Thailand Statistician
July 2022; 20(3): 562-574
http://statassoc.or.th
Contributed paper

Stratified Path Sampling

Mena Lao*
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand.
*Corresponding author; e-mail: mena_patummasut@yahoo.com

Received: 15 July 2020
Revised: 8 April 2021
Accepted: 24 August 2021

Abstract

This paper proposes the employment of path sampling when the population is stratified into smaller regions, called stratified path sampling, resulting in more precision estimation. An estimator of the population mean and its variance including the variance estimator are provided. Simulation study was performed to examine the efficiency of stratified path sampling compared to stratified random sampling by using real-world population data. The benefit and disadvantage of stratified path sampling are described. From the simulation result, stratified path sampling is less efficient in non-rare population with low variation of study variable values, but it is more efficient in rare population and non-rare population with high variation of study variable values, particularly when the starting and ending units are on the columns containing large values of study variable.

Keywords: Stratified sampling, path sampling, unequal probability sampling, Horvitz-Thompson estimator.

1. Introduction

Path sampling is a sampling design implemented in a spatial population, and it enables researcher to sample all units in a path, created by researcher, while sampling. Thus, it is an economical or cost-effective sampling design when the sampling cost is considered as the number of units travelled to observe all units in the sample (Patummasut and Dryver 2012). In path sampling, a path or paths are selected by simple random sampling without replacement from all paths in the population. To observe all units in the sample, a researcher can just travel along a path, which is time-saving and practical. On the other hand, if a researcher uses simple random sampling, the sample may consist of units aggregating all over the population, which makes sampling time consuming and difficult to observe all units in the sample. For the same sample size, path sampling has less cost of travelling for observing all units in the sample than simple random sampling (Patummasut and Nidsunkid 2021).

Stratified sampling usually provides more precise estimates for population total and mean. A population can be stratified into subgroups, called strata, according to the adjacency of units if a geographical area is considered (Lohr 2010). For a large population region, it can be stratified into smaller regions to make sure that the sample is dispersed all over the region, leading to better estimates (Thompson 2012). There are many stratified sampling designs developed, for example, stratified random sampling, stratified adaptive cluster sampling (Thompson 1991), stratified capture-recapture sampling (Plante et al. 1998), stratified adaptive cluster sampling with spatially clustered secondary

units (Patummasut and Borkowski 2014). Therefore, this paper proposes stratified path sampling, when the population is stratified into smaller areas, potentially leading to more precision in estimation.

2. Path Sampling

Consider a rectangular population area divided into W rows and L columns resulting in the population size of WL units as shown in Figure 1. The symbol (w,l) represents a unit in row w and column l for $w=1,2,3,...,W$ and $l=1,2,3,...,L.$ Let $x_{(w,l)}$ denote the value of the study variable of unit $(w,l),$ and the population total and population mean are respectively

$$T=\sum_{w=1}^{W}\sum_{l=1}^{L}x_{(w,l)} \text{ and } \mu=T/WL=\frac{1}{WL}\sum_{w=1}^{W}\sum_{l=1}^{L}x_{(w,l)}. \quad (1)$$

A path sample is a sample in which a path or paths are selected by simple random sampling without replacement from all paths in the population. The idea of creating a path is traveling from a starting unit and returning to such starting unit, which leads to convenience and saved cost when observing sampled units (Patummasut and Dryver 2012).

Let the starting unit be $(1,ls)$ a unit on the edge of a region in row 1 and column ls because it is practical and less expensive to start sampling at the edge of a region than starting inside or in the middle of a region. From that starting unit, path i is constructed to go down column ls to rows i and $i+1,$ then return the starting point via column $ls+1,$ and finally ending at is a unit $(1,ls+1)$ as shown in Figure 1. The ordered units in path i from the starting unit to ending units are $(1,ls),(2,ls),(3,ls),$ $\ldots,(i,ls),(i,ls-1),(i,ls-2),\ldots,(i,1),(i+1,1),(i+1,2),\ldots,(i+1,L),(i,L),(i,L-1),(i,L-2),\ldots,$ $(i,ls+1),(i-1,ls+1),(i-2,ls+1),\ldots,(1,ls+1).$ Consequently, there are $W-1$ possible paths in the set of W rows, and all paths have overlapping units in the starting column ls and ending column $ls+1.$ Path $i-1$ and path i overlap in row i.

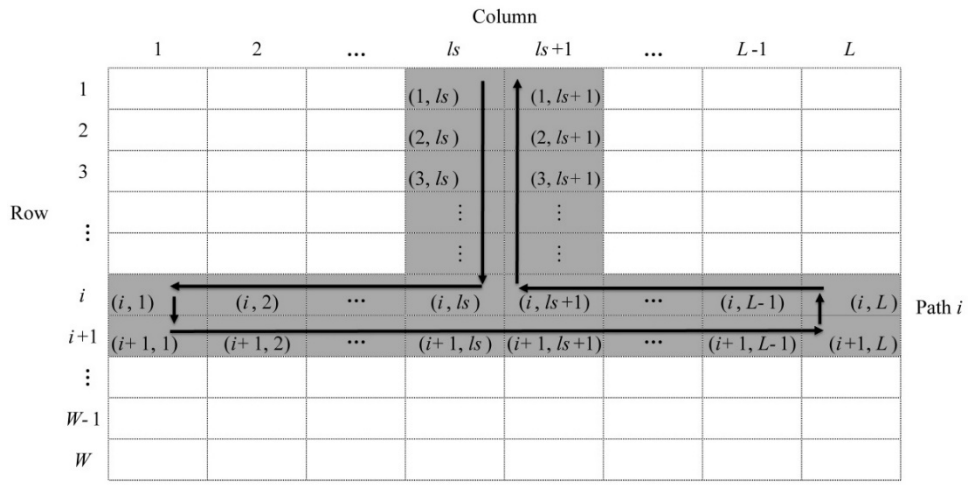


Figure 1 Path i with a starting unit $(1,\ ls)$ and ending unit $(1,\ ls+1)$ in the population partitioned into W rows and L columns

In path sampling, p paths are selected by simple random sampling without replacement from $W-1$ paths in the population. Let a_i denote a path i in the sample for $i=1,2,3,...,p$ and the sample can be denoted as $s_0=(a_1,a_2,a_3,...,a_p)$, and let s be the set of distinct units in the sample s_0. An unbiased estimator of the population total, which is the Horvitz-Thompson estimator is

$$\hat{T}=\sum_{(w,l)\in s}\frac{x_{(w,l)}}{\alpha_{(w,l)}}, \tag{2}$$

where $\alpha_{(w,l)}$ is the probability that unit (w,l) is included in a sample, known as an inclusion probability. Because of the overlapping of paths, units have unequal inclusion probabilities, which can be calculated from the formula:

$$\alpha_{(w,l)}=\begin{cases}1-\binom{w-2}{p}\Big/\binom{W-1}{p} & \text{if } (w,\ l)\in V_1,\\ 1-\binom{w-3}{p}\Big/\binom{W-1}{p} & \text{if } (w,\ l)\in V_2,\\ 1-\binom{w-2}{p}\Big/\binom{W-1}{p} & \text{if } (w,\ l)\in V_3,\end{cases} \tag{3}$$

where $V_1=\{\text{unit } (w,\ l)\mid w=1,2,3,\ldots,W \text{ and } l=ls \text{ and } ls+1\}$,

$V_2=\{\text{unit } (w,\ l)\mid w=2,3,\ldots,W-1 \text{ and } l=1,2,3,\ \ldots,\ ls-1,\ ls+2,\ ls+3,\ \ldots,\ L\}$,

$V_3=\{\text{unit } (w,l)\mid w=1,\ W \text{ and } l=1,2,3,\ldots,ls-1,ls+2,ls+3,\ldots,L\}$.

Let $\alpha_{(w,l),(w',l')}$ be the probability that both units (w,l) and unit (w',l') are included in the sample, known as the joint inclusion probability, and it is calculated by the formula:

$$\alpha_{(w,l),(w',l')}=\alpha_{(w,l)}+\alpha_{(w',l')}-\left[1-\binom{b}{p}\Big/\binom{W-1}{p}\right], \tag{4}$$

where b is the number of paths not having either units (w,l) or (w',l'). The variance of $\hat{T}_p$ is

$$v(\hat{T})=\sum_{w=1}^{W}\sum_{l=1}^{L}\left(\frac{1-\alpha_{(w,l)}}{\alpha_{(w,l)}}\right)x_{(w,l)}^2+\sum_{w=1}^{W}\sum_{l=1}^{L}\sum_{\substack{w'=1\\(w,l)\neq(w',l')}}^{W}\sum_{l'=1}^{L}\left(\frac{\alpha_{(w,l),(w',l')}-\alpha_{(w,l)}\alpha_{(w',l')}}{\alpha_{(w,l)}\alpha_{(w',l')}}\right)x_{(w,l)}x_{(w',l')}. \tag{5}$$

The estimator of $v(\hat{T}_p)$ is

$$\hat{v}(\hat{T})=\sum_{(w,l)\in s}\left(\frac{1}{\alpha_{(w,l)}^2}-\frac{1}{\alpha_{(w,l)}}\right)x_{(w,l)}^2+\sum_{\substack{(w,l)\ and\ (w',l')\in s\\(w,l)\neq(w',l')}}\left(\frac{1}{\alpha_{(w,l)}\alpha_{(w',l')}}-\frac{1}{\alpha_{(w,l),(w',l')}}\right)x_{(w,l)}x_{(w',l')}. \tag{6}$$

Hence, an unbiased estimator of the population mean is $\hat{\mu}=\hat{T}/WL$. The variance of $\hat{\mu}$ is $v(\hat{\mu})=v(\hat{T})/(WL)^2$ and the estimator of this variance is $\hat{v}(\hat{\mu})=\hat{v}(\hat{T})/(WL)^2$.

3. Stratified Path Sampling

Suppose a population region of W rows and L columns, which the population size is WL units, is partitioned into H strata. Let N_h be the number of units in the stratum h for $h=1,2,3,...,H$ where

$\sum_{h=1}^{H} N_h = WL.$ Stratum h contains a rectangle of adjacent W_h rows and L_h columns, thus $N_h = W_h L_h$, $\sum_{h=1}^{H} W_h = W$ and $\sum_{h=1}^{H} L_h = L.$ Let $x(h)_{(w,l)}$ denote the value of the study variable of unit (w,l) in stratum $h.$ The population total of stratum h is $T(h) = \sum_{w=1}^{W_h} \sum_{l=1}^{L_h} x(h)_{(w,l)}$, so the population total is

$$T = \sum_{h=1}^{H} T(h) = \sum_{h=1}^{H} \sum_{w=1}^{W_h} \sum_{l=1}^{L_h} x(h)_{(w,l)}. \tag{7}$$

For stratum $h,$ let the starting unit be a unit on the edge of the stratum in row 1 and column ls_h, which is unit $(1, ls_h),$ because it is more practical and less expensive to start sampling at the edge of the stratum than starting inside or in the middle of the stratum. The ordered units in path i from the starting unit to ending units are $(1,ls_h),(2,ls_h),(3,ls_h),\ldots,(i,ls_h),(i,ls_h-1),(i,ls_h-2),\ldots,(i,1),(i+1,1),$ $(i+1,2),\ldots,(i+1,L_h),(i,L_h),(i,L_h-1),(i,L_h-2),\ldots,(i,\ ls_h+1),\ (i-1,\ ls_h+1),(i-2,ls_h+1),...,(1,\ ls_h+1)$ as shown in Figure 2. There are all possible $W_h - 1$ paths, and all paths overlap in the starting column ls_h and ending column $ls_h + 1.$ Path $i-1$ and path i overlap in row $i.$

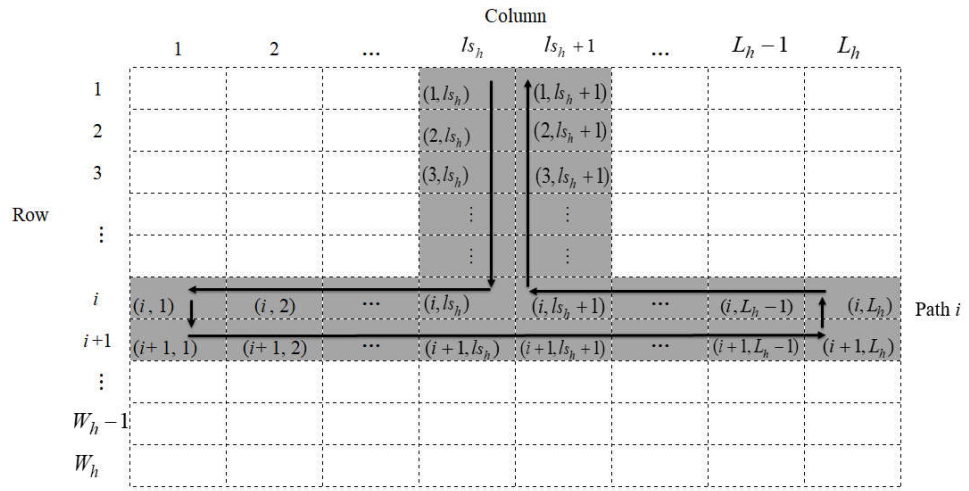


Figure 2 Path i with a starting unit $(1,ls_h)$ and ending unit $(1,ls_h+1)$ in stratum h containing W_h rows and L_h columns

Stratified path sampling design is a sampling design in which path sampling is used in each stratum, that is, within stratum $h,$ p_h paths are selected by simple random sampling without replacement from $W_h - 1$ paths. Let $a_i(h)$ denotes a path i in the sample for $i = 1,2,3,...,p_h,$ and the sample of paths can be denoted as $s_0(h) = (a_1(h), a_2(h), a_3(h), ..., a_p(h)),$ and let $s(h)$ be the set of distinct units in the sample $s_0(h).$ By Equation (2), an unbiased estimator of $T(h)$ is

$$\hat{T}(h)=\sum_{(w,l)\in s(h)}\frac{x(h)_{(w,l)}}{\alpha(h)_{(w,l)}}, \tag{8}$$

where $\alpha(h)_{(w,l)}$ is the probability that unit (w,l) in stratum h is included in a sample $s(h)$, which can be calculated from the formula

$$\alpha(h)_{(w,l)}=\begin{cases}1-\binom{w-2}{p_h}\Big/\binom{W_h-1}{p_h} & \text{if } (w,\ l)\in V_1(h),\\ 1-\binom{w-3}{p_h}\Big/\binom{W_h-1}{p_h} & \text{if } (w,\ l)\in V_2(h),\\ 1-\binom{w-2}{p_h}\Big/\binom{W_h-1}{p_h} & \text{if } (w,\ l)\in V_3(h),\end{cases} \tag{9}$$

where $V_1(h)=\{\text{unit }(w,l)\mid w=1,2,3,\ldots,W_h \text{ and } l=ls_h \text{ and } ls_h+1\}$,

$V_2(h)=\{\text{unit }(w,l)\mid w=2,3,\ldots,W_h-1 \text{ and } l=1,2,3,\ \ldots,\ ls_h-1, ls_h+2, ls_h+3,\ \ldots, L_h\}$

$V_3(h)=\{\text{unit }(w,l)\mid w=1, W_h \text{ and } l=1,2,3,\ldots,ls_h-1, ls_h+2, ls_h+3,\ldots, L_h\}$

By Equation (5), the variance of $\hat{T}(h)$ is

$$v(\hat{T}(h))=\sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\left(\frac{1-\alpha(h)_{(w,l)}}{\alpha(h)_{(w,l)}}\right)x_{(w,l)}^2+\sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\sum_{\substack{w'=1\\(w,l)\neq(w',l')}}^{W_h}\sum_{l'=1}^{L_h}\left(\frac{\alpha(h)_{(w,l),(w',l')}-\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}{\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}\right)x(h)_{(w,l)}x(h)_{(w',l')}, \tag{10}$$

where $\alpha(h)_{(w,l),(w',l')}$ is the probability that both unit (w,l) and unit (w',l') are included in the sample, and it is also called the joint inclusion probability of unit $(w,\ l)$ and unit (w',l'). By Equation (4), it is calculated with the formula

$$\alpha(h)_{(i,j),(w',l')}=\alpha(h)_{(w,l)}+\alpha(h)_{(w',l')}-\left[1-\binom{b_h}{p_h}\Big/\binom{W_h-1}{p_h}\right], \tag{11}$$

where b_h is the number of paths not having either units (w,l) or (w',l') in stratum h. By Equation (6), an unbiased estimator of $v(\hat{T}(h))$ is

$$\hat{v}(\hat{T}(h))=\sum_{(w,l)\in s(h)}\left(\frac{1}{\alpha(h)_{(w,l)}^2}-\frac{1}{\alpha(h)_{(w,l)}}\right)x_{(w,l)}^2+\sum_{\substack{(w,l)\ and\ (w',l')\in s(h)\\(w,l)\neq(w',l')}}\left(\frac{1}{\alpha(h)_{(w,l)}\alpha_{(w',l')}}-\frac{1}{\alpha(h)_{(w,l),(w',l')}}\right)x(h)_{(w,l)}x(h)_{(w',l')}. \tag{12}$$

Then, an unbiased estimator of the overall population total in stratified path sampling is the sum of the estimators from all strata (Thompson 2012)

$$\hat{T}_{sps}=\sum_{h=1}^{H}\hat{T}(h)=\sum_{h=1}^{H}\sum_{(w,l)\in s(h)}\frac{x(h)_{(w,l)}}{\alpha(h)_{(w,l)}}. \tag{13}$$

The selections in different strata are independent, thus the variance of $\hat{T}_{sps}$ is the sum of the variances from all strata (Thompson 2012)

$$v(\hat{T}_{sps}) = \sum_{h=1}^{H} v(\hat{T}(h)). \tag{14}$$

An unbiased estimator of that variance is the sum of individual stratum estimators

$$\hat{v}(\hat{T}_{sps}) = \sum_{h=1}^{H} \hat{v}(\hat{T}(h)). \tag{15}$$

Hence, the stratified estimator for the population mean is $\hat{\mu}_{sps} = \hat{T}_{sps} / WL$ with the variance $v(\hat{\mu}_{sps}) = v(\hat{T}_{sps})/(WL)^2$. An unbiased estimator of that variance is $\hat{v}(\hat{\mu}_{sps}) = \hat{v}(\hat{T}_{sps})/(WL)^2$.

4. Simulation Study

To examine the performance of stratified path sampling compared to stratified simple random sampling based on the estimated variance of the mean estimator, rare and non-rare populations are used in a simulation study using R statistical program.

4.1. Simulation study on rare population

Ring-necked ducks data in a 50×100 square kilometers area in central Florida (Smith et al. 1995) partitioned into W =5 rows and L =10 columns forming WL =50 units is considered in the simulation study of rare population. The study variable x is the number of the ring-necked ducks in a unit. Notice that most units have zero ring-necked ducks, so it is a rare population. This population is stratified into H =2 strata; each stratum has W_h =5 rows and L_h =5 columns as shown in Figure 3. The number of all possible paths in each stratum is $W_h - 1$ =4. In stratified path sampling, path sample is selected from each stratum with a given starting unit. Let the starting unit be a unit on the edge of the stratum in row 1 and column ls_h, which is unit $(1, ls_h)$, because it is more practical and less expensive than starting inside or in the middle of the stratum. Note that if a starting and ending units are on high x-value columns, path sampling is more efficient. Thus, if possible, a starting and ending units are set be on high x-value columns (Patummasut and Dryver 2012; Patummasut and Nidsunkid 2021). Here, for Ring-necked ducks population, column 1 and 2 of stratum 1 contain high x-values, and column 3, 4 and 5 of stratum 2 contain high x-values. Thus, in this simulation, the starting unit in stratum 1 and 2 are set to be unit $(1, ls_1)$ =(1,1) and unit $(1, ls_2)$ =(1,4), respectively. In addition, to investigate the performance of stratified path sampling, another 2 pairs of starting units in stratum 1 and 2, which are unit (1,1) and (1,3), and unit (1,2) and (1,3), are used as shown in Table 1.

Under stratified path sampling, the simulation is conducted for 2,000 iterations for each value of p and the given starting unit. For each iteration, a stratified path sample was selected, and the mean estimate $\hat{\mu}_{sps}$ and the final sample size were obtained. The estimated variance $\hat{v}(\hat{\mu}_{sps})$ is estimated from the sample variance of the estimates

$$\hat{v}(\hat{\mu}_{sps}) = \frac{1}{1,999} \sum_{i=1}^{2,000} (\hat{\mu}_{sps(i)} - \bar{\mu}_{sps})^2, \tag{16}$$

where $\bar{\mu}_{sps}$ is the sample mean of the 2,000 $\hat{\mu}_{sps(i)}$ values (Dryver and Thompson 2005).

For efficiency comparison purposes, the sample size of stratified simple random sampling is set equal to the expected final sample size of stratified path sampling, denoted by υ, which is the sum of

inclusion probabilities. For each iteration, a stratified simple random sample was selected, and the estimator of the population mean, $\hat{\mu}_{ssrs} = \sum_{h=1}^{L} N_h \bar{x}_h \Big/ WL$ where $\bar{x}_h$ is a sample mean in stratum h, was calculated. Then, the average variance $\hat{v}(\hat{\mu}_{ssrs})$ is calculated from the mean of the sample variances of the estimates

$$\hat{v}(\hat{\mu}_{ssrs}) = \frac{1}{1,999} \sum_{i=1}^{2,000} (\hat{\mu}_{srss(i)} - \bar{\mu}_{ssrs})^2, \tag{17}$$

where $\bar{\mu}_{ssrs}$ is the sample mean of the 2,000 $\hat{\mu}_{ssrs(i)}$ values (Dryver and Thompson 2005). Finally, the relative efficiency of stratified path sampling compared to stratified simple random sampling is calculated by the formula

$$re = \frac{\hat{v}(\hat{\mu}_{ssrs})}{\hat{v}(\hat{\mu}_{sps})}. \tag{18}$$

		Column									
		1	2	3	4	5	1	2	3	4	5
Row	1	0	200	200	75	0	0	0	0	675	0
	2	4000	13500	234	1335	4	0	0	35	0	55
	3	0	0	0	0	97	0	4	0	1815	0
	4	0	0	0	0	0	0	0	0	0	0
	5	0	0	1	0	0	0	0	0	0	1283
		Stratum 1					Stratum 2				

Figure 3 Ring-necked ducks data in the population region divided into 2 strata, 25 units each

The simulation results for the ring-necked ducks population presented in Table 1 shows that stratified path sampling is more efficient than stratified simple random sampling for all pairs of starting units because the relative efficiencies are greater than 1.

It can be seen that stratified path sampling with the starting unit (1,1) in stratum 1 and unit (1, 4) in stratum 2 is the most efficient because it has the largest relative efficiency for all p. Notice that with these starting units, the starting and ending columns have the largest values of study variable compared to other columns

Similarly, stratified path sampling with the starting unit (1,1) in stratum 1 and unit (1, 3) in stratum 2 has the larger relative efficiencies than stratified path sampling with the starting unit (1,2) in stratum 1 and unit (1, 3) in stratum 2. Notice that the starting and ending columns with the former starting units have the larger values of study variable than those with the later starting units.

It can be concluded that the larger values in starting and ending columns, the more efficient of stratified path sampling for rare population. Moreover, there is high variation among values of the study variable in this population, making stratified simple random sampling less efficient than stratified path sampling.

Table 1 Estimated variances and relative efficiencies of stratified path sampling for different starting units and stratified simple random sampling from the simulation study on ring-necked ducks data population p

p	υ	$\hat{v}(\hat{\mu}_{sps})$			$\hat{v}(\hat{\mu}_{ssrs})$	re		
		$(1,ls_1)=(1,2)$ $(1,ls_2)=(1,3)$	$(1,ls_1)=(1,1)$ $(1,ls_2)=(1,3)$	$(1,ls_1)=(1,1)$ $(1,ls_2)=(1,4)$		$(1,ls_1)=(1,2)$ $(1,ls_2)=(1,3)$	$(1,ls_1)=(1,1)$ $(1,ls_2)=(1,3)$	$(1,ls_1)=(1,1)$ $(1,ls_2)=(1,4)$
1	26	14,540.62	4,343.61	1,423.28	73,312.72	5.04	16.88	51.51
2	38.34	2,667.06	899.16	295.82	23,412.95	8.78	26.04	79.15
3	46	225.36	207.34	10.26	7,701.50	34.17	37.15	750.96

Note: $(1,ls_1)$ and $(1,ls_2)$ are the starting unit in stratum 1 and 2.

4.2. Simulation study on non-rare population

4.2.1. Kimboto trees data

For a non-rare population, Kimboto trees data in a region 400 by 525 meters at Paracou in French Guiana (Flores 2005) is considered, and it is divided into W =20 rows and L =20 columns forming WL =400 units as shown in Figure 4. The study variable x is the number of Kimboto trees in a unit. This population is partitioned into H =2 strata, each with $W_h = 10$ rows and L_h =20 columns as shown in Figure 4. The number of all possible paths in each stratum is $W_h - 1 = 9$. In stratified path sampling, the starting units in stratum 1 and 2 are set as shown in Table 2. The simulation procedure is the same as the simulation of the ring-necked ducks data in previously described.

The simulation results from the Kimboto trees population presented in Table 2 shows that stratified path sampling is less efficient than stratified simple random sampling for all pairs of starting units and for every p because the values of relative efficiencies are less than 1.

Row	Column 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	7	2	5	5	4	1	0	1	1	1	1	1	0	2	1	2	2	0	2	1	
2	1	0	0	3	3	8	1	2	2	3	1	0	1	3	0	1	3	0	3	0	
3	1	3	0	5	1	0	4	0	2	1	6	2	1	0	5	1	2	1	1	0	
4	2	1	3	2	1	0	4	1	2	2	8	4	0	0	1	0	3	0	1	1	
5	2	2	0	0	0	3	1	4	2	2	5	3	0	0	1	1	2	1	0	2	Stratum 1
6	1	1	3	5	0	0	0	0	0	2	10	0	2	3	1	2	3	2	0	3	
7	0	0	1	1	3	4	3	1	0	4	3	3	3	5	7	0	1	0	2	2	
8	2	3	3	1	4	7	1	1	0	1	0	2	2	7	1	1	1	0	0	4	
9	3	11	3	3	5	4	4	2	2	1	0	1	3	3	2	1	0	0	0	4	
10	0	4	3	1	2	3	2	4	1	1	0	2	1	3	3	8	1	1	4	4	
1	3	1	5	2	0	0	2	6	5	5	2	1	1	6	7	1	3	1	5	12	
2	2	4	7	3	0	1	1	1	3	4	3	3	0	1	3	1	0	2	2	3	
3	3	5	4	4	3	3	0	1	2	6	8	2	0	3	6	3	1	1	1	6	
4	3	0	2	4	2	1	7	1	3	8	2	1	0	7	4	3	2	0	0	2	
5	1	1	4	2	3	3	1	4	2	5	8	2	1	3	14	10	5	1	8	2	Stratum 2
6	1	1	3	1	3	2	1	1	4	18	4	0	1	0	0	4	0	1	4	1	
7	1	0	4	0	0	1	2	1	3	15	1	2	2	4	1	0	2	1	1	1	
8	2	1	1	0	0	1	2	3	2	2	2	1	4	1	1	2	1	0	2	0	
9	0	1	1	3	2	1	0	2	5	2	0	4	0	1	0	0	2	0	1	0	
10	2	3	1	0	0	1	1	3	0	1	2	4	3	1	3	2	1	2	0	4	

Figure 4 Kimboto trees data in the population region divided into 2 strata, 200 units each

Table 2 Estimated variances and relative efficiencies of stratified path sampling and stratified simple random sampling from the simulation study on the Kimboto trees data population

		$\hat{v}(\hat{\mu}_{sps})$				re		
p	υ	$ls_1:(1,3)$ $ls_2:(1,1)$	$ls_1:(1,5)$ $ls_2:(1,5)$	$ls_1:(1,10)$ $ls_2:(1,10)$	$\hat{v}(\hat{\mu}_{ssrs})$	$ls_1:(1,3)$ $ls_2:(1,1)$	$ls_1:(1,5)$ $ls_2:(1,5)$	$ls_1:(1,10)$ $ls_2:(1,10)$
1	96	0.3051	0.2208	0.2782	0.0620	0.20	0.28	0.22
2	166.67	0.1369	0.0984	0.1343	0.0314	0.23	0.32	0.23
3	226	0.0898	0.0677	0.0787	0.0228	0.25	0.34	0.29
4	276	0.0554	0.0407	0.0568	0.0144	0.26	0.35	0.25
5	317.33	0.0418	0.0320	0.0366	0.0111	0.27	0.35	0.30
6	350.29	0.0323	0.0262	0.0239	0.0091	0.28	0.35	0.38
7	375	0.0239	0.0207	0.0130	0.0070	0.29	0.34	0.54

4.2.2. Modified Kimboto trees data

Kimboto trees population data are modified by adding the larger values of the study variable in column 5 and 10 as shown in Figure 5. Using the same simulation procedure as the simulation of the original Kimboto trees including setting the same starting units, the simulation results for the modified Kimboto trees data are presented in Table 3.

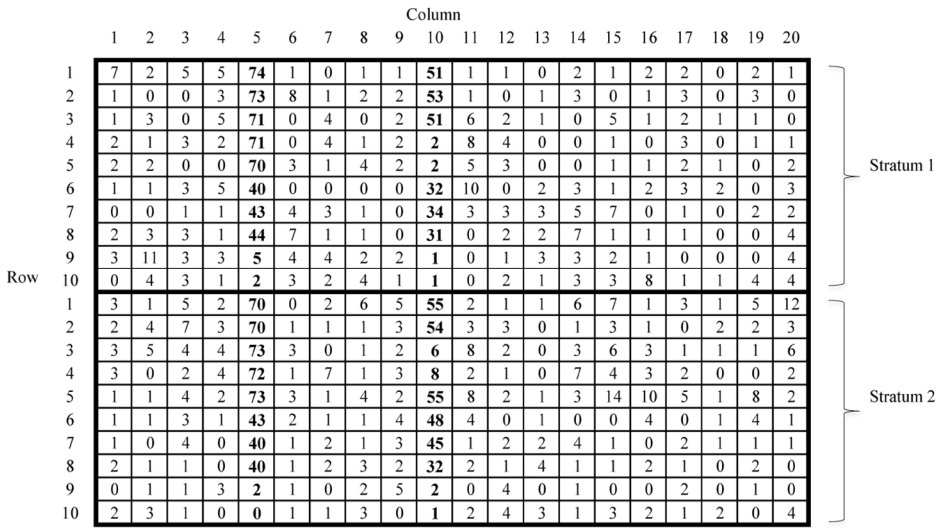


Figure 5 The modified Kimboto trees data

From Table 3, for starting units (1, 5) in stratum 1 and 2, stratified path sampling is more efficient than stratified simple random sampling for p equal to 1 to 4 due to the relative efficiencies greater than 1; however, stratified path sampling is less efficient than stratified simple random sampling for p equal to 5 to 7 because the relative efficiencies are less than 1.

Table 3 Estimated variances and relative efficiencies of stratified path sampling for different starting units and stratified simple random sampling from the simulation study on the modified Kimboto trees data population

		$\hat{v}(\hat{\mu}_{sps})$				re		
p	υ	$ls_1:(1,3)$ $ls_2:(1,1)$	$ls_1:(1,5)$ $ls_2:(1,5)$	$ls_1:(1,10)$ $ls_2:(1,10)$	$\hat{v}(\hat{\mu}_{ssrs})$	$ls_1:(1,3)$ $ls_2:(1,1)$	$ls_1:(1,5)$ $ls_2:(1,5)$	$ls_1:(1,10)$ $ls_2:(1,10)$
1	96	2.6080	0.6306	0.5960	1.5329	0.59	2.43	2.57
2	166.67	1.3127	0.4809	0.5075	0.7054	0.54	1.47	1.39
3	226	0.8949	0.3043	0.3658	0.3780	0.42	1.24	1.03
4	276	0.5335	0.1990	0.2710	0.2115	0.40	1.06	0.78
5	317.33	0.3540	0.1318	0.1922	0.1298	0.37	0.98	0.68
6	350.29	0.2371	0.0839	0.1078	0.0714	0.30	0.85	0.66
7	375	0.1176	0.0460	0.0600	0.0304	0.26	0.66	0.51

Similarly, for starting units (1,10) in stratum 1 and 2, stratified path sampling is more efficient than stratified simple random sampling for p equal to 1 to 3 because of the values of relative efficiency greater than 1, but path sampling is less efficient than stratified simple random sampling for p equal to 4 to 7 due to the relative efficiencies less than 1.

For starting units (1,3) and (1,1) in stratum 1 and 2, respectively, stratified path sampling is less efficient than stratified simple random sampling for every p because the relative efficiencies are less than 1.

Notice that, with the starting unit (1,5) and (1,10), the starting and ending columns contain large values of the study variable, and stratified path sampling is more efficient. On the other hand, with the starting units (1,3) and (1,1) in stratum 1 and 2, the starting and ending columns contain small values of the study variable, and stratified path sampling is less efficient. Therefore, it can be concluded that if starting and ending columns contain larger values of the study variable, stratified path sampling can be more efficient. In addition, there is high variation among values of the study variable in this population, thus stratified simple random sampling can be less efficient than stratified path sampling.

5. Discussion

Path sampling uses Horvitz-Thompson estimator, which is more efficient when the values of the study variable are proportional to the inclusion probabilities (Horvitz and Thompson 1952), that is, a unit with the larger (or smaller) value of the study variable has the larger (or smaller) inclusion probability. Therefore, if the units in starting and ending columns have the larger values of study variable than other units in the population, path sampling is more efficient because these units have the larger inclusion probabilities than the others as a consequence of overlapping of paths (Patummasut and Dryver 2012; Patummasut and Nidsunkid 2021). This explains why the simulation results for the Ring-necked ducks population presented in Table 1 show that stratified path sampling with the starting unit (1,1) in stratum 1 and unit (1, 4) in stratum 2 is the most efficient. With these starting units, the units in starting and ending columns of stratum 1 and stratum 2, which are the first two columns of stratum 1 and last two columns of stratum 2, have the larger values of study variable than other columns, thus stratified path sampling with these starting units has the largest relative efficiency compared to other pairs of starting units.

Likewise, the estimator of the overall population mean in stratified path sampling is the sum of Horvitz-Thompson estimator from all strata, so it is more efficient when units in the starting and ending columns in each stratum have the larger values of study variable than other units. Accordingly, in the simulation study of non-rare populations, stratified path sampling performed better on the modified Kimboto trees population than the original Kimboto trees population because the starting and ending columns of the modified Kimboto trees data contain the larger values of the study variable than the original data, which encourages more precise estimates. It can be seen that from the simulation result on the modified Kimboto trees population in Table 3, the starting units (1,5) and (1,10) are on the columns with large values of the study variable, allowing stratified path sampling to become more efficient.

For a rare population, there is high variation among values of the study variable, thus simple random sampling is not efficient and not appropriate for this type of population (Levy and Lemeshow 1999). Consequently, from the simulation study on a rare population, the Ring-necked ducks data, stratified path sampling is more efficient than stratified simple random sampling.

Hence, it can be concluded that stratified path sampling is appropriate for a rare population and for a non-rare population with high variation among values of the study variable, especially when the starting and ending columns have large values of study variable.

6. Conclusions

Stratified path sampling is proposed in this research. It is a sampling design in which a sample is selected by path sampling in each stratum when a population region is partitioned into strata. An unbiased estimator of the population mean and its variance are obtained. To examine the efficiency of stratified path sampling compared to stratified simple random sampling the simulation study was conducted on rare and non-rare populations. From the simulation study results, it is found that stratified path sampling can be efficient and appropriate for a rare population. For a non-rare population, it is efficient when there is high variation among values of the study variable and when the starting and ending columns contain large values of study variable. However, it is not efficient in a non-rare population with low variation among values of the study variable.

Appendix

1. Prove that the variance of $\hat{T}$ is

$$v(\hat{T}) = \sum_{w=1}^{W}\sum_{l=1}^{L}\left(\frac{1-\alpha_{(w,l)}}{\alpha_{(w,l)}}\right)x_{(w,l)}^2 + \sum_{w=1}^{W}\sum_{l=1}^{L}\sum_{\substack{w'=1\\(w,l)\neq(w',l')}}^{W}\sum_{l'=1}^{L}\left(\frac{\alpha_{(w,l),(w',l')}-\alpha_{(w,l)}\alpha_{(w',l')}}{\alpha_{(w,l)}\alpha_{(w',l')}}\right)x_{(w,l)}x_{(w',l')}.$$

Proof: The variance of $\hat{T}$ can be found by using the variance of the Horvitz-Thompson estimator since $\hat{T}$ is the Horvitz-Thompson estimator. Details of the Horvitz-Thompson estimator and its variance for any sampling design are explained as follows. For any sampling design, let π_i be the inclusion probability of unit i for $i=1,2,3,...,N,$ which is the probability that unit i is included in the sample, and let π_{ij} be the joint inclusion probability of unit i and unit $j,$ which is the probability that both unit i and unit j are included in the sample. An unbiased estimator of the population total τ is $\hat{\tau} = \sum_{i=1}^{\eta}\frac{x_i}{\pi_i}$ where η the number of distinct units in the sample. This estimator $\hat{\tau}$ is called the Horvitz-Thompson estimator (Horvitz and Thompson 1952). The variance of $\hat{\tau}$ is

$$v(\hat{\tau}) = \sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)x_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)x_i x_j. \quad \text{(A1)}$$

For path sampling, $\alpha_{(w,l)}$ is the inclusion probability of unit (w,l) for $w=1,2,3,...,W$ and $l=1,2,3,...,L,$ and $\alpha_{(i,j),(w',l')}$ is the joint inclusion probability of unit (w,l) and unit (w',l'). Thus, based on the Horvitz-Thompson estimator, an unbiased estimator of the population total is $\hat{T} = \sum_{(w,l)\in s}\frac{x_{(w,l)}}{\alpha_{(w,l)}}$. Since $\hat{T}$ is the Horvitz-Thompson estimator, the variance of $\hat{T}$ can be obtained by using Equation (A1). Hence, the variance of $\hat{T}$ is

$$v(\hat{T}) = \sum_{w=1}^{W}\sum_{l=1}^{L}\left(\frac{1-\alpha_{(w,l)}}{\alpha_{(w,l)}}\right)x_{(w,l)}^2 + \sum_{w=1}^{W}\sum_{l=1}^{L}\sum_{w'=1}^{W}\sum_{l'=1}^{L}\left(\frac{\alpha_{(w,l),(w',l')}-\alpha_{(w,l)}\alpha_{(w',l')}}{\alpha_{(w,l)}\alpha_{(w',l')}}\right)x_{(w,l)}x_{(w',l')}.$$
$$(w,l)\neq(w',l')$$

2. Prove that the variance of $\hat{T}(h)$ is

$$v(\hat{T}(h)) = \sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\left(\frac{1-\alpha(h)_{(w,l)}}{\alpha(h)_{(w,l)}}\right)x_{(w,l)}^2 +$$
$$\sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\sum_{w'=1}^{W_h}\sum_{l'=1}^{L_h}\left(\frac{\alpha(h)_{(w,l),(w',l')}-\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}{\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}\right)x(h)_{(w,l)}x(h)_{(w',l')}.$$
$$(w,l)\neq(w',l')$$

Proof: In stratified path sampling, path sampling is used in each stratum. For stratum h, $\alpha(h)_{(w,l)}$ is the probability that unit (w,l) for $w=1,2,3,...,W_h$ and $l=1,2,3,...,L_h$ in stratum h for $h=1,2,3,...,H$ is included in the sample $s(h)$, and $\alpha(h)_{(w,l),(w',l')}$ is the probability that both unit (w,l) and unit (w',l') are included in the sample. The probability $\alpha(h)_{(w,l)}$ is also called the inclusion probability of unit (w,l), and the probability $\alpha(h)_{(w,l),(w',l')}$ is called the joint inclusion probability of unit (w,l) and unit (w',l'). Based on the Horvitz-Thompson estimator, an unbiased estimator of the total in stratum h is

$$\hat{T}(h) = \sum_{(w,l)\in s(h)}\frac{x(h)_{(w,l)}}{\alpha(h)_{(w,l)}}.$$

Since $\hat{T}(h)$ is the Horvitz-Thompson estimator, the variance of $\hat{T}(h)$ can be obtained by using Equation (A1). Hence, the variance of $\hat{T}(h)$ is

$$v(\hat{T}(h)) = \sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\left(\frac{1-\alpha(h)_{(w,l)}}{\alpha(h)_{(w,l)}}\right)x_{(w,l)}^2 +$$
$$\sum_{w=1}^{W_h}\sum_{l=1}^{L_h}\sum_{w'=1}^{W_h}\sum_{l'=1}^{L_h}\left(\frac{\alpha(h)_{(w,l),(w',l')}-\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}{\alpha(h)_{(w,l)}\alpha(h)_{(w',l')}}\right)x(h)_{(w,l)}x(h)_{(w',l')}.$$
$$(w,l)\neq(w',l')$$

References

Dryver AL, Thompson SK. Improving unbiased estimators in adaptive cluster sampling. J R Stat Soc Ser B Stat Methodol. 2005; 67(1): 157-166.

Flores O. Determinism of regeneration in fifteen tropical tree species in the Guiana forest: the effects of environment and limitation by dispersal. PhD [dissertation]. Montpellier: University of Montpellier II; 2005.

Borkowski MM. Infant sleep and feeding: a telephone survey of Hispanic Americans. PhD [dissertation]. Mount Pleasant (MI): Central Michigan University; 2002.

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

Levy PS, Lemeshow S. Sampling of populations: methods and applications, New York: John Wiely & Sons; 1999.

Lohr SL. Sampling: design and analysis. California: Duxbury Press Publishing; 2010.

Patummasut M, Borkowski JJ. Adaptive cluster sampling with spatially clustered secondary units. J Appl Sci. 2014; 14(20): 2516-2522.

Patummasut M, Dryver AL. A new sampling design for a spatial population: path sampling. J Appl Sci. 2012; 12(13): 1355-1363.

Patummasut M, Nidsunkid, S. A ratio estimator in path sampling. Commun Stat - Simul Comput. 2021; 50(2): 376-387.

Plante N, Rivest LP, Tremblay G. Stratified capture-recapture estimation of the size of a closed population. Biometrics. 1998; 54(1): 47-60.

Smith DR, Conroy MJ, Brakhage DH. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics. 1995; 51(2): 777-788.

Thompson SK. Adaptive cluster sampling. J Amer Stat Assoc. 1990; 85(412): 1050-1059.

Thompson SK. Stratified adaptive cluster sampling. Biometrika. 1991b; 78(2): 389-397.

Thompson SK. Sampling. New York: John Wiely & Sons; 2012.