Environ Ecol Stat (2018) 25:305–323
https://doi.org/10.1007/s10651-018-0406-6

Halton iterative partitioning: spatially balanced sampling via partitioning

Blair Robertson[1] · **Trent McDonald**[2] · **Chris Price**[1] · **Jennifer Brown**[1]

Received: 15 November 2017 / Revised: 14 March 2018 / Published online: 4 May 2018

Abstract A new spatially balanced sampling design for environmental surveys is introduced, called Halton iterative partitioning (HIP). The design draws sample locations that are well spread over the study area. Spatially balanced designs are known to be efficient when surveying natural resources because nearby locations tend to be similar. The HIP design uses structural properties of the Halton sequence to partition a resource into nested boxes. Sample locations are then drawn from specific boxes in the partition to ensure spatial diversity. The method is conceptually simple and computationally efficient, draws spatially balanced samples in two or more dimensions and uses standard design-based estimators. Furthermore, HIP samples have an implicit ordering that can be used to define spatially balanced over-samples. This feature is particularly useful when sampling natural resources because we can dynamically add spatially balanced units from the over-sample to the sample as non-target or inaccessible units are discovered. We use several populations to show that HIP sampling draws spatially balanced samples and gives precise estimates of population totals.

Keywords BAS · Environmental sampling · Halton sequence · Over-sampling · Spatial balance

Handling Editor: Bryan F. J. Manly.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10651-018-0406-6) contains supplementary material, which is available to authorized users.

✉ Blair Robertson
blair.robertson@canterbury.ac.nz

[1] School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

[2] Western EcoSystems Technology, Inc., Cheyenne, WY 82001, USA

1 Introduction

A spatial sampling design determines where sample locations are placed in a study area. The main objective of a spatial design is to draw sample locations in such a way that valid scientific inferences can be made to all regions of a study area (McDonald 2014). A common feature in natural resource sampling is that nearby locations tend to be similar because they interact with one another and are influenced by the same set of factors (Stevens and Olsen 2004). This means sample efficiency can be increased by spreading sample locations over the study area. Stevens and Olsen (2004) called well spread samples "spatially balanced samples" and measured spatial balance using the Voronoi tessellation of a sample. A sample is considered spatially balanced if the total inclusion probability in each Voronoi polygon is approximately equal to one. Grafström and Lundström (2013) provide the theoretical motivation for spatially balanced samples and show that they are effective in a variety of situations.

Spatially balanced sampling designs are commonly used for sampling natural resources (Grafström and Lundström 2013) and a variety of designs have been proposed. Stevens and Olsen (2004) introduced a spatially balanced design that is frequently used in environmental monitoring, called Generalized Random Tessellation Stratified (GRTS) design. GRTS recursively divides the study area into nested square boxes of equal size until the total inclusion probability in each box is less than or equal to one. All the boxes that contain at least one sampling unit are then given a one-dimensional address using a base four hierarchical numbering scheme. These addresses are then placed in order on the real line and a systematic sample is drawn using Brewer and Hanif's (1983) design. The sampled addresses are then mapped back to their respective two-dimensional locations, to yield the sample locations. GRTS can be applied to point, linear and continuous resources and can draw unequal probability samples. Another useful property of GRTS is its reverse hierarchical ordering, which orders an observed sample so that contiguous sub-samples are also spatially balanced. This makes the design particularly useful in environmental monitoring because we can dynamically add units from a spatially balanced over-sample to the sample as non-target or inaccessible units are discovered (Stevens and Olsen 2004; Larsen et al. 2008). The true inclusion probabilities of this inverse sampling strategy may not be known, but inference can be based on inclusion probabilities conditional on the achieved sample size (Stevens and Olsen 2004). This feature does not eliminate the non-response or the bias of an inference, but it is popular with field researchers because the largest sample that their budget permits can be analysed.

The Local Pivotal Method (LPM), introduced by Grafström et al. (2012), is an application of the Pivotal Method (Deville and Tillé 1998) that gives spatially balanced samples. LPM obtains a sample of size n from N points by iteratively updating each point's inclusion probability until n points have inclusion probabilities equal to one. At each iteration, a point competes with its neighbour for inclusion in the sample. The winning point has its inclusion probability increased and the losing point has its decreased. These competitions make it very unlikely for a sample to simultaneously include neighbouring points and thus forces the sample to be spatially balanced. LPM can be applied to point resources in multiple dimensions and can draw unequal probability samples. LPM can also draw samples from continuous resources (Grafström

et al. 2017a, b), but the authors are not aware of an LPM over-sampling approach. A drawback of the original implementation of the second LPM version, called LPM2, was that its computational complexity was $O(N^2)$, making it computationally prohibitive on large point resources. However, the linear searches in LPM2 can be replaced by k-d trees to reduce the average complexity of LPM2 to $O(N \log N)$ (see lpm2_kdtree in Grafström and Lisic (2016)). Alternatively, a rapid implementation of the LPM, called suboptimal LPM, can be used (Grafström et al. 2014). This method reduces the search space for finding a point's nearest neighbour. By considering h possible neighbours, rather than all possible points, the complexity of the algorithm is reduced to $O(hN)$ making it computationally feasible on large point resources.

Another spatially balanced design is called Balanced Acceptance Sampling (BAS) (Robertson et al. 2013) and its modified version (Robertson et al. 2017). BAS uses a quasi-random number sequence, called the random-start Halton sequence (Halton 1960; Wang and Hickernell 2000), to draw spatially balanced samples from point and continuous resources. BAS can draw equal and unequal probability samples in multiple dimensions, is conceptually simple and performs well on continuous resources (Robertson et al. 2013). However, BAS has two drawbacks when sampling point resources. First, acceptance/rejection sampling is required to draw its sample and hence, targeted inclusion probabilities are not necessarily achieved. Robertson et al. (2017) provided a simple modification to BAS that reduced the differences between targeted and actual inclusion probabilities. However, the danger of not achieving targeted inclusion probabilities is inflated variance of estimators.

The second drawback of BAS is its sampling frame for point resources. BAS constructs its frame by replacing the point resource in $[0, 1)^d$ with N non-overlapping boxes of equal size, with one point in each box (see the left panel of Fig. 1). The BAS sample is then drawn as follows. First, a random-start Halton sequence is defined

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\} \subset [0, 1)^d, \tag{1}$$

where k is chosen so that n boxes contain at least one random-start Halton point. The points from the resource within these n boxes define the BAS sample (see the right panel of Fig. 1). However, if the point resource is large or lacks grid structure, BAS becomes inefficient or computationally prohibitive because the boxes tend to be small and the number of points in (1) is enormous. This drawback is illustrated for small N in the right panel of Fig. 1 and is discussed further in Sect. 4.

In this article we introduce a new spatially balanced sampling design, called Halton Iterative Partitioning (HIP). Our design is an alternative to BAS that shares its desirable properties, without the previously mentioned drawbacks. We achieve this by using structural properties of the Halton sequence, rather than points from the sequence, to draw our sample. HIP iteratively partitions a resource into nested boxes using a quasi-periodic property of the Halton sequence. Points are then drawn from particular boxes in a specific order to achieve a spatially balanced sample. The partition can be defined in two or more dimensions if spatial balance over more than geographic locations is sought. Our design is conceptually simple, computationally efficient on large N point resources with computational complexity $O(N \log N)$, has a rapid implementation for equal probability point resources and is embarrassingly parallel. It can be applied to

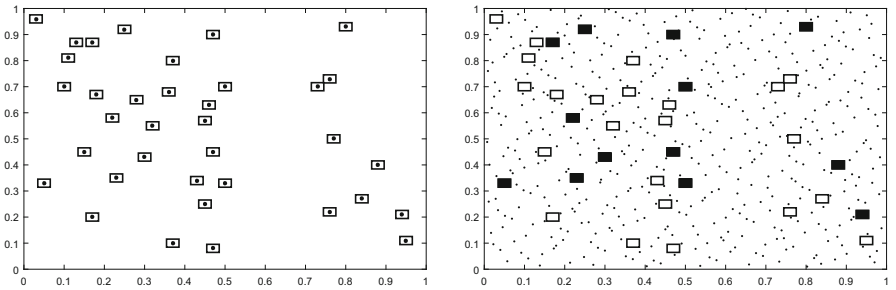

Fig. 1 Left: An $N = 36$ point resource without grid structure and the corresponding frame of boxes constructed by BAS. Right: A BAS sample of $n = 13$ (points from the shaded boxes) and the required $k = 500$ random-start Halton points needed to draw the sample

continuous and point resources and achieves targeted inclusion probabilities/density. Linear resources, for example rivers, can be sampled if they are discretized (see for example Robertson et al. 2017). A HIP sample has a specific ordering that ensures contiguous sub-samples are spatially balanced. This feature makes HIP particularly useful for spatially balanced over-sampling if non-target or inaccessible units are discovered.

The rest of the article is organized as follows. In Sect. 2 the Halton sequence is defined and its quasi-periodicity is explained. The HIP design is introduced in Sect. 3 and its spatial balance is illustrated in Sect. 4. Section 5 considers design-based estimation and variance estimation techniques, which are applied to three populations. Concluding remarks are given in Sect. 6.

2 Halton sequence

The Halton sequence, defined for vectors $\{\mathbf{x}_k\}_{k=0}^{\infty}$ in $[0, 1)^d$ (Halton 1960), is a quasi-random number sequence that distributes points evenly in low dimensions (e.g. $d \leq 10$). The ith coordinate of each point in the sequence has an associated base, b_i, and all bases are pairwise co-prime. The co-prime condition means that the only positive integer that evenly divides b_i and b_j is one, for all $i, j = 1, 2, \ldots, d$ $(i \neq j)$. In this article, $b_1 = 2$, $b_2 = 3$, and b_d is the dth prime number. The ith coordinate of the kth point in the sequence is Price and Price (2012), Robertson et al. (2017)

$$x_k^{(i)} = \phi_{b_i}(k) = \sum_{j=0}^{\infty} \left\{ \left\lfloor \frac{k}{b_i^j} \right\rfloor \bmod b_i \right\} \frac{1}{b_i^{j+1}}, \tag{2}$$

where $\lfloor x \rfloor$ is the floor function — the largest integer that is less than or equal to x. The 4th point in the three-dimensional Halton sequence, for example, is

$$\mathbf{x}_4 = (\phi_2(4), \phi_3(4), \phi_5(4)) = (1/8, 4/9, 4/5).$$

Each point's subscript denotes the integer that is mapped using (2) to a point in $[0, 1)^d$. We call these subscripts Halton indices.

2.1 Properties of the Halton sequence

We now describe properties of the Halton sequence that are pertinent to the HIP sampling design. Let $B = \prod_{i=1}^{d} b_i^{J_i}$, where J_i is any non-negative integer. It can be shown that B consecutive points from a Halton sequence with bases b_i, will have exactly one point in each of the boxes defined by

$$\prod_{i=1}^{d} \left[m_i b_i^{-J_i}, (m_i + 1) b_i^{-J_i} \right), \tag{3}$$

where m_i is an integer satisfying $0 \leq m_i < b_i^{J_i}$, for all $i = 1, 2, \ldots, d$ (Halton 1960; Price and Price 2012). We call these boxes Halton boxes and examples are given in the left panels of Fig. 2. If $\mathbf{x}_k$ is in a particular Halton box, then k must take a specific mod B value from the set $\{0, 1, \ldots, B-1\}$ (Halton 1960; Price and Price 2012). Therefore, the points $\mathbf{x}_k, \mathbf{x}_{k+B}, \mathbf{x}_{k+2B}, \ldots$ must be in the same Halton box and hence, each box is associated with a particular Halton index (mod B). In this sense, the Halton sequence is quasi-periodic with period B. The Halton index k for a particular box with lower bounds $\ell_1, \ldots, \ell_d$ is obtained by solving the system of d congruences

$$k = a_i \pmod{b_i^{J_i}},$$

where

$$a_i = \sum_{j=1}^{J_i} \left\{ \lfloor \ell_i b_i^j \rfloor \mod b_i \right\} b_i^{j-1}$$

and $i = 1, 2, \ldots, d$. For example, the Halton index for $[1/4, 1/2) \times [1/3, 4/9)$ with $B = 36$ requires solving

$$\begin{aligned} k &= 2 \pmod 4 \\ k &= 1 \pmod 9, \end{aligned}$$

which gives $k = 10 \pmod{36}$.

The form of (3) means that small Halton boxes are nested in larger parent boxes. The indices of the nested boxes must be congruent mod B, where B is the number of parent boxes, because each box has a specific mod B index. This nested structure is illustrated in the left panels of Fig. 2 and Web Table 2, where $B = 36$ Halton boxes are shown. The pairs of Halton boxes with indices congruent mod 12 are nested in the mod 12 Halton boxes with $(J_1, J_2) = (2, 1)$ and those with indices congruent mod 6 are nested in the six Halton boxes with $(J_1, J_2) = (1, 1)$.

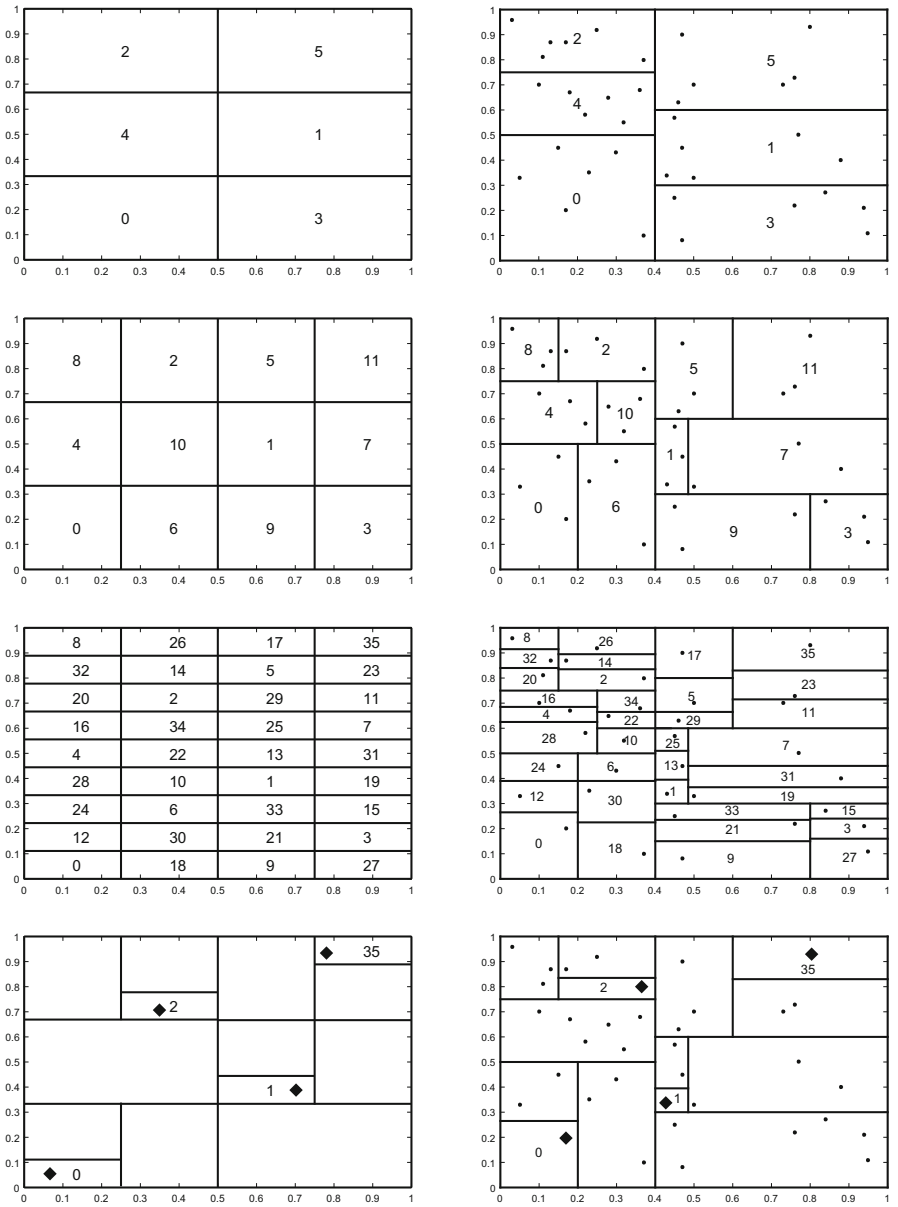

Fig. 2 Left panels: Halton boxes for $B = 2 \times 3 = 6$, $B = 2^2 \times 3 = 12$ and $B = 2^2 \times 3^2 = 36$. A point $\mathbf{x}_k$ from the Halton sequence will be in the box numbered k mod B. For example, $\mathbf{x}_{59}$ will be in the boxes numbered 5 (mod 6), 11 (mod 12) and 23 (mod 36). Right panels: Nested partitions of an equal probability point resource containing $N = 36$ points. Halton indices from $B = 6$, 12 and 36 Halton boxes are mapped to their corresponding boxes in the nested partition. The bottom panels show equal probability HIP samples from the two resources with $B = 36$, $n = 4$ and $k = 35$, which gives $\mathcal{S}_{35} = \{35, 0, 1, 2\}$. The necessary splits to draw the HIP sample are shown and the diamonds indicate the sampled units. In the continuous case (left), each point was drawn from uniform distribution over its box

3 Halton iterative partitioning sampling

The Halton Iterative Partitioning (HIP) design draws its sample of n points by partitioning the resource into $B = \prod_{i=1}^{d} b_i^{J_i} \geq n$ boxes. The partition is constructed iteratively so that the B boxes have the same nested structure as Halton boxes. However, the size of each box is chosen so that the inclusion density/probability of each box is constant. This means boxes will tend to be larger in regions where the inclusion density/probability is low and smaller where the inclusion density/probability is high. Each box in the partition is uniquely numbered using the Halton index of its corresponding Halton box (mod B) (see Fig. 2) and the HIP sample is obtained by drawing n points from consecutively numbered (mod B) boxes in the partition. Nested partitions of this form can be constructed for point and continuous resources with equal or unequal inclusion density/probability. We discuss each case in the subsections that follow. For simplicity, HIP designs are explained for two-dimensional resources.

3.1 Continuous resources

Consider drawing a HIP sample of n points from a continuous resource $\Omega \subset [0, 1)^2$ with an inclusion density function of the form

$$\pi(\mathbf{x}) = nf(x_1, x_2),$$

where Ω has positive Lebesgue measure and $f(x_1, x_2) : \Omega \to \mathbb{R}_{\geq 0}$ is a bounded probability density function. The HIP design iteratively partitions the continuous resource into $B = 2^{J_1} \times 3^{J_2}$ nested boxes, where J_1 and J_2 are chosen so that $2^{J_1} \approx 3^{J_2}$ to ensure each dimension has approximately the same number of splits. Furthermore, the number of spatially different samples is $O(B)$, so $B \gg n$ is necessary. We recommend a fine-grained partition with $B = 2^8 \times 3^5 = 256 \times 243 = 62{,}208$. The partition itself is implicit and only n of the B boxes need to be computed, requiring no more than $n(J_1+2J_2)$ data splits. Hence, choosing a large B value does not substantially increase the computational effort required to draw a HIP sample. The partitioning strategy is illustrated for $B = 36$ boxes in the left panels of Fig. 2, where $\Omega = [0, 1)^2$ and $\pi(\mathbf{x})$ is the uniform inclusion density function.

Before the partition is formed, the indices of the boxes to be sampled are selected

$$\mathcal{S}_k = \{k, (k + 1) \bmod B, \ldots, (k + n - 1) \bmod B\}, \tag{4}$$

where k is a random integer from the set $\{0, 1, \ldots, B - 1\}$. The partition is then iteratively formed to find the boxes with indices in $\mathcal{S}_k$. Initially, $[0, 1)^2$ is split into two boxes along x_1 at $0 < q < 1$

$$H_0 = [0, q) \times [0, 1) \quad \text{and} \quad H_1 = [q, 1) \times [0, 1), \tag{5}$$

such that the inclusion density in each box is equal. These boxes are then partitioned using x_2 to give six boxes of the form

$$\begin{aligned} H_0 &= [0,q) \times [0,r) & H_3 &= [q,1) \times [0,r') \\ H_4 &= [0,q) \times [r,s) & H_1 &= [q,1) \times [r',s') \\ H_2 &= [0,q) \times [s,1) & H_5 &= [q,1) \times [s',1), \end{aligned} \tag{6}$$

where H_k denotes the box with Halton index k. Each box's index is computed using its corresponding Halton box by setting $q = 1/2, r = r' = 1/3$ and $s = s' = 2/3$ (see Sect. 2.1). The values $r < s$ and $r' < s'$ are chosen so that the inclusion density in each box is constant

$$\int_{H_k} \pi(\mathbf{x}) d\mathbf{x} = \frac{n}{6},$$

for $k = 0, 1, \ldots, 5$.

The method then repeats on each box in (6). Each box is split into two boxes using x_1, followed by a partition into three boxes using two splits along x_2, where each split is chosen so that the inclusion density is evenly divided among the boxes. These 36 boxes have unique mod 36 Halton indices and the boxes nested in a particular H_k from (6) have indices congruent to k mod 6 (see the left panels of Fig. 2). This iterative partitioning and numbering continues until the number of boxes in the partition is greater than n. For the remaining iterations, it is computationally wasteful to split every box, and so we only split the boxes that contain sub-boxes with indices in $\mathcal{S}_k$. This is illustrated in the bottom right panel of Fig. 2. Here 13 splits were made to define the $n = 4$ boxes with indices in $\mathcal{S}_{35}$, rather than using the 35 splits needed to define all $B = 36$ boxes.

The HIP sample is obtained by sampling the boxes numbered by $\mathcal{S}_k$. For each $k \in \mathcal{S}_k$, one point is drawn from box H_k using acceptance/rejection sampling with the density function

$$f_k(\mathbf{x}) = \begin{cases} \frac{B}{n}\pi(\mathbf{x}) & \text{if } \mathbf{x} \in H_k \\ 0 & \text{otherwise.} \end{cases}$$

If $\Omega = [0,1)^2$ and $\pi(\mathbf{x})$ is the uniform inclusion density, a HIP sample will be similar to n consecutive points from a Halton sequence, because both samples draw points from the same boxes and B is large. Otherwise, HIP draws more points from regions where the inclusion density is high because the boxes are larger in low density regions.

We now show that a HIP sample achieves the targeted inclusion density. Let $\omega \subset \Omega$ be a Lebesgue measurable set and let n_ω denote the number of points in ω from a HIP sample of size n. The probability of selecting a particular $\mathcal{S}_k$ is $1/B$ and each Halton index is included in n of the $\mathcal{S}_k$'s. Hence, the expected sample size in ω is

$$\begin{aligned} \mathrm{E}(n_\omega) &= \frac{1}{B} \sum_{k=0}^{B-1} \sum_{j \in \mathcal{S}_k} \int_{H_j \cap \omega} f_j(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{n} \sum_{k=0}^{B-1} \sum_{j \in \mathcal{S}_k} \int_{H_j \cap \omega} \pi(\mathbf{x}) d\mathbf{x} \end{aligned}$$

$$= \frac{1}{n} \sum_{k=0}^{B-1} n \int_{H_k \cap \omega} \pi(\mathbf{x}) d\mathbf{x}$$
$$= \int_{\omega} \pi(\mathbf{x}) d\mathbf{x},$$

as required.

To simplify sampling, f_k can be replaced with a uniform density over $\Omega \cap H_k$. In this case, the expected sample size in ω is

$$\mathrm{E}(n_\omega) = \frac{n}{B} \sum_{k=0}^{B-1} \frac{m(\omega \cap H_k)}{m(\Omega \cap H_k)} \approx \int_{\omega} \pi(\mathbf{x}) d\mathbf{x},$$

for large B, where $m(.)$ denotes the Lebesgue measure.

3.2 Point resources

Consider drawing a HIP sample of size n from N points in $[0, 1)^2$, where the ith point has an inclusion probability $0 < \pi_i < 1$ such that $\sum_{i=1}^{N} \pi_i = n$. Before the partition is formed, a small amount of random noise is added to the point resource so that the probability of two points sharing the same coordinate value is zero. The partition is similar to the continuous resource partition in Sect. 3.1. Rather than keeping the inclusion density in each box constant, this partition keeps the inclusion probability in each box as similar as possible. The iterative partition is illustrated in the right panels of Fig. 2 for an $N = 36$ equal probability point resource.

Initially the points are partitioned into two boxes H_0 and H_1, of the form (5), where q minimizes

$$\left(\sum_{i:\mathbf{x}_i \in H_0} \pi_i - \sum_{j:\mathbf{x}_j \in H_1} \pi_j \right)^2.$$

These boxes are then partitioned using splits along x_2 to give six boxes of the form (6) such that

$$\sum_{k \in \{0,\ldots,5\}} \left(\sum_{i:\mathbf{x}_i \in H_k} \pi_i - n/6 \right)^2$$

is minimized. The method then repeats on each box to define a nested sequence of B boxes with unique Halton indices mod B, such that the total inclusion probability of each box is as similar as possible and less than or equal to one. The optimal number of boxes $B \geq n$, is found by considering candidate values from Web Table 1 and choosing the value with the best average spatial balance for the resource.

Each box H_k is then assigned an inclusion interval

$$I_k = \left[\sum_{i=0}^{k-1} d_i, \sum_{i=0}^{k} d_i \right) \quad (k > 0),$$

where $I_0 = [0, d_0)$ and

$$d_k = \sum_{i:\mathbf{x}_i \in H_k} \pi_i,$$

for $k = 0, 1, \ldots, B-1$. The boxes whose intervals contain a point from

$$\{(s + \lambda\alpha) - n\lfloor (s + \lambda\alpha)/n \rfloor : \lambda = 0, 1, \ldots, n-1\}, \tag{7}$$

have one point drawn from them using selection probabilities δ_i, where s is randomly chosen from $\in [0, n)$, $\alpha = \max\{d_k\}$ and

$$\delta_i = \pi_i / d_k : \mathbf{x}_i \in H_k.$$

Because $\alpha \geq d_k$ for all k and $B \geq n$, n different boxes have one point drawn from them. The inclusion probability of $\mathbf{x}_i \in H_k$ is $d_k \delta_i = \pi_i$, as required.

3.3 Rapid HIP design for equal probability point resources

If an equal probability point sample is required, it is not necessary to construct all $B = 2^{J_1} \times 3^{J_2}$ boxes as in the previous section. We call this approach rapid HIP and recommend its use for equal probability point resources. The partitioning method is similar to the approach described in Sect. 3.1, but instead of keeping the inclusion density of each box constant, the number of points in each box is held constant. To achieve this, some points may be removed from the point resource during the construction. We choose the number of boxes in the partition as the largest B value in Web Table 1 that satisfies $n \leq B \leq N$. However, several candidate B values could be tested for a particular resource, where the value that yields the best average spatial balance is chosen.

Before the partition is formed, the set of indices $\mathcal{S}_k$ of the boxes to be sampled are selected. The partition is then iteratively constructed to find the boxes with indices in $\mathcal{S}_k$. If N is odd, randomly remove one point from the resource. Partition the remaining N_1 points into two boxes H_0 and H_1, of the form (5), where q is chosen to give $N_1/2$ points in each box. Randomly remove $N_1/2 \bmod 3$ points from H_0 and from H_1, leaving a total of N_2 points in the resource so that $N_2/2 = 0 \pmod 3$. Partition H_0 and H_1 to give six boxes of the form (6), where r, s, r' and s' are chosen to give $N_2/6$ points in each box.

The method then repeats on each box in (6). If $N_2/6$ is odd, randomly remove one point from each box leaving a total of N_3 points in the resource so that $N_3/6 = 0 \pmod 2$. Split each box using x_1 to give $N_3/12$ points in each box. For each of these

12 boxes, randomly remove $N_3/12$ mod 3 points leaving a total of N_4 points in the resource so that $N_4/12 = 0$ (mod 3). Partition each box using two splits along x_2 to give 36 boxes with $N_4/36$ points in each. This iterative partitioning and numbering continues until the number of boxes is greater than n. For the remaining iterations, only boxes that contain sub-boxes with indices in $\mathcal{S}_k$ are split. This is illustrated in the bottom right panel of Fig. 2. Here 13 splits were made to construct the $n = 4$ boxes with indices in $\mathcal{S}_{35}$, rather than using the 35 splits needed to construct all $B = 36$ boxes.

The HIP sample is obtained by randomly drawing one point from each box with a Halton index in $\mathcal{S}_k$. Let λ denote the number of points in each box so that $N_j = \lambda B \leq N$ is the number of the points in the final partition. The inclusion probability of the ith point is

$$\begin{aligned}\pi_i &= \left(\frac{N_1}{N}\right)\left(\frac{N_2}{N_1}\right)\cdots\left(\frac{N_j}{N_{j-1}}\right)\left(\frac{n}{B}\right)\left(\frac{1}{\lambda}\right)\\ &= \frac{n}{N}\end{aligned}$$

as required, where $N_i \geq N_{i+1}$ is the number of points remaining after the ith iteration of partitioning.

3.4 Permutation of Halton indices

The HIP design draws points from particular boxes in its nested partition. Equation (4) determines which boxes can be sampled and there are B different choices. Hence, regardless of the number of points in each box, the number of spatially different HIP samples can be relatively low, $O(B)$. To increase the robustness of HIP sampling, the number of spatially different samples is substantially increased to $O(B^2)$ by permuting the Halton indices before the sample is drawn. The permutation method is given in Web Section 1 and does not affect the spatial balance or theoretical properties of HIP sampling. Numerical results in the following sections were generated using the rapid HIP design with permuted Halton indices.

4 Spatial balance

We measured the spatial balance of a point sample of size n using the Voronoi polygon approach introduced by Stevens and Olsen (2004). For a sample, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset [0, 1)^2$, the Voronoi polygon for $\mathbf{x}_i$ is

$$\omega_i = \{\mathbf{x} \in [0, 1)^2 : \|\mathbf{x} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_j\| \text{ for all } j = 1, 2, \ldots, n\}.$$

A sample is spatially balanced if

$$v_i = \sum_{j:\mathbf{x}_j \in \omega_i} \pi_j \approx 1,$$

for $i = 1, 2, \ldots, n$. If a point is on the boundary of multiple polygons, its inclusion probability is divided evenly among the polygons. The measure of spatial balance is expressed as the mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (v_i - 1)^2 .$$

To compare different sampling designs, the average MSE was computed using 1000 samples. We report relative spatial balance by dividing the average MSE of the proposed design by the average MSE under simple random sampling. Values less than one indicate better spatial balance in the proposed design when compared with simple random sampling. Random and grid point resources in $[0, 1)^2$ were considered with N ranging from 900 to 100,000.

We compared the spatial balance of (rapid) HIP with four competitive spatially balanced designs, LPM2 and suboptimal LPM in the R package BalancedSampling (Grafström and Lisic 2016), GRTS and GRTS $2n$ over-sampling in the R package spsurvey (Kincaid and Olsen 2016). For the large point resources with $N \geq 90,000$, we used LPM2 with k-d trees and the suboptimal LPM design. The parameter $h = 500$ (number of candidate neighbours/competitors) was used in suboptimal LPM (Grafström et al. 2014). In the GRTS $2n$ over-sampling approach, we selected a $2n$ over-sample and calculated the spatial balance of the first n points in reverse hierarchical order. Results for BAS are not presented. It was efficient to implement BAS on the grid point resources and similar results to HIP were obtained. However, it was computationally prohibitive to implement BAS on the random point resources. The fraction of $[0, 1)^2$ that defined the acceptance region for these resources was less than 1e-5, making the acceptance/rejection sampling technique that BAS uses extremely inefficient. For example, the number of random-start Halton points needed to draw 20 units from the 1000 random point resource was approximately 2.5 million points. Results for the other designs are given in Fig. 3.

The designs we considered have relative spatial balance values less than one meaning these designs had better spatial balance than simple random sampling on both random and grid point resources. The GRTS $2n$ over-sampling approach produced worse results than GRTS. Hence, observing the first n points of a GRTS $2n$ over-sample in reverse hierarchical order was not as effective as observing a GRTS sample of size n. We investigate this further in the following subsection.

LPM2 and HIP had better or similar spatial balance when compared with GRTS. The results for HIP and LPM2 were similar on the small grid structure of $N = 900$ points. However, for the $N = 1000$ random points resource, LPM2 had the best spatial balance. Because neighbouring points compete for inclusion in an LPM2 sample, the method is extremely effective when N is small (Grafström et al. 2012). For the $N = 10,000$ point resources, LPM2 performed slightly better than HIP and when $N \geq 90,000$ their performances were similar. Although LPM2 (with k-d trees) and HIP have complexities $O(N \log N)$, HIP is embarrassingly parallel meaning computationally efficient parallel implementations on large point resources are possible.

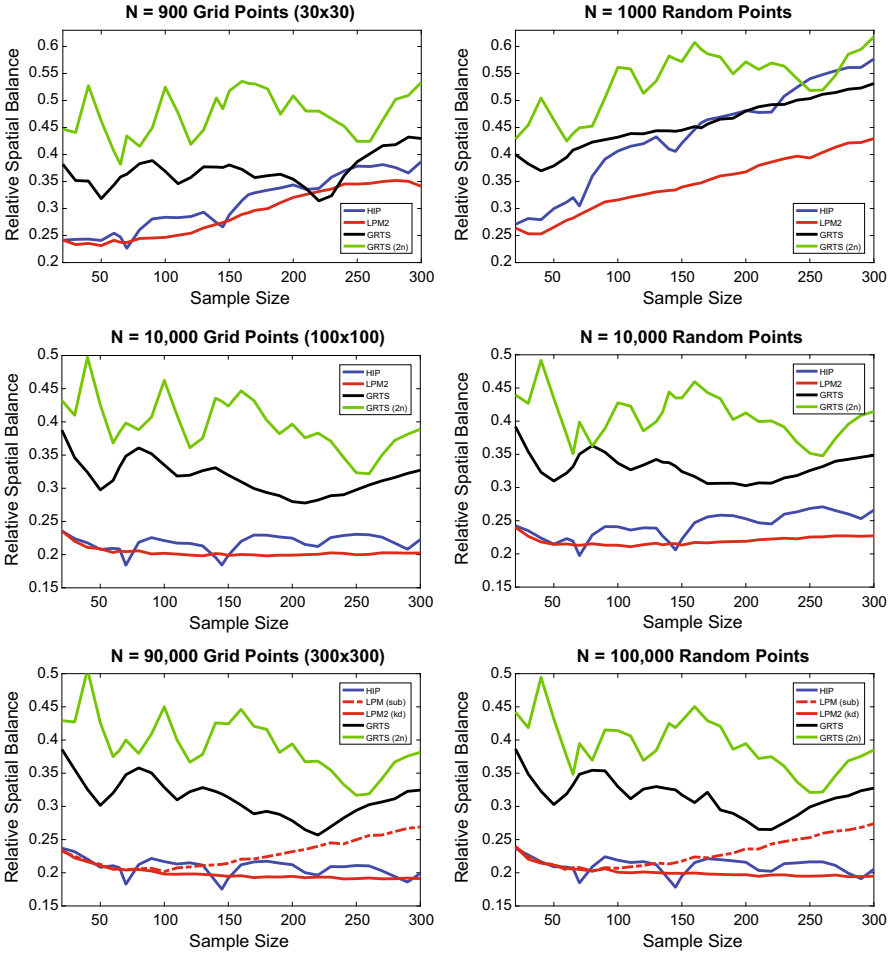

Fig. 3 Relative spatial balance of designs for sampling random and grid point resources as a function of sample size

For the two large point resources with $N \geq 90,000$, suboptimal LPM was also used. HIP, suboptimal LPM and LPM2 (with *k-d* trees) had similar performances for $n < 180$, but for larger sample sizes HIP and LPM2 performed noticeably better. Substantially increasing the parameter h (number of candidate neighbours/competitors) in suboptimal LPM may give better spatial balance for these larger sample sizes, but this made the method too computationally demanding to generate the required results.

4.1 Over-sampling

A practical problem in environmental sampling is the difficulty in obtaining an accurate sampling frame and in some instances available frames include many non-target

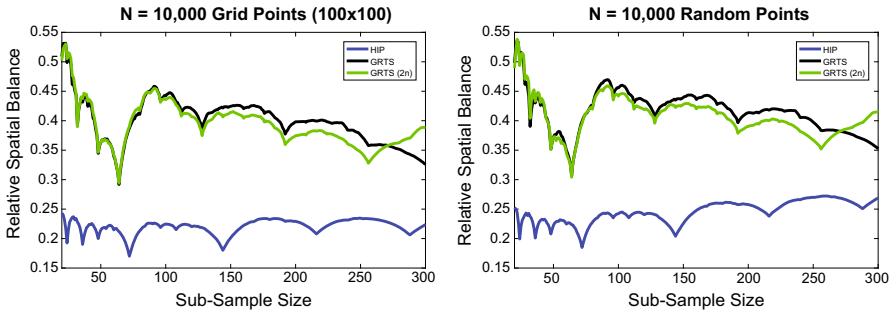

Fig. 4 Relative spatial balance of sub-samples of the form $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n^*)}\}$, where $\mathbf{x}^{(i)}$ is the ith point in an ordered over-sample of size 300 and n^* is the sub-sample size. In the GRTS $2n$ approach, an ordered over-sample of 600 points was drawn and the first 300 points were considered

units (Stevens and Olsen 2004). Another problem is that parts of the resource may be inaccessible because of physical location, safety or denied access from the land owner (Stevens and Olsen 2004). Both of these problems result in fewer units being observed than planned by the researcher. To achieve the desired sample size, an over-sampling (usually $2n$) strategy can be used (Stevens and Olsen 2004; Larsen et al. 2008), where we dynamically add units from the over-sample to the sample as non-target or inaccessible units are discovered. Inference is then based on inclusion probabilities conditional on the achieved sample size (Stevens and Olsen 2004). This inverse sampling approach does not eliminate the non-response or the bias of an inference, but it does allow for adaptive sample sizes and researchers can obtain the largest spatially balanced sample that their budget permits.

For a spatially balanced over-sampling approach, we require the over-sample to be ordered so that sub-samples of the form

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n^*)}\} \tag{8}$$

are spatially balanced, where $\mathbf{x}^{(i)}$ is the ith point in an ordered over-sample of size $2n$ and n^* is the sub-sample size. In the GRTS design, units are observed in reverse hierarchical order to ensure sub-samples of the form (8) are spatially balanced (Stevens and Olsen 2004). An ordered HIP over-sample is implicitly defined because the same hierarchical partition is used to draw a sample of size n^* or $2n$ provided $2n \leq B$, with a specific ordering determined by the Halton indices. The authors are not aware of an over-sampling strategy for LPM. Results for (rapid) HIP, GRTS and GRTS $2n$ over-sampling are given in Fig. 4.

The relative spatial balance values for HIP, GRTS and GRTS $2n$ were less than one indicating that each sub-sample drawn from the ordered over-sample was spatially balanced. The GRTS and GRTS $2n$ methods performed similarly, showing that the reverse hierarchical ordering strategy was effective for most sub-sample sizes. However, at the full over-sample size $n^* = 300$, GRTS performed better than GRTS $2n$, which agrees with the results in Fig. 3. HIP sampling had far better spatial balance than both GRTS approaches and is more flexible because the over-sample size does

not need to be specified prior to sampling. If a GRTS over-sample is too small, no additional points can be drawn using GRTS.

5 Estimation

Consider a point resource, $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, and let y_i denote the response value at the ith point. The Horvitz–Thompson estimator (Horvitz and Thompson 1952) 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 a sample and π_i is the inclusion probability of the ith point. The variance of $\widehat{\tau}$ can be estimated using the Sen-Yates-Grundy estimator (Sen 1953; Yates and Grundy 1953), but this estimator is biased and tends to be unstable because spatially balanced designs' second order inclusion probabilities of neighbouring points are close to zero (Grafström et al. 2012; Robertson et al. 2013). The local mean variance estimator (Stevens and Olsen 2003) is commonly used for spatially balanced designs and can be used for HIP. This estimator is

$$\widehat{\mathrm{V}}_{\mathrm{NBH}}(\widehat{\tau}) = \sum_{i \in \mathcal{S}} \sum_{j \in D_i} w_{ij} \left(\frac{y_j}{\pi_j} - \widehat{\tau}_{D_i} \right)^2, \tag{9}$$

where D_i is a neighbourhood containing at least four nearest neighbours to the ith point, $\widehat{\tau}_{D_i}$ is an estimate of the population total on D_i and w_{ij} are weights. Details on how to define neighbourhoods and compute weights can be found in Stevens and Olsen (2003).

Grafström and Schelin (2014) proposed a variance estimator for their LPM design,

$$\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau}) = \frac{1}{2} \sum_{i \in \mathcal{S}} \left(\frac{y_i}{\pi_i} - \frac{y_{j_i}}{\pi_{j_i}} \right)^2, \tag{10}$$

where $j_i \in \mathcal{S}$ is the index of the nearest neighbour in the sample to the ith point. This estimator does not require neighbourhood definitions or computation of weights and hence, is much simpler than $\widehat{\mathrm{V}}_{\mathrm{NBH}}(\widehat{\tau})$. In the following subsection we show that $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$ is also an appropriate variance estimator for HIP.

5.1 Numerical simulations and discussion

We demonstrate the performance of HIP sampling using two point resources: $N = 900$ points on a 30×30 grid and $N = 10,000$ points on a 100×100 grid. The population total of three artificial populations was estimated for each point resource, where the response value for each grid point was calculated via a continuous function. Three functions with different spatial structure were considered (see Web Figure 1) and are given in Web Section 2. Comparisons are made with simple random sampling (SRS),

Table 1 Results for populations 1, 2 and 3 with $N = 900$ for different sample sizes, n

Pop	n	SRS	GRTS		GRTS ($2n$)		LPM2			HIP		
		V_{SRS}	V_{SIM}	$\widehat{V}_{NBH}$	V_{SIM}	$\widehat{V}_{NBH}$	V_{SIM}	$\widehat{V}_{NBH}$	$\widehat{V}_{LPM}$	V_{SIM}	$\widehat{V}_{NBH}$	$\widehat{V}_{LPM}$
1	20	0.1022	0.0194	0.0281	0.0253	0.0276	0.0133	0.0305	0.0170	**0.0130**	0.0296	0.0159
	50	0.0395	0.0038	0.0053	0.0086	0.0050	0.0026	0.0057	0.0029	**0.0022**	0.0054	0.0026
	100	0.0186	0.0012	0.0014	0.0035	0.0013	0.0007	0.0015	0.0007	**0.0006**	0.0011	0.0006
	150	0.0116	0.0006	0.0006	0.0017	0.0006	**0.0003**	0.0007	0.0003	**0.0003**	0.0007	0.0003
	200	0.0081	0.0003	0.0004	0.0008	0.0003	**0.0002**	0.0004	0.0002	**0.0002**	0.0004	0.0002
2	20	0.1754	0.0797	0.1135	0.0747	0.1125	**0.0679**	0.1153	0.1131	0.0682	0.1135	0.1077
	50	0.0678	0.0172	0.0315	0.0270	0.0300	0.0137	0.0325	0.0234	**0.0129**	0.0318	0.0217
	100	0.0319	0.0069	0.0101	0.0112	0.0100	0.0042	0.0109	0.0063	**0.0041**	0.0105	0.0057
	150	0.0199	0.0033	0.0051	0.0071	0.0049	0.0021	0.0053	0.0029	**0.0019**	0.0052	0.0026
	200	0.0139	0.0016	0.0031	0.0039	0.0029	0.0014	0.0032	0.0016	**0.0013**	0.0030	0.0015
3	20	72.084	46.074	44.344	46.496	44.859	**30.164**	47.194	45.979	30.780	47.774	44.909
	50	27.851	9.450	13.084	12.318	12.603	**6.171**	13.367	10.057	7.754	13.378	9.563
	100	13.106	3.006	4.287	5.295	4.188	2.071	4.537	2.842	**1.913**	4.485	2.593
	150	8.191	1.316	2.165	3.016	2.083	**0.943**	2.226	1.316	1.025	2.212	1.179
	200	5.734	0.727	1.317	1.711	1.253	**0.613**	1.357	0.733	0.678	1.312	0.675

The reported values are averages using 1000 different samples, where V_{SIM} is the simulated MSE (11), $\widehat{V}_{NBH}$ is the local mean variance estimator (9) and $\widehat{V}_{LPM}$ is given by (10). Exact values are shown for V_{SRS}. The lowest simulated variance for each problem is shown in bold

GRTS and LPM2. For each design, the variance of $\widehat{\tau}$ was estimated using the simulated mean square error

$$\mathrm{V}_{\mathrm{SIM}} = \frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\tau}_i - \tau)^2, \tag{11}$$

where $\widehat{\tau}_i$ is the Horvitz–Thompson estimate for the ith sample (exact values are given for SRS). The average performances (over 1000 runs) of $\widehat{\mathrm{V}}_{\mathrm{NBH}}(\widehat{\tau})$ and $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$ were also computed. Results for $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$ are not presented for GRTS because this estimator substantially under-estimated the simulated variance for GRTS. The authors attribute this to the poorer spatial balance of GRTS and would not recommend using $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$ for GRTS. There was little difference between the results for $N = 900$ and $N = 10,000$ (see Web Table 3) so we focus our discussion on the $N = 900$ case. These results are presented in Table 1.

Table 1 shows that spatially balanced sampling is effective on the problems considered, with GRTS, LPM2 and HIP producing estimates of the population total with better precision than SRS. The best gains in precision were made on population 1, where a strong spatial trend is present (see Web Figure 1(a)).

HIP and LPM2 performed better than GRTS on the problems considered, but there was negligible difference between HIP and LPM2. The spatial balance of LPM2 was slightly better than HIP on these problems (see Fig. 3), but these small gains in spatial balance did not increase LPM2's precision. The local mean variance estimator provided a conservative estimate of the variance of $\widehat{\tau}$ for these designs. A tighter estimate for LPM2 and HIP was given by $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$. This estimator performed well for most sample sizes, but was conservative for some of the smaller sample sizes. The authors recommend using $\widehat{\mathrm{V}}_{\mathrm{LPM}}(\widehat{\tau})$ as an estimate of $\mathrm{V}(\widehat{\tau})$ for HIP. Researchers interested in model-based estimators for spatially based designs are referred to Foster et al. (2017). A fuller discussion of model-based designs is beyond the scope of this article.

The local mean variance estimator was optimistic for the GRTS $2n$ over-sampling approach, with an average estimate smaller than the simulated variance. The authors attribute this to the poor spatial balance of the first n points from an ordered GRTS $2n$ over-sample (see Fig. 3). Hence (9) should be used with caution for GRTS $2n$ over-sampling. Provided $B \geq 2n$, the variance estimators (9) and (10) can be used for HIP over-sampling because an ordered HIP over-sample of size $n^* \leq 2n$ is also a HIP sample of size n^*.

6 Conclusion

In this article we introduced the HIP sampling design. This design uses structural properties of the Halton sequence to iteratively partition a resource into nested boxes. Sample points are then drawn from specific boxes to give a spatially balanced sample. The HIP design can be applied to point and continuous resources with equal or unequal inclusion probability/density. We have shown that HIP draws spatially balanced samples and performs well when compared with competing designs. The main advantages of HIP sampling over existing designs is that it is computationally efficient, embar-

rassingly parallel on large point resources and can be used for spatially balanced over-sampling. This makes HIP particularly useful for sampling natural resources because imperfect sampling frames and accessibility problems result in fewer units being observed than planned. Although the spatially balanced over-sampling strategy achieves the desired sample size and is popular with field researchers, it will not eliminate the non-response or the bias of an inference. Design based estimators were provided and were effective on three populations with different spatial structures.

Acknowledgements We thank two anonymous referees and the editors for valuable comments that led to an improved article.

References

Brewer KRW, Hanif M (1983) Sampling with unequal probabilities. Lecture notes in statistics. Springer, New York

Deville JC, Tillé Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85:89–101

Foster SD, Hosack GR, Lawrence E, Przeslawski R, Hedge P, Caley MJ, Barrett NS, Williams A, Li J, Lynch T et al (2017) Spatially-balanced designs that incorporate legacy sites. Methods Ecol Evol 8:1433–1442

Grafström A, Lisic J (2016) BalancedSampling: balanced and spatially balanced sampling. *R* package version 1.5.1. https://www.antongrafstrom.se/balancedsampling

Grafström A, Lundström NLP (2013) Why well spread probability samples are balanced. Open J Stat 3:36–41

Grafström A, Schelin L (2014) How to select representative samples. Scand J Stat 41:277–290

Grafström A, Lundström NLP, Schelin L (2012) Spatially balanced sampling through the pivotal method. Biometrics 68:514–520

Grafström A, Saarela S, Ene LT (2014) Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. Can J Forest Res 44:1156–1164

Grafström A, Schnell S, Saarela S, Hubbell SP, Condit R (2017a) The continuous population approach to forest inventories and use of information in the design. Environmetrics 28(8):1–12

Grafström A, Zhao X, Nylander M, Petersson H (2017b) A new sampling strategy for forest inventories applied to the temporary clusters of the Swedish national forest inventory. Can J For Res 47:1161–1167

Halton JH (1960) On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numer Math 2:84–90

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685

Kincaid TM, Olsen AR (2016) spsurvey: spatial survey design and analysis. *R* package version 3.3. https://CRAN.R-project.org/package=spsurvey

Larsen DP, Olsen AR, Jr Stevens D L (2008) Using a master sample to integrate stream monitoring programs. J Agric Biol Environ Stat 13:243–254

McDonald T (2014) Sampling designs for environmental monitoring. In: Manly BFJ, Navarro Alberto JA (eds) Introduction to ecological sampling. CRC Press, Taylor and Francis Group, Boca Raton

Price CJ, Price CP (2012) Recycling primes in Halton sequences: an optimization perspective. Adv Model Optim 14:17–29

Robertson BL, Brown JA, McDonald T, Jaksons P (2013) BAS: balanced acceptance sampling of natural resources. Biometrics 3:776–784

Robertson BL, McDonald T, Price CJ, Brown JA (2017) A modification of balanced acceptance sampling. Stat Probab Lett 129:107–112

Sen AR (1953) On the estimate of variance in sampling with varying probabilities. J Indian Soc Agric Stat 7:119–127

Stevens DL Jr, Olsen AR (2003) Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14:593–610

Stevens DL Jr, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99:262–278

Wang X, Hickernell FJ (2000) Randomized Halton sequences. Math Comput Model 32:887–899

Yates F, Grundy PM (1953) Selection without replacement from within strata with probability proportional to size. J R Stat Soc Ser B 15:235–261

Blair Robertson received a Ph.D. in Mathematics from the University of Canterbury, New Zealand in 2010. He was an assistant professor at the University of Wyoming, United States, from 2014 to 2015 and is currently a Senior Lecturer at the School of Mathematics and Statistics, University of Canterbury, New Zealand. He won the New Zealand Statistical Association Worsley Early Career Research Award in 2015. His research interests are statistical sampling designs and numerical optimization.

Trent McDonald graduated with Honors from the University of Wyoming in 1988 with a degree in Statistics and Computer Science, from New Mexico State University in 1990 with a M.S. degree in Experimental Statistics, and in 1996 completed a Ph.D. in Statistics at Oregon State University. He joined Western EcoSystems Technology Inc. (WEST) in 1996 and has been a Statistician and Project Manager ever since. He specializes in sample design, capture–recapture analyses, habitat selection, and linear model applications to species from bowhead whales to Aleutian terns. He has received the Outstanding Article, Outstanding Edited Book, and a Certificate of Excellence from The Wildlife Society. He also possesses a Unit Award for Excellence of Service issued by the U.S. Secretary of the Interior.

Chris Price first graduated in electrical and electronic engineering, and then went on to gain a Ph.D. in mathematics in 1992. He is currently an Associate Professor in Mathematics at the University of Canterbury, specializing in direct search methods for optimization.

Jennifer Brown is a Professor in Statistics at University of Canterbury. She graduated from Otago University with a Ph.D., and has an undergraduate degree in Forestry (University of Canterbury) and a postgraduate qualification in statistics (Massey University). She has been Head of the School of Mathematics and Statistics at University since 2009, and is a past president of the NZ Statistical Association. Her research interests are in environmental monitoring, survey sampling and health sciences.