Thailand Statistician
January 2021; 19(1): 125-139
http://statassoc.or.th
Contributed paper

Using Balanced Incomplete Block Designs to Generate New Sampling Designs

Paramaporn Sangpara [a]*, John Joseph Borkowski [b] and Patchanok Srisuradetchai [a]
[a] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand.
[b] Department of Mathematical Sciences, Montana State University, Bozeman, Montana, USA.
*Corresponding author; e-mail: para.sangpara@gmail.com

Received: 21 November 2019
Revised: 6 January 2020
Accepted: 28 January 2020

Abstract

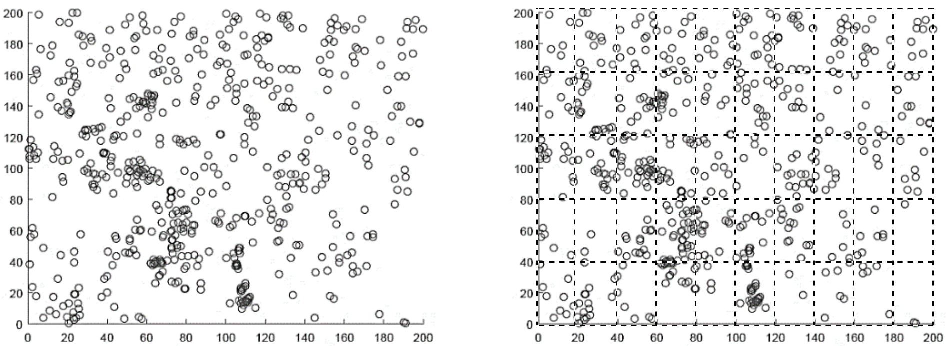
Consider a studies of biological species or communities in which the abundance measurements tend to be spatially correlated within the study region. When sampling spatially, the geographical study region is often partitioned into quadrat. The goals of the survey center on the sampling designs which ensure the sample is spread over the study region. In this paper we introduce a new probability sampling design which we call a balanced incomplete block sampling design (BIBSD). The goal is estimation of the population total or mean using the Horvitz-Thompson estimators. The BIBSD-based designs can provide precise estimates of the population mean or total when strong spatial patterns or trends exist in the population. It is shown that the BIBSD-based designs when compared the efficiencies of the new Horvitz-Thompson BIBSD estimators of population total or mean to the estimators using simple random sampling will provide estimators with smaller variability.

Keywords: Horvitz-Thompson estimators, BIBSD-based designs, spatial sampling, inclusion probabilities.

1. Introduction

Sampling consists of selecting and observing some part of a population so that a researcher may estimate and make inferences regarding a characteristic associated with the whole population. The primary statistical problems are how best to obtain the sample, and then how to estimate the population characteristic from the sample data (See Cochran 1997, Lohr 2010 and Thompson 2012). A population of interest that is spatially distributed over a study region is often partitioned into quadrats which represent the sampling units. For example, Figure 1(a) shows the 584 locations of individual long-leaf pine trees in a 200m × 200m study region in Thomas Country, Georgia, USA (Rathbun and Cressie 1994). In Figure 1(b), the study region is divided into fifty 20m × 40m quadrats. With a spatially distributed population, the variable of interest can be measured for each quadrat, and that information can be used to estimate a population parameter (such as a population mean, total, or proportion). Common applications of spatial sampling include biological, ecological, agricultural and sociological studies.

If spatial correlation is present, sampling designs which ensure the sample is well-distributed over the study region will, in general, improve estimation of population parameters relative to designs that do not (Borkowski 2003). In this research we introduce a new probability sampling design which we call a balanced incomplete block sampling design (BIBSD). The samples collected from this design will provide greater coverage of the study region than samples of the same size collected using simple random sampling (SRS), and, therefore, will have estimators that are more efficient than estimators obtained from the SRS designs when positive spatial autocorrelation is suspected or known to exist. Theoretically, to use a BIBSD the study region must be partitioned into a rectangular grid of a columns and b rows. The layout of a BIBSD is based on the structure of a balanced incomplete experimental design (BIBED) having a treatments in b blocks in which it is only possible to assign a subset of k treatments $(k < a)$ within each of the b blocks. The k treatments within each block are balanced such that every pair of treatments appears together in the same number (λ) of blocks. With a BIBSD, the b rows and a columns of the study region correspond, respectively, to the b blocks and a treatments in a BIBED. The k population units to be sampled by a BIBSD in each row correspond to the subset of k treatments appearing in each of the b blocks of the BIBED (See John 1980, Street and Street 1987, Stinson 2004, Raghavarao and Padgett 2005).



(a) (b)

Figure 1 Location map of a tree population (Rathbun and Cressie, 1994)

Suppose N is the number of sampling units (quadrats) in the population of interest. Horvitz-Thompson (HT) estimation (Horvitz and Thompson 1952) can be applied to the proposed BIBSDs in this research because HT estimation can be used with any probability sampling design, with or without replacement, as long as the researcher knows (i) the first-order inclusion probability π_i that sampling unit u_i (the i^{th} quadrat) is included by the design for $i = 1,...,N$ and (ii) the second-order joint inclusion probability π_{ij} that a pair of sampling units u_i and u_j (the i^{th} and j^{th} quadrats) are both included by the design (Hedayat and Sinha 1991, Thompson 2012). For BIBSDs, the main problem is to derive the first-order and second-order inclusion probabilities. The second-order inclusion probabilities depend on whether units u_i and u_j are in the same row, the same column, or are in different rows and columns.

Sampling designs are defined in Sections 2 and 3. Horvitz-Thompson estimators are presented in Section 4. Examples of the application of BIBSDs are presented in Section 5.

2. Balanced Incomplete Block Sampling Designs (BIBSDs)

Before defining the properties of a BIBSD, we must describe the properties of a balanced incomplete block experimental design (BIBED). A BIBED has design parameters a, b, k, r and λ where a is the number of experimental treatments, b is the number of blocks, k is number of distinct treatments occurring in each block, r is number of blocks in which each treatment appears and λ is number of blocks that each pair of treatments appears together (Dey 2010). Related to the design parameters are n and N, where n is number of units in the BIBED $(n = bk = ar)$ and N is the number of potential block and treatment combinations $(N = a \times b)$.

In a sampling framework, suppose the population units form a grid of $a \times b$ quadrats. The two dimensions will be referred to as "column" and "row" and correspond to the treatments and blocks of a BIBED, respectively. To generate a random BIBSD, start with any BIBED with design parameters a, b, k, r and λ. Then simply replace each block and treatment combination in the BIBED with a ● symbol to indicate a unit to be sampled from the population of quadrats. Then randomize the rows of the initial BIBSD to form a random BIBSD.

Example: Consider a study region that has 6 rows and 4 columns. We start by taking a BIBED with $a = 4$ treatments (A, B, C, D) and $b = 6$ blocks. The starting BIBED is shown in Figure 2(a). Each treatment appears in exactly $r = 3$ blocks and each pair of treatments appears together in exactly $\lambda = 1$ block. To generate the initial BIBSD, just replace the treatment letters with a ● symbol. The initial BIBSD appears in Figure 2(b). Then randomly permute the rows to generate a random BIBSD (Figure 2(c)).

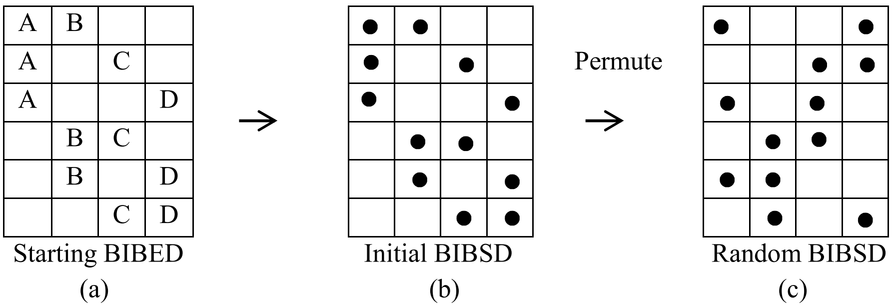


Figure 2 Example of forming a BIBSD from a BIBED

3. Inclusion Probabilities

For each unit in the sample, we need to derive the general forms for first-order inclusion probabilities (π_i) and the second-order (π_{ij}) inclusion probabilities for all $i, j = 1, 2, ..., N$ to be used when calculating Horvitz-Thompson estimates of the mean $\hat{\mu}_{BIB}$ and the population total $\hat{\tau}_{BIB}$, estimators of their variances $\mathbf{V}(\hat{\mu}_{BIB})$ and $\mathbf{V}(\hat{\tau}_{BIB})$, and estimated variances $\hat{\mathbf{V}}(\hat{\mu}_{BIB})$ and $\hat{\mathbf{V}}(\hat{\tau}_{BIB})$ when using a BIBSD. Because the general forms depend only on the design parameters a, b, k, r and λ they can be applied to any BIBSD. The inclusion probabilities will also depend on the number of possible BIBSDs that can be generated from a starting BIBSD by randomly permuting the rows of the starting BIBSD. The resulting BIBSD will indicate which n units to select from the population of N units.

3.1. First-order inclusion probabilities for BIBSDs

For a BIBSD, the first-order inclusion probability (π_i) of any unit u_i is the probability that unit u_i is in the BIBSD or, equivalently, the probability that unit u_i is a ● quadrat in the grid with b rows and a columns. Because of the row and column balance of units in a BIBSD, each unit has the same probability of being selected. And because there are r sampled rows in the column containing u_i, there are $r(b-1)!$ possible ways to select unit u_i in the grid with b rows and a columns from the $b!$ possible BIBSDs. Therefore,

$$\pi_i = \frac{r(b-1)!}{b!} = \frac{r}{b}. \tag{1}$$

3.2. Second-order inclusion probabilities for BIBSDs

The second-order inclusion probability (π_{ij}) of a pair of sampling units u_i and u_j in a BIBSD is the probability that unit u_i and unit u_j are in a sample or, equivalently, the probability that unit u_i and unit u_j are both a ● quadrat in the grid with b rows and a columns. For distinct u_i and u_j, calculation of π_{ij} depends on one of the following three cases.

Case 1: Unit u_i and unit u_j are in the same column but different rows.

Suppose u_i and u_j are both in column c. Then there are $r(r-1)$ possible ordered ways to select u_i and u_j, and after this selection, there are $(b-2)!$ different BIBSDs that contain u_i and u_j in column c from the set of all possible $b!$ BIBSDs. Therefore,

$$\pi_{ij} = \frac{r(r-1)(b-2)!}{b!} = \frac{r(r-1)}{b(b-1)}. \tag{2}$$

Case 2: Unit u_i and unit u_j are in the same row but different columns.

Suppose u_i is in column c_i and u_j is in column c_j $(c_i \neq c_j)$, but both units are in the same row. There are λ possible pairs for u_i and u_j to be in the same row, and there are $(b-1)!$ different BIBSDs that contain u_i and u_j in the same row from the set of all possible $b!$ BIBSDs. Therefore,

$$\pi_{ij} = \frac{\lambda(b-1)!}{b!} = \frac{\lambda(b-1)!}{b(b-1)!} = \frac{\lambda}{b}. \tag{3}$$

Case 3: Unit u_i and unit u_j are in different rows and different columns.

Suppose u_i and u_j are in different rows and columns. There are two cases to consider:

Case 3a: Suppose the unit in the same row as unit u_i and the same column as unit u_j or the unit in the same column as u_i and the same row as u_j is also in the BIBSD. Then there are λ pairs to select u_i and $r-1$ possible ways for u_j and $(b-2)!$ ways to permute the remaining rows of the BIBSD from all possible $b!$ BIBSDs. Therefore,

$$\pi_{ij} = \frac{\lambda(r-1)(b-2)!}{b!} = \frac{\lambda(r-1)(b-2)!}{b(b-1)(b-2)!} = \frac{\lambda(r-1)}{b(b-1)}. \tag{4}$$

Case 3b: Suppose the unit in the same row as u_i and the same column as u_j and the unit in the same column as u_i and the same row as u_j are both not in the BIBSD. Then there are $r(r-\lambda)$ possible ways to select u_i and u_j and $(b-2)!$ ways to permute the remaining rows of the BIB sample from all possible $b!$ BIBSDs. Therefore,

$$\pi_{ij} = \frac{r(r-\lambda)(b-2)!}{b!} = \frac{r(r-\lambda)(b-2)!}{b(b-1)(b-2)!} = \frac{r(r-\lambda)}{b(b-1)}. \tag{5}$$

Combining the results from Case 3a and Case 3b when u_i and u_j are in different rows and different columns yields

$$\pi_{ij} = \frac{\lambda(r-1)}{b(b-1)} + \frac{r(r-\lambda)}{b(b-1)} = \frac{r^2-\lambda}{b(b-1)}. \tag{6}$$

The results that $\sum_{i=1}^{N}\pi_i = n$ and $\sum_{i=1}^{N}\sum_{i\neq j}^{N}\pi_{ij} = \binom{n}{2}$ are required for any probability sampling design (Hedayat and Sinha 1991). This will be verified for a BIBSD having a sample size of $n = ar = bk$ units. Note that

$$\sum_{i=1}^{N}\pi_i = \sum_{i=1}^{ab}\frac{r}{b} = \frac{abr}{b} = ar = n, \text{ and}$$

$$\sum_{i=1}^{N}\sum_{i\neq j}^{N}\pi_{ij} = \sum_{i=1}^{a}\sum_{i\neq j}^{\binom{n}{2}}\pi_{ij} + \sum_{i=1}^{b}\sum_{i\neq j}^{\binom{n}{2}}\pi_{ij} + \sum_{i=1}\sum_{i\neq j}\pi_{ij} = \sum_{i=1}^{a}\sum_{i\neq j}^{\binom{b}{2}}\frac{r(r-1)}{b(b-1)} + \sum_{i=1}^{b}\sum_{i\neq j}^{\binom{a}{2}}\frac{\lambda}{b} + \sum_{i=1}\sum_{i\neq j}\frac{r^2-\lambda}{b(b-1)}$$

$$= a\binom{b}{2}\frac{r(r-1)}{b(b-1)} + b\binom{a}{2}\frac{\lambda}{b} + \left(\binom{ab}{2} - a\binom{b}{2} - b\binom{a}{2}\right)\frac{(r^2-\lambda)}{b(b-1)}$$

$$= \frac{1}{2}ar(ar-1) = \binom{ar}{2} = \binom{n}{2}.$$

An example showing the values for the first-order inclusion probabilities and the second-order inclusion probabilities for Cases 1, 2, 3a and 3b is now presented. The units indicated by O are the units for which inclusion probabilities will be calculated.

Example: Suppose the goal is to generate a random BIBSD with $a = 4$, $b = 4$, $k = 3$, $r = 3$ and $\lambda = 2$. There are $4! = 24$ possible BIBSDs generated from an initial BIBSD, and which are shown Figure 3. The first-order inclusion probability of unit $u_i = \frac{\text{\#of BIBSDs with unit O}}{\text{Total \#of BIBSDs}} = \frac{18}{24} = \frac{3}{4}$.

R1	○	●	●	
R2	●	●		●
R3	●		●	●
R4		●	●	●

And from Equation (1): $\pi_i = \frac{r}{b} = \frac{3}{4}$.

The second-order inclusion probability π_{ij} depends on one of the following three cases:

Case 1: Units u_i and u_j are in the same column but different rows.

The second-order inclusion probability $= \dfrac{\text{\# of BIBSDs with both O units}}{\text{Total \# of BIBSDs}} = \dfrac{12}{24} = \dfrac{1}{2}.$

R1	○	●	●	
R2	○	●		●
R3	●		●	●
R4		●	●	●

And from Equation (2): $\pi_{ij} = \dfrac{r(r-1)}{b(b-1)} = \dfrac{3(3-1)}{4(4-1)} = \dfrac{6}{12} = \dfrac{1}{2}.$

Case 2: Units u_i and u_j are in the same row but different columns.

The second-order inclusion probability $= \dfrac{\text{\# of BIBSDs with both O units}}{\text{Total \# of BIBSDs}} = \dfrac{12}{24} = \dfrac{1}{2}.$

And

R1	○	○	●	
R2	●	●		●
R3	●		●	●
R4		●	●	●

from Equation (3): $\pi_{ij} = \dfrac{\lambda}{b} = \dfrac{2}{4} = \dfrac{1}{2}.$

Case 3: Units u_i and u_j are in different rows and different columns.

Case 3a: The X unit in the same row as unit u_i and the same column as unit u_j is also in the BIBSD.

The second-order inclusion probability $= \dfrac{\text{\# of BIBSDs with both O units and unit X}}{\text{Total \# of BIBSDs}} = \dfrac{8}{24} = \dfrac{1}{3}.$

And

R1	○	X	●	
R2	●	○		●
R3	●		●	●
R4		●	●	●

from Equation (4): $\pi_{ij} = \dfrac{\lambda(r-1)}{b(b-1)} = \dfrac{2(3-1)}{4(4-1)} = \dfrac{4}{12} = \dfrac{1}{3}.$

Case 3b: The unit in the same row as u_i and the same column as u_j is not in the BIBSD. This corresponds to the - unit neighboring the O units.

The second-order inclusion probability $= \dfrac{\text{\# of BIBSDs with both O units and not the - unit}}{\text{Total \# of BIBSDs}} = \dfrac{6}{24} = \dfrac{1}{4}.$

And

R1	○	-	●	●
R2	●	○	●	
R3	●	●		●
R4		●	●	●

from Equation (5): $\pi_{ij} = \dfrac{r(r-\lambda)}{b(b-1)} = \dfrac{3(3-2)}{4(4-1)} = \dfrac{3}{12} = \dfrac{1}{4}.$

Combining the results from Case 3a and Case 3b when u_i and u_j are in different rows and different columns yields $\pi_{ij} = \dfrac{r^2-\lambda}{b(b-1)} = \dfrac{3^2-2}{4(3-1)} = \dfrac{7}{12} = \dfrac{1}{3} + \dfrac{1}{4}.$

4. The Horvitz-Thompson Estimators

Goals of this research include deriving unbiased estimators of the population total τ and population mean μ, deriving formulas for the true variances of these estimators, and deriving unbiased estimators of the true variances when using BIBSD. Horvitz-Thompson estimation will be used to generate formulas for these three estimation goals.

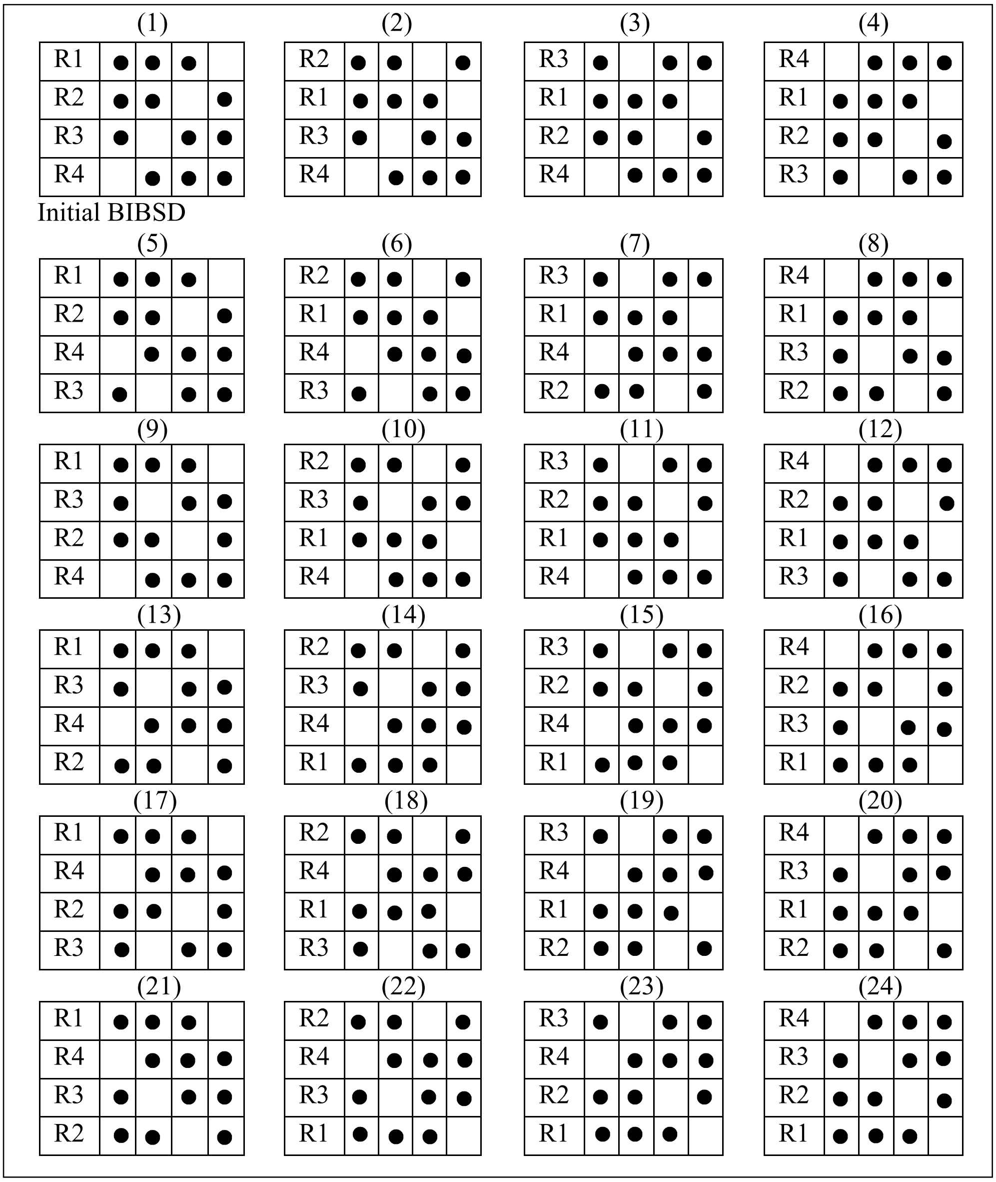


Figure 3 The 24 possible BIBSDs with $a = 4,\ b = 4,\ k = 3,\ r = 3$ and $\lambda = 2$

Because inclusion probabilities in Equations (1) to (6) are known for BIBSDs, they can be used to estimate a population total τ or a population mean μ. The general forms of the Horvitz-Thompson (HT) estimators (Horvitz and Thompson 1952) of τ and μ when the sample of size n is selected without replacement from a population of units are $\hat{\tau} = \sum_{i=1}^{N} \frac{y_i}{\pi_i}$ and $\hat{\mu} = \hat{\tau} / N$. These estimators are also unbiased. Given $\pi_i = r / b$ and $n = bk$, the HT estimators $\hat{\tau}_{BIB}$ and $\hat{\mu}_{BIB}$ for a BIBSD are:

$$\hat{\tau}_{BIB} = b \sum_{i=1}^{bk} \frac{y_i}{r} \text{ and } \hat{\mu}_{BIB} = \frac{1}{k} \sum_{i=1}^{bk} \frac{y_i}{r}. \tag{7}$$

The variance of HT estimator $\hat{\tau}$ is

$$\mathbf{V}(\hat{\tau})=\sum_{i=1}^{N}\left(\frac{1}{\pi_i}-1\right)y_i^2+2\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)y_iy_j \quad (8)$$

and an estimator of this variance is

$$\hat{\mathbf{V}}(\hat{\tau})=\sum_{i=1}^{n}\left(\frac{1}{\pi_i}-1\right)y_i^2+2\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)y_iy_j. \quad (9)$$

It follows directly that the variance $\mathbf{V}(\hat{\mu})=\dfrac{\mathbf{V}(\hat{\tau})}{N^2}$ and the estimated variance $\hat{\mathbf{V}}(\hat{\mu})=\dfrac{\hat{\mathbf{V}}(\hat{\tau})}{N^2}$.

The general form for the variance of HT estimator $\hat{\tau}_{BIB}$ is

$$\mathbf{V}(\hat{\tau}_{BIB})=\sum_{i=1}^{ab}\left(\frac{b}{r}-1\right)y_i^2+2\sum_{i=1}^{ab-1}\sum_{j>i}^{ab}\left(\frac{b^2\pi_{ij}}{r^2}-1\right)y_iy_j \quad (10)$$

and an estimator of this variance is

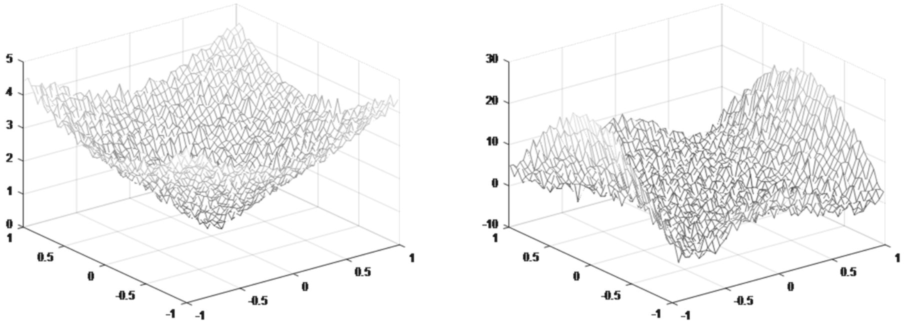
$$\hat{\mathbf{V}}(\hat{\tau}_{BIB})=\sum_{i=1}^{bk}\left(\frac{b(b-r)}{r^2}\right)y_i^2+2\sum_{i=1}^{bk-1}\sum_{j>i}^{bk}\left(\frac{b^2}{r^2}-\frac{1}{\pi_{ij}}\right)y_iy_j, \quad (11)$$

where the double summations in (10) and (11) are over all there are $\binom{ab}{2}$ and $\binom{bk}{2}$ pairs of units in the population and sample, respectively. For estimation of the population mean, it follows directly that $\hat{\mu}_{BIB}=\dfrac{\hat{\tau}_{BIB}}{ab}$, so the variance $\mathbf{V}(\hat{\mu}_{BIB})=\dfrac{\mathbf{V}(\hat{\tau}_{BIB})}{(ab)^2}$ and the estimated variance $\hat{\mathbf{V}}(\hat{\mu}_{BIB})=\dfrac{\hat{\mathbf{V}}(\hat{\tau}_{BIB})}{(ab)^2}$. Because $\pi_{ij}>0$ for all $i,j=1,2,...,ab$, the HT estimators $\hat{\mathbf{V}}(\hat{\tau}_{BIB})$ and $\hat{\mathrm{v}}(\hat{\mu}_{BIB})$ will be unbiased estimators in the design sense (Thompson 2012). That is, $\mathbf{E}[\hat{\mathbf{V}}(\hat{\tau}_{BIB})]=\mathbf{V}(\hat{\tau}_{BIB})$ and $\mathbf{E}[\hat{\mathbf{V}}(\hat{\mu}_{BIB})]=\mathbf{V}(\hat{\mu}_{BIB})$ where the expectations are taken over all possible BIBSDs.

5. Examples of Populations and Application of BIBSDs

The efficiencies of the Horvitz-Thompson BIBSD estimators of the population mean μ will be compared to the estimators using simple random sampling for two hypothetical populations and one real population across different BIBSDs that can be applied to those populations. The comparisons will be based on calculated values of the variances and the confidence interval coverage percentages and widths for BIBSDs and SRSs for several combinations of sample size n and population type.

Let $N=a\times b$ be the number of quadrats in the population of interest. The variances of SRS estimators (Cochran 1977) are $\mathbf{V}(\hat{\mu}_{SRS})=((N-n)/N)\sigma^2$, where σ^2 is the population variance. The value of the variance of the HT estimator of the mean $(V(\hat{\mu}_{BIB})=V(\hat{\tau}_{BIB})/(ab)^2)$ can be calculated from the variance formula in (10). Then, generate 10,000 random BIBSDs with sample of size $n=bk$ and calculate the estimator variances $(\hat{\mathbf{V}}(\hat{\mu}_{BIB})=\hat{\mathbf{V}}(\hat{\tau}_{BIB})/(ab)^2)$ for each of these 10,000 samples.



(a) Population 1 (b) Population 2

Figure 4 Two hypothetical populations, Population 1 in (a) has a trend towards a center minimum, Population 2 in (b) exhibits a saddle point trend

First consider the two hypothetical populations in Figure 4 that exhibit two types of spatial correlation. Population 1 in Figure 4(a) is partitioned into a 50×50 grid with a center minimum trend. Population 2 in Figure 4(b) is also partitioned into a 50×50 grid but with a saddle point. The vertical axis represents the population response for that grid unit.

The 2,500 units in the 50×50 grids for Populations 1 and 2 were partitioned into two population sizes: grids of $N = 10 \times 5 = 50$ and $25 \times 25 = 625$ quadrats. The sample sizes considered in Table 1 correspond to the parameters of BIBEDs with $n = ar = bk$. The average estimated variances of the BIBSD estimator over 10,000 samples are only slightly different from the true variances. This must be the case because the variance estimators are unbiased for both sampling plans.

Table 1 A comparison of the variances and the average estimated variances for estimation of the population mean using BIBSDs and SRSs taken from the two populations in Figure 4

Population 1: Spatial Type 1: a population with a center minimum trend

					BIBSD			SRS	
N	n	a	b	k	$V(\hat{\mu}_{BIB})$	$\hat{V}(\hat{\mu}_{BIB})$	#Neg	$V(\hat{\mu}_{SRS})$	$\hat{V}(\hat{\mu}_{SRS})$
50	20	5	10	2	2.2019	2.3333	4135	52.2385	52.1600
50	30	5	10	3	0.9786	1.0231	3955	23.2171	23.1977
625	225	25	25	9	0.0016	0.0016	2869	0.0345	0.0345
625	400	25	25	16	0.0005	0.0005	857	0.0109	0.0109

Population 2: Spatial Type 2: a population with a saddle point trend

					BIBSD			SRS	
N	n	a	b	k	$V(\hat{\mu}_{BIB})$	$\hat{V}(\hat{\mu}_{BIB})$	#Neg	$V(\hat{\mu}_{SRS})$	$\hat{V}(\hat{\mu}_{SRS})$
50	20	5	10	2	632.5634	633.5841	89	1167.7161	1167.1981
50	30	5	10	3	281.1393	281.9979	0	518.9849	518.8796
625	225	25	25	9	0.3958	0.3964	0	1.1624	1.1583
625	400	25	25	16	0.1252	0.1252	0	0.3678	0.3675

Because of the strong spatial pattern in Population 1, the variance of the estimator is much smaller than the variance under SRS. Thus, the average estimated variances of BIBSDs were also much smaller than those for SRS for all population sizes. Similarly, for Population 2 with the strong saddle point pattern, the variances of the BIBSD estimators are much smaller than the variances under SRS for all population sizes. Thus, BIBSDs were much more efficient for estimating the population mean (or total) than using SRSs. The differences in precision are reflected in the estimated sampling distributions of $\hat{\mu}_{BIB}$ and $\hat{\mu}_{SRS}$ that are summarized by the histograms in Figures 5 and 6. The empirical distributions of the 10,000 BIBSD, $\hat{\mu}_{BIB}$ estimates are much narrower than the distributions of the $\hat{\mu}_{SRS}$ estimates under SRS.

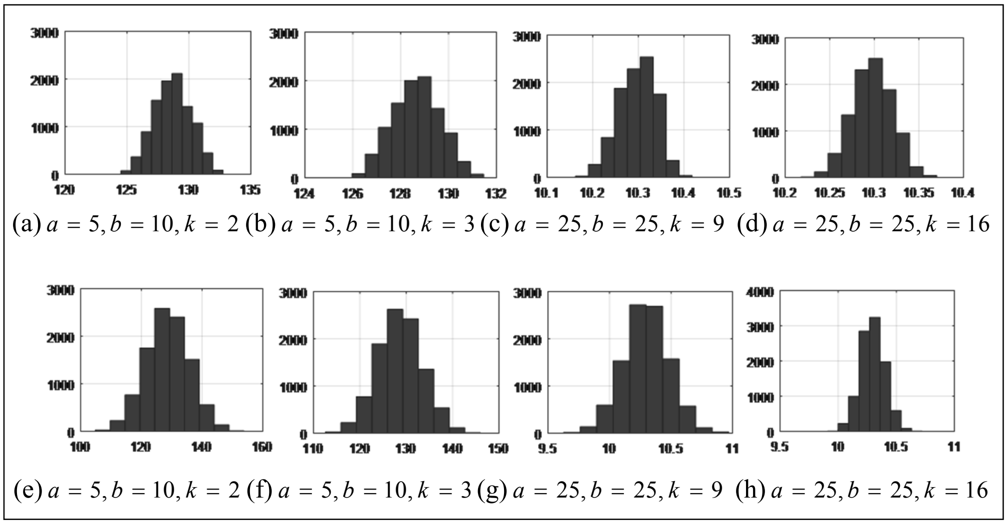


Figure 5 The empirical sampling distributions of $\hat{\mu}_{BIB}$ in (a), (b), (c), (d) and $\hat{\mu}_{SRS}$ in (e), (f), (g), (h) are based on 10,000 initial BIBSDs and SRSs for the hypothetical Population1

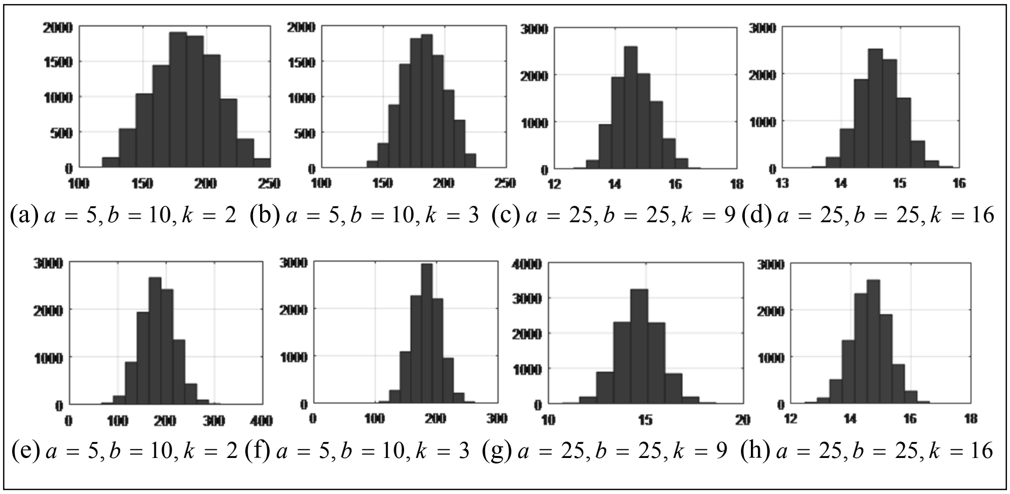


Figure 6 The empirical sampling distributions of $\hat{\mu}_{BIB}$ in (a), (b), (c), (d) and $\hat{\mu}_{SRS}$ in (e), (f), (g), (h) are based on 10,000 initial BIBSD and SRS for the hypothetical Population 2

Although the HT variance estimator in (11) is unbiased, it is unstable in the sense that BIBSDs can produce negative variance estimates. Table 1 indicates that this was the case when large numbers of negative variance estimates were generated when sampling from Population 1. However, for Population 2, there was only one case having negative estimated variances. Negative variance estimates prevent the use of normal or t-based confidence intervals because a standard error cannot be calculated. For this situation, the smoothing method of Munholland and Borkowski (1996) was applied. For each BIBSD sample from Populations 1 and 2, the smoothing method was applied to generate a smoothed region from which the estimated variances (and standard errors) were calculated from 5,000 BIBSD resamples taken from the smoothed region. The resulting empirical sampling distributions of $\hat{\mathbf{V}}_{sm}(\hat{\mu}_{BIB})$ are displayed in Figure 7. Finally, t-based confidence intervals were generated.

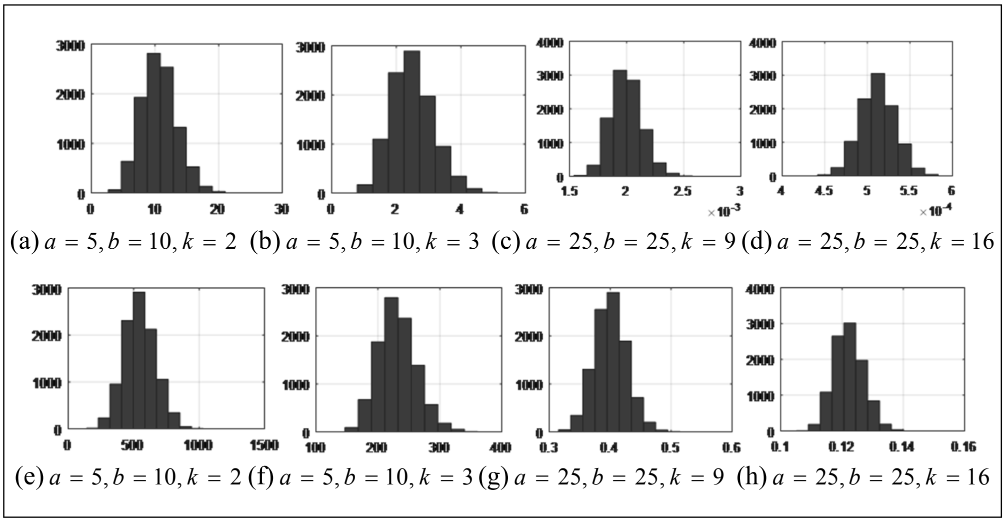


Figure 7 The empirical sampling distributions of $\hat{V}_{sm}(\hat{\mu}_{BIB})$ was generated from each of the smoothed regions associated with 5,000 BIBSDs selected from hypothetical Population 1 shown in (a), (b), (c), (d) and from hypothetical Population 2 shown in (e), (f), (g), (h)

Table 2 summarizes the coverage percentages for approximate 95% confidence intervals and the average confidence interval widths when sampling from Populations 1 and 2. For the spatial minimum trend in Population 1, the BIBSD coverage rates are all close to 100% for the smallest population size $N = 5 \times 10 = 50$ while the rates are closer to the nominal 95% for the largest population size $N = 25 \times 25 = 625$. For the spatial saddle point trend in Population 2, the BIBSD coverage rates are all close to 98% for the smallest population size $N = 50$ while the rates are closer to the nominal 95% for the largest population size $N = 625$. For both populations, the SRS coverage rates are relatively close to the nominal 95%.

Next, a real population having five subpopulations exhibiting spatial patterns will be studied. The Lansing Woods population has multiple species of trees in the 924 ft×924 ft study plot (Gerrard 1969). These species include hickories, maples, red oaks, white oaks, and black oaks. Figure 6 contains scatterplots of the tree locations of these five species in the Lansing dataset.

Because of the varying tree abundances and strengths of spatial correlations exhibited in Figure 8, BIBSDs and SRSs will be applied for the multivariate response of five tree counts per quadrat. Thus, the average estimated variances of trees per quadrat will vary across the tree species. The estimation goal is to compare the estimated variance of trees per quadrat for the different tree species.

Table 2 A comparison of the coverage rates and average widths of confidence intervals width for BIBSDs and SRSs. The smoothing method was used with BIBSDs

Population 1: Spatial Type 1: a population with a center minimum trend

					BIBSD		SRS	
N	n	a	b	k	Cover	Width	Cover	Width
50	20	5	10	2	100%	15.82	95.12%	30.05
50	30	5	10	3	99.98%	6.63	94.88%	19.65
625	225	25	25	9	97.90%	0.18	94.86%	0.73
625	400	25	25	16	95.35%	0.09	95.01%	0.41

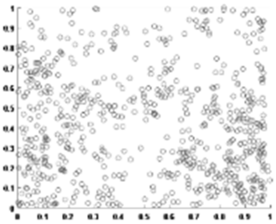
Population 2: Spatial Type 2: a population with a saddle point trend

					BIBSD		SRS	
N	n	a	b	k	Cover	Width	Cover	Width
50	20	5	10	2	97.82%	113.70	94.08%	141.95
50	30	5	10	3	97.38%	64.32	95.19%	92.90
625	225	25	25	9	95.55%	2.50	94.29%	4.23
625	400	25	25	16	95.08%	1.37	94.59%	2.38

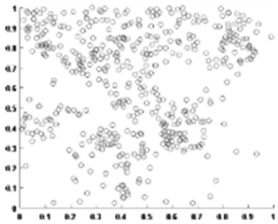
The average estimated variances of BIBSDs are all smaller than those for SRSs. The average estimated BIBSD and SRS variances, however, become closer for the larger population and sample sizes. The closeness of BIBSD and SRS average variance estimates also depends on the strength of the spatial patterns. In Figure 8, red oak and white oak (and to a lesser degree, black oak) tree locations appear randomly scattered. Thus, there is little gain in efficiency using BIBSDs resulting in the average estimated variances of BIBSD to be very close to those of SRS. This is seen in the closeness of the estimates for the three oaks in Figure 9. However, the hickory and maple locations have regions of high and low density, thus the presence of stronger spatial patterns. Thus, the estimated variances of BIBSDs were smaller than those of SRSs for the hickory and maple tree species. This is seen by the separation of curves in Figure 9 for hickories and maples.

6. Discussion

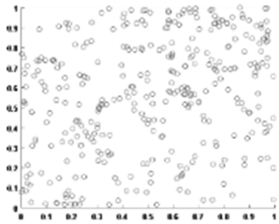
The examples presented in this research show that BIBSD-based designs can provide precise estimates of the population mean or total when strong spatial patterns or trends exist in the population. The difference in variance depends on both type and strength of the spatial pattern. That is, for the population with the very strong center minimum spatial trend, the BIBSD estimators have much smaller variances than those of the SRS estimators. There are also clear differences in variances for population with the strong saddle point trend, with the BIBSD estimators again having smaller variances than the SRS estimators. However, those hypothetical populations having the strong spatial patterns can also lead to negative estimated variances which will prevent generation of the classical t-based confidence intervals. The negative variance estimation problem was addressed by using an alternate method for generating confidence intervals. For both hypothetical spatial types, the BIBSD confidence interval coverage rates are greater than SRS coverage rates while the average interval widths are narrower than those for SRS. This indicates the BIBSDs have good estimation properties due to their better coverage of the study region in comparison to simple random sampling. For the Lansing dataset in Figure 8, the cases in which the tree locations appear to be non-random (hickories and maples), the BIBSD estimators have much smaller variances compared to the SRS estimators. However, when the tree locations are randomly dispersed throughout the study region (such as red oak, white oak and black oak), then there is only a slight difference in estimator variances between BIBSD-based designs and those using simple random sampling.



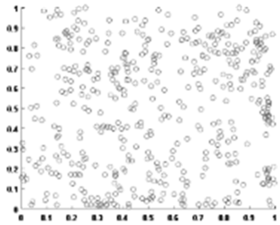
(a) Hickory Locations



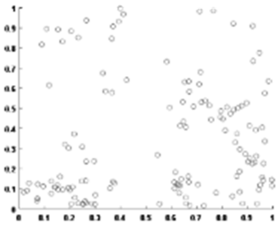
(b) Maple Locations



(c) Red oak Locations



(d) White oak Locations

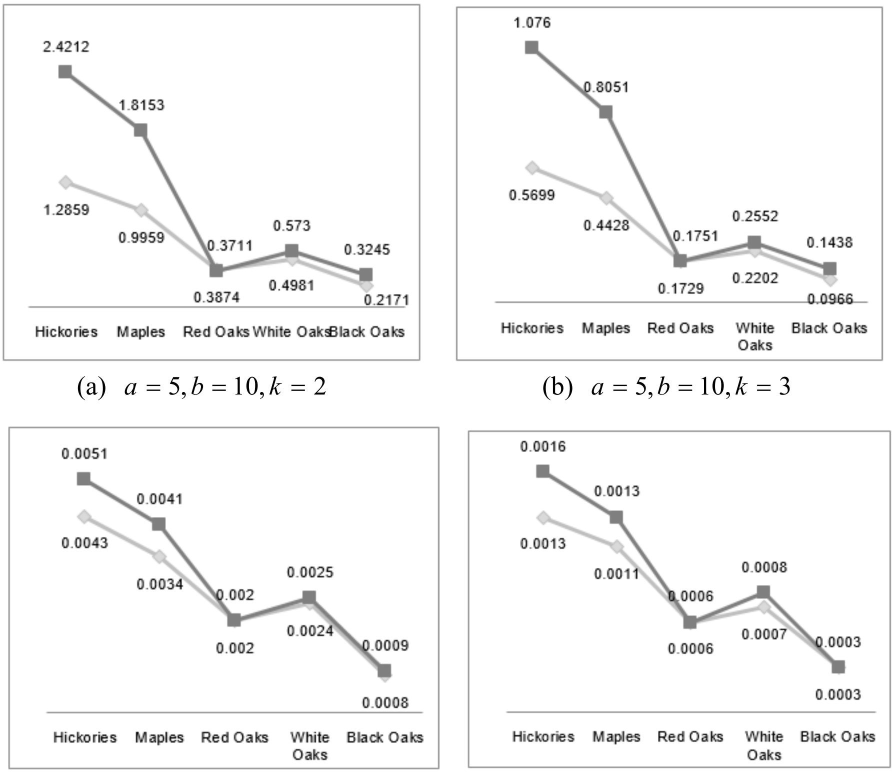


(e) Black oak Locations

Figure 8 A 924 ft × 924 ft plot study area in Lansing Woods, Clinton Country, Michigan USA. The dataset contains the locations of 2251 trees and their botanical classifications into hickories, maples, red oaks, white oaks and black oaks trees. (Gerrard 1969)

The new BIBSD-based sampling designs ensure that the sample is well-distributed over the study region. Therefore, the samples collected from BIBSD-based designs will provide greater coverage of the study region than samples of the same size collected using designs based on simple random sampling, and they will have estimators that are more efficient than estimators obtained from simple random sampling designs. This happens because the BIBSD construction does not permit clustering of sampling units in sub-regions as is possible using SRS which can contain subsets of similar neighboring units and leave large sub-regions of the study region unsampled. Thus, local clusters of units containing similar small responses or similar large response is more likely to occur using SRS. The BIBSD-based designs also ensure unbiased estimation using the Horvitz-Thompson estimators of the mean and its variance.

Balanced incomplete block sampling designs (BIBSDs) can be applied to any study region that can partitioned into a rectangular grid of a columns and b rows of quadrats. That is, the BIBSD has the flexibility in sample size for any study region that can be partitioned into a $a \times b$ grid of quadrats with corresponding BIBED parameters a, b, k, r and λ for any existing BIBED.



(a) $a = 5, b = 10, k = 2$ (b) $a = 5, b = 10, k = 3$

(c) $a = 25, b = 25, k = 9$ (d) $a = 25, b = 25, k = 16$

Figure 9 A comparison of the average estimated variances of the mean estimators using BIBSDs and SRSs across four $a \times b$ grid sizes for the five tree species in the Lansing Woods study area. (■ SRS and ◆ BIBSD)

Acknowledgements

The research of Dr. Srisuradetchai is currently supported by the Thammasat University Research Unit in Theoretical and Computational Statistics. We gratefully acknowledge financial support from Thammasat University. Also, Paramaporn Sangpara would like to thank the Department of Mathematical Sciences, Montana State University–Bozeman and the Ministry of Science and Technology (MOST) for making a research visit possible. The authors are grateful to two referees for their helpful suggestions.

References

Borkowski JJ. Simple Latin square sampling $\pm k$ designs. Commu Stat - Theory Methods. 2003; 32(1): 215-237.

Cochran WG. Sampling techniques. New York: John Wiley & Sons; 1977.

Dey A. Incomplete block designs. New Jersey: World Scientific Publishing; 2010.

Gerrard DJ. Competition quotient: a new measure of the competition affecting individual forest trees. Research Bulletin 20. Agricultural Experiment Station. Michigan: Michigan State University; 1969.

Hedayat AS, Sinha BK. Design and inference in finite population sampling. New York: John Wiley & Sons; 1991.

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

John PWM. Incomplete block designs. New York: Dekker; 1980.

Lohr S L. Sampling: design and analysis. Boston: Brooks/Cole, Cengage Learning; 2010.

Munholland P L, Borkowski J J. Simple Latin square sampling +1: A spatial design using quadrats. Biometrics. 1996; 52: 125-136.

Raghavarao D, Padgett LV. Block designs: analysis combinatorics and applications. Series on Applied Mathematics, 17. World Scientific Publishing, Hackensack, NJ; 2005.

Rathbun SL, Cressie N. A space-time survival point process for a longleaf pine forest in southern Georgia. J Am Stat Assoc. 1994; 89: 1164-1174.

Stinson DR. Combinatorial designs constructions and analysis. New York: Springer; 2004.

Street AP, Street DJ. Combinatorics of experimental design. Oxford: Oxford Science; 1987.

Thompson S K. Sampling. New York: John Wiley & Sons; 2012.