مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۱۱، شماره‌ی ۱، بهار و تابستان ۱۳۹۳، صص ۱–۲۳

J. Statist. Res. Iran **11** (2014): 1–23

A Comparative Study of Performance of Adaptive Web Sampling and General Inverse Adaptive Sampling in Estimating Olive Production in Iran

Ada Afshar* and Hamidreza Navvabpour

Allameh Tabataba'i University

Abstract. Nowadays, there is an increasing use of sampling methods in network and spatial populations. Although the most common link-tracing designs such as adaptive cluster sampling and snowball sampling have advantages over conventional sampling designs such as simple random sampling and cluster sampling, these designs still present many drawbacks. Adaptive web sampling is a new link-tracing design that overcomes some of the drawbacks. In this paper, after introducing adaptive web sampling design, an application of this method to estimate total production of olive in Iran using 1382 census data of agriculture is presented. Then by conducting a simulation study the performance of adaptive web sampling in estimating total olive production will be compared with that of general inverse adaptive sampling.

Keywords. Roa-Blackwell estimator; Markov chain Monte Carlo; link-tracing designs; adaptive web sampling; general inverse adaptive sampling.

MSC 2010: 62D05, 62F40.

1 Introduction

With increasing social concerns such as infectious disease outbreaks like AIDS, and its environmental issues such as extinction of species, sampling

* Corresponding author

methods need to be developed to study such target populations more efficiently. Sampling in such network and spatial settings has therefore received increasing attention in recent years.

Most current link-tracing designs like adaptive cluster sampling and snowball sampling have many advantages over conventional designs such as simple random sampling and cluster sampling. Some of these advantages include a potential reduction in time, effort, and expenses required to obtain samples of equal size (Frank and Snijders, 1994; Thompson, 1990). Additional advantages consist of an increase in the overall yield of units of higher interest in the sample (Frank and Snijders, 1994; Thompson, 1990, 2006b), and a potential reduction in variance of the estimates (Thompson, 1990). However, many of the current designs still present many drawbacks. Adaptive Web Sampling (AWS) is a new link-tracing design that overcomes some of these drawbacks. The AWS design is said to be adaptive since selection probabilities of new members to be included in the sample depend on the observed variables of interest in the current sample and also it set to be web sampling because to reflect their ability to reach web like into interesting areas of the target population.

A new type of design introduced by Thompson (2004) gains over link-tracing type sampling. Thompson (2006a) applied this method in HIV/AIDS, at-risk population (Potterat et al., 1993). Xu (2007) presented new resampling strategies for inference based on adaptive web sampling designs and produced designs with practical restrictions to minimize the cost and maximize the sampling simultaneously, and model based estimation for non-responses. Vincent (2008) introduced three various designs and compared them in a simulated network population, an empirical population at risk for HIV/AIDS (Potterat et al., 1993), a simulated spatial population, and an empirical bird population (Smith et al., 1995).

AWS begins with selection of an initial sample by some conventional sampling design. New members can be added to the sample by either tracing links from the members in the current sample or a conventional sampling design. This design is more flexible in controlling how far/deep the sampling procedure could go, how the sample could be spread out, how large the sample size could be, and etc.

In Section 2 we introduce notations used in AWS designs (Thompson, 2006a). In sections 3-6 we present sampling setup for AWS, three strategies for inference based on AWS design, its estimators, and the Markov chain Monte Carlo (MCMC) methods for estimating the Rao-Blackwellized estima-

tors, respectively. In Section 7 we present setup for general inverse adaptive sampling. Finally in Section 8 by conducting simulations studies performances of adaptive web sampling and general inverse adaptive sampling in estimating total olive production using 1382 Iran census of agriculture data will be compared.

2 Notations

2.1 Network Setting

In a network setting, population units are labeled $1, 2, \ldots, N$. Corresponding to every i, there is an observable variable y_i. In the general network setting, y_i can take any numerical value. In more specific network settings, y_i can be an indicator variable where

$$Y_i = \begin{cases} 1 & \text{if unit } i \text{ is a unit of interest} \\ 0 & \text{otherwise.} \end{cases}$$

"Unit of interest" means to have (or not) some specific characteristic. For every ordered pair of individuals (i, j), there is an observable variable w_{ij} that represents the existence or strength of the relationship between units i and j, and determines the graph structure of the units. In most general network settings, w_{ij} can also take any numerical value. In more specific network settings, w_{ij} is an indicator variable where

$$w_{ij} = \begin{cases} 1 & \text{if there is a link from unit } i \text{ to unit } j \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we define $w_{ii} = 0$ for all $i = 1, 2, ..., N$.

2.2 Spatial Setting

A spatial setting is shown as geographic area that is divided into separate units. In the spatial setting, the y_i variable takes on the count of the number of point-object in the ith square and w_{ij} is defined as follows:

$$w_{ij} = \begin{cases} 1 & \text{if units } i \text{ and } j \text{ are adjacent and unit } i \text{ is a unit of interest} \\ 0 & \text{otherwise.} \end{cases}$$

Units i and j are considered adjacent if unit i is directly above, below, left, or right of unit j. In the spatial setting, symmetry of links only holds if two units both possess the characteristic of interest and are adjacent.

3 Adaptive Web Sampling

An adaptive web sample is selected as follows: at first, an initial sample of size n_0 is selected from population with probability p_0 by a conventional sampling design. New unit selection for inclusion in the sample is called to occur in wave. For each wave k, new unit selection depends on a current active set $a_k \subseteq s_{ck}$ that s_{ck} is current sample at wave k so far selected. Active set selection is flexible and it may consist of all the units selected so far, or the most recently selected units, or other possibilities such as the last two steps units or sequences of units.

In link tracing designs, since the following nodes links for inclusion in the sample are used, therefore new unit selection may be easily influenced by the nodes that are in the initial sample. To help overcome this issue, AWS introduces the use of a mixture distribution in the selection probabilities of new units, so that with probability d (dampening value) the next unit is selected adaptively using a distribution based on the unit values or graph structure of the active set and with probability $1 - d$ it is selected conventionally using a distribution based on the sampling frame or spatial structure of the population. The value of d is not always constant and may depend on the sampling wave and the active set.

An example of this design in a spatial setting is illustrated in Figure 1. A sample of 20 nodes is selected from the graph using the following procedure. First, an initial sample of 12 units is chosen by random sampling without replacement. It turns out that none of the initial sample units contain any of the objects of interest, but, since the total sample size of 20 has been decided on in advance, random sampling without replacement continues for another step. The 13th selected unit contains some of the objects, and so has four links out from it. With $d = 0.9$, one of these links will be selected at random and followed, while with $d = 0.1$ probability of another unit will be selected at random from the units in the population not already selected. As it happens, it is determined to select a link, and the one chosen leads us to the 14th selected unit, which also contains objects of interest. Now, with $d = 0.9$ probability that we can select a link from any of the six links, originating at the 13th and 14th units (in Figure 1 on spatial population,

they are marked with numbers 1-6), which lead out from the current sample. This selection could be made at random but, because the 14th unit contains a lot more of the objects than does the 13th unit, we will use a design variation instead that selects the next link with probability proportional to the value of the originating node. This gives higher probability to following a link from node 14, and indeed that is what leads us to node 15. Continuing with this design, the remaining five units are selected. The active set in this example consists of the entire current sample, and selections depend adaptively on both the node values and the link indicator values associated with this set (Thompson, 2004).

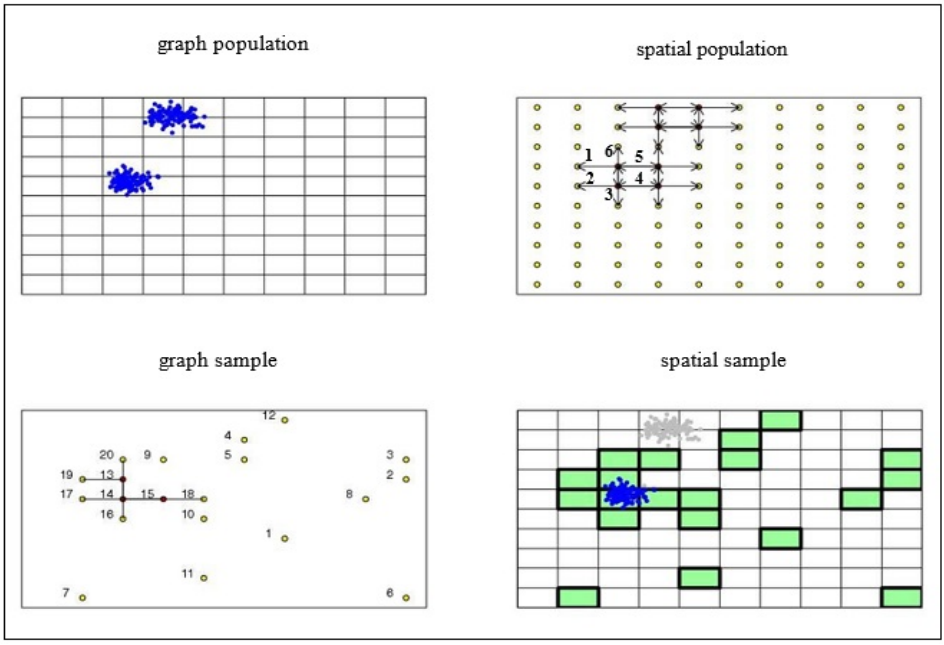

Figure 1. An example of adaptive web sampling design.

4 Strategies of Adaptive Web Sampling

Thompson (2006a) proposed three new methods for estimating the population mean or proportion that second and third strategies provide additional flexibility in the sample selection.

4.1 Strategy 1: Random Choice of Links

In this strategy an initial sample is selected and in each wave a link from a node in active set is followed with a constant probability d. Suppose that in wave k, current sample $s_{ck} = \bigcup_{i=0}^{k-1} s_i$ includes some active set a_k, then we show the selection probability of the ith unit selected at sub-step t in wave k as q_{kti} and in sampling without-replacement is

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_{kt}+}} + (1-d)\frac{1}{N - n_{s_{ckt}}}$$

where $w_{a_k i}$ is the number of links from the nodes in the active set out to a node i not in the current sample $w_{a_{kt}+}$ is the number of links from nodes in the active set out to members not in the current sample $n_{s_{ckt}}$ is the size of the current sample in sub-step t, and N is population size. When sampling is with-replacement then

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_{kt}+}} + (1-d)\frac{1}{N}$$

where $w_{a_k i}$ is the number of links from the nodes in the active set out to a node i, $w_{a_{kt}+}$ is the number of links from nodes in the active set out to any members of the population.

Hence, the overall probability of selecting a sample $\mathbf{s} = (s_{1_1}, s_{1_2}, \ldots, s_{1_{n_1}}, \ldots, s_{K_1}, s_{K_2}, \ldots, s_{K_{n_K}})$ is

$$\Pr(\mathbf{s}) = p_0 \prod_{k=1}^{K} \prod_{t=1}^{n_k} q_{kti}$$

where K is the number of sampling waves.

4.2 Strategy 2: Changing Dampening Values

In this strategy, dampening value (d) can change in each wave. After selection of initial sample with a conventional design, selection probability of a new

node i in wave k is

$$q_{ki} = d(k, a_k, y_{a_k}, w_{a_k}) \frac{w_{a_k i}}{w_{a_k +}} + \{1 - d(k, a_k, y_{a_k}, w_{a_k})\} \frac{1}{N - n_{s_{ck}}}$$

where $d(k, a_k, y_{a_k}, w_{a_k})$ depends on values related to nodes and links in the active set or changing as sample selection progresses.

4.3 Strategy 3: Following Links that Originate from Nodes of High Interest with a Pre Specified Probability

In this strategy, value of $\theta_{H|d} \in [0, 1]$ is introduced tow eight the chances of tracing links that originate from nodes of high interest in the active set given that a node is to be traced from the active set. In network setting where nodes values are either 0 or 1, active set a_k is partitioned into the two set

$$a_{k0} = \{i \in a_k : y_i = 0\}, \qquad a_{k1} = \{i \in a_k : y_i = 1\}$$

After selection of the initial sample by a conventional design, new node i in wave k is selected with probability

$$q_{ki} = d \left\{ (\theta_{H|d}) \frac{w_{a_{k1} i}}{w_{a_{k1} +}} + (1 - \theta_{H|d}) \frac{w_{a_{k0} i}}{w_{a_{k0} +}} \right\} + (1 - d) \frac{1}{N - n_{s_{ck}}}$$

where $w_{a_{k1} i}(w_{a_{k0} i})$ is the number of links from nodes of high interest (low interest) in the current active set out to node i, $w_{a_{k1} +}(w_{a_{k0} +})$ is the total number of links from nodes of high interest (low interest) in the current active set out to members not in the current sample and $n_{s_{ck}}$ is the size of current sample. In the event that there are only nodes of one type in the current active set, a random jump is taken.

In the spatial setting, link tracing is followed with probability proportional to the originating value and $\theta_{H|d} = 1$.

5 Estimators for Adaptive Web Sampling

5.1 Preliminary Estimators

i) Population Mean Estimator Based on Initial Sample Mean

Suppose an initial sample s_0 is selected and inclusion probability of unit i in the initial sample is π_i , then the Horvitz-Thompson (1952)

unbiased estimator for population mean is

$$\hat{\mu}_{01} = \frac{1}{N} \sum_{i \in s_0} \frac{y_i}{\pi_i}.$$

ii) Population Mean Estimator Based on Conditional Selection Probabilities

This estimator is a type of Hansen-Hurwitz (1943) estimator and is expressed as a weighted mean from the initial sample mean and selection probabilities of nodes value step by step in each wave. If $\hat{\tau}_{s_0} = \sum_{i \in s_0} \frac{y_i}{\pi_i}$ be the population total estimator based on the initial sample and $z_i = \sum_{j \in s_{ck}} y_j + \frac{y_i}{q_{ki}}$ bean unbiased estimator of population total that is defined for each node selected after the initial sample, then an unbiased estimator of population mean for $0 \leqslant d < 1$ is

$$\hat{\mu}_{02} = \frac{1}{Nn} \left[n_0 \hat{\tau}_{s_0} + \sum_{i=n_0+1}^{n} z_i \right]$$

where n_0 is the initial sample size. This estimator is not unbiased when $d = 1$.

iii) Composite Conditional Generalized Ratio Estimator

A composite conditional generalized ratio estimator is formed as the ratio of the two conditional probability-based estimators:

$$\hat{\mu}_{03} = \frac{N}{\hat{N}} \hat{\mu}_{02}$$

where $\hat{N}$ is the weighted mean of $\hat{N}_0 = \sum_{i \in s_0} \frac{1}{\pi_i}$ and $\hat{N}_i = n_{s_{ck}} + \frac{1}{q_{ki}}$ and computed as follows:

$$\hat{N} = \frac{1}{n} \left[n_0 \hat{N}_0 + \sum_{i=n_0+1}^{n} \hat{N}_i \right].$$

iv) Composite Conditional Mean of Ratio Estimator

This estimator is equal to:

$$\hat{\mu}_{04} = \frac{1}{n}\left[\frac{n_0}{\hat{N}_0}\hat{\tau}_{s_0} + \sum_{i=n_0+1}^{n}\frac{z_i}{\hat{N}_i}\right]$$

where the symbols defined as before.

5.2 Improved Estimators

Based on the Rao-Blackwell theorem, preliminary estimators can be improved by finding the conditional expectation of this estimator given the minimal sufficient statistics $d_r = \{(i, y_i, w_{i+}, w_{ij}) : i, j \in s\}$. Let μ_0 be the preliminary estimator, then the improved exact Rao-Blackwell estimator is

$$\hat{\mu}_i = E(\hat{\mu}_{0i}\,|d_r\,) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{0i}\Pr(\mathbf{s}\,|d_r\,) \qquad i = 1, 2, 3, 4$$

where the conditional distribution is

$$\Pr(\mathbf{s}\,|d_r) = \frac{\Pr(\mathbf{s})}{\sum\limits_{\mathbf{s}:r(\mathbf{s})=s}\Pr(\mathbf{s})}$$

and r is the function that reduced reordering of the sample s to the set s of distinct elements.

5.3 Variance Estimators and Confidence Intervals

The variance estimator of first preliminary estimator, while the initial sample of size n_0 is selected by simple random sample without replacement, is

$$\hat{V}(\hat{\mu}_{01}) = \frac{(N - n_0)v_0}{Nn_0}$$

where v_0 is the sample variance of the initial sample and $\frac{(N-n_0)}{Nn_0}$ is finite population correction.

The variance estimator of second preliminary estimator if $v_1 = \frac{(N-n_0)v_0}{Nn_0}$ and $v_2 = \sum\limits_{i=n_0+1}^{n}\frac{(z_i-\bar{z}_2)^2}{(n-n_0)(n-n_0-1)N^2}$ where $\bar{z}_2 = \sum\limits_{i=n_0+1}^{n}\frac{z_i}{(n-n_0)}$ is

$$\hat{V}(\hat{\mu}_{02}) = \left(\frac{n_0}{n}\right)^2 v_1 + \left(\frac{n - n_0}{n}\right)^2 v_2.$$

In general, variances of improved estimators $\hat{\mu}_i = E(\hat{\mu}_{0i} \,|d_r\,)$ with using the conditional decomposition of variances is

$$V(\hat{\mu}_i) = V(\hat{\mu}_{0i}) - E\left[V(\hat{\mu}_{0i} \,|d_r)\right] \qquad i = 1, \ldots, 4$$

An unbiased estimator of this variance is

$$\hat{V}(\hat{\mu}_i) = E\left[\hat{V}(\hat{\mu}_{0i}) \,|d_r\right] - V(\hat{\mu}_{0i} \,|d_r)\,.$$

Approximate $(1-\alpha)100\%$ confidence intervals for the estimator μ can be constructed with the familiar formula

$$\hat{\mu} \pm t_{m-1,\alpha/2}\sqrt{\hat{V}(\hat{\mu})}$$

Where $t_{m-1,\alpha/2}$ is the upper $\alpha/2$ point of the Student t distribution with $(m-1)$ degrees of freedom (according to the central limit theorem).

6 Markov Chain Monte Carlo for Resampling Estimators

For small sample sizes, Rao Blackwell estimators are computationally feasible. For larger sample sizes, the numbers of permutations or combinations of potential selection sequences in the conditional sample space become prohibitively large for the exact, enumerative calculation. For this reason, a resampling approach is utilized as a general method for obtaining estimates with designs of these types. Suppose x be a point in the conditional sample space. In this case, x is typically a permutation of the n units selected for the sample and the sample space consists of all possible permutations. The estimator $\hat{\mu} = E(\hat{\mu}_0 \,|d_r)$ can be written as $\hat{\mu} = \sum_x \hat{\mu}_0(x) Pr(x \,|d_r)$, where x is a point in the sample space, $\Pr(x \,|d_r)$ is the probability of selecting x with the given sampling design conditional on the realized value d_r of the minimal sufficient statistic, and the sum is over all points in the sample space. One way to obtain a sample s_r of permutations x from the conditional distribution $\Pr(x \,|d_r)$ is through a Markov chain accept/reject procedure (Hastings, 1970).

The object is to achieve a Markov chain $x_0, x_1, x_2, \ldots$ having stationary distribution $\Pr(x \,|d_r) = \Pr(\mathbf{s} \,|d_r)$. The Markov chain is obtained as follows:

Step 0: The chain is started by setting $x_0 = s$, where s is the original ordered sample. Assume that at the previous step, $(k-1)$, the value for x_{k-1} is some permutation j, so that j denotes the current permutation of the sample data in the chain.

Step 1: An experimental permutation t_k is produced from the candidate distribution, where the candidate distribution p_t consists of all permutations of the original sample s obtained by applying the same sampling design with sample size n, as if the population was comprised only of these members $(N = n)$.

Step 2: Let $\alpha = \min\left\{\left[\frac{p(t_k)}{p(x_{k-1})}\right]\left[\frac{p_t(x_{k-1})}{p_t(t_k)}\right], 1\right\}$. With probability α, t_k is accepted and take $x_k = t_k$ and with probability $(1-\alpha)$, t_k is rejected and take $x_k = x_{k-1}$. Return to step 1.

After creating a large number of resampled permutations (denoted n_r), let $\hat{\mu}_{0j}^{(k)}$ denote the value of the jth estimator on the kth resample. An enumerative estimator of the form $\hat{\mu} = E(\hat{\mu}_0)$ is replaced by the resampling estimator

$$\tilde{\mu}_j = \frac{1}{n_r}\sum_{k=0}^{n_r-1}\hat{\mu}_{0j}^{(k)} \qquad j = 1, 2, 3, 4$$

$$\tilde{V}(\hat{\mu}_{0j}\,|d_r) = \frac{1}{n_r}\sum_{k=0}^{n_r-1}\left\{\hat{\mu}_{0j}^{(k)} - \tilde{\mu}_j\right\}^2$$

$$\tilde{E}\left\{\hat{V}(\hat{\mu}_{0j})\,|d_r\right\} = \frac{1}{n_r}\sum_{k=0}^{n_r-1}\hat{V}\left\{\hat{\mu}_{0j}^{(k)}\right\}$$

Then variance estimation of $\tilde{\mu}_j$ is

$$\hat{V}(\tilde{\mu}_j) = \tilde{E}\left\{\hat{V}(\hat{\mu}_{0j})\,|d_r\right\} - \tilde{V}(\hat{\mu}_{0j}\,|d_r)$$

7 General Inverse Adaptive Sampling

Given a population $U = \{u_1, \ldots, u_N\}$ of N units, where N is known, let y_i denote the y-value associated with u_i, $i = 1, \ldots, N$. The population is

divided into two subpopulations by $P_M = \{u : y_i \in C, i = 1, \ldots, N\}$ and $P_{N-M} = \{u : y_i \notin C, i = 1, \ldots, N\}$ according to whether the y-values satisfy, $C = \{u_i : y_i \geqslant c\}$ where $M = |P_M|$ is the unknown number of units, or cardinality, of P_M. If a sample unit satisfies condition C, the network to which it belongs is sampled. The final sample consists of the original sample taken by general inverse, together with the units belonging to the network of the k sequentially sample units that satisfied condition C, and the edge units of those networks. The stopping criteria of reaching k units in C or hitting the ceiling of n_2 units in total are just applied to the initial sample. In this design, an unbiased estimator of total population is

$$\hat{Y}_{IA} = \begin{cases} \frac{N}{n_0} \sum_{i=1}^{n_0} \bar{y}_i^* & n = n_0 \\ \hat{M}\bar{y}_M^* + (N - \hat{M})\bar{y}_{N-M}^* & n_0 < n < n_2, |S_M| = k \\ \frac{N}{n_2} \sum_{i=1}^{n_2} \bar{y}_i^* & n = n_2, |S_M| < k \end{cases}$$

where $\bar{y}_i^* = \frac{1}{m_i} \sum_{j \in A_i} y_j$, A_i is the set of index in a network that unit i belong to this, m_i is the number of units in A_i. S_M and S_{N-M} are the index set of units in sample of s which are members of P_M and P_{N-M}, respectively.

The unbiased estimator of variance is given by

$$\hat{V}(\hat{Y}_{IA}) = \begin{cases} N^2 \left(1 - \frac{n_0}{N}\right) \frac{s_0^{2*}}{n_0} & n = n_0 \\ a s_M^{2*} + \hat{V}(\hat{M}) \left(\bar{y}_M^* - \bar{y}_{N-M}^*\right)^2 + b s_{N-M}^{2*} & n_0 < n < n_2, |S_M| = k \\ N^2 \left(1 - \frac{n_2}{N}\right) \frac{s_2^{2*}}{n_2} & n = n_2, |S_M| < k \end{cases}$$

where

$$s_r^{2*} = \frac{1}{n_r - 1} \sum_{i=1}^{n_r} (\bar{y}_i^* - \bar{y}_r^*)^2, \qquad \bar{y}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} y_i^*$$

$$s_M^{2*} = \frac{1}{k - 1} \sum_{i \in S_M} (\bar{y}_i^* - \bar{y}_M^*)^2,$$

$$s_{N-M}^{2*} = \frac{1}{n - k - 1} \sum_{i \in S_{N-M}} (\bar{y}_i^* - \bar{y}_{N-M}^*)^2$$

$$a = \frac{\hat{M}^2}{k}\left\{\frac{(N-n+1)(nk-n-k)-N(n-2)}{N(n-2)(k-1)}\right\}$$

$$b = \left\{\frac{N(N-n+1)(n-k-1)}{(n-1)(n-2)}\right\}$$

$$\hat{V}(\hat{M}) = N\frac{(n-k)(k-1)(N-n+1)}{(n-1)^2(n-2)} = \left(1-\frac{n-1}{N}\right)\frac{\hat{M}(N-\hat{M})}{n-2}$$

for $r = 0, 2$ (Salehi and Siber, 2004).

8 An Application

According to the 1382 Iran census data of agriculture, which was conducted by the Statistical Center of Iran, the number of olive producing townships are small relative to the total number of townships in Iran. Regarding olive as a rare crop, we are estimating its total production in Iran in 1382 using adaptive web sampling design. To compare adaptive web sampling with one of the designs used of rare populations sampling, we are conducting general inverse adaptive sampling on same data.

In this section, a simulation study is used to estimate total olive production in Iran in 1382 by different designs of adaptive web sampling and general inverse adaptive sampling. Simulations and calculations of this section are performed using SAS software.

8.1 Survey Design

To conduct adaptive web sampling and general inverse adaptive sampling to estimate total olive production in Iran in 1382, Iran map is divided into the 684 units of equal-sized rectangles. Each of the 684 units is a sampling unit in this design; therefore population size is $N = 684$. Each unit consists of one or more township. Area of each sampling unit is about 2613 km^2. Study variable (y) is olive production (in thousand tons) and estimated quantity is total olive production in Iran in 1382. Total olive production for each population unit is specified on the map (Figure 2). Figure 2, shows that 54 units of 684 units in the population have the characteristic of interest (olive production). Consequently, $p = \frac{54}{684} = 0.0789$ i.e. about %8 of the population units had olive production.

Sampling method in this study is as follows: 30 provinces in Iran (at that time) are considered as 30 strata and then an initial sample of rectangle areas with stratified random sampling of size n_0 is selected so that $n_0 = \sum_{h=1}^{30} n_h$, where n_h is the number of selected sample units from stratum h. After selecting the initial sample, at every sampling wave, a unit is selected at random with probability d amongst all links and with probability $(k-1)$ from the remaining units that are not in the sample and is added to the initial sample till final sample size equals n. We generated 1000 bootstrap samples and total olive production in Iran in 1382 is estimated for two methods, preliminary and MCMC, by four estimators, population mean estimator based on initial sample mean, population mean estimator based on conditional selection probabilities, composite conditional generalized ratio estimator, and composite conditional mean of ratio estimator.

In a pre-test, to determine the best initial sample size in strata by strategy 1, stratified variance estimation of total olive production obtained and then the initial sample size in strata was calculated assuming equal cost of gathering information for each sampling unit. Also estimate of total olive production is obtained using general inverse adaptive sampling, with $k = 7$ and final sample sizes are used in adaptive web sampling design.

8.2 Comparison Criteria

8.2.1 Bootstrap Estimate of Absolute Relative Bias

In each strategy, a stratified random sample of size $n_0 = \sum_{h=1}^{30} n_h$ was selected from the population and then final sample size was set to n. From the sample, bootstrap samples were generated and four each bootstrap sample the total olive production and it's bias were calculated. Suppose $\hat{t}_b$ is estimate of total olive production in replication b, $b = 1, 2, \ldots, 1000$. We calculated the bootstrap total olive production, variance, and the mean squared error as follows:

$$\bar{\hat{t}} = \frac{1}{1000} \sum_{b=1}^{1000} \hat{t}_b$$

$$\hat{V}(\hat{t}) = \frac{1}{1000-1} \sum_{b=1}^{1000} \left(\hat{t}_b - \bar{\hat{t}} \right)^2$$

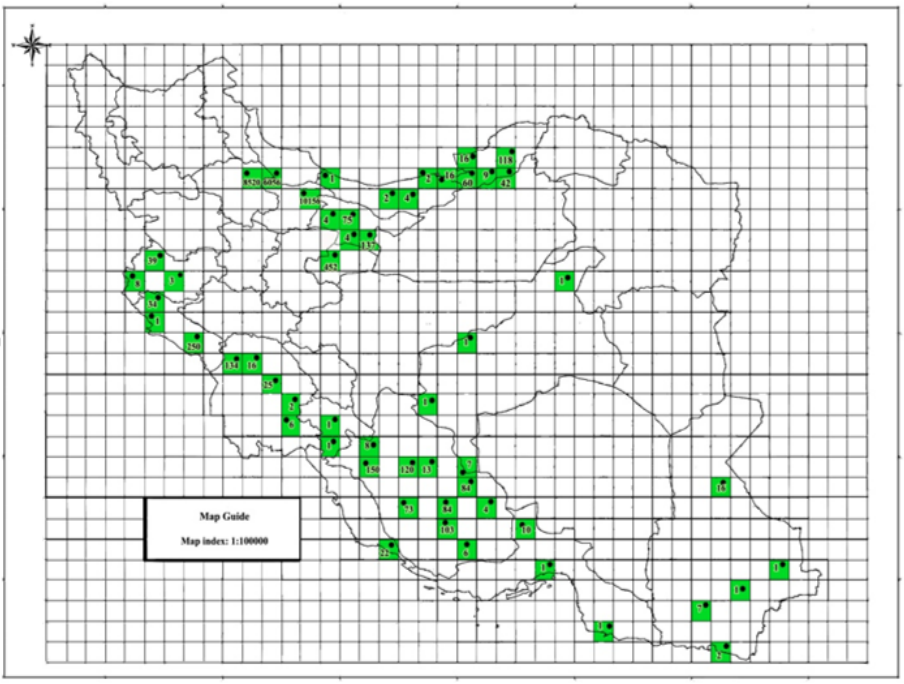

Figure 2. Scattering of productive units of olive in Iran.

$$\widehat{MSE}(\bar{\hat{t}}) = \hat{V}(\hat{t}) + \left\{Bias(\bar{\hat{t}})\right\}^2$$

where $Bias(\bar{\hat{t}}) = \bar{\hat{t}} - t$ and $t = 27.087$ thousand tons.

Therefore bootstrap estimate of absolute relative bias of total olive production is

$$\left|ARB(\bar{\hat{t}})\right| = \frac{\left|Bias(\bar{\hat{t}})\right|}{t} = \frac{\left|\bar{\hat{t}} - t\right|}{t}$$

where t is the known total production of olive in Iran in 1382.

8.2.2 Estimated Relative Mean Squared Error

To compare four estimators in three strategies, we use estimated relative mean squared error:

$$\widehat{RMSE} = \frac{\widehat{MSE}(\bar{\hat{t}})}{\hat{t}^2}.$$

8.3 Determining the Best Sample Size

To determine the best initial sample size in strata, %20 of sampling units in each stratum were randomly selected in a pre-test. With the initial sample size of $n_0 = \sum_{h=1}^{30} n_h = 151$, using strategy 1, $d = 0.9$ by selecting a sampling unit in each wave we come up with final sample size of 165 in pre-test. After selecting pre-test sampling units, variance of olive production was estimated by MCMC method for each four estimator and then the initial sample size was obtained by

$$n = \frac{z^2_{\alpha/2} s^2}{e^2 + \frac{z^2_{\alpha/2} s^2}{N}}$$

for $\alpha = 0.05$ and the tolerable error $e = \sqrt{0.3}$ (in thousand tons). Table 1 shows variance estimate (in 1 million tons) and the initial sample size for each of the four estimators.

Table 1. The sample size using each of the four estimators in pre-test

	Variance of Estimator	**Sample Size**
Estimator 1	1970	666
Estimator 2	2.71	34
Estimator 3	6.61	78
Estimator 4	11.2	122

To select the desired sample size, we considered the greatest value of Table 1 i.e. $n = 666$ but this was not an appropriate sample size (population size is 684). Therefore 122 was selected as the best sample size. Since variances were changing over strata and by assumption of equal survey cost in all strata, we allocated the total sample size $n = 122$ to strata using the following formula:

$$n_h = \frac{N_h s_h}{\sum_{\ell=1}^{H} N_\ell s_\ell} n.$$

Table 2 shows sample sizes in strata.

Table 2. Sampling strata with the stratum size and the sample size

State	Stratume Size	Sample Size
Elam	11	10
Tehran	8	5
Zanjan	9	9
Sistan and Balochestan	77	31
Fars	44	40
Ghazvin	5	5
Kermanshah	10	6
Golestan	11	11
Mazandaran	13	5

It should be noted that in pre-test we removed those selected sampling units in strata that their olive productions were zero.

8.4 Findings of the Simulation Study in Adaptive Web Sampling

8.4.1 Strategy 1 (Random Choice of Links)

As it was mentioned, in this strategy initially a stratified random sample of size $n_0 = 122$ is selected and then one node is added to the initial sample at each wave until the final sample size equals $n = 135$. The dampening values is held constant at 0.9.

Table 3 shows the preliminary and MCMC estimates of Iran total olive production (in thousand tons) in 1382 using this strategy for each of the four estimators.

Table 3 shows using MCMC method performs better than the preliminary method. The fourth estimator showed a larger amount of bias when compared to the other three estimators. The second estimator has the least $\sqrt{\widehat{MSE}}$ and $\widehat{RMSE}$ then performs better than the other three estimators. Therefore the best estimate of total olive production using the first strategy was obtained by the second estimator.

Table 3. Total production estimate of olive (in thousand tons), Bootstrap estimates of absolute relative bias, $\sqrt{\widehat{MSE}}$ and $\widehat{RMSE}$ of strategy 1 for $n_0 = 122$ and $n = 135$ using the preliminary and MCMC methods

		Strategy 1			
		Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary method	**Total production estimate of olive**	20.38	20.32	28.55	18.59
	Bootstrap estimate of absolute relative bias	0.25	0.25	0.05	0.31
	$\boldsymbol{\sqrt{\widehat{MSE}}}$	14.55	14.35	19.43	14.40
	$\boldsymbol{\widehat{RMSE}}$	0.51	0.49	0.46	0.60
MCMC method	**Total production estimate of olive**	20.58	20.39	28.59	18.57
	Bootstrap estimate of absolute relative bias	0.24	0.24	0.05	0.31
	$\boldsymbol{\sqrt{\widehat{MSE}}}$	14.52	13.21	17.63	13.69
	$\boldsymbol{\widehat{RMSE}}$	0.50	0.40	0.40	0.54

8.4.2 Strategy 2 (Changing Dampening Values)

In this strategy, as first strategy, one node is added to the initial sample ($n_0 = 122$) at each wave until the final sample size is set to $n = 135$. In strategy 2, the dampening values are changing at each wave.

Figure 3 shows two vectors of dampening values one for each strategy. The flat line at the value of zero corresponds with the selection of the initial sample.

Table 4 displays preliminary and MCMC estimates of Iran total olive production (in thousand tons) in 1382 using the second strategy for each of the four estimators.

According to Table 4 we see there is no improvement using strategy 2 in estimates compared to the first strategy because $\sqrt{\widehat{MSE}}$'s are greater than those for strategy 1. In this strategy, Using MCMC method performs better than the preliminary method and second estimator has the least value of $\sqrt{\widehat{MSE}}$ and $\widehat{RMSE}$ among four estimators, and shows better performance compared to the other estimators.

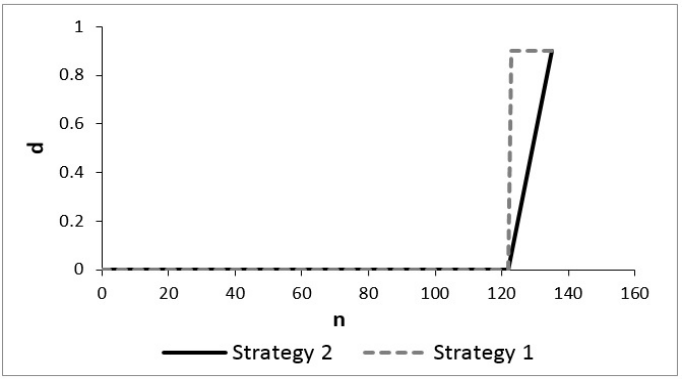

Figure 3. Dampening values of strategy 1 and 2 in different waves.

Table 4. Total production estimate of olive (in thousand tons), Bootstrap estimates of absolute relative bias, $\sqrt{\widehat{MSE}}$ and $\widehat{RMSE}$ of strategy 2 for $n_0 = 122$ and $n = 135$ using the preliminary and MCMC methods

		Strategy 2			
		Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary method	**Total production estimate of olive**	20.67	20.61	34.60	18.99
	Bootstrap estimate of absolute relative bias	0.23	0.24	0.28	0.29
	$\sqrt{\widehat{\boldsymbol{MSE}}}$	15.08	14.83	74.54	14.74
	$\widehat{\boldsymbol{RMSE}}$	0.53	0.51	0.46	0.60
MCMC method	**Total production estimate of olive**	20.87	20.62	34.03	19.01
	Bootstrap estimate of absolute relative bias	0.23	0.24	0.26	0.29
	$\sqrt{\widehat{\boldsymbol{MSE}}}$	15.03	13.50	22.02	13.83
	$\widehat{\boldsymbol{RMSE}}$	0.51	0.43	0.43	0.52

8.4.3 Strategy 3 (Following Links that Originate from Nodes of High Interest with a Pre Specified Probability)

In the third strategy, one node follows links with probability proportional to the originating node value and units with the greatest observed values are added to the initial sample. In this strategy similar to the first strategy sampling begins with the choice of a stratified random sample of size $n_0 = 122$ and then one node is added to the initial sample at each wave until the final sample size is set to $n = 135$. The dampening value is held constant at 0.9.

Table 5 shows the preliminary and MCMC methods of estimating Iran total olive production (in thousand tons) in 1382 using the third strategy for each of the four estimators.

Table 5. Total production estimate of olive (in thousand tons), Bootstrap estimates of absolute relative bias, $\sqrt{\widehat{MSE}}$ and $\widehat{RMSE}$ of strategy 3 for $n_0 = 122$ and $n = 135$ using the preliminary and MCMC methods

		Strategy 3			
		Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary method	**Total production estimate of olive**	20.22	20.24	18.96	17.31
	Bootstrap estimate of absolute relative bias	0.25	0.25	0.30	0.36
	$\boldsymbol{\sqrt{\widehat{MSE}}}$	14.82	14.58	15.29	14.86
	$\boldsymbol{\widehat{RMSE}}$	0.53	0.51	0.65	0.73
MCMC method	**Total production estimate of olive**	20.38	20.31	17.03	17.31
	Bootstrap estimate of absolute relative bias	0.24	0.25	0.37	0.36
	$\boldsymbol{\sqrt{\widehat{MSE}}}$	14.76	13.48	14.35	14.74
	$\boldsymbol{\widehat{RMSE}}$	0.52	0.44	0.56	0.72

Table 5 shows improvement in estimators using MCMC method. The fourth estimator showed a larger amount of bias when compared to the other three estimators. In this strategy, the second estimator has the least $\sqrt{\widehat{MSE}}$

and $\widehat{RMSE}$. We conclude that the second estimator has better performance relative to the other estimators.

8.5 Finding of Simulation Study using General Inverse Adaptive Sampling

In this sampling scheme, a stratified random sample of size $n_0 = 122$ was selected from the population. If number of units with olive production in initial sample were greater or equals 7, the sampling would be stop otherwise sampling was continuing until k units are selected with $C = \{y : y \geqslant 1\}$ or final sample size was set to $n = 135$. Then results are compared with results of adaptive web sampling and information of the 1382 agriculture census.

Table 6 shows the estimating Iran total olive production (in thousand tons) in 1382 using the general inverse adaptive sampling with estimates of best estimators in three strategies of AWS.

Table 6. Comparing performances of general inverse adaptive sampling with adaptive web sampling in estimating total olive production (in thousand tons)

	General Inverse	**Adaptive Web Sampling**		
	Adaptive Sampling	**Strategy 1**	**Strategy 2**	**Strategy 3**
Total production of olive estimate	19.62	20.39	20.36	20.21
Bootstrap estimate of absolute relative bias	0.29	0.24	0.24	0.25
$\sqrt{\widehat{MSE}}$	16.97	13.21	13.50	13.21
$\widehat{RMSE}$	0.78	0.40	0.43	0.40

Table 6 shows performances of the general inverse adaptive sampling estimates are not as good as those of AWS strategies.

9 Conclusion

When a sampling frame is not available and members of the population are stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames, that is the population is rare,

hidden or hard to reach, it is sometimes difficult to obtain required information. Link-tracing sampling designs are commonly used to draw samples from these populations. Adaptive web sampling is one of these designs.

In the study, we estimate total production of olive in Iran in 1382 by considering three sampling strategies and two estimation methods (preliminary and MCMC), 80 of total sample size was first selected and in each wave a unit was added to initial sample in all strategies until we achieved the final sample size. Total production of olive and relative mean squared error were estimated by each four estimators in each strategy.

In this study, MCMC method gave better estimators than the preliminary estimators in the sense of $\widehat{RMSE}$.

The use of dampening value vectors in strategy 2 improved some of the estimators. Estimators from the third strategy showed no significant improvements over those in strategy 1.

Findings of the simulations show that the performances of the AWS strategies in estimating 1382 total olive production in Iran are better than those computed from the general inverse adaptive sampling.

References

Frank, O. and Snijders, T. (1994). Estimating the Size of Hidden Populations using Snowball Sampling. *Journal of Official Statistics*, **10**, 53-67.

Hansen, M.M. and Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, **14**, 333-362.

Hastings, W.K. (1970). Monte Carlo Sampling Methods using Markov Chains and Their Applications. *Biometrics*, **57**, 97–109.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663–685.

Potterate, J.J., Woodhouse, D.E., Rothenberg, R., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993). AIDS in Colorado Springs: Is There Epidemic? *ADIS 7*, 1517-1521.

Salehi M., M. and Siber, G.A.F. (2004). A General Inverse Sampling Scheme and It's Application to Adaptive Cluster Sampling. *Australian and Zealand Journal of Statistics*, **43**, 483-494.

Smith, D.R., Conroy ,M.J., and Brakhage, D.H. (1995). Efficiency of Adaptive Cluster Sampling for Estimating Density of Wintering Waterfowl. *Biometrics*, **51**, 777-788.

Thompson, S.K. (1990). Adaptive Cluster Sampling. *JASA*, **85**, 1103-1115.

Thompson, S.K. (2004). Adaptive Web Sampling Difficult-to-reach Populations. *Proceedings of Statistics Canada Symposium*.

Thompson, S.K. (2006a). Adaptive Web Sampling. *Biometrics*, **62**, 1224-1234.

Thompson, S.K. (2006b). Target Random Walk Designs. *Survey Methodology*, **32**, 11-24

Vincent, K.S. (2008). *Design Variations in Adaptive Web Sampling*. A Project Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Department of Statistics and Actuarial Science, Simon Fraser University.

Xu, H. (2007). *Contributions to Adaptive Web Sampling Designs*. A Project Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Department of Statistics, Pennsylvania State University.

Ada Afshar
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *adaafa@gmail.com*

Hamidreza Navvabpour
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *h.navvabpour@srtc.ac.ir*