1, b_2, \ldots, b_d\}$, are required to be pair-wise co-prime. In this article b_i is the ith prime number. The ith coordinate of the kth point in this sequence is (Price & Price, 2012)

$$x_k^{(i)} = \sum_{j=0}^{\infty} \left\{ \left\lfloor \frac{u_i + k}{b_i^j} \right\rfloor \mod b_i \right\} \frac{1}{b_i^{j+1}},$$

where u_i is a random non-negative integer and $\lfloor x \rfloor$ is the floor function — the largest integer that is less than or equal to x. The random-start Halton sequence is

$$\{\mathbf{x}_k\}_{k=1}^{\infty} = \left\{ x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(d)} \right\}_{k=1}^{\infty}. \tag{1}$$

Choosing the first d prime numbers as bases and setting $u_i = 0$ for all i gives the classical Halton sequence (Halton, 1960).

To observe an equi-probable BAS sample from a continuous resource such as a polygon or geographic region, which we label Ω, a random-start Halton sequence is defined over a minimal bounding box that encloses Ω. If $\mathbf{x}_1 \in \Omega$, the point is included in the sample, otherwise the candidate point is rejected. The next point in the sequence is then considered. If $\mathbf{x}_2 \in \Omega$, the point is included, otherwise it is rejected. This process of testing successive points for membership in Ω is repeated until the required number of points has been accepted. Unequal probability samples can be achieved by adding a dimension and using an acceptance/rejection sampling strategy (Robertson et al., 2013).

Sampling point resources using BAS, for example, a collection of coordinate locations on a map or a grid defined over a polygon, is similar. First, however, 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 $\mathbf{x}_1$ is within a unit's box, that unit is included in the sample. Otherwise, no unit is selected. The next point in the sequence, $\mathbf{x}_2$, is then considered and the method repeats. A BAS sample is realized when the required number of distinct units has been accepted.

BAS has several desirable properties including spatial balance in two or more dimensions, spatially balanced over-samples and admittance of standard design-based estimators (Robertson et al., 2013). However, when sampling point resources, targeted inclusion probabilities are not necessarily achieved because BAS uses an acceptance/rejection sampling technique. The acceptance/rejection technique changes the actual inclusion probabilities and they need to be estimated (or calculated if the population is sufficiently small) for unbiased estimation of population parameters (Robertson et al., 2013). The dangers of not achieving targeted inclusion probabilities include bias and increased variance of the Horvitz-Thompson estimator.

In this article we present a modification of BAS which causes inclusion probabilities to better approximate targeted inclusion probabilities and introduce a new sample frame constructor for point and continuous resources, called Halton frames, which allow exact equi-probable BAS samples to be drawn from these resources. We begin by discussing properties of the Halton sequence that are pertinent to our modification and Halton frames. In Section 4 we present the modification of BAS and show how exact equi-probable BAS samples and cluster samples are drawn from Halton frames. Design-based estimators are given in Section 4.3 and concluding remarks are given in Section 5.

2. Properties of the Halton Sequence

It can be shown that for any set of positive integers J_i, any $B = \prod_{i=1}^{d} b_i^{J_i}$ consecutive points from a Halton sequence (1) with co-prime bases b_i, will have exactly one point in each of the boxes determined by

$$\prod_{i=1}^{d} \left[m_i b_i^{-J_i}, (m_i + 1) b_i^{-J_i} \right), \tag{2}$$

where m_i is an integer satisfying $0 \leq m_i < b_i^{J_i}$, for all $i = 1, 2, \ldots, d$ (Price & Price, 2012; Halton, 1960). Indeed, this property ensures the Halton sequence is well-spread over the unit box. We call these boxes Halton boxes and they are illustrated in Figure 1 for $(J_1, J_2) = (1, 1), (2, 1)$ and $(2, 2)$. The size and shape

of these boxes can be altered by choosing different co-prime bases and different J_i values. Increasing J_i reduces the size of the boxes in the ith dimension and varying J_i for different dimensions changes the shape of the boxes.

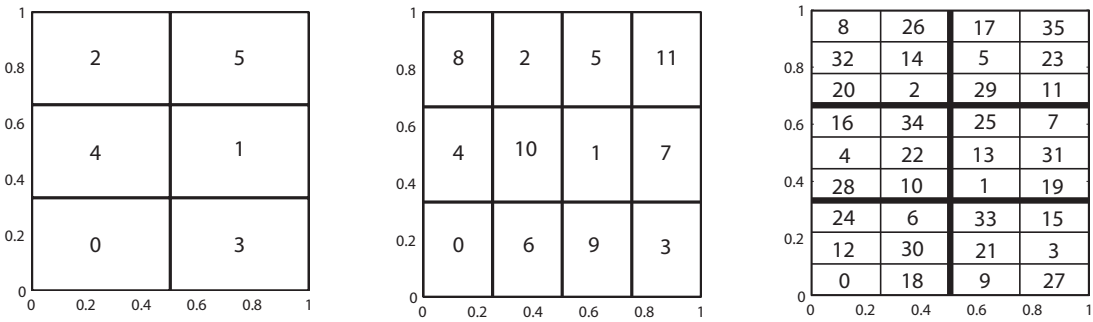

Figure 1: Halton boxes with $b_1 = 2$ and $b_2 = 3$ for different J_i values. Left: $J = (1, 1)$ and $B = 2\times3 = 6$; Center: $J = (2, 1)$ and $B = 2^2\times3 = 12$; Right: $J = (2, 2)$ and $B = 2^2\times3^2 = 36$. Any B consecutive points from a random-start Halton sequence will have exactly one point in each of the B boxes. Points from the classical Halton sequence ($u_1 = u_2 = 0$) with the same k values (mod B) will be in the Halton box labeled k mod B. For example, $\mathbf{x}_{71}$ will be in the box numbered 5 (left), 11 (center) and 35 (right).

Another property of the Halton sequence is that it is quasi-periodic. If $\mathbf{x}_k$ is in a specific Halton box, then k mod $b_i^{J_i}$ must take a specific value from the set $\{0, 1, \ldots, b_i^{J_i} - 1\}$ for each $i = 1, 2, \ldots, d$ (Price & Price, 2012; Halton, 1960). To determine the value of k (mod B) for points in each box, a system of congruences is solved using the Chinese Remainder Theorem (CRT). For example, if $b_1 = 2$, $b_2 = 3$, $J_i = 2$ (giving $B = 2^2 \times 3^2 = 36$) and k satisfies the following congruences

$$\begin{aligned} k &= 1 \pmod 4 \\ k &= 4 \pmod 9, \end{aligned}$$

then $k = 13$ (mod 36) by the CRT. Hence, any $\mathbf{x}_j$ with $j = 13$ (mod 36) will be in the same Halton box as $\mathbf{x}_{13}$. The sequence is quasi-periodic in the sense that for any $\mathbf{x}_k$, the points $\mathbf{x}_{k+B}, \mathbf{x}_{k+2B}, \ldots$ will be in the same Halton box as $\mathbf{x}_k$. The mod B values for the classical Halton sequence ($u_1 = u_2 = 0$)

with different J_i values are shown in Figure 1. For random-start Halton sequences, these mod B values are permuted. For example, if $u_1 = 1$ and $u_2 = 0$, the mod B values in Figure 1 (left) are permuted in the following way $\sigma(0, 1, 2, 3, 4, 5) = (3, 4, 5, 0, 1, 2)$.

The quasi-periodic property also means consecutive points from the Halton sequence cyclically visit the Halton boxes in a specific order determined by the mod B values. The boxes are also nested giving a hierarchical structure. For example, if two points have different mod 36 values but are congruent mod 6, they will be in the same box in Figure 1 (left) and in different boxes in Figure 1 (right). This explains why contiguous Halton subsequences of varying lengths are spatially balanced and why BAS over-samples are spatially balanced.

3. Halton Frames

A Halton frame is defined as the set of Halton boxes that intersect a point or continuous resource, where each box is uniquely numbered using the Halton sequence. That is, if $\mathbf{x}_k$ is in a particular box, that box is numbered k mod B. In Section 4, we show that implementing BAS on a Halton frame produces exact equi-probable samples.

The Halton frame for a continuous study area, for example a geographic map, is the N boxes that intersect the study area (a discretization of the study area). For linear (1-dimensional) resources like rivers, the Halton frame is the N boxes that intersect the linear resource. In both cases, B should be chosen large so that a fine-grained discretization of the resource is achieved. For example, with $b_1 = 2$, $b_2 = 3$ and choosing $J_1 = 8$ and $J_2 = 5$, the unit box is discretized into $B = 62,208$ approximately square Halton boxes with dimensions $1/256$ by $1/243$. Examples of coarse-grained Halton frames for the unit box are shown in Figure 1 and for continuous and linear resources in Figure 2.

The Halton frame for point resources, for example geographic coordinates, is the N boxes that contain at least one point from the resource. Rather than forcing each box to contain no more than one point, it can be more efficient to

allow some boxes to contain multiple points and implement a cluster sampling design (see Subsection 4.2). An example of a Halton frame with one point per box is given in Figure 2 and multiple points per box in Figure 4.

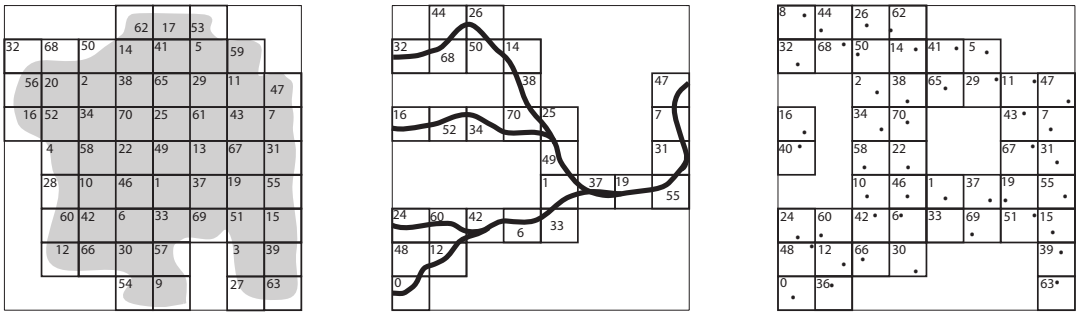

Figure 2: Halton frames for continuous (shaded region), linear (bold lines) and point resources using $b_1 = 2$, $b_2 = 3$ and $J = (3, 2)$. Points from the Halton sequence with the same k values (mod 72) will be in the Halton box labeled k mod 72.

If the resource is highly correlated, we suggest rotating the resource to better fill the bounding box before the Halton frame is constructed. For point resources, the eigenvectors of the estimated covariance matrix are aligned with the coordinate axes (see Web Figure 1). Rotating the resource decreases the number of Halton boxes excluded from the frame while preserving nearest neighbor relationships. Sampling the rotated resource tends to give better spatial balance because fewer boxes are skipped during BAS sampling (see Web Figure 2).

4. Modification to BAS

In this section we modify the original BAS design of Robertson et al. (2013) and show how the modified design achieves exact equi-probable samples from Halton frames. For simplicity, the modified design uses a random-start in the classical Halton sequence ($u_1 = u_2 = \ldots = u_d$).

Targeted inclusion probabilities are not necessarily achieved with BAS because a variable number of BAS points are rejected for not selecting a unit at the very beginning of the sequence. When an initial point is rejected, the sequence

'fast-forwards' to the first un-rejected point, thus giving it higher probability of inclusion than others. We correct this by requiring the first random-start Halton point $\mathbf{x}_1$ to select a unit. To achieve better approximations to targeted inclusion probabilities we suggest the following modification to BAS:

> *If the first point in the random-start Halton sequence does not select a sampling unit, discard the sequence and generate another.*

The advantage of this modification is best illustrated using a Halton frame, where exact probabilities can be calculated. Consider selecting an *equi-probable* BAS sample of $n = 2$ boxes from $N = 5$, rather than 6, boxes numbered 0,1,2,4 and 5 (exclude box 3) in the Halton frame given in Figure 1 (left). First, a non-negative integer u_i must be selected. For example, let u_i be randomly selected from $U = \{0, 1, \ldots, 6^7 - 1\}$. Using the properties of the Halton sequence from Section 2, any two consecutive points from the Halton sequence will select two unique boxes from our frame. If a point lands in the box numbered 3, BAS simply rejects the box and the next point in the sequence selects the box numbered 4. Hence, the possible samples are

$$\{0,1\}, \{1,2\}, \{2, \not{3}, 4\}, \{\not{3}, 4, 5\}, \{4,5\}, \{5,0\}. \tag{3}$$

Note that $\{4, 5\}$ is selected twice because sequences whose first point is in the box numbered 3 or in the box numbered 4 select this sample. Because $u_1 = u_2 \in U$ and $|U| = 6^7$, there are 6^7 possible sequences of the form $\{\mathbf{x}_1, \mathbf{x}_2\}$ that can be used to select the BAS sample. Of these sequences, 6^6 of the $\mathbf{x}_1$ points will be in the same Halton box, because U has the following mod B structure

$$U = \{0, 1, \ldots, 5, 0, 1, \ldots, 5, \ldots, 0, 1, \ldots, 5\} \pmod 6.$$

Thus, each of the BAS samples in (3) has exactly the same probability of being selected, giving the following inclusion probabilities

$$\pi_0 = \pi_1 = \pi_2 = 1/3 \quad \text{and} \quad \pi_4 = \pi_5 = 1/2,$$

which are not equal.

If the modified BAS design is applied to this problem, the repeated sample in (3) is removed, because sequences with $\mathbf{x}_1$ in the box numbered 3 are discarded. This leaves five unique samples that are equally probable. Hence, the targeted equal inclusion probabilities are achieved using the modified design, $\pi_i = 2/5$ for all i. We discuss the modified BAS approach with Halton frames further in subsections 4.1 and 4.2.

If the Halton frame is not used, targeted inclusion probabilities may not be achieved exactly using the modified BAS approach, but can be estimated using the approaches described in (Robertson et al., 2013). Consider drawing equi-probable BAS samples from a discretization (see Figure 3) of the northern part of New Zealand's South Island (this is not a Halton frame). The inclusion probability of the ith unit can be estimated by counting the fraction of samples containing the ith unit over many repeated samples. We computed the estimated inclusion probability, $\widehat{\pi}_i$, for each unit using 100,000 samples and reported the mean square error

$$\text{MSE} = \sum_{i=1}^{N} (\widehat{\pi}_i - n/N)^2$$

using BAS and modified BAS relative to the MSE under SRS using the sample function in R (R Development Core Team, 2017). Results are given in Figure 3 for various sample sizes, where values close to one indicate little difference between the estimated and targeted inclusion probabilities. The modified approach produced better estimates of the targeted values for all sample sizes considered. By removing random-start Halton sequences whose first points fail to select a sampling unit from the BAS design, better approximations to targeted inclusion probabilities were obtained. The authors' recommend using this modification in all BAS applications.

4.1. Modified BAS with Halton Frames

When the modified BAS design is used with a Halton frame, exact equi-probable spatially balanced samples are observed. For continuous and linear resources, the Halton boxes should be small (large B) and approximately square

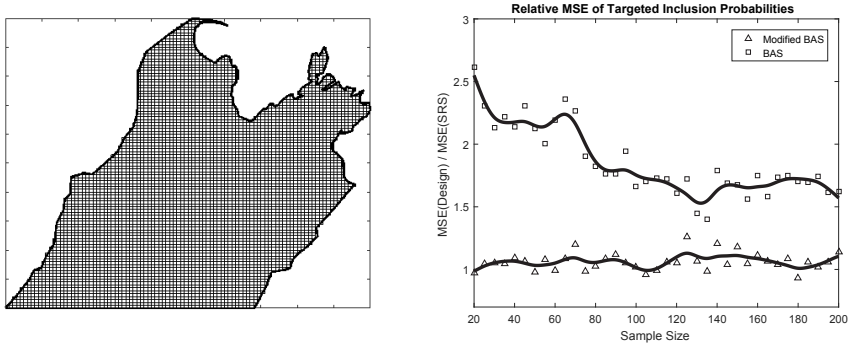

Figure 3: Left: a discretization ($N = 5434$ boxes) of the northern part of New Zealand's South Island (this is not a Halton frame). Right: mean square error of targeted inclusion probabilities using BAS and modified BAS relative to the MSE under SRS as a function of sample size. The curves are fitted using a smoothing spline.

to yield a sufficient discretization. For point resources, we define a Halton frame by sequentially increasing B until there is no more than one point per box (multiple points per box is considered in the next subsection).

Let $\{a_0, a_1, \ldots, a_{N-1}\}$ be the set of numbered Halton boxes in increasing order. A modified BAS sample is realized observing the units in the boxes numbered

$$\{a_i : i = j, j+1 \bmod N, \ldots, j+n-1 \bmod N\},$$

where j is randomly chosen from $\{0, 1, \ldots, N-1\}$. Note that the indices 'wrap' to the beginning of the frame when they exceed B. Observing units in these boxes is analogous to applying the modified BAS design with the Halton frame. The possible samples are

$$\begin{array}{lllll}
\{a_0, & a_1, & \ldots, & a_{n-2}, & a_{n-1}\} \\
\{a_1, & a_2, & \ldots, & a_{n-1}, & a_n\} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\{a_{N-n}, & a_{N-n+1}, & \ldots, & a_{N-2}, & a_{N-1}\} \\
\{a_{N-n+1}, & a_{N-n+2}, & \ldots, & a_{N-1}, & a_0\} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\{a_{N-1}, & a_0, & \ldots, & a_{n-3}, & a_{n-2}\},
\end{array}$$

where each a_j value appears in exactly n of the N samples. Each sample is equally likely because j is randomly chosen from $\{0, 1, \ldots, N-1\}$, giving $\pi_i = n/N$ — an equal-probable sample. If a point sample is required from a continuous resource, the center points (within the continuous resource) of sufficiently small Halton boxes can be used.

4.2. Cluster Sampling with Halton Frames

In cluster sampling (single-stage) the population is partitioned into clusters, with each cluster containing secondary units. Cluster sampling designs select a sample of clusters and measure the response at all the secondary units within the sampled clusters. These designs are usually used for reasons of convenience or practicality (Thompson, 1992). We present a BAS cluster sampling design for point resources for three reasons. First, a Halton frame with multiple points per box provides a way to define clusters of nearby points. Second, when sampling natural resources, it can be more cost effective to sample clusters of nearby points rather than individual points (Lohr, 2010). When sampling point resources, BAS replaces the points with a collection of non-overlapping equally sized boxes, where each box contains exactly one point. This frame can be difficult to construct if some points are close together and can make the acceptance/rejection sampling strategy that BAS uses computationally prohibitive. A Halton frame can also be difficult to construct in this setting. One can reduce effort by choosing sufficiently small Halton boxes so that each box contains no more than a few points and using cluster samples.

Consider a point resource consisting of M points. First, a minimal bounding box that contains all the points is defined and partitioned into B approximately square Halton boxes with each box containing no more than a few points. The $N \leq M$ boxes containing at least one point are clusters and the points themselves are secondary units. A modified BAS sample of $n < N$ clusters is then selected, giving an equi-probable spatially balanced sample of clusters. This approach is illustrated in Figure 4.

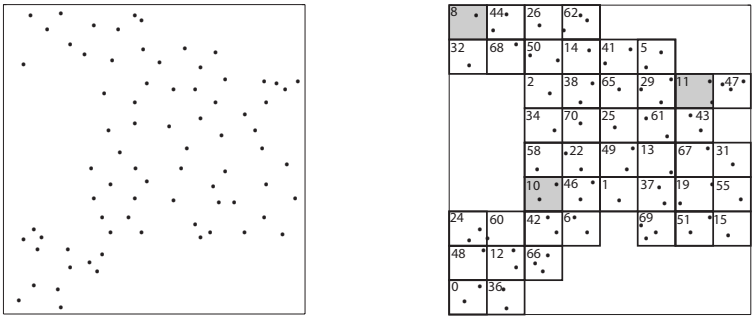

Figure 4: Left: a population of $M = 76$ points. Right: $N = 45$ clusters using a Halton frame with $J = (3, 2)$ and $B = 2^3 \times 3^2 = 72$. To observe a cluster sample, the modified BAS approach is used to select $n < N$ clusters (boxes) and every secondary unit within the sampled clusters is observed. For example, if $n = 3$ and $a_j = 8$, the boxes numbered 8, 10 and 11 are selected and 5 secondary units are observed.

4.3. Estimation

The estimation and variance estimation techniques in (Robertson et al., 2013) can be used for the modified BAS approach and are given here for completeness. The cluster sampling approach described in the previous subsection is single-stage, meaning every secondary unit is observed in the sampled clusters. Hence, these estimation techniques are also applicable to the cluster sampling approach by simply replacing 'unit' with 'cluster' in the following description.

Let the sampling units be numbered $1, 2, \ldots, N$ and let y_i denote the response value at the ith unit. The Horvitz-Thompson estimator of the population total τ is

$$\widehat{\tau} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i},$$

where $\mathcal{S} \subset \{1, 2, \ldots, N\}$ is the sample and π_i is the inclusion probability of the ith unit. The variance of $\hat{\tau}$ can be estimated using the Sen-Yates-Grundy estimator (Yates & Grundy, 1953), but this estimator is biased and tends to be unstable for spatially-balanced designs (Grafström et al., 2012; Robertson et al., 2013). The local mean variance estimator (Stevens & Olsen, 2003) is commonly used for spatially balanced designs and is recommended for the modified BAS

approach. The estimator is

$$\widehat{\mathrm{V}}_{\mathrm{NBH}}(\widehat{\tau}) = \sum_{i \in \mathcal{S}} \sum_{j \in D_i} w_{ij} \left(\frac{y_j}{\pi_j} - \widehat{\tau}_{D_i} \right)^2,$$

where D_i is a neighborhood containing at least four nearest neighbors to the ith unit, $\widehat{\tau}_{D_i}$ is an estimate of the population total on D_i, and w_{ij} are weights. The weights decrease as the distance between the ith unit and jth unit increases and $\sum_i w_{ij} = \sum_j w_{ij} = 1$. Details on how to compute neighborhoods and weights can be found in (Stevens & Olsen, 2003).

5. Conclusion

This article has presented a modification for BAS and introduced Halton frames for sampling continuous and point resources. By removing random-start Halton sequences whose first points fail to select a sampling unit from the BAS design, better approximations to targeted inclusion probabilities are obtained. If Halton frames are pertinent to the sampling problem, exact equi-probable spatially balanced samples are drawn using the modified BAS approach. These frames can also be used to cluster point resources into groups of nearby points, providing a simpler frame construction for BAS. Spatially balanced equi-probable cluster samples from point resources that are grouped in this way can be observed using the modified BAS approach.

Acknowledgements

We thank an anonymous referee and the editor for valuable comments that led to an improved article.

References

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

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

Lohr, S. L. (2010). Sampling: Design and Analysis. Brooks/Cole, Boston, USA.

Price, C. J., & Price, C. P. (2012). Recycling primes in Halton sequences: an optimization perspective. *Advanced Modeling and Optimization*, *14*, 17–29.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `http://www.R-project.org` ISBN 3-900051-07-0.

Robertson, B. L., Brown, J. A., McDonald, T., & Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, *3*, 776–784.

Stevens, D. L. J., & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, *14*, 593–610.

Stevens, D. L. J., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*, 262–278.

Thompson, S. K. (1992). Sampling. New York: Wiley.

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

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