Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 12 (13): 1355-1363, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.1355.1363

A New Sampling Design for a Spatial Population: Path Sampling

[1]Mena Patummasut and [2]Arthur L. Dryver
[1]Graduate School of Applied Statistics,
[2]Graduate School of Business Administration,
National Institute of Development Administration, Bangkapi District, Bangkok, 10240, Thailand

Abstract: This study proposed a new cost-effective and convenient sampling design for a spatial population, called "path sampling" and which offers the ability to sample all of the units in the researcher's path traversed during the sampling. Path sampling is a design in which the researcher selects a path or paths from start to finish, as opposed to selecting units. Path sampling offers unbiased estimators for both mean and variance. This paper covers the pros and cons of path sampling in comparison to simple random sampling and cluster sampling.

Key words: Cluster sampling, Horvitz-Thompson estimator, spatial population

INTRODUCTION

A spatial setting can be represented as a geographical area partitioned into single units. To estimate the population total or mean in an area, the population study area is divided into spatial units generally of the same size and the numbers of objects are counted on a selection of the units (Vincent, 2008). In sampling in a spatial population, there are many designs that can be used, for example, simple random sampling, stratified sampling, cluster sampling and systematic sampling or adaptive sampling in the case of a rare or clustered population. Thompson (2002) illustrated the application of those sampling designs to spatial populations. In cluster sampling, a primary unit which is a sampling unit, consists of a cluster of secondary units, usually in close proximity to each other. In the spatial setting, primary units include spatial arrangements as square collections of adjacent units. A simple random sample of m primary units is taken from M primary units in the population. Thompson (1990) introduced adaptive cluster sampling and this was compared to simple random sampling using simulation study on the spatial population. Dryver and Thompson (2005) and Dryver and Chao (2007) proposed more efficient estimators for adaptive cluster sampling and their illustrative examples were applied to spatial populations. Thompson (2006) proposed adaptive web sampling for sampling a population in network and spatial settings. However, it tends to be more efficient when used with many spatial populations (Thompson, 2011). Borkowski (2003) proposed simple Latin square sampling ±k designs which was a new class of probability sampling design that ensured that the sample was well-distributed over the study region when a spatial correlation was present.

Many factors often go into choosing a sampling strategy to implement. Such factors often include ease of implementation, cost, efficiency, etc. (Thompson, 2002; Mier and Picquelle, 2008). For example, simple random sampling is more efficient, given the same number of data points sampled as in cluster sampling; often, however, cluster sampling will be implemented, as it is easier to implement and may cost less (Lohr, 1999).

By applying simple random sampling and cluster sampling, a sample may cover all of the regions since each sampling unit has an equal chance of selection. Thus, traveling from place to place to observe every unit selected for sampling can be costly, as the distance traveled can be quite long (Hansen *et al*., 1953). One of the difficulties is that of collecting quantities of data dispersed over a large area. The new sampling design, path sampling, introduced in this paper also addresses this issue, especially when the distance travelled is a large part of the sampling cost.

PATH SAMPLING AND TECHNICAL NOTATION

This section deals with defining all possible paths in the spatial population, the path sampling scheme and estimation. Suppose the researcher's goal is to estimate the population total or mean. Initially, it will be assumed that the study region can be partitioned into an r×c (r: rows and c: columns) grid of rc quadrats or units. The population consists of rc spatial units. Each population

Corresponding Author: Mena Patummasut, Graduate School of Applied Statistics,
National Institute of Development Administration, Bangkapi District, Bangkok, 10240, Thailand

unit is labeled with 2 coordinates, say (i, j) which are the row and column of the unit, respectively, for i = 1, 2, 3,…, r and j = 1, 2, 3,…, c. Associated with each unit (i, j), the value of the population variable of interest is denoted as $y_{(i,j)}$. The parameter of interest in this study is the population mean:

$$\mu = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} y_{(i,j)} = \frac{1}{rc}\sum_{all(i,j)} y_{(i,j)} \quad (1)$$

Path sampling design is a sampling design in which p distinct paths are selected by simple random sample without replacement from q paths in the population and the sample consists of all units in the selected paths. Thus, a path(s) is chosen instead of units. In this study, we use path sampling for spatial population.

Define all possible paths in a spatial population: A path is basically the path or route taken from start to finish. Let q be the number of all possible paths. Let P_k denote a path k for k = 1, 2, 3,…, q. A path will be defined to start from row 1 and column j*; that is, a unit labeled (1, j*) is a starting unit and end at a unit (1, j*+1). We began sampling at an edge, at unit (1, j*), of a region because it was assumed to be more convenient and less expensive than beginning inside or in the middle of a region. The path k taken will begin from such a starting unit and then go to a particular row, say row k, to the end of the row on the left and then go along row k+1 and comes back to the starting unit. That is, path k taken will be from (1, j*) to (2, j*) then to (k, j*) to (k, j*-1) to (k, j*-2) to (k, 1) to (k+1, 1) to (k+1, 2) to (k+1, c) to (k, c) to (k, c-1) to (k, c-2) to (k, j*+1) to (k, -1 j*+1) to (k, -2 j*+1) and to (1, j*+1). Thus, for a spatial population of r rows, there are q = r-1 possible paths. In general, a path k in the spatial setting population of r rows and c columns can be written as: P_k = ((1, j*), (2, j*), (3, j*),..., (k, j*), (k, j*-1), (k, j*-2),..., (k, 1), (k+1, 1), (k+1, 2),..., (k+1, c), (k, c), (k, c-1), (k, c-2), (k, j-+1), (k-1, j*+1), (k-2, j*+1),..., (1, j*+1)) for k = 1, 2, 3,…, q = r-1.

The number of units belonging to path P_k is 2c+2(k-1). All possible paths are shown in Fig. 1. Notice that the numbers of units in each path are not the same. We can see that the paths overlap in column j* and j*+1 which are the going-out and coming-back column, respectively. Also, the paths next to each other overlap with the row between them. Thus, it can be written that path k-1 and path k overlap in row k for k = 2, 3,…, q = r-1. We assume that we sample the units in a logical manner such that all units will only be observed once. Finally, the researcher can define the rows and columns arbitrarily; thus, path sampling is not limited in its starting and ending position even written as is.

Path sampling design: The spatial population of r rows and c columns consists of units labeled (i,j) for i = 1, 2, 3,…, r and j = 1, 2, 3,…, c. There are q = r-1 possible paths in the population denoted by P_1, P_2, P_3,..., P_q. By

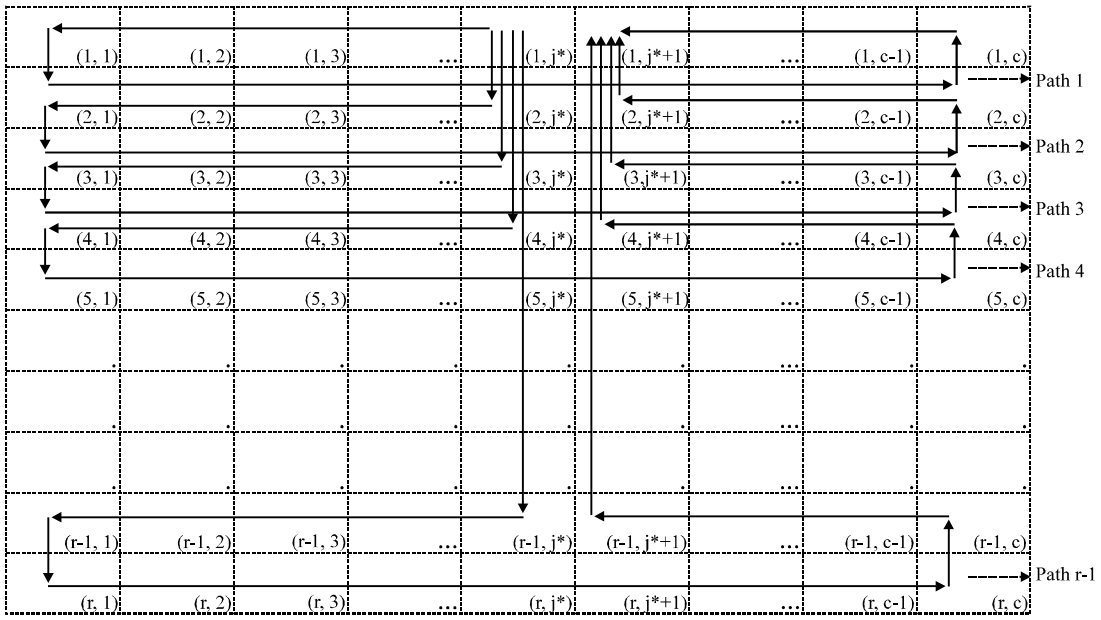

Fig. 1: All possible paths with a starting unit (1, j*) and all units labeled with two coordinates in a spatial population

SRSWOR, p paths are selected from q possible paths in the population. Let P_k denote a path k in the sample for k = 1, 2, 3,..., p. The sample consists of all units in the selected paths. The sample is represented as $P_S = (p_1, p_2, p_3,..., p_p)$. The probability of selecting a sample is:

$$P(s) = \frac{1}{\binom{q}{p}} = \frac{1}{\binom{r-1}{p}}$$

since paths are selected by SRSWOR and the inclusion probability of path k is:

$$\pi_k = \frac{p}{q} = \frac{p}{r-1}$$

There is an overlapping of paths, so, there are repeat observations. Although, each path has an equal probability of selection, the units do not have an equal probability of selection, as the same unit may be in one or more paths. The inclusion probability of each unit is the probability that a unit is included in the sample. In path sampling, the inclusion probability of unit (i, j) is denoted as $\pi_{(i,j)}$. It is defined as:

$$\begin{aligned}\pi_{(i,j)} &= P(\text{unit } (i,j) \text{ is in the sample})\\ &= 1 - P(\text{unit } (i,j) \text{ is not in the sample})\\ &= 1 - \frac{\text{The No. of samples not containing unit } (i,j)}{\text{The No. of all possible samples}}\end{aligned}$$

Since paths overlap in rows and columns, the probabilities that units are included in the sample are not equal. That is, the inclusion probabilities of each unit in a path are not equal. All paths overlap in column j* and j*+1 and some paths overlap in a row. Thus, the inclusion probabilities can be divided into three cases due to overlapping of paths.

$$\pi_{(i,j)} = \begin{cases} 1 - \dfrac{\binom{i-2}{p}}{\binom{q}{p}} & i = 1,2,3,...,r \text{ and } j = j^* \text{and } j^*+1 \\ 1 - \dfrac{\binom{q-2}{p}}{\binom{q}{p}} & i = 2,3,...,r-1 \text{ and } j = 1,2,3,..., j^*-1, j^*+2, j^*+3,...,c \\ 1 - \dfrac{\binom{q-1}{p}}{\binom{q}{p}} & i = 1, r \text{ and } j = 1,2,3,..., j^*-1, j^*+2, j^*+3,...,c \end{cases} \quad (2)$$

Note: Some of the combinations in the numerator of Eq. 2 can equal to 0.

Let the probability that both units (i, j) and (i',j') are included in the sample be denoted by $\pi_{(i,j),(i',j')}$, also called the joint inclusion probability. It is defined as:

$$\pi_{(i,j),(i',j')} = \frac{\text{The No. of samples containing both unit } (i,j) \text{ and } (i',j')}{\text{The No. of all possible samples}}$$

The probability that the sample does not contain either units (i, j) or (i', j') is:

$$\frac{\binom{f}{p}}{\binom{q}{p}} = \frac{\text{The No. of sample not containing either units } (i,j) \text{ or } (i',j')}{\text{The No. of all possible sample}}$$

where, f = the number of paths not containing either units (i, j) or (i', j'). Thus:

$$\pi_{(i,j),(i',j')} = \pi_{(i,j)} + \pi_{(i',j')} - \left(1 - \frac{\binom{f}{p}}{\binom{q}{p}}\right) \quad (3)$$

f can be found as follows. Let U_1 be a set of all units in column j* and j*+1 (units type 1). $U_1 = \{(i_1, j_1) | i_1 = 1, 2, 3,..., r \text{ and } j_1 = j^* \text{ and } j^*+1\}$. Let U_2 be a set of all units not in column j* and j*+1 and not in the first row or the last row (unit type 2). $U_2 = \{(i_2, j_2) | i_2 = 2, 3,..., r\text{-}1 \text{ and } j_2 = 1, 2, 3,..., j^*-1, j^*+2, j^*+3,..., c\}$. Let U_3 be a set of all units in the first row and the last row but not in column j* or j*+1 (unit type 3). $U_3 = \{(i_3, j_3) | i_3 = 1, r \text{ and } j_3 = 1, 2, 3,..., j^*-1, j^*+2, j^*+3,..., c\}$. A formula of f is shown in Eq. 4:

$$f = \begin{cases} \min(i,i') - 2 & \text{for } (i,j) \text{ and } (i'j') \in U_1 \\ i\text{-}2 & \text{for } (i,j) \in U_1 \text{ and } (i',j') \in (U_2 \cup U_3) \text{ and } i \le i' \\ i\text{-}3 & \text{for } (i,j) \in U_1 \in \text{and } (i',j') \in U_2 \text{ and } i - i' = 1 \\ & \text{or for } (i,j) \in U_1 \text{ and } (i',j') \in U_3 \text{ and } i \ge i' \\ i\text{-}4 & \text{for } (i,j) \in U_1 \text{ and } (i',j') U_2 \text{ and } i\text{-}i' \ge 2 \\ q\text{-}1 & \text{for } (i,j) \text{ and } (i',j') \in U_3 \text{ and } i\text{-}i' \\ q\text{-}2 & \text{for } (i,j) \text{ and } (i',j') \in U_2 \text{ and } |i\text{-}i'| = 0 \\ & \text{or for } (i,j) \in U_2 \text{ and } (i',j') \in U_3 \text{ and } |i\text{-}i'| = 1 \\ & \text{or for } (i,j) \text{ and } (i',j') \in U_3 \text{ and } i \ne i' \\ q\text{-}3 & \text{for } (i,j) \text{ and } (i',j') \in U_3 \text{ and } |i-j| = 1 \\ & \text{or for } (i,j) \in U_2 \text{ and } (i',j') \in U3 \text{ and } |i-j| \ge 2 \\ q\text{-}4 & \text{for } (i,j) \text{ and } (i',j') \in U_2 \text{ and } |i-j| \ge 2 \end{cases} \quad (4)$$

Note that if f < 0, then f is set equal to 0.

Estimation: Let $p_s = (p_1, p_2, p_3,..., p_p)$ denote the sample of paths selected. Let s denote the set of distinct units in the sample. By using the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), an unbiased estimator of the population mean under path sampling is:

$$\hat{\mu}_{ps} = \frac{1}{rc}\sum_{(i,j)\in s}\frac{y_{(i,j)}}{\pi_{(i,j)}} \quad (5)$$

Let $I_{(i,j)}$ be the indicator function taking the value one if unit (i,j) is selected in the sample and 0 otherwise. It can be written as:

$$I_{(i,j)} = \begin{cases} 1 & \text{if unit } (i,j) \text{ is included in the sample} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $\hat{\mu}_{ps}$ can be written in the alternative form:

$$\hat{\mu}_{ps} = \frac{1}{rc}\sum_{all(i,j)}\frac{y_{(i,j)}I_{(i,j)}}{\pi_{(i,j)}} \quad (6)$$

$\hat{\mu}_{ps}$ is the unbiased estimator for the population mean μ.

The variance of $\hat{\mu}_{ps}$ is:

$$v(\hat{\mu}_{ps}) = \frac{1}{(rc)^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{1-\pi_{(i,j)}}{\pi_{(i,j)}}\right)y_{(i,j)}^2 + \sum_{i=1}^{r}\sum_{i'\neq i}\sum_{j=1}^{c}\sum_{j'\neq j}\left(\frac{\pi_{(i,j),(i',j')} - \pi_{(i,j)}\pi_{(i',j')}}{\pi_{(i,j)}\pi_{(i',j')}}\right)y_{(i,j)}y_{(i',j')}\right) \quad (7)$$

and the estimator of this variance is:

$$\hat{v}(\hat{\mu}_{ps}) = \frac{1}{(rc)^2}\left(\sum_{(i,j)\in s}\left(\frac{1}{\pi_{(i,j)}^2} - \frac{1}{\pi_{(i,j)}}\right)y_{(i,j)}^2 + \sum_{\substack{(i,j),(i',j')\in s \\ (i,j)\neq(i',j')}}\left(\frac{1}{\pi_{(i,j)}\pi_{(i',j')}} - \frac{1}{\pi_{(i,j),(i',j')}}\right)y_{(i,j)}y_{(i',j')}\right) \quad (8)$$

The estimate of variance may be negative.

The spatial population of 4 rows and 6 column as shown in Fig. 2 is considered. The population mean and variance are 8.208 and 549.6, respectively. The objective is to estimate the population mean by using path sampling. First, all possible paths are created. The number of rows in this population is r = 4 and the number of columns is c = 6. Thus, the number of all possible paths is q = r-1 = 4-1 = 3. In general, a path k in the spatial setting population of r rows and c columns with starting unit (I, j*) is written as: P_k = ((1, j*), (2, j*), (3, j*),..., (k, j*), (k, j*-1), (k, j *-2),..., (k, 1), (k+1, 1), (k+1, 2),..., (k+1, c), (k, c), (k, c-1), (k, c-2),..., (k, j*+1), (k-1, j*+1), (k-2, j*+1),..., (1, j*+1)) for k = 1, 2, 3,..., q = r-1.

Let the starting unit be (1, 3), so, j* = 3. Thus, we have all possible paths with their labeled units as follows:

P_1 = ((1, 3), (1, 2), (1, 1), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (1, 6), (1, 5), (1, 4))

P_2 = ((1, 3), (2, 3), (2, 2), (2, 1), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (2, 6), (2, 5), (2, 4), (1, 4))

P_3 = ((1, 3), (2, 3), (3, 3), (3, 2), (3, 1), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (3, 6), (3, 5), (3, 4), (2, 4), (1, 4))

Since the number of units belonging to P_k is 2c+2(k-1), the number of units belonging to P_1 is 2(6)+2(1-1) = 12 units, the number of units belonging to P_2 is 2(6)+2(2-1) = 14 units and the number of units belonging to P_3 is 2(6)+2(3-1) = 16 units. Suppose the number of sample paths is 2, so, by using SRSWOR, p = 2 sample paths are selected. There are 3 possible samples which are $p_{s1} = (P_1, P_2)$, $p_{s2} = (P_1, P_3)$ and $p_{s3} = (P_2, P_3)$.

p_{s1} = (P_1, P_2) reduces to s_1 = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)}

p_{s2} = (P_1, P_3) reduces to s_2 = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)}

p_{s3} = (P_2, P_3) reduces to s_3 = {(1, 3), (1, 4), (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)}

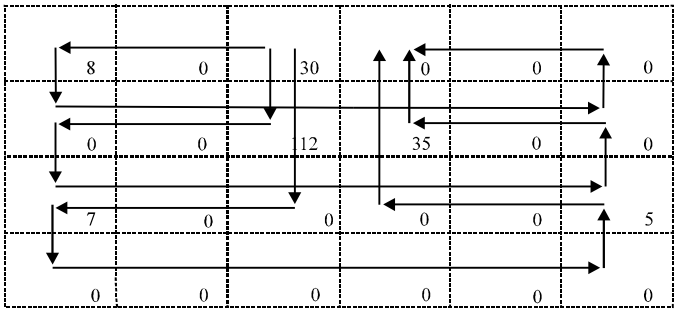

Fig. 2: All possible paths of the spatial population for 4 rows and 6 columns with a y-value of each unit

Next, the inclusion probabilities are calculated by the formula Eq. 2. First, the inclusion probabilities for units in column 3 and 4 (unit type 1) will be calculated. For i = 1, 2, 3, 4 and j = 3 and 4, we have:

$$\pi_{(i,j)} = 1 - \frac{\binom{i-2}{p}}{\binom{q}{p}}$$

Then, we get:

$$\pi_{(1,3)} = 1 - \frac{\binom{1-2}{2}}{\binom{3}{2}} = 1 - 0 = 1 = \pi_{(1,4)} \qquad \pi_{(2,3)} = 1 - \frac{\binom{2-2}{2}}{\binom{3}{2}} = 1 - 0 = 1 = \pi_{(2,4)}$$

$$\pi_{(3,3)} = 1 - \frac{\binom{3-2}{2}}{\binom{3}{2}} = 1 - 0 = 1 = \pi_{(3,4)} \qquad \pi_{(4,3)} = 1 - \frac{\binom{4-2}{2}}{\binom{3}{2}} = 1 - \frac{1}{3} = \frac{2}{3} = \pi_{(4,4)}$$

Next, the inclusion probabilities for units not in column 3 and 4 and not in the first row or last row (unit type 2) will be calculated. For i = 2, 3 and j = 1, 2, 5, 6:

$$\pi_{(i,j)} = 1 - \frac{\binom{q-2}{p}}{\binom{q}{p}}$$

Then:

$$\pi_{(2,1)} = 1 - \frac{\binom{3-2}{2}}{\binom{3}{2}} = 1 - 0 = 1 = \pi_{(2,2)} = \pi_{(2,5)} = \pi_{(2,6)} = \pi_{(3,1)} = \pi_{(3,2)} = \pi_{(3,5)} = \pi_{(3,6)}$$

Finally, the inclusion probabilities for units in the first row and the last row but not in column 3 or 4 (unit type 3) are calculated. For i = 1 and 4 and j = 1, 2, 5, 6:

$$\pi_{(i,j)} = 1 - \frac{\binom{q-1}{p}}{\binom{q}{p}}$$

Then:

$$\pi_{(1,1)} = 1 - \frac{\binom{3-1}{2}}{\binom{3}{2}} = 1 - \frac{1}{3} = \frac{2}{3} = \pi_{(1,2)} = \pi_{(1,5)} = \pi_{(1,6)} = \pi_{(4,1)} = \pi_{(4,2)} = \pi_{(4,5)} = \pi_{(4,6)}$$

The inclusion probabilities are shown in Fig. 3. Estimates of the mean for all possible samples are shown in Table 1. It can be seen that $\hat{\mu}_{ps}$ is an unbiased estimator since its bias is zero.

Recall that $p_{s1} = (P_1, P_2)$ reduce to $s_1 = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)\}$ corresponding to y = {8, 7, 30, 24, 6, 5, 0, 10, 112, 35, 5, 8, 7, 7, 32, 0, 0, 5}. By using Eq. 5:

$$\hat{\mu}_{ps} = \frac{1}{rc}\sum_{(i,j)\in s_1}\frac{y_{(i,j)}}{\pi_{(i,j)}} = \frac{1}{4(6)}\left[\frac{y_{(1,1)}}{\pi_{(1,1)}} + \frac{y_{(1,2)}}{\pi_{(1,2)}} + ... + \frac{y_{(3,6)}}{\pi_{(3,6)}}\right] = \frac{1}{24}\left[\frac{8}{2/3} + \frac{0}{2/3} + ... + \frac{5}{1}\right]$$
$$= \frac{201}{24} = 8.375$$

Similarly, the estimate of variance is calculated using Eq. 8.

Table 1: Estimates of the mean and variance estimator for all possible samples

Sample	$\hat{\mu}_{ps}$	Sample size	$\hat{v}(\hat{\mu}_{ps})$
$p_{s1} = (P_1, P_2)$	8.375	18.00	0.083
$p_{s2} = (P_1, P_3)$	8.375	24.00	0.083
$p_{s3} = (P_2, P_3)$	7.875	20.00	0.000
Mean	8.208	20.67	0.056
Bias	0.000		0.000
Variance	0.056		

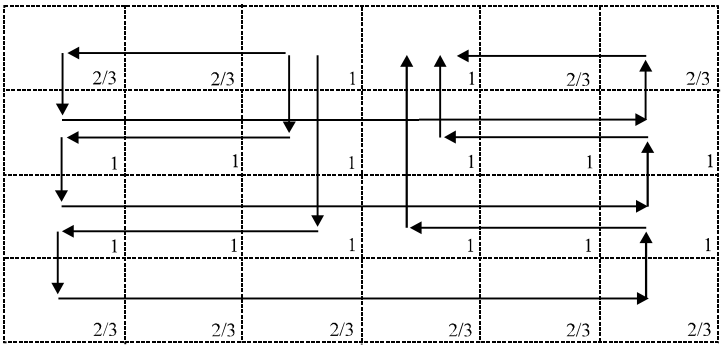

Fig. 3: The inclusion probabilities of the population of 4 rows and 6 columns

SIMULATION STUDY

Rare and non-rare population data are used in a simulation to examine the performance of path sampling compared to a comparable sampling design which in this research are SRSWOR and cluster sampling. The simulation consists of 1000 iterations. The formula used to estimate the variance is:

$$\hat{v}(\hat{\mu}) = \frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\mu}_i - \bar{\mu})^2 \quad (9)$$

where, $\hat{\mu}_i$ is the value for the relevant estimator for sample i and $\bar{\mu}$ is the average of the $\hat{\mu}_i$ (Dryver and Thompson, 2005).

Simulation study for rare population: The authors used blue-winged teal data (Smith *et al*., 1995) in Fig. 4 for part of the simulation study, as it is a rare population. In cluster sampling, let a cluster be an entire column, consisting of 10 units, as shown in Fig. 4. This population data have high variation among clusters with CV of 4.26. The expected sample size will be denoted E(υ) and the sample size used in the other designs was set equal the ceiling of the E(υ) for path sampling. For cluster sampling, the number of clusters sampled was set equal to the ceiling of $\frac{E(\upsilon)}{10}$. In SRSWOR, the sample size was set equal to E(υ) in order to compare it to path sampling.

The results from the simulations are shown in Table 2. From these results, for starting unit (1, 1) and (1, 10), path sampling was more efficient than cluster sampling since the relative efficiency was greater than 1. Noticeably, the y-values in column 17, 18 and 19 were higher than others, so, there was high variation among the clusters in this population. This made cluster sampling less efficient. However, path sampling was less efficient than SRSWOR since the relative efficiency was less than 1. Notice that when the starting unit is in a high-valued column which is unit (1,17), path sampling was more efficient than SRSWOR since the relative efficiency was greater than 1 and much more efficient than cluster sampling since the relative efficiency was greater than 4.

Simulation study for non-rare population: Two simulated data were considered. First, we used the simulated data in Fig. 5. Each unit was Poisson distributed with a mean of 50. To compare path sampling to cluster sampling, let a cluster be a cluster of an entire column. In this population, the CV among the clusters is 0.04. The simulation results are shown in Table 3.

From the simulation results in Table 3, it can be seen that path sampling was less efficient than both cluster sampling and SRSWOR because the relative efficiency was less than 1. Noticeably, there was a small variation of y-values, so there was low variation among clusters (CV among clusters is 0.04) in this population. This makes cluster sampling more efficient.

Table 2: Results from the simulations on blue-winged teal data

			$\hat{v}(\hat{\mu}_{ps})$					(1,1)		(1, 10)		(1,17)*	
p	E (υ)	m_c	(1,1)	(1, 10)	(1,17)*	$\hat{v}(\hat{\mu}_{cls})$	$\hat{v}(\hat{\mu}_{srs})$	R.E.cls	R.E.srs	R.E.cls	R.E.srs	R.E.cls	R.E.srs
1	48	50 (5)	10977.94	10753.86	2783.91	13771.3	7021.76	1.25	0.64	1.28	0.65	4.95	2.52
2	83.33	90 (9)	4415.87	4115.08	579.54	5525.01	3312.64	1.25	0.75	1.34	0.81	9.53	5.72
3	113	120 (12)	2146.11	2105.84	96.29	3023.43	1691.70	1.41	0.79	1.44	0.80	31.40	17.57
4	138	140(14)	1072.83	1102.51	9.37	1960.50	1015.52	1.83	0.95	1.78	0.92	209.23	108.38
5	158.6	160 (16)	521.31	541.39	0.25	1232.52	623.476	2.36	1.20	2.28	1.15	4930.08	2493.90

The number in parentheses is the number of clusters selected in cluster sampling, m_c is the No. of units in a cluster sample, * means that such a starting unit is on a high y-value column j* or has high y-value column j*+1. R.E.cls = $\hat{v}(\hat{\mu}_{cls})/\hat{v}(\hat{\mu}_{ps})$, R.E.srs = $\hat{v}(\hat{\mu}_{srs})/\hat{v}(\hat{\mu}_{ps})$

cls1	cls2	cls3	cls4	cls5	cls6	cls7	cls8	cls9	cls10	cls11	cls12	cls13	cls14	cls15	cls16	cls17	cls18	cls19	cls20
0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	20	4	2	12	0	0	0	0	0	10	103	0	0	0
0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	150	7144	1	0
0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	6	6339	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	122	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	60
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0

Fig. 4: Clusters in blue-winged teal data

cls1	cls2	cls3	cls4	cls5	cls6	cls7	cls8	cls9	cls10	cls11	cls12	cls13	cls14	cls15	cls16	cls17	cls18	cls19	cls20
43	51	40	40	55	56	49	43	61	61	49	39	38	42	53	61	57	50	48	47
47	51	51	53	51	43	60	55	50	60	61	49	56	50	57	55	59	49	40	47
60	52	46	49	54	51	58	45	48	44	43	54	61	50	63	50	57	45	47	50
55	56	51	51	47	38	55	50	51	51	61	42	48	35	50	41	67	48	47	48
49	55	55	44	39	61	47	54	60	55	67	43	41	50	52	55	44	45	54	54
60	50	49	46	57	49	57	49	44	56	37	44	47	47	46	48	46	42	29	52
53	57	35	52	43	51	49	65	54	51	55	52	55	68	39	44	39	48	68	56
55	51	56	34	50	57	49	58	52	64	41	49	47	61	52	50	55	57	41	47
51	62	41	45	41	55	43	51	46	33	49	54	56	41	51	46	61	55	43	35
47	47	41	46	47	56	47	43	61	44	43	59	39	52	46	37	48	59	49	61

Fig. 5: Simulated data, each unit is Poisson distributed with a mean of 50 with CV among clusters of 0.04

cls1	cls2	cls3	cls4	cls5	cls6	cls7	cls8	cls9	cls10	cls11	cls12	cls13	cls14	cls15	cls16	cls17	cls18	cls19	cls20
43	51	40	40	55	556	49	43	61	1610	49	39	38	42	553	61	57	50	48	47
47	51	51	53	51	643	60	55	50	55	61	49	56	50	657	55	59	49	40	47
60	52	46	49	54	651	58	45	48	1404	43	54	61	50	563	50	57	45	47	50
55	56	51	51	47	638	55	50	54	55	61	42	48	35	689	41	67	48	47	48
49	55	55	44	39	561	47	54	60	67	67	43	41	50	552	55	44	45	54	54
60	50	49	46	57	665	57	49	44	155	37	44	47	47	546	48	46	42	29	52
53	57	35	52	43	651	49	65	54	1501	55	52	55	68	639	44	39	48	68	56
55	51	56	34	50	457	49	58	52	64	41	49	47	61	457	50	55	57	41	47
51	62	41	45	41	555	43	51	46	33	49	54	56	41	551	46	61	55	43	35
47	47	41	46	47	356	47	43	61	133	43	59	39	52	446	37	48	59	49	61

Fig. 6: Simulated data with CV among clusters of 1.46

Table 3: Results from the simulation on a non-rare population with low CV among clusters

			$\hat{v}(\hat{\mu}_{ps})$				(1, 10)		(1, 17)	
p	E (υ)	m_c	(1, 10)	(1, 17)	$\hat{v}(\hat{\mu}_{cls})$	$\hat{v}(\hat{\mu}_{srs})$	R.E. cls	R.E. srs	R.E. cls	R.E. srs
1	48	50 (5)	89.38	86.18	0.74	0.81	0.0083	0.0091	0.0086	0.0094
2	83.33	90 (9)	77.70	65.57	0.29	0.31	0.0037	0.0040	0.0044	0.0047
3	113	120 (12)	54.57	53.37	0.16	0.20	0.0029	0.0034	0.0030	0.0040
4	138	140(14)	44.92	49.11	0.10	0.12	0.0022	0.0027	0.0020	0.0024
5	158.6	160 (16)	34.08	31.24	0.06	0.07	0.0018	0.0021	0.0019	0.0022

Table 4: Results from simulation on non-rare population with high CV among clusters

			$\hat{v}(\hat{\mu}_{ps})$							(1, 2)		(1, 5)*		(1, 10)*		(1, 15)*		(1, 17)	
p	E (υ)	m_c	(1, 2)	(1, 5)*	(1, 10)*	(1, 15)*	(1, 17)	$\hat{v}(\hat{\mu}_{cls})$	$\hat{v}(\hat{\mu}_{srs})$	R.E.cls	R.E.srs	R.E.cls	R.E.srs	R.E.cls	R.E.srs	R.E.cls	R.E.srs	R.E.cls	R.E.srs
1	48	50 (5)	1125.40	797.54	294.37	802.52	1015.10	5108.45	869.33	4.54	0.77	6.41	1.09	17.35	2.95	6.37	1.08	5.03	0.86
2	83.33	90 (9)	652.52	561.49	280.53	504.01	628.53	1955.48	403.27	2.99	0.62	3.48	0.72	6.97	1.44	3.88	0.80	3.11	0.64
3	113	120 (12)	491.86	374.83	219.67	374.17	484.41	1087.56	206.43	2.21	0.42	2.90	0.55	4.95	0.94	2.91	0.55	2.25	0.43
4	138	140 (14)	350.09	289.09	170.70	276.28	356.62	722.55	114.66	2.06	0.33	2.50	0.40	4.23	0.67	2.62	0.42	2.03	0.32
5	158.6	160 (16)	248.91	188.63	129.05	203.40	250.73	430.94	73.01	1.73	0.29	2.28	0.39	3.34	0.57	2.12	0.36	1.72	0.29

The number in parentheses is the number of clusters selected in cluster sampling, m_c is the No. of units in a cluster sample, * means that such a starting unit is on a high y-value column j* or has high y-value column j*+1. R.E.cls = $\hat{v}(\hat{\mu}_{cls})/\hat{v}(\hat{\mu}_{ps})$, R.E.srs = $\hat{v}(\hat{\mu}_{srs})/\hat{v}(\hat{\mu}_{ps})$

Next, simulated data, as shown in Fig. 6 is used. All units were the same as the population data in Fig. 5, except column 6, 10 and 15. The y-values in these 3 columns were replaced with a higher value. To compare path sampling with cluster sampling, let a cluster be a cluster of a column. This population data had high variation among the clusters with CV among clusters of 1.46. The simulation results are shown in Table 4.

According to the simulation results in Table 4, for starting unit (1, 2) and (1, 17), path sampling was more efficient than cluster sampling because the relative efficiency was greater than 1. Noticeably, the y-values in column 6, 10 and 15 were very higher than the others, so there was high variation among clusters (CV of 1.46) in this population. This made cluster sampling less efficient. Notice that when the starting unit is in a high-valued

column which are unit (1, 5), (1, 10) and (1, 15), path sampling was much more efficient than cluster sampling since the relative efficiency was greater than 2.

For starting unit (1, 2) and (1, 17), path sampling was less efficient than SRSWOR since the relative efficiency was less than 1 for any p. However, for the starting unit in a high-valued column which are unit (1, 5), (1, 10) and (1, 15), path sampling was more efficient than SRSWOR for p = 1 since the relative efficiency was greater than 1 but it was less efficient than SRSWOR for p>2 because the relative efficiency was less than 1.

DISCUSSION

Path sampling can be very cost-effective for sampling many units. This is true when cost is mainly a function of distance travelled, as the number of units sampled equals the number of units travelled. In path sampling, the researcher can sample all of the consecutive units in a path traversed during the sampling. On the other hand, for cluster sampling, the cost of traveling between clusters will be higher the more widespread the sample (Hansen *et al*., 1953). In situations with budget constraints it is possible that a researcher could sample more units with path sampling, thus giving it an added advantage in this respect. Unfortunately, for path sampling the number of units in the final sample is random and can vary a lot as a result of the number of units in each path vary. Therefore, the expense of sampling when cost is a function of distance travelled would also be random, possibly creating budget problems. However, the expected sample size in path sampling can be obtained as with adaptive cluster sampling (Thompson, 1990). It is the sum of the inclusion probabilities.

As a result of the way in which the paths were formed, path sampling is a type of unequal probability sampling and the authors used the Horvitz-Thompson estimator for estimation of the population mean. Similarly in path sampling, much of the literature has applied the Horvitz-Thompson estimator (Birnbaum and Sirken, 1965; Thompson, 1990; Nafiu and Adewara, 2007) because of the unequal probability of selection. For the Horvitz-Thompson estimator, it is desirable to have the y-values proportional to the probability of selection in order to obtain a relatively small variance which is observed by Horvitz and Thompson (1952). This limitation is clear when comparing path sampling to simple random sampling in the simulation results in Table 2, 3 and 4. If there is an auxiliary variable correlated with the variable of interest it is desirable, when possible, to select a starting and ending point for the paths which would have high y-value units having a high probability of selection and vice-versa for low-valued units.

In addition, when the CV from cluster to cluster in cluster sampling is high, then path sampling may be a viable alternative to cluster sampling, as can be seen in Table 2 and 4. Correspondingly, Chih (2011) mentioned that cluster sampling is less efficient when the between-cluster variability is large.

Path sampling should be implemented when two conditions are met-when the cost of the sampling is mainly a function of distance travelled and when it is believed that the y-values are positively correlated with the probability of selection. It is known that the ratio estimator is often more precision (Dryver and Chao, 2007). Therefore, if there is an auxiliary variable known to be correlated with the variable of interest, then perhaps a ratio estimator for path sampling should be considered. Finally, for rare and hidden populations, further research should be carried out that investigates combining adaptive cluster sampling and path sampling.

CONCLUSION

In this study, sampling in a spatial population was studied. Sampling a spatial population by applying previous sampling designs, such as simple random sampling and cluster sampling, was inconvenient because the researcher had to travel from place to place to observe every unit in a sample. Thus, path sampling was proposed and compared to simple random sampling and cluster sampling. Path sampling is more convenient and cost-effective but less efficient in some circumstances. According to the simulation results for a rare population and a non-rare population with high variation of y-values among clusters, path sampling is more efficient than cluster sampling but less efficient than SRSWOR. However, for a non-rare population with a low variation of y-values among clusters, path sampling is less efficient than cluster sampling and SRSWOR. An illustrative example was offered by applying path sampling to a spatial population of 4 rows and 6 columns. The calculation of the estimate of the mean and variance was also shown. Finally, all possible paths in this study are created in a certain way, so that, inclusion probabilities and joint inclusion probabilities can be obtained and the Horvitz-Thompson estimator can be applied. Another form of path could be created that is more convenient and cost-effective. Moreover, other estimators could be created to improve the precision. In a rare and clustered population, adaptive path sampling could be of interest.

ACKNOWLEDGMENT

We are grateful to the Commission on Higher Education, Thailand, for financial support through a grant under the Strategic Scholarships Fellowships Frontier Research Networks.

REFERENCES

Birnbaum, Z.W. and M.G. Sirken, 1965. Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. Vital Health Statistics Series 2.

Borkowski, J.J., 2003. Simple Latin square sampling ±K designs. Commun. Stat. Theory Methods, 32: 215-237.

Chih, C.P., 2011. The design effects of cluster sampling on the estimation of mean lengths and total mortality of reef fish. Fish. Res., 109: 295-302.

Dryver, A.L. and C.T. Chao, 2007. Ratio estimators in adaptive cluster sampling. Environmetrics, 18: 60-620.

Dryver, A.L. and S.K. Thompson, 2005. Improved unbiased estimators in adaptive cluster sampling. J. R. Stat. Soc. Ser. B: Stat. Methodol., 67: 157-166.

Hansen, H.M., N.W. Hurwitz and G.W. Madow, 1953. Sample Survey Methods and Theory. John Wiley, New York, USA.

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

Lohr, S.L., 1999. Sampling: Design and Analysis. 1st Edn., Duxbury Press, California, USA., ISBN-13: 978-0534353612, Pages: 512.

Mier, K.L. and S.J. Picquelle, 2008. Estimating abundance of spatially aggregated populations: Comparing adaptive sampling with other survey designs. Can. J. Fish. Aquat. Sci., 65: 176-197.

Nafiu, L.A. and A.A. Adewara, 2007. On the use of network sampling in diabetic J. Res. National Dev., 5: 5-9.

Smith, D.R., M.J. Conroy and D.H. Brakhage, 1995. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics, 51: 777-788.

Thompson, S.K., 1990. Adaptive cluster sampling. J. Am. Stat. Assoc., 85: 1050-1059.

Thompson, S.K., 2002. Sampling. 2nd Edn., John Wiley and Sons, New York, USA., ISBN-13: 9780471291169, Pages: 367.

Thompson, S.K., 2006. Adaptive web sampling. Biometrics, 62: 1224-1234.

Thompson, S.K., 2011. Adaptive network and spatial sampling. Surv. Mehtodol., 37: 183-196.

Vincent, K.S., 2008. Design variations in adaptive web sampling. M.S. Thesis, Simon Fraser University, Burnaby, BC Canada.