DESIGN VARIATIONS IN ADAPTIVE WEB SAMPLING

by

Kyle Shane Vincent

B.Sc (Hons), University of Winnipeg, 2006

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
Summer, 2008

APPROVAL

Name: Kyle Shane Vincent

Degree: Master of Science

Title of Project: Design Variations in Adaptive Web Sampling

Examining Committee: Dr. Brad McNeney
Chair

Dr. Steve Thompson
Senior Supervisor
Simon Fraser University

Dr. Charmaine Dean
Supervisor
Simon Fraser University

Dr. Derek Bingham
External Examiner
Simon Fraser University

Date Approved: July 25, 2008

Abstract

There is an increasing body of literature related to sampling for network and spatial settings. Although current link-tracing methods like adaptive cluster sampling, snowball sampling, and targeted random walk designs have advantages over conventional designs, some of the following drawbacks remain evident: there is a lack of flexibility in sample placement; there is an inability to control over sample sizes; and efficiency gains over conventional sampling designs for estimating population parameters may not be achievable. Adaptive web sampling (AWS) is a recently developed link-tracing design that overcomes some of these issues. Furthermore, the flexibility inherent to the AWS method permits many design variations. Using a simulated network population, an empirical population at risk for HIV/AIDS, a simulated spatial population, and an empirical population of birds, this project performs a simulation study to compare the performance of three variations of AWS strategies.

Keywords: Adaptive sampling, Link-tracing designs, Markov chain Monte Carlo, Network sampling, Rao-Blackwellization, Spatial sampling

Subject Terms: Hidden populations, Link-tracing designs, Markov chain Monte Carlo, Sampling

Acknowledgments

I have been extremely fortunate to have an excellent supervisor for the duration of my studies at Simon Fraser University. I would like to thank Steve Thompson for all of his guidance and patience. Without his support, this project would not have been possible.

I would like to thank all of the new friends I have made in the Department of Statistics at Simon Fraser University. The needed encouragement from all of my peers has given me the motivation to keep moving forward. In particular, I would like to thank Simon Bonner and Ryan Lekivetz, for all of the help in getting me prepared for my presentation.

I have two close friends that helped edit my project. The contributions that Kevin Georgison and Matt Richard have made has helped to increase the clarity of the ideas presented. Thank you for the time that you both put in to reviewing the write up.

Finally, a thank you to my parents for all of your support, both in the past and future.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With increasing societal concerns such as the prevalence of epidemics like HIV and environmental issues such as species extinction, improved sampling 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, snowball sampling, and targeted random walk designs 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; for instance, there may be little flexibility in where to allocate sampling effort, limited control over how much effort is allocated to adaptive selections, and an inability to fix the final sample sizes in advance. Adaptive web sampling (AWS) is a new adaptive link-tracing design based method that overcomes some of these drawbacks. The AWS design is said to be adaptive since selection distributions of new members to be included in the sample depends on the observed variables of interest in the current sample.

AWS starts with the selection of an initial sample through some conventional sampling design. New members can be added to the sample by either tracing links from the members in the current sample or through a conventional sampling design. The choice of which links that are traced may depend on the current sample size, the status of the members in the

current sample, or the behavior of their relationships with members in the sample and/or to members outside of the sample, just to name a few possibilities. In many populations, relationships will have a tendency to only form between members that share similar characteristics. Hence, with a link-tracing design the probability of including new members of the population into the sample is likely to be distributed unevenly, and at face value the final sample will not be representative of the population. Estimators used in AWS designs compensate for uneven selection probabilities, which is achieved by making use of the inclusion probabilities of the sampled units at the time they were selected, to form unbiased or consistent estimators.

Rao-Blackwellization (RB) of the estimators involves averaging over paths that are consistent with the minimal sufficient statistic (m.s.s.). For small sample sizes, these estimators are computationally feasible. For large sample sizes, the number of possible permutations becomes prohibitively large for exact calculations. A Markov chain resampling method makes these computations feasible.

Most work in AWS has been done using a general design that randomly selects links to trace in order to add new units to the sample. This project considers some new variations of the AWS design that provide more flexibility over the general AWS design. With the new design variations, more flexibility in the sample selection will come from using two new features. The first allows for the choice of which sample selection steps will tend to use adaptive or conventional sampling. The second allows the choice for assigning different probabilities of following links that originate from different types of nodes in the current sample. This project is a simulation study which compares the efficiency of the estimators of the current versus the new designs.

The remainder of this chapter introduces some of the notation used in AWS designs as well as an example of a network and spatial population. In Chapter 2, the sampling setup for AWS and its estimators will be introduced, as well as the Markov chain Monte Carlo (MCMC) methods for estimating the Rao-Blackwellized estimators. Chapter 3 yields simulation studies comparing the various designs for a simulated network population, an empirical population at risk for HIV/AIDS, a simulated spatial population, and an empirical bird population. Chapter 4 summarizes the studies, provides conclusions, and a discussion for future research.

1.1 The Network setting

In a network setting, the units of the population are labelled $1, 2, ..., N$. For every unit i, there is an observable variable of interest y_i. In the general network setting, y_i can take on any numerical value. In more specific network settings, y_i can be an indicator variable where

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

One may declare a unit i to be a "unit of interest" if the unit possesses (or does not possess) some specific characteristic or trait. 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 (nodes). The variable w_{ij} may be a measure on the distance between the two individuals, how often they come into contact, or indicate if they are mutual friends.

In most general network settings, w_{ij} can also take on 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 shall define $w_{ii} = 0$ for all $i = 1, 2, ..., N$.

When collecting an adaptive web sample, it is assumed that for all units i and j in the sample, y_i, w_{ij}, and w_{i+}, are recorded. The variable w_{i+} is the number of links out from the node, and is referred to as the out-degree of node i.

Figure 1.1 illustrates a graphical representation of a simulated population with two types of members, in which all links between nodes are symmetric. One could think of the simulated population as being comprised of injection and non-injection drug users, where dark nodes represent the injection drug users. Links could potentially exist between members if they share drug using paraphernalia or come into sexual contact. Researchers have reported that drug using populations share many of the features that are seen in this type of simulated population (i.e. patterns of cluster activity, tendency of relationships between similar nodes, etc. (Hoff et al. (2002))). This simulated population, and an empirical population, will be further investigated in Chapter 3.

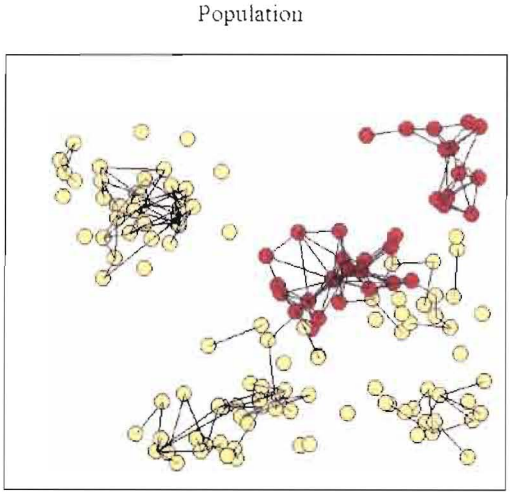


Figure 1.1: A simulated network population

1.2 The Spatial setting

A spatial setting can be depicted as a geographical area partitioned into single units. For example, in the simulated spatial population presented in Figure 1.2, each unit is represented by a square, and the y_i variables take on the count of the number of point-objects in the square.

For the spatial distribution, w_{ij} is defined to be

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

According to the structure of the population plots in Figure 1.2, 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.

In Figure 1.2, the graph representation of the simulated spatial population is presented on the right. This provides a visual representation of the units of interest and where one-way relationships exist. Units of interest consist of the plots that contain at least one point-object, and are represented by the dark nodes in the graph. One can think of the

spatial distribution of this simulated population as a population of species that exhibits clustering characteristics (like plants, fish, deer, or even human beings). Each square may represent an area of land, and the y_i values take on the count of the number of animals within the square. This simulated population, and an empirical population, will also be further investigated in Chapter 3.

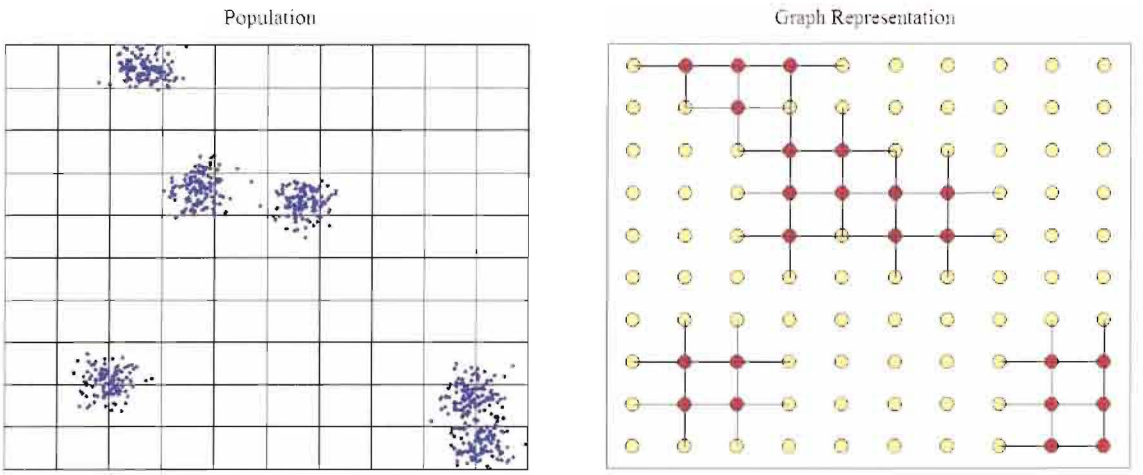


Figure 1.2: A simulated spatial population

Chapter 2

New Variations in Adaptive Web Sampling Designs

2.1 Adaptive Web Sampling

An adaptive web sample is selected in steps that begins by conventionally selecting an initial sample of size n_0 with probability p_0. Selection of new units to be included in the sample is said to occur in waves. For each wave k, selection of a new unit depends on a current active set $a_k \subseteq s_{ck}$, where s_{ck} is the current sample at wave k. Choice of the active set is flexible and may consist of all sampled nodes, or only nodes of higher interest, with links to units outside of the current sample. With the flexible choice of an active set, adaptive web sampling designs have an advantage over random walk designs in that the active set is not confined to the most previous unit selected.

Upon using a link tracing design, selection probabilities of new units to be included in the sample may be easily influenced by the nodes, and the behavior of their linkage tendencies, that were selected for the initial sample and in earlier waves. Consider, for example, a population under a network setting. Suppose an initial sample consists primarily of nodes that cluster together. If one were to always follow links to build up the sample, then this sample would consist entirely, or almost entirely, of nodes within the cluster(s) that were initially sampled from. This may be undesirable to the sampler and could potentially force the estimators to deviate away from their expected values. To help overcome this issue, AWS introduces the use of a mixture distribution in the selection probabilities of new units

to be included in the sample.

The mixture distribution is the convex combination of two probability distributions. The first is the probability distribution of adaptively selecting a new node as a function of the observed values in the active set. The second is the probability distribution of selecting a new node through a conventional design. The two component distributions of the mixture distribution are weighted with values d and $1 - d$, respectively. The value of d is referred to as the dampening value since its purpose is to "dampen" the out-degree of nodes in the active set. Values of d do not always have to be constant, but can depend on the wave or the active set. In AWS, flexibility comes not only from the choices of initial sample size and active set, but also the mixture distribution which gives the choice of how much effort is allocated towards tracing links.

Initial choices such as sample size and dampening values greatly influence the composition of the final sample. For initial samples that are of a large enough size, wide coverage of the population will tend to be immediately available and this will help render the final sample to be unbiasedly representative of the population. In contrast, with smaller initial sample sizes, larger choices of d will tend to make the final sample consist of more clustered nodes. For such cases, smaller choices of d may be necessary to obtain more coverage of the population, if desired. These two cases are exemplified with the two samples in Figure 2.1, where dampening values are held constant at 0.9 and the final samples are of size 20.

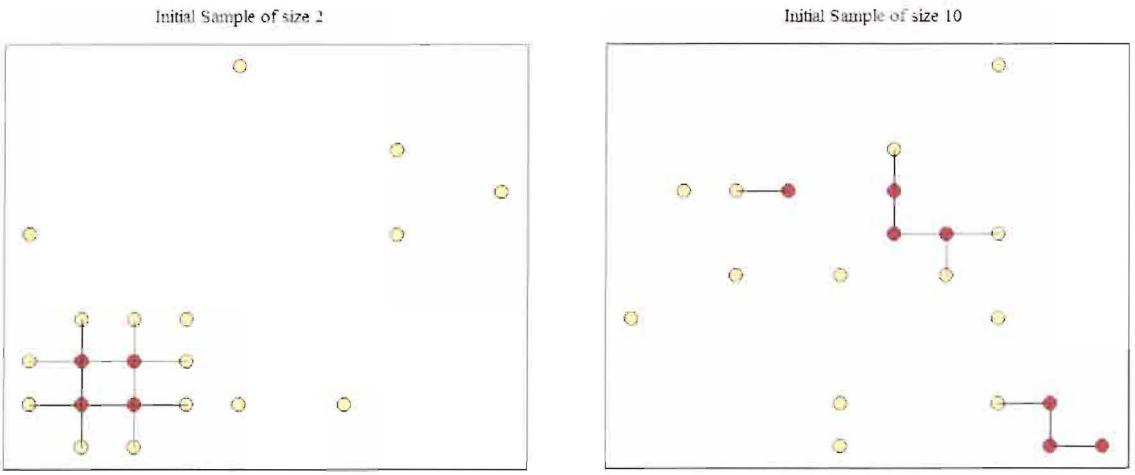


Figure 2.1: Two adaptive web samples from simulated spatial population

As can be seen in Figure 2.1, the sample obtained with the smaller initial sample size has selected every node in the cluster at the bottom left corner, while leaving all other five clusters unsampled. In contrast, the sample obtained with the larger initial sample size has observed at least one node from four of the six clusters. Hence, the AWS approach provides the advantage of giving the user the choice to either penetrate deep into the population by following many links, or going wide by following few links.

We now introduce some of the mathematical formulas behind the AWS design. Suppose that at wave k, the current sample s_{ck} contains some active set a_k. We shall let $q_k(s_k|a_k, y_{a_k}, w_{a_k})$ denote the design probability of selecting a new node at wave k. One general design makes use of selecting one node, i say, in each wave with probability proportional to the number of links from the active set out to node i (Thompson (2006a)). For this design, the conventional part of the mixture distribution is based on simple random sampling (SRS).

If we let q_{ki} denote the probability of selecting node i at step k, then when using this general design and sampling without-replacement, this selection probability is

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

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, $w_{a_k +}$ is the number of links from nodes in the active set out to members not in the current sample, and $n_{s_{ck}}$ is the size of the current sample. When sampling with-replacement

$$q_{ki} = d\frac{w_{a_k i}}{w_{a_k +}} + (1-d)\frac{1}{N},$$

where $w_{a_k i}$ is the number of links from the nodes in the active set out to node i, and $w_{a_k +}$ is the number of links from nodes in the active set out to any members of the population. In the event that there are no links at all out from the active set, then when sampling without-replacement

$$q_{ki} = \frac{1}{N-n_{ck}},$$

and when sampling with-replacement

$$q_{ki} = \frac{1}{N}.$$

To clarify the design previously described, suppose the adaptive web sample in Figure 2.2 is chosen without-replacement and is presently at some wave k. Then for the indicated node i, $q_{ki} = d\frac{2}{4} + (1-d)\frac{1}{N-6}$.

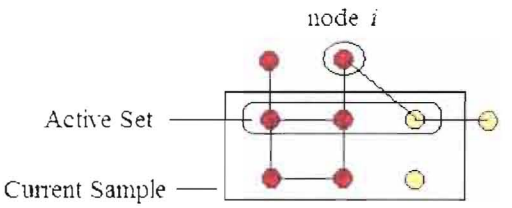


Figure 2.2: Example of inclusion probability of a node given the current active set

The overall selection probability of the ordered adaptive web sample $\mathbf{s}$, where one node is selected at each wave, is therefore

$$p(\mathbf{s}) = p_0 \prod_{i=n-n_0}^{n} q_i,$$

where n is the final sample size. Since only one node is selected at each wave, notation for the overall selection probability involving k is redundant since we have just reordered the n nodes in the sample to be the first n nodes in the population.

If one wishes to choose $n_k > 1$ nodes at some wave k, then we can denote the inclusion probability of the tth unit selected at sub-step t in wave k as q_{kt}. The selection probability at this time when sampling without-replacement for a node i not in the current sample is

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

and when sampling with replacement the selection probability for any node i in the population is

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

where the active links are similarly redefined for the sampling without-replacement or with-replacement cases as were presented earlier. Hence, the overall probability of selecting a sample $\mathbf{s} = (s_{1_1}, s_{1_2}, ..., s_{1_{n_1}}, ..., s_{K_1}, s_{K_2}, ..., s_{K_{n_K}})$ is

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

where K is the number of waves.

Since sampling can stop at anytime, K does not necessarily have to be fixed in advance. If the sampler feels that adequate coverage and information of the population has been obtained, they have the option to choose to stop sampling at the present wave. Adaptive web sampling therefore has an advantage over adaptive cluster and snowball sampling designs by not requiring every link from a sampled node to be followed, and hence sample sizes can be fixed in advance.

In AWS, selection probabilities for links to be traced can also be a function of observed link weights or depend on auxiliary variables associated with the sampled nodes in the active set. More specific variations like these may prove to be very useful for especially hard to reach populations (like those infected with HIV), where any pertinent information can be exploited to recruit members of higher interest to be included in the sample, when desired.

2.2 Description of New Variations of Adaptive Web Sampling

Three specific variations of AWS designs are examined in this study, the second two of which are new. Apart from being intuitively appealing, the two new strategies provide additional flexibility in the sample selection.

Strategy 1: Random choice of links

Most work in AWS has been done using constant dampening values in conjunction with Strategy 1. Strategy 1 was outlined in the previous section; an initial sample is selected conventionally and, at every wave, the choice to follow a link from a node in the active set is made with a constant probability d. If a link is to be traced to add a new node to the sample then it is chosen with uniform probability amongst all links that stem out from the active set.

Strategy 2: Changing dampening values

In this new strategy we allow for changing dampening values at each wave.

After selecting the initial sample with a conventional design, a new node i is chosen to be in the sample at wave k with probability

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

This strategy gives the user the ability to choose in advance which waves will tend to be used for adaptive or conventional sampling.

Strategy 3: Following links that originate from nodes of high interest with a pre specified probability

In this new strategy a pre specified value $\theta_{H|d} \epsilon (0, 1)$ is introduced to flexibly weight 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 the network setting where all nodes take on a value of either 0 or 1, we can partition the active set a_k into the two subsets a_{0k} and a_{1k}, where $a_{0k} = \{i \epsilon a_k : y_i = 0\}$ and $a_{1k} = \{i \epsilon a_k : y_i = 1\}$. After selecting the initial sample with a conventional design, a new node i is chosen to be in the sample at wave k with probability

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

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, and $w_{a_{k1}+}$ ($w_{a_{k0}+}$) 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. In the event that there are only nodes of one type in the current active set, a random jump is taken (i.e. a new node is selected completely at random).

In the spatial setting, a similar strategy has been developed by Thompson (2006a) where one follows links with probability proportional to the originating node value. In the network setting, this would be equivalent to setting $\theta_{H|d} = 1$. For example, suppose the adaptive web sample in Figure 2.3 is chosen without replacement, and the current active set consists of units 1, 3, and 6 at some current wave k. Then the probability of selecting unit 2 is $q_{k2} = d\frac{5+3}{5+5+3+1+1} + (1 - d)\frac{1}{9-3}$.

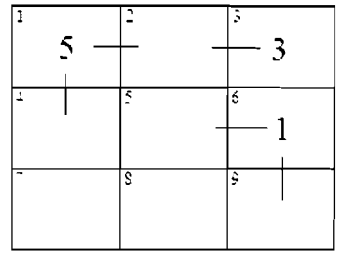


Figure 2.3: Example of Strategy 3 inclusion probability of a node in the spatial setting given the current active set

The motivation behind developing this strategy is to take advantage of the "like sticks with like" phenomenon (i.e. dark colored nodes tend to link to dark colored nodes and light colored nodes tend to link to light colored nodes) to build up a sample whose proportion of nodes of high interest has some relationship with the value of $\theta_{H|d}$.

2.3 The Minimal Sufficient Statistic for the Rao-Blackwell method

In the network setting, sample data consists of a set of nodes $s^{(1)}$ and a set of pairs of nodes $s^{(2)}$. In an adaptive web sample $s^{(1)}$ consists of the units i, in the order selected, and their observed y_i values. Note that, in the case of sampling with replacement that nodes may be repeated. The set $s^{(2)}$ consists of all w_{ij} values for each unit i in the sample and unit j in the population.

In the graph setting, the minimal sufficient statistic (m.s.s.) consists of the reduced data $d_r = \{(i, y_i, w_i+, w_{ij}) : i, j \epsilon s\}$ (Thompson and Seber (1996)). The minimal sufficient statistic will be used in conjunction with the Rao-Blackwell Theorem to form improved estimators. These improved estimators are described in the following section.

2.4 Estimators for Adaptive Web Samples

Four estimators have been developed for inference upon using an AWS design (Thompson (2006a)), and are described below. The primary motivation for developing the four estimators rests on the property that none produce uniformly lower mean square errors, due to the incompleteness of the m.s.s. for design-based sampling in the finite population setting (Thompson (2002)).

1) Estimator Based on Initial Sample Mean

Suppose an initial sample s_0 is chosen where each unit i has some probability π_i of being included in the initial sample. Then the Horvitz-Thompson estimator

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

is an unbiased estimator for the population mean μ for all values of $0 \leq d \leq 1$. If the initial sample is chosen through a SRS design then $\hat{\mu}_{01}$ is the initial sample mean $\bar{y}_0$.

The Rao-Blackwell estimator can be formed by weighting estimates from all samples consistent with the m.s.s. against the conditional probability

$$p(\mathbf{s}|d_r) = \frac{p(\mathbf{s})}{\sum_{\mathbf{s}:r(\mathbf{s})=s} p(\mathbf{s})},$$

where r is the function that reduces reorderings of the sample $\mathbf{s}$ to the set s of distinct elements. The improved estimator can be expressed as

$$\hat{\mu}_1 = E(\hat{\mu}_{01}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{01}(\mathbf{s})p(\mathbf{s}|d_r).$$

For sampling without-replacement, the summation consists of the sample estimates from all $n!$ reorderings of the sampled nodes. For initial samples chosen with a simple random sample design, the expected value is over all $\binom{n}{n_0}$ combinations for the initial sample and the $(n - n_0)!$ reorderings of the nodes selected after the initial sample.

2) Estimator Based on Conditional Selection Probabilities

The second estimator is a Hansen-Hurwitz type estimator since it takes into account the selection probability for each node, at the wave it was selected, to be in the sample. We shall let $\hat{\tau}_{s_0} = \sum_{i \epsilon s_0} \frac{y_i}{\pi_i}$ be the unbiased estimator of the population total $\tau = \sum_{i=1}^{N} y_i$ based on the initial sample. For each node selected after the initial sample, we shall define $z_i = \sum_{j<i} y_j + \frac{y_i}{q_{ki}}$. Each z_i is also an unbiased estimator of the population total for values of $0 \leq d < 1$, and hence the composite estimator

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

is an unbiased estimator for the population mean μ for values of $0 \leq d < 1$. In the event that one chooses to only follow links ($d = 1$), $\hat{\mu}_{02}$ is not unbiased since not all the conditional selection probabilities are greater than 0. In this case, the z_i's will only estimate the total of the node values in the sample and those that are accessible, for which it is unbiased.

The second improved estimator

$$\hat{\mu}_2 = E(\hat{\mu}_{02}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{02}(\mathbf{s})p(\mathbf{s}|d_r),$$

is formed by carrying out estimates over every sample path consistent with the m.s.s.

In a specific network setting where y values only take on values of 0 or 1, the second estimator can take on values greater than 1. Ratio estimators can be used to help reduce the occurrence of such extreme values, of which two are described below.

3) Composite Conditional Generalized Ratio Estimator

A Horvitz-Thompson type unbiased estimator for the population size N can be formed by replacing each y_i value by 1 in the population total estimator $\hat{\tau}_{s_0} = \sum_{i \epsilon s_0} \frac{y_i}{\pi_i}$ to get $\hat{N}_0 = \sum_{i \epsilon s_0} \frac{1}{\pi_i}$. If the initial sample is chosen with a SRS design, then $\hat{N}_0$ reduces to N since $\pi_i = \frac{n}{N}$ for all units i in the initial sample.

Similarly, if we replace each y value found in the z_i's by 1, then these form unbiased estimates for the population size for values of $0 \leq d < 1$, and the expressions are equivalent to $\hat{N}_i = n_{ck} + \frac{1}{q_{ki}}$. Taking a weighted average of these estimates yields an unbiased estimator $\hat{N}$ for the population size where

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

and using $\hat{N}$ we can form the ratio estimator

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

Since $\hat{\mu}_{03}$ is a ratio estimator, there may be some bias in the estimates. However, the improved estimator

$$\hat{\mu}_3 = E(\hat{\mu}_{03}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{03}(\mathbf{s}) p(\mathbf{s}|d_r),$$

will have the same bias but variance as small or smaller than that for $\hat{\mu}_{03}$.

4) Composite Conditional Mean-of-Ratios Estimator

Ratio estimates for the population mean μ can be formed by dividing the initial sample estimator $\hat{\tau}_{s_0}$ by $\hat{N}_0$, and each z_i by $\hat{N}_i$. The mean of these ratio estimates form the estimator $\hat{\mu}_{04}$, where

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

The improved estimator is

$$\hat{\mu}_4 = E(\hat{\mu}_{04}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{04}(\mathbf{s}) p(\mathbf{s}|d_r).$$

2.5 Markov Chain Monte Carlo for resampling estimators

For estimating the RB estimators previously described, a Markov chain accept/reject algorithm is used (Hastings (1970),Thompson (2006a)). The resampling approach proves to be especially useful for large sample sizes, where the number of permutations is prohibitively large for exact calculation of the improved estimators.

We desire to obtain a Markov chain $x_0, x_1, x_2, ...$, having the stationary distribution $p(x|d_r) = p(\mathbf{s}|d_r)$ that was presented in the previous section. When sampling without replacement, x is a permutation of the original ordered sample s from the population. When sampling with replacement, x is a vector of length n whose elements reduce to the set that is consistent with the reduced set of the original sampled units.

A Markov chain that remains in the stationary distribution is obtained as follows:

Step 0: Start the chain in its stationary distribution by setting $x_0 = s$, where s is the original ordered sample. Suppose that at the previous step $k-1$, the value for x_{k-1} is some permutation j.

Step 1: Form a tentative permutation t_k from the candidate distribution, where the candidate distribution (denoted p_t) consists of all permutations of the original sample s obtained by applying the same sampling design to the n sampled members as if the population were comprised only of these members.

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

After making a large number of resampled permutations (n_r say), let $\hat{\mu}_{0j}^{(k)}$ be the value of the jth estimator on the kth resample. We can then estimate the population mean with the enumerative estimator

$$\tilde{\mu_j} = \frac{1}{n_r} \sum_{k=0}^{n_r-1} \hat{\mu}_{0j}^{(k)},$$

for each of $j = 1, 2, 3, 4$.

2.6 Variance estimators and confidence intervals

A recommended approach (Thompson (2006a)) to estimating the variance and obtaining confidence intervals of adaptive web sampling estimators is to start by selecting m independent samples. With an estimate of $\hat{\mu}_k$ for the population mean from sample k, we can estimate μ with $\hat{\mu} = \sum_{k=1}^{m} \frac{\hat{\mu}_k}{m}$. An unbiased estimator for the variance of μ is

$$\hat{V}(\hat{\mu}) = \sum_{k=1}^{m} \frac{(\hat{\mu}_k - \hat{\mu})^2}{m(m-1)}.$$

Approximate $100(1-\alpha)\%$ confidence intervals for the estimator $\hat{\mu}$ 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's t distribution with $m-1$ degrees of freedom.

Chapter 3

Simulation Experiments

The previous chapter covered the sampling strategies and their estimators. In this chapter, simulation studies are compared between the AWS designs previously presented for a simulated network population, a population at risk for HIV/AIDS, a simulated spatial population, and a bird population.

3.1 Network populations

3.1.1 Simulated network population (Population 1)

The population in Figure 3.1 (Population 1) was presented in Chapter 1, and was simulated using the Adaptive Network Sampling package in R. The links between individuals follow a logistic distribution that is based on their status and distance. Many researchers have reported that modelling network populations with this type of setup has helped to capture the true network structure of the population. For a further discussion on network modelling, see Chow and Thompson (2003), Frank and Thompson (2000), Hoff et al. (2002), and Linkletter (2007).

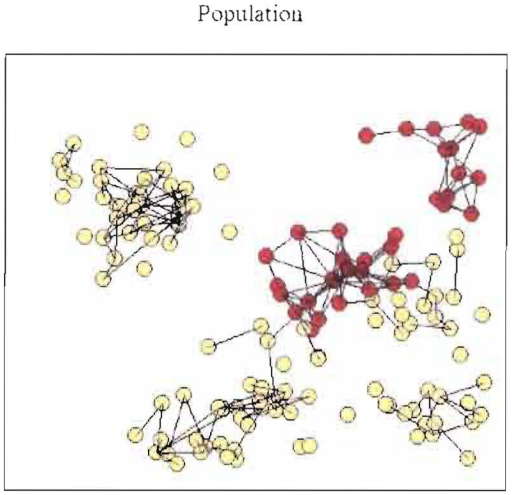


Figure 3.1: Population 1

Population 1 has 147 members. The dark colored nodes are classified as nodes of interest and take on a value of 1, and the light colored nodes take on a value of 0. Links between nodes are symmetric and take on a value of 1. Estimates of the proportion of dark nodes in the population (i.e. the population mean) are found in the following sections.

Strategy 1

The simulations of sampling with Strategy 1 used a without-replacement design. An initial sample of size 10 was selected through a SRS design, and the final sample size was set to 20. One node was selected at each wave, and the dampening values were held constant at 0.9. An illustrative example of a sample of this type can be found in Figure 3.2.

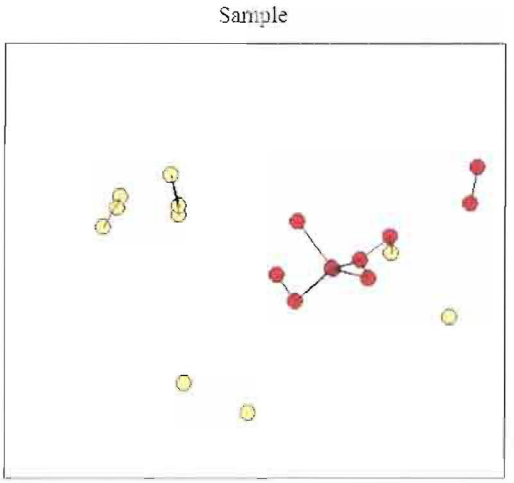


Figure 3.2: A Strategy 1 sample from Population 1

Table 3.1 gives the MSE scores for the AWS Strategy 1 design. A series of 2000 simulation runs were used. For each run, 10,000 resamples were used to obtain the MCMC estimates.

Table 3.1: MSE scores of Strategy 1 estimators for Population 1

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	0.0170	0.0324	0.0241	0.0294
Improved estimate	0.0136	0.0147	0.0147	0.0293

Properties of the sampling strategy and its estimators can be found in the histograms in Figure 3.3. The true population mean $\mu = 0.231$ is indicated by the solid triangle, and the approximate expectation of each estimator is indicated by the transparent triangle.

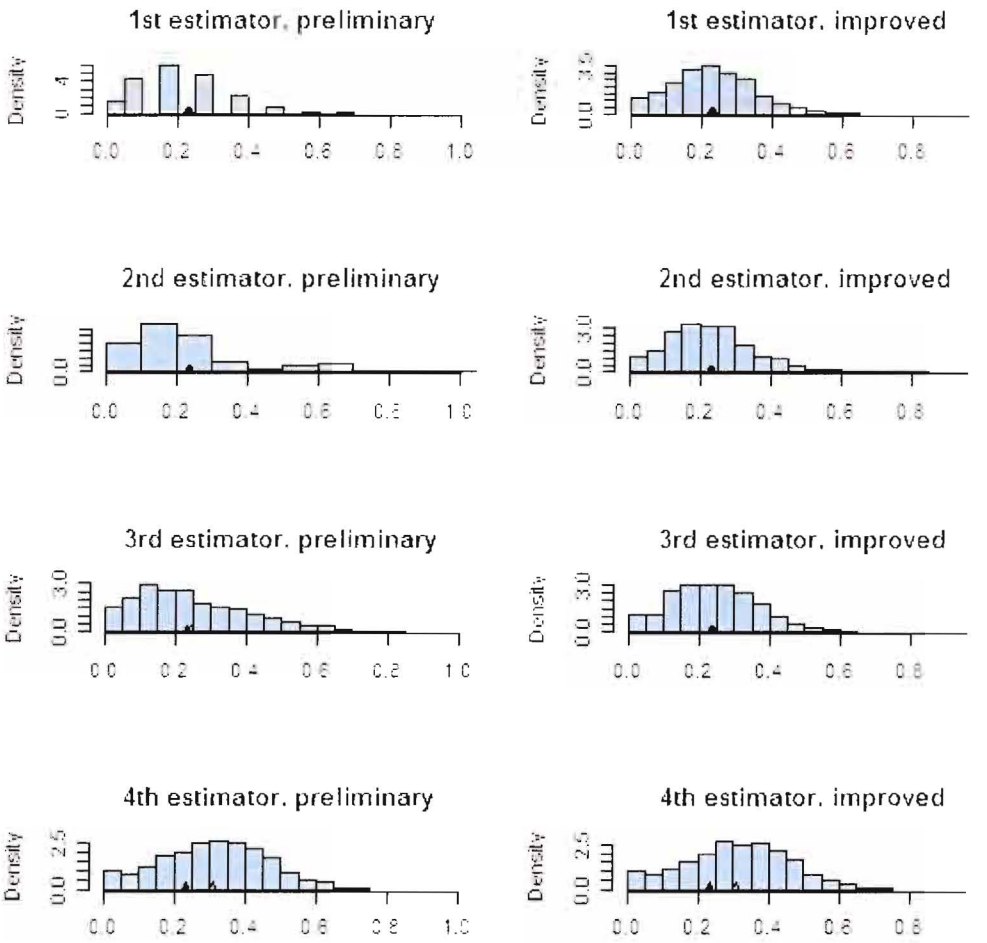


Figure 3.3: Histograms of Strategy 1 estimators for Population 1

Table 3.1 indicates that significant improvements have been made for the first, second, and third Rao-Blackwellized estimators, and the histograms show that there is more symmetry in their distributions. As expected, the first and second estimators appear to be unbiased. The fourth estimator shows a large amount of bias when compared to the third estimator, and this bias accounts for most of its MSE. Extreme values were taken on by the second estimator, and in some cases it took on values greater than one. Rao-Blackwellization of the second estimator helped minimize its range, while the ratio estimators show further improvement. The first estimator performed the best overall.

Strategy 2

The AWS Strategy 2 design makes use of changing dampening values. For this project, the pre chosen dampening vectors all summed to the same value. This ensured that the number of nodes that were randomly chosen, via the initial sample or a random jump at any wave after the initial sample, would be expected to be approximately equal.

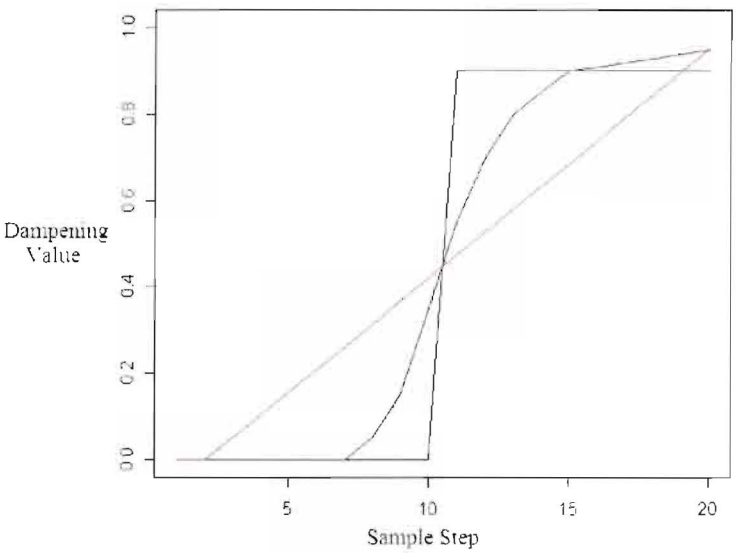


Figure 3.4: Strategy 2 dampening vectors used for Population 1

Three vectors were of interest, and are illustrated in Figure 3.4. The vectors associated with the black, blue, and red line correspond to Vectors 1, 2, and 3, respectively. The flat line at a value of zero corresponds with the simple random selection of the initial sample. Vector 1 was used in the Strategy 1 simulation study. The motivation behind choosing the values for Vector 2 was to allow for a smoother transition in the extreme values that are similarly used in Vector 1. Values for Vector 3 were chosen to allow for the freedom of potentially following links early in the sample selection.

Table 3.2 gives the MSE scores for the preliminary and improved estimators for the AWS Strategy 2 design. Bar charts for the MSE scores of the improved estimators for each vector are presented in Figure 3.5.

Table 3.2: MSE scores of Strategy 2 estimators for Population 1

		Vector 1	Vector 2	Vector 3
Estimator 1	Preliminary estimate	0.0170	0.0248	0.0836
	Improved estimate	0.0136	0.0138	0.0173
Estimator 2	Preliminary estimate	0.0324	0.0285	0.0190
	Improved estimate	0.0147	0.0130	0.0111
Estimator 3	Preliminary estimate	0.0241	0.0209	0.0159
	Improved estimate	0.0147	0.0135	0.0120
Estimator 4	Preliminary estimate	0.0294	0.0293	0.0316
	Improved estimate	0.0293	0.0291	0.0314

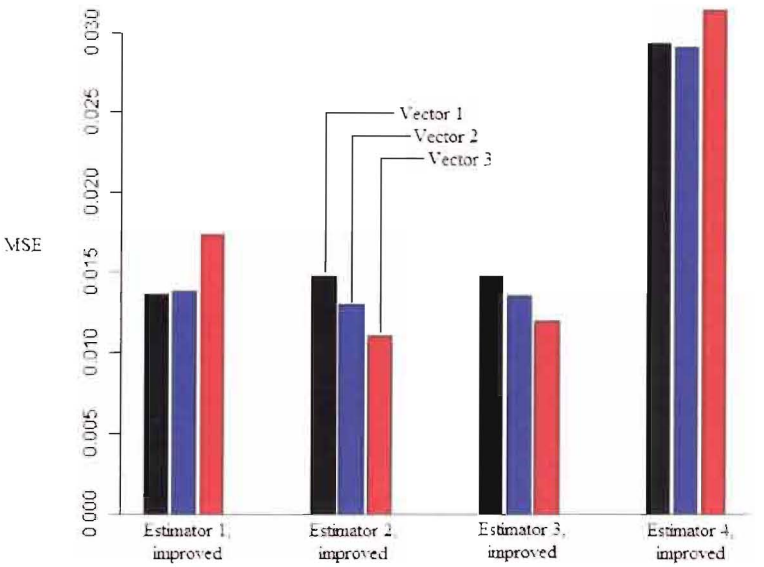


Figure 3.5: Bar charts of MSE scores of Strategy 2 improved estimators for Population 1

Table 3.2 and Figure 3.5 show that using Strategy 2 with the two changing dampening vectors improved the efficiency scores of the second and third estimators over those found in the Strategy 1 variation. Vector 3 was shown to perform the best overall. The larger MSE scores from the first estimators could be attributed to the smaller initial sample sizes that

were used with Vectors 2 and 3. As can be seen, there is almost no change in efficiency for the fourth estimator.

The most efficient estimate from the Strategy 1 design came from the first estimate, while the most efficient estimate from the Strategy 2 design came with the second estimate when Vector 3 was used. The relative efficiency of these two estimators was found to be $\frac{0.0111}{0.0136} = 0.816$.

Table 3.3 on the following page gives the estimated bias of the estimators and semi-length of confidence intervals, as well as the estimated coverage rates of the estimators from the Strategy 2 design for a final sample comprised of six independent samples. The bias scores have been rescaled for easier comparison purposes.

Table 3.3 indicates that most coverage rates were close to 95%. The weak coverage rate from the fourth estimator was accounted for by its large amount of bias. Coverage rates were slightly stronger for the Rao-Blackwellized estimates, which could be attributed to their more symmetric distributions upon comparison to their preliminary counterparts. As can be seen in the table, the coverage rate for the second preliminary estimator became successively stronger when using the dampening values from the second and third estimators. Bias scores for each estimator are approximately equal for all vectors and semi-lengths of the confidence intervals for the second and third estimators from Vector 3 were significantly smaller than those found in the Strategy 1 design.

Table 3.3: Bias, CI semi-length, and coverage scores of Strategy 2 estimators for Population 1 upon using six independent samples

Vector 1

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.101	0.130	0.947
	Improved estimate	-0.016	0.114	0.952
Estimator 2	Preliminary estimate	-0.188	0.164	0.889
	Improved estimate	-0.125	0.117	0.951
Estimator 3	Preliminary estimate	0.903	0.153	0.943
	Improved estimate	0.869	0.120	0.951
Estimator 4	Preliminary estimate	8.320	0.148	0.789
	Improved estimate	8.341	0.148	0.790

Vector 2

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	0.112	0.156	0.951
	Improved estimate	0.032	0.117	0.952
Estimator 2	Preliminary estimate	0.011	0.155	0.915
	Improved estimate	0.035	0.114	0.949
Estimator 3	Preliminary estimate	0.933	0.145	0.945
	Improved estimate	0.989	0.116	0.949
Estimator 4	Preliminary estimate	8.512	0.150	0.783
	Improved estimate	8.502	0.149	0.781

Vector 3

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.327	0.295	0.957
	Improved estimate	-0.103	0.133	0.951
Estimator 2	Preliminary estimate	-0.140	0.127	0.946
	Improved estimate	-0.139	0.107	0.949
Estimator 3	Preliminary estimate	0.417	0.125	0.947
	Improved estimate	0.394	0.110	0.954
Estimator 4	Preliminary estimate	8.492	0.158	0.803
	Improved estimate	8.500	0.157	0.800

Strategy 3

The Strategy 3 design provides more flexibility over the previous design by allowing the user to choose values of $\theta_{H|d}$ to weight the chances of selecting links that originate from nodes of higher interest. With this method, the user has some control over how much of the sample will be comprised of higher units of interest.

Figure 3.6 below provides plots of the MSE values for each estimator for values of $\theta_{H|d} = 0.05, 0.10, 0.15, ..., 0.90, 0.95$. Preliminary estimators are represented by the dashed line, and the improved estimators are represented by the solid lines. MSE scores for the Strategy 1 improved estimators are represented by the solid blue line.

Figure 3.6: MSE scores for Strategy 3 estimators for Population 1

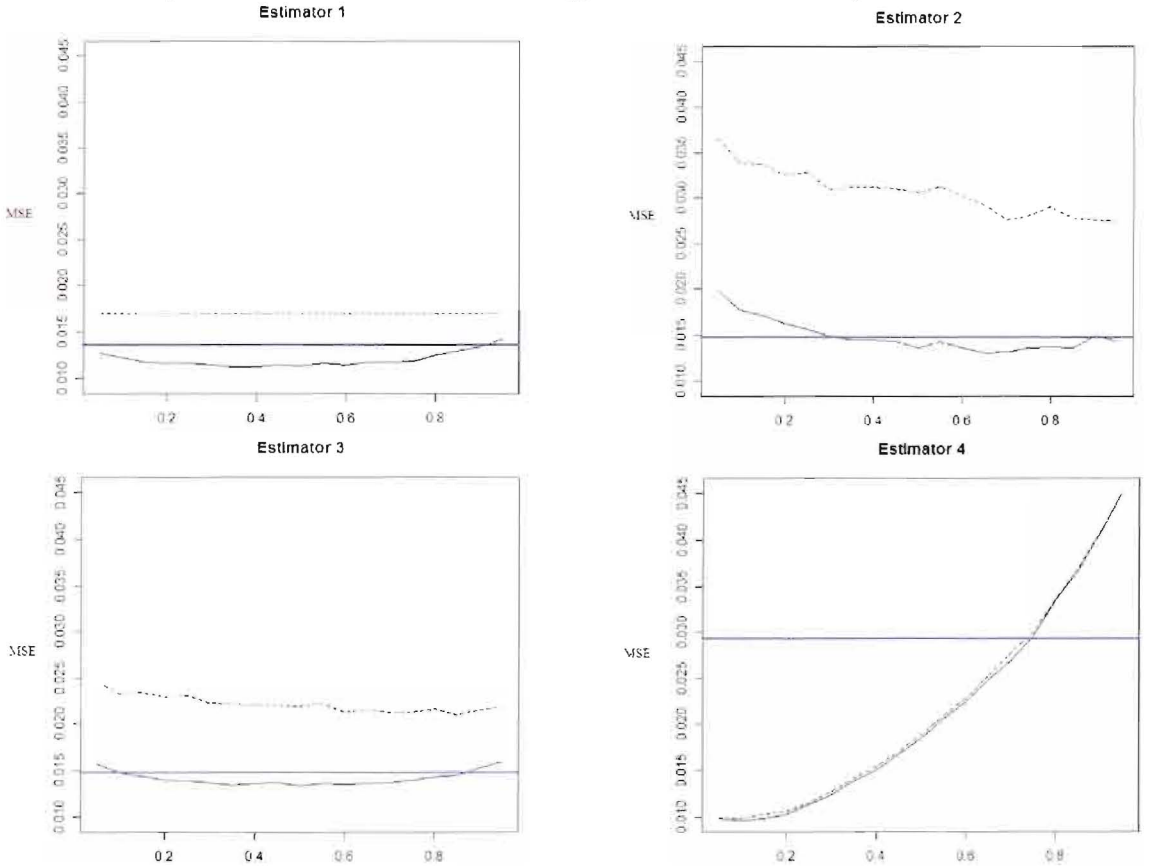


Figure 3.6 shows that some improvements have been made for improved estimators one and three for values of $\theta_{H|d}$ within close proximity of values slightly greater than $\mu = 0.231$, while little to no improvement was found for all values of $\theta_{H|d}$ for estimator two. Although Rao-Blackwellization of the fourth estimator provided very little improvement, significant improvements have been achieved over that found with the corresponding Strategy 1 estimator for values of $\theta_{H|d}$ within close proximity of the population mean $\mu = 0.231$. The smallest MSE score when using Strategy 3 came with the fourth estimator and a value of $\theta_{H|d} = 0.10$. The relative efficiency of the best estimate from Strategy 1 and this estimate was found to be $\frac{0.0095}{0.0136} = 0.699$.

The solid line in Figure 3.7 gives the expectation of the fourth estimator for the corresponding values of $\theta_{H|d}$. The sample mean is represented by the dashed line and the approximate expectation of the fourth estimator from the Strategy 1 design is represented by the blue line.

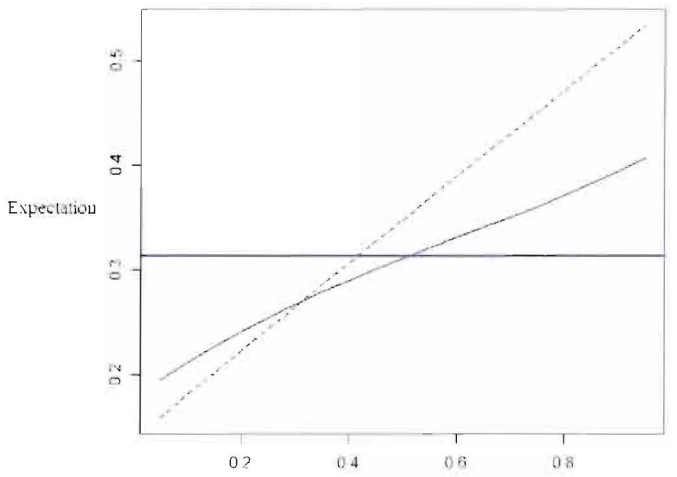


Figure 3.7: Expected values of fourth estimator from Strategy 3 for Population 1

As seen in Figure 3.7, the smaller MSE scores for estimator four, for values of $\theta_{H|d}$ close to the population mean $\mu = 0.231$, come from a large reduction in the amount of bias and some reduction in variability. The relationship between the expectation of the fourth estimator and the sample mean is visibly evident, and the correlation of the observed simulated values was found to be 0.991. The graph also shows that the expectation of the fourth estimator and the sample mean intersect at a value close to the population mean.

A simple property of this estimator, and adaptive web sampling in general, is seen in the graph: the values of the expectation of the estimator are almost always closer to the true mean than the sample mean, hence giving the intersection at this value.

One alternative strategy is to set $\theta_{H|d}$ equal to the initial sample mean. This may prove to be especially useful if one wishes to use Strategy 3 with $\theta_{H|d}$ as close to the true population mean as possible, but has no insight on what this value may be. Table 3.4 gives the estimated bias and semi-length of confidence intervals, as well as the estimated coverage rates of the estimators from setting $\theta_{H|d} = \mu$ and the initial sample mean. The final sample is comprised of six independent samples, and the bias has again been rescaled for comparison purposes.

Table 3.4: Bias, CI semi-length, and coverage scores of Strategy 3 estimators for Population 1 upon using six independent samples

$\theta_{H|d} = \mu$

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.002	0.130	0.944
	Improved estimate	0.495	0.107	0.947
Estimator 2	Preliminary estimate	-0.133	0.165	0.854
	Improved estimate	-0.004	0.124	0.942
Estimator 3	Preliminary estimate	0.740	0.148	0.935
	Improved estimate	0.684	0.119	0.950
Estimator 4	Preliminary estimate	1.821	0.104	0.942
	Improved estimate	1.878	0.102	0.939

$\theta_{H|d} = \bar{y}_{s_0}$

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.003	0.130	0.944
	Improved estimate	0.200	0.108	0.947
Estimator 2	Preliminary estimate	-0.234	0.164	0.863
	Improved estimate	-0.627	0.122	0.939
Estimator 3	Preliminary estimate	0.714	0.151	0.926
	Improved estimate	0.382	0.121	0.944
Estimator 4	Preliminary estimate	2.030	0.129	0.934
	Improved estimate	1.972	0.126	0.940

Table 3.4 suggests that bias, confidence interval semi-length, and coverage rate scores for the first, second, and third estimators are approximately the same for both choices of $\theta_{H|d}$. For the fourth estimator, very little bias is introduced with setting $\theta_{H|d}$ equal to the sample mean, and semi-lengths of the confidence intervals are about twenty-five percent larger. The coverage rates from both cases show they are almost identical. Further improvements could possibly be made if one were able to set $\theta_{H|d}$ to the mean of the initial sample means of the independent samples that will make up the final sample, if possible.

In comparison to the bias, confidence interval semi-lengths, and coverage rate scores from the Strategy 1 estimators, the scores seem to parallel those found with the first, second, and third estimators. Significant improvements have been made for each score with the fourth estimator from Strategy 3.

3.1.2 At risk for HIV/AIDS population (Population 2)

Population 2 is presented in Figure 3.8, and has 595 members. The empirical data set comes from the Colorado Springs Study on the heterosexual transmission of HIV/AIDS (Potterat et al. (1993), Darrow et al. (1999)). Dark nodes represent injection drug users, and the symmetric links between the nodes indicate drug-using relationships.

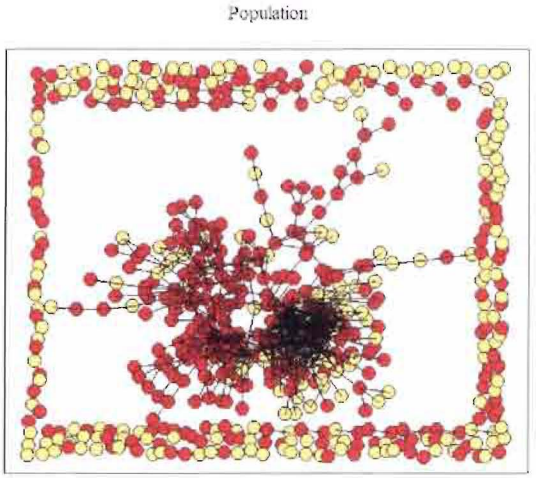


Figure 3.8: Population 2

For hidden populations of this type, the difficulty and cost of sampling with conventional methods is usually much greater than with a link-tracing design like adaptive web sampling. Using the AWS strategies, we wish to estimate the proportion of injection drug users (i.e. the population mean).

Strategy 1

The simulations of sampling from this population with Strategy 1 used a without-replacement design. An initial sample of size 15 was selected through a SRS design, and the final sample size was set to 30. One node was selected at each wave, and the dampening values were held constant at 0.9.

Table 3.5 gives the MSE scores for the AWS Strategy 1 design estimators. Again, a series of 2000 simulation runs were used, and 10,000 resamples were used on each run for the MCMC estimates.

Table 3.5: MSE scores of Strategy 1 estimators for Population 2

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	0.0160	0.0775	0.0306	0.0114
Improved Estimate	0.0154	0.0725	0.0281	0.0113

Properties of the sampling strategy and its estimators can be found in the histograms in Figure 3.9. The true population mean $\mu = 0.575$ is indicated by the solid triangle.

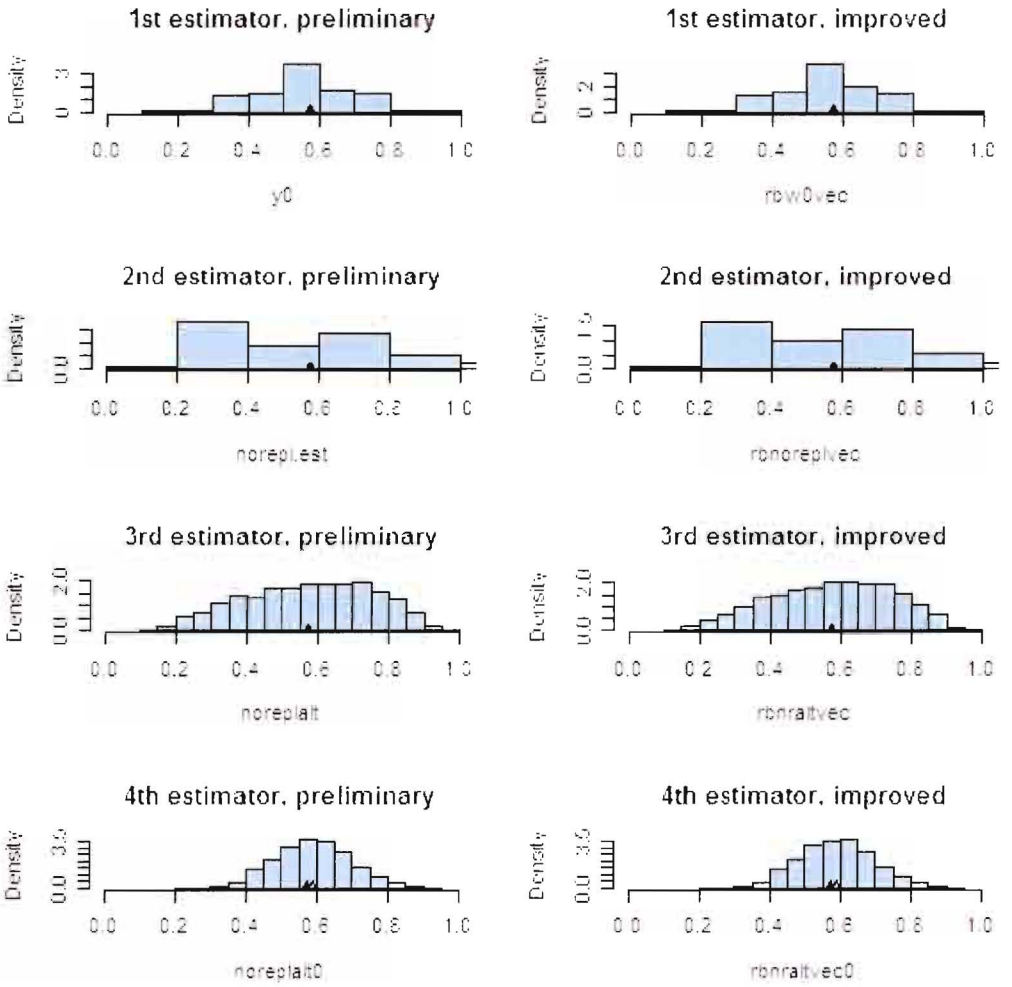


Figure 3.9: Histograms of Strategy 1 estimators for Population 2

Table 3.5 and Figure 3.9 show that Rao-Blackwellization of the preliminary estimators has slightly improved the MSE scores. Again, the bias is evident in the fourth estimator, and accounts for most of its MSE. For the ratio estimators, significant reductions in the MSE scores are evident over the second estimator. Recall that the fourth estimator performed the worst in Population 1 when using Strategy 1, and for Population 2 the fourth estimator has performed the best overall.

Strategy 2

Once again, three damp vectors are of interest for the AWS Strategy 2 design for Population 2. The plots of these vectors can be found in Figure 3.10, and they all sum to the same value.

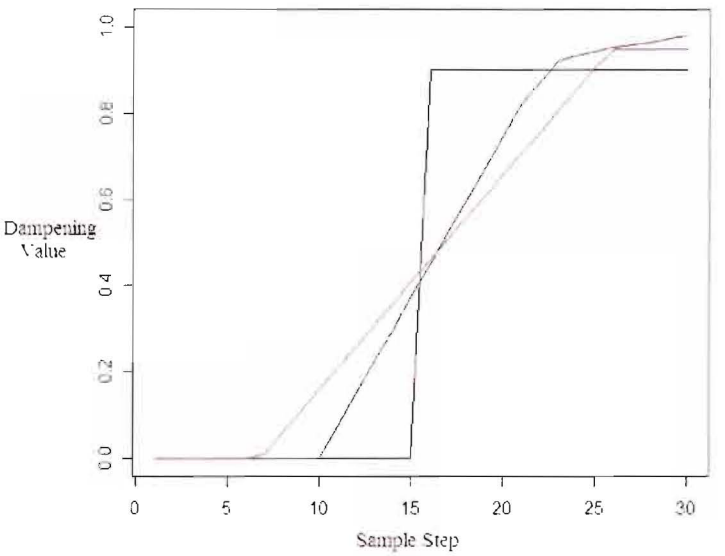


Figure 3.10: Strategy 2 dampening vectors used for Population 2

Larger initial sample sizes for the three vectors, in comparison to those used for the Strategy 2 design when sampling from the simulated network population, ensured that there was a strong probability that at least one node in the initial sample would have a link out to some member not in the initial sample. This way, the dampening vector would not be "cheated", and it was found that these higher initial sample sizes were sufficient enough to significantly reduce the chances of adding more randomness to the sample selection than was desired.

Table 3.6 gives the MSE scores for the preliminary and improved estimators, and bar charts for the MSE scores of the improved estimators for each vector are presented in Figure 3.11.

Table 3.6: MSE scores of Strategy 2 estimators for Population 2

		Vector 1	Vector 2	Vector 3
Estimator 1	Preliminary estimate	0.0160	0.02416	0.0409
	Improved Estimate	0.0154	0.0233	0.0369
Estimator 2	Preliminary estimate	0.0775	0.1195	0.0695
	Improved Estimate	0.0725	0.1115	0.0615
Estimator 3	Preliminary estimate	0.0306	0.0274	0.0235
	Improved Estimate	0.0281	0.0257	0.0214
Estimator 4	Preliminary estimate	0.0114	0.0091	0.0093
	Improved Estimate	0.0113	0.0091	0.0093

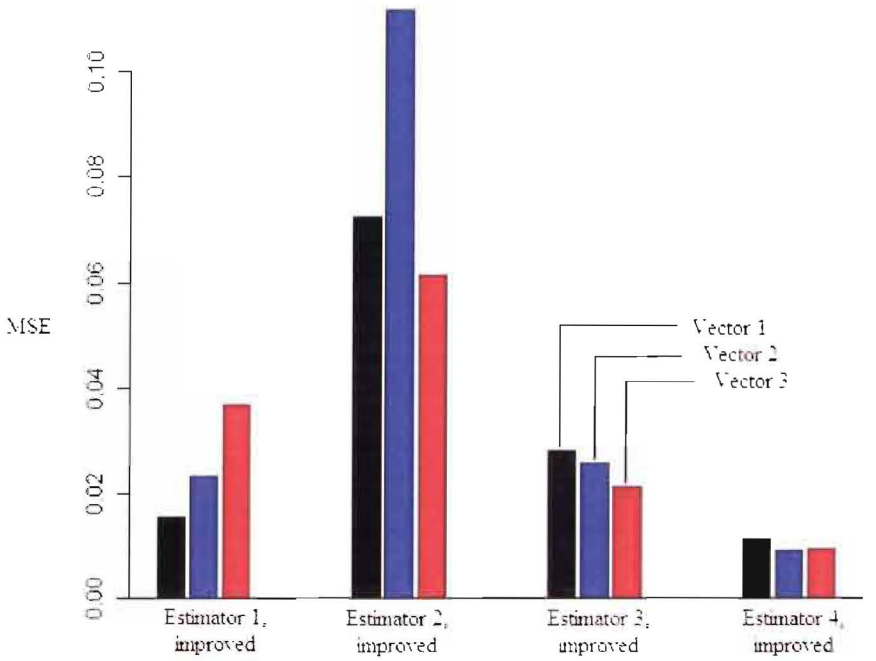


Figure 3.11: Bar charts of MSE scores of Strategy 2 improved estimators for Population 2

Table 3.6 and Figure 3.11 show that some improvements were made for the third estimator when using both Vectors 2 and 3, while the second estimator has shown a significant decrease in efficiency with the use of Vector 2. The best estimates from using Strategy 1 and Strategy 2 designs both came from the fourth estimator, and the relative efficiency of these two estimators was found to be $\frac{0.0091}{0.0113} = 0.805$.

Table 3.7 gives the estimated bias and semi-length of confidence intervals, as well as the estimated coverage rates of the estimators for a final sample comprised of six independent samples. Again, the bias scores have been rescaled for easier comparison purposes.

Table 3.7 shows that coverage rate scores for the second estimator were weakest with Vectors 2 and 3. For all three scores, estimator three performed very well with Vector 3. The improvement in the coverage rate scores for the RB estimators over the preliminary estimators are not as good as those seen in Population 1. This may possibly be due to a smaller improvement, in magnitude, of the RB estimators for Population 2 when comparing against those found in Population 1.

Table 3.7: Bias, CI semi-length, and coverage scores of Strategy 2 estimators for Population 2 upon using six independent samples

Vector 1

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.012	0.126	0.953
	Improved estimate	-0.307	0.123	0.962
Estimator 2	Preliminary estimate	-0.370	0.276	0.934
	Improved estimate	-0.236	0.265	0.938
Estimator 3	Preliminary estimate	0.078	0.178	0.942
	Improved estimate	0.094	0.170	0.939
Estimator 4	Preliminary estimate	1.072	0.105	0.943
	Improved estimate	1.065	0.105	0.942

Vector 2

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	0.121	0.156	0.954
	Improved estimate	0.094	0.151	0.955
Estimator 2	Preliminary estimate	-0.126	0.313	0.840
	Improved estimate	0.067	0.296	0.864
Estimator 3	Preliminary estimate	0.083	0.172	0.948
	Improved estimate	0.147	0.162	0.945
Estimator 4	Preliminary estimate	1.617	0.106	0.939
	Improved estimate	1.616	0.106	0.942

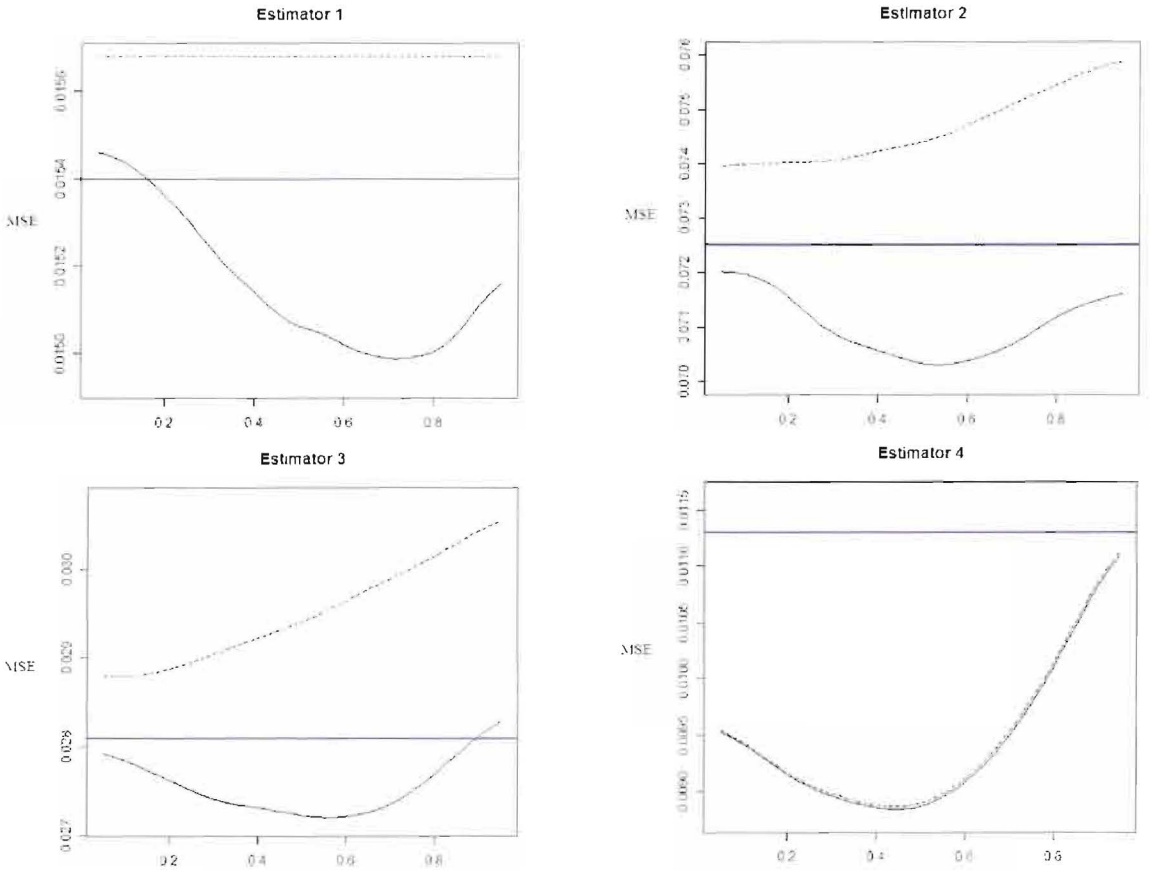
Vector 3

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.122	0.201	0.952
	Improved estimate	-0.942	0.182	0.939
Estimator 2	Preliminary estimate	-0.270	0.254	0.880
	Improved estimate	-0.005	0.234	0.911
Estimator 3	Preliminary estimate	-0.006	0.157	0.941
	Improved estimate	-0.003	0.145	0.949
Estimator 4	Preliminary estimate	1.480	0.108	0.946
	Improved estimate	1.484	0.108	0.945

Strategy 3

Figure 3.12 below gives plots of the MSE scores for each estimator for values of $\theta_{H|d} = 0.05, 0.10, 0.15, ..., 0.90, 0.95$. Due to the large difference in the MSE scores, the range of the MSE axes are adjusted for each estimator to capture the behavior of the estimates at the varying choices of $\theta_{H|d}$. The plotted values were gently smoothed with a normal kernel smoothing method to better illustrate the trend in the estimators. The preliminary estimates are represented by the dashed lines, and the improved estimates are represented by the solid lines. The MSE score for each estimator when using the Strategy 1 design is represented by the solid blue line.

Figure 3.12: MSE scores for Strategy 3 estimators for Population 2



The graphs in Figure 3.12 indicate that for all estimates, efficiency improvements were made over those found in the Strategy 1 estimators for almost all values of $\theta_{H|d}$. It appears that minimum values for the MSE scores for improved estimators two, three, and four were made at values very close to the population mean $\mu = 0.575$. The best estimate that came with the Strategy 3 design came with the fourth estimator at a value of $\theta_{H|d} = 0.45$, and the relative efficiency of this estimator to the best estimator from Strategy 1 was found to be $\frac{0.0088}{0.0113} = 0.779$.

Figure 3.13 gives the expectation of the fourth estimator for the corresponding values of $\theta_{H|d}$. Again, the sample mean is represented by the dashed line and the approximate expectation of the fourth estimator from the Strategy 1 design is represented by the blue line.

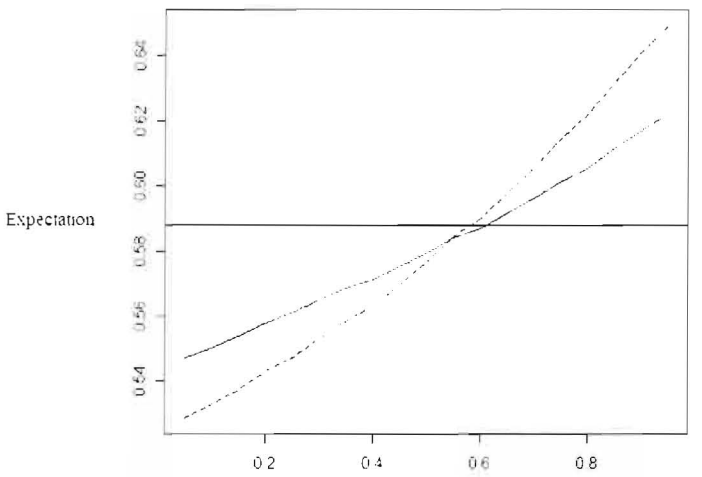


Figure 3.13: Expected values of fourth estimator from Strategy 3 for Population 2

Figure 3.13 shows a crossover in the expected value of the fourth estimator with the sample mean at a value that is approximately equal to the population mean $\mu = 0.575$, and the correlation of the observed values was found to be 0.999. The graph also shows that, when comparing the expected values of the fourth estimator from the Strategy 3 design with the Strategy 1 design, a lot of the reduction in the MSE scores came from a significant decrease in its variance. As can be seen, for values of $\theta_{H|d}$ within close proximity of 0.45, the bias was at a minimum.

Table 3.8 gives the estimated bias of the estimators and semi-length of confidence intervals, as well as the estimated coverage rates of the estimators from setting $\theta_{H|d} = \mu$ and

the initial sample mean. The final sample is comprised of six independent samples, and the bias has again been rescaled for comparison purposes.

Table 3.8: Bias, CI semi-length, and coverage scores of Strategy 3 estimators for Population 2 upon using six independent samples

$\theta_{H|d} = \mu$

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.020	0.126	0.953
	Improved estimate	-0.005	0.123	0.957
Estimator 2	Preliminary estimate	-0.060	0.269	0.937
	Improved estimate	0.039	0.260	0.935
Estimator 3	Preliminary estimate	0.051	0.174	0.940
	Improved estimate	0.022	0.166	0.942
Estimator 4	Preliminary estimate	1.079	0.094	0.941
	Improved estimate	1.077	0.094	0.940

$\theta_{H|d} = \bar{y}_{s_0}$

		Bias$\times 10^2$	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.196	0.126	0.951
	Improved estimate	-0.302	0.122	0.954
Estimator 2	Preliminary estimate	-0.364	0.270	0.927
	Improved estimate	-0.121	0.263	0.936
Estimator 3	Preliminary estimate	-0.387	0.176	0.942
	Improved estimate	-0.338	0.168	0.947
Estimator 4	Preliminary estimate	1.210	0.104	0.946
	Improved estimate	1.230	0.104	0.946

Similar to the results from Population 1, Table 3.8 indicates that the scores for estimators one, two, and three are approximately equal for both cases. For the fourth estimator, the semi-length of the confidence interval from setting $\theta_{H|d}$ equal to the initial sample mean was about ten percent larger than that found with setting $\theta_{H|d}$ equal to the population mean, while slightly more bias was introduced.

In comparison with estimates from the Strategy 1 design, semi-lengths of the confidence intervals were slightly smaller for estimators two and three from the Strategy 3 design, as well for the fourth estimator when setting $\theta_{H|d} = \mu$, and this seems to correspond with

the MSE plots found in Figure 3.12. The bias and coverage rate scores were approximately equal in both Strategies 1 and 3 for these estimators.

3.2 Spatial populations

3.2.1 Simulated spatial population (Population 3)

The population in Figure 3.14 (Population 3) was presented in Chapter 1, and was simulated using the Adaptive Network Sampling package in R. Population 3 was generated using six parent locations whose values followed a Poisson distribution with mean 120. The central location of these clusters was randomly selected in the unit square, and their dispersion followed a symmetric Gaussian distribution with standard deviation 0.03.

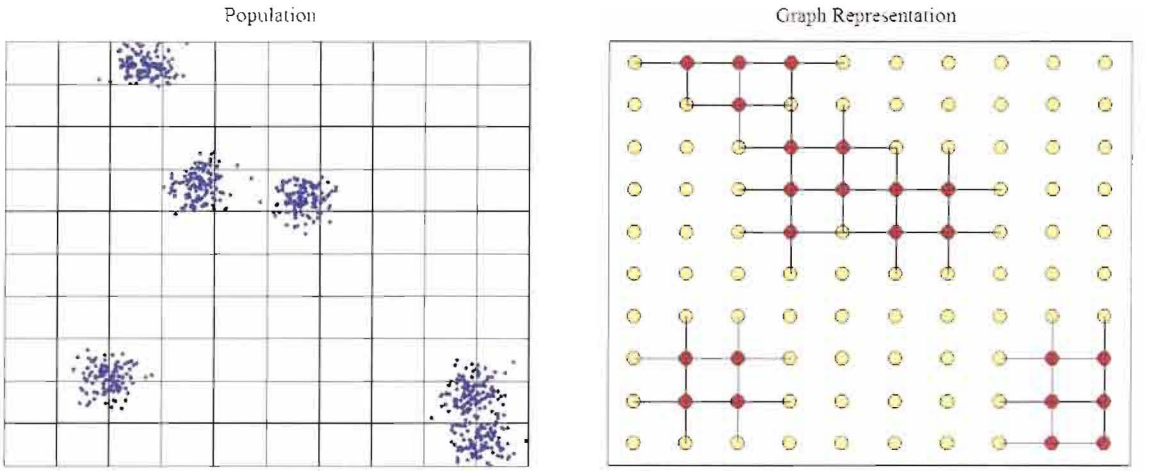


Figure 3.14: Population 3

The study area was divided into 100 plots, and those of which contained at least one point-object were classified as units of interest. Estimates of the mean number of point-objects in each plot are found in the following sections.

Strategy 1

The simulations of sampling from this population with Strategy 1 used a without-replacement design. An initial sample of size 15 was selected through a SRS design, and the final sample size was set to 30. One node was selected at each wave, and the dampening values were held constant at 0.9. A sample of this type can be found in Figure 3.15.

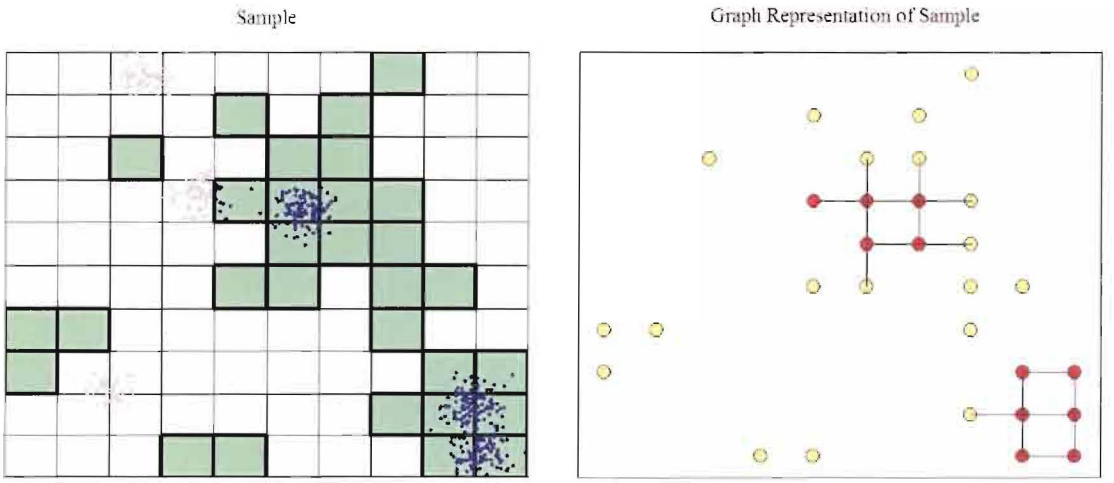


Figure 3.15: A Strategy 1 sample from Population 3

Table 3.9 gives the MSE scores for the AWS Strategy 1 design estimators. A series of 2000 simulation runs were used, and 10,000 resamples were used on each run for the MCMC estimates.

Table 3.9: MSE scores of Strategy 1 estimators for Population 3

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	21.62	27.92	23.56	16.75
Improved Estimate	15.96	21.22	18.09	15.56

Properties of the sampling strategy and its estimators can be found in the histograms in Figure 3.16. The true population mean $\mu = 7.15$ is indicated by the solid triangle, and the approximate expectation of each estimator is indicated by the transparent triangle.

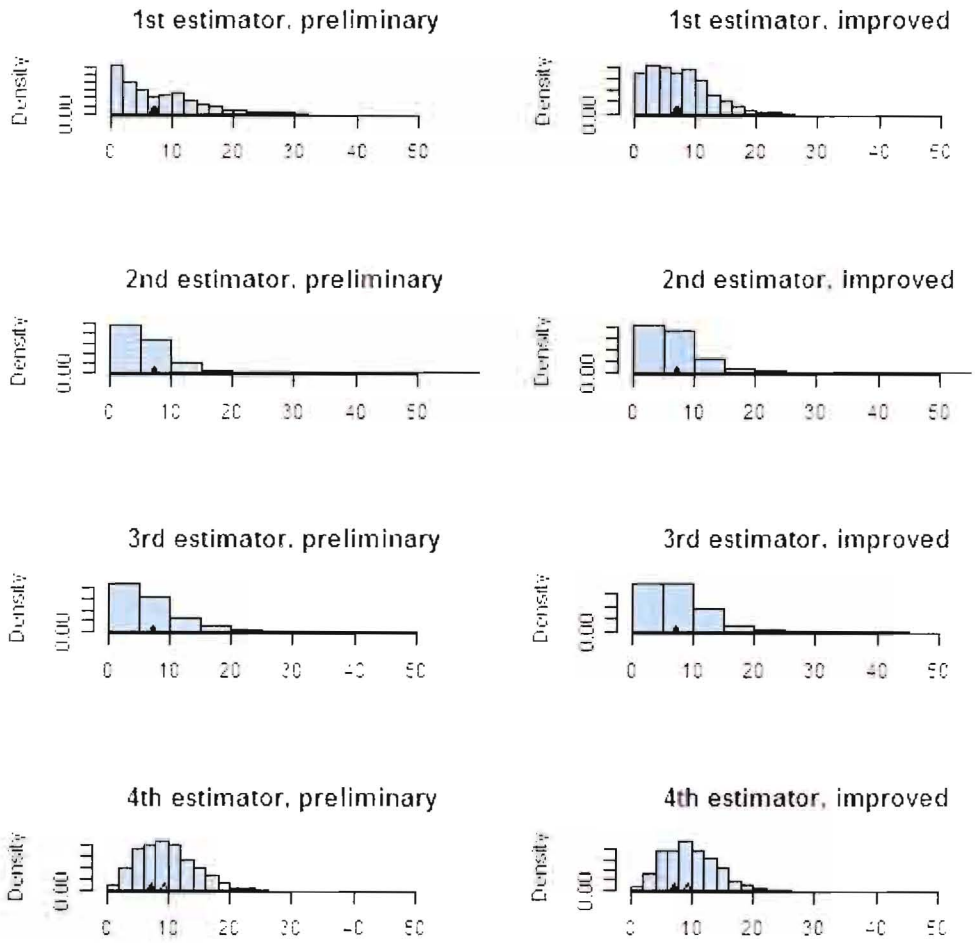


Figure 3.16: Histograms of Strategy 1 estimators for Population 3

Table 3.9 shows that significant gains in efficiency were made for the first, second, and third improved estimators over the preliminary estimators. The histogram plots reveal that some symmetry has been introduced with Rao-Blackwellization of the preliminary estimators. Similar to the simulation results seen in the network populations, the histograms show that the second estimator took on extreme values. Rao-Blackwellization and the use of the ratio estimators has once again helped to reduce the occurrence of the extreme values. The fourth estimator showed a large amount of bias when compared to the other three estimators. However, this estimator still performed the best overall.

Strategy 2

The three damp vectors that were used for the Strategy 2 design are the same as those that were used for sampling from Population 2.

Table 3.10 gives the MSE scores for the preliminary and improved estimators, and bar charts for the MSE scores of the improved estimators for each vector are presented in Figure 3.17.

Table 3.10: MSE scores of Strategy 2 estimators for Population 3

		Vector 1	Vector 2	Vector 3
Estimator 1	Preliminary estimate	21.619	32.967	57.020
	Improved Estimate	15.962	23.171	32.571
Estimator 2	Preliminary estimate	27.923	42.417	25.163
	Improved Estimate	21.217	25.825	17.741
Estimator 3	Preliminary estimate	23.563	23.469	19.928
	Improved Estimate	18.086	17.934	14.240
Estimator 4	Preliminary estimate	16.748	17.085	17.000
	Improved Estimate	15.555	16.284	15.933

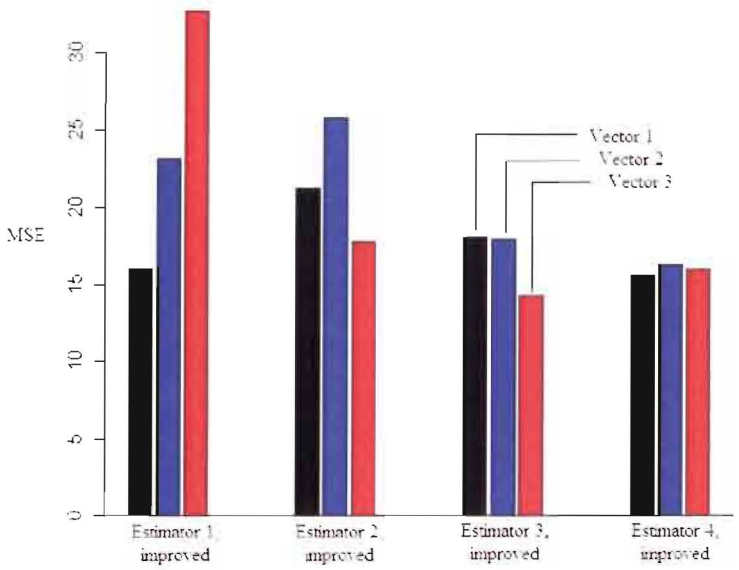


Figure 3.17: Bar charts of MSE scores of Strategy 2 improved estimators for Population 3

Table 3.10 and Figure 3.17 show that efficiency gains were made when using the AWS Strategy 2 design with Vector 3 for the second and third estimators. With Vector 2, it appears that no improvements were made over the Strategy 1 design. Similar to the network populations, the higher MSE values for the first estimator when using Vectors 2 and 3 may be attributed to their smaller initial sample sizes, while no change in efficiency was found with the fourth estimator.

The most efficient estimate from the Strategy 1 design was with the fourth estimate, and the most efficient estimate from Strategy 2 was with the third estimate from Vector 3. The relative efficiency of these two estimators was found to be $\frac{14.240}{15.962} = 0.892$.

Table 3.11 on the following page gives the estimated bias of the estimators and semi-length of the confidence intervals, as well as the estimated coverage rate of the estimators used with the corresponding vectors for one final sample comprised of six independent samples.

Table 3.11 shows that the coverage rates are better with the first, second, and third improved estimators over those found with the preliminary estimators. Also, for the second and third estimators, coverage rates are better when using Vectors 2 and 3 in comparison to Vector 1. The bias scores from all three dampening vectors were approximately the same for all four estimators. The weak coverage from the fourth estimator reflects a consequence of the bias that is present with this estimator, and Rao-Blackwellization shows little to no improvement in the behavior of the confidence intervals constructed with it. Overall, significant gains in efficiency were found for estimators two and three when using Vector 3.

Table 3.11: Bias, CI semi-length, and coverage scores of Strategy 2 estimators for Population 3 upon using six independent samples

Vector 1

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	0.011	4.608	0.938
	Improved estimate	0.058	4.116	0.953
Estimator 2	Preliminary estimate	-0.078	4.537	0.885
	Improved estimate	-0.049	4.165	0.906
Estimator 3	Preliminary estimate	0.117	4.434	0.920
	Improved estimate	0.146	4.050	0.936
Estimator 4	Preliminary estimate	2.223	3.423	0.783
	Improved estimate	2.245	3.283	0.750

Vector 2

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	-0.024	5.735	0.929
	Improved estimate	-0.020	4.936	0.935
Estimator 2	Preliminary estimate	-0.037	4.563	0.894
	Improved estimate	-0.069	4.158	0.918
Estimator 3	Preliminary estimate	0.195	4.362	0.936
	Improved estimate	0.184	3.945	0.944
Estimator 4	Preliminary estimate	2.313	3.410	0.737
	Improved estimate	2.332	3.291	0.715

Vector 3

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	0.041	7.505	0.899
	Improved estimate	0.077	5.930	0.935
Estimator 2	Preliminary estimate	0.024	4.189	0.916
	Improved estimate	0.004	3.789	0.938
Estimator 3	Preliminary estimate	0.158	4.071	0.931
	Improved estimate	0.184	3.677	0.942
Estimator 4	Preliminary estimate	2.461	3.406	0.718
	Improved estimate	2.459	3.246	0.691

Strategy 3

In this strategy, links are traced with probability proportional to the originating node values. With this variation the user has the advantage of adding units that possess higher observed values into the sample, but perhaps at the expense of more effort required to count the number of point-objects within the plot.

The MSE scores for each estimator when using Strategy 3 are presented in Table 3.12.

Table 3.12: MSE scores of Strategy 3 estimators for Population 3

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	21.02	31.43	24.69	16.47
Improved estimate	15.65	23.52	18.92	15.50

As can be seen, the estimates show no improvement over those found in Strategy 1. When using the changing dampening vectors in conjunction with this strategy, it was found that the estimates did not perform nearly as well as those that were presented in the previous section.

3.2.2 Wintering Waterfowl population (Population 4)

Figure 3.18 illustrates an empirical blue-winged teal bird population (Population 4) on a wildlife refuge (Smith et al. (1995)). The study area was divided into 50 plots, and the number of birds within each plot is represented in the population box. The graph representation version indicates plots of interest (those with a count of at least one bird), and one way relationships exist from plots of interest to adjacent boxes. Estimates of the mean number of birds in each plot are found in the following sections.

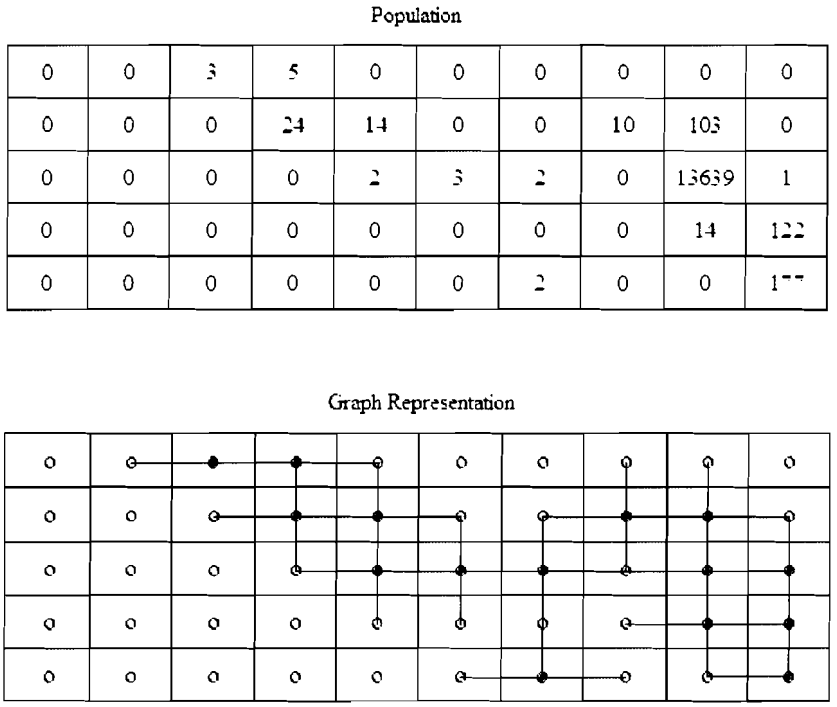


Figure 3.18: Population 4

Strategy 1

The simulations of sampling from this population with Strategy 1 used a without-replacement design. An initial sample of size 15 was selected through a SRS design, and the final sample size was set to 30. One node was selected at each wave, and the dampening values were held constant at 0.9.

Table 3.13 gives the MSE scores for the AWS Strategy 1 design estimators. Again, a series of 2000 simulations were used for each design, and 10,000 resamples were used on each run for the MCMC estimates.

Table 3.13: MSE scores of Strategy 1 estimators for Population 4

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	174431.02	84922.88	84972.93	63469.09
Improved Estimate	36093.01	20238.76	22132.54	26827.33

Properties of the four estimators can be found in the histograms in Figure 3.19. The true population mean $\mu = 282.42$ is indicated by the solid triangle.

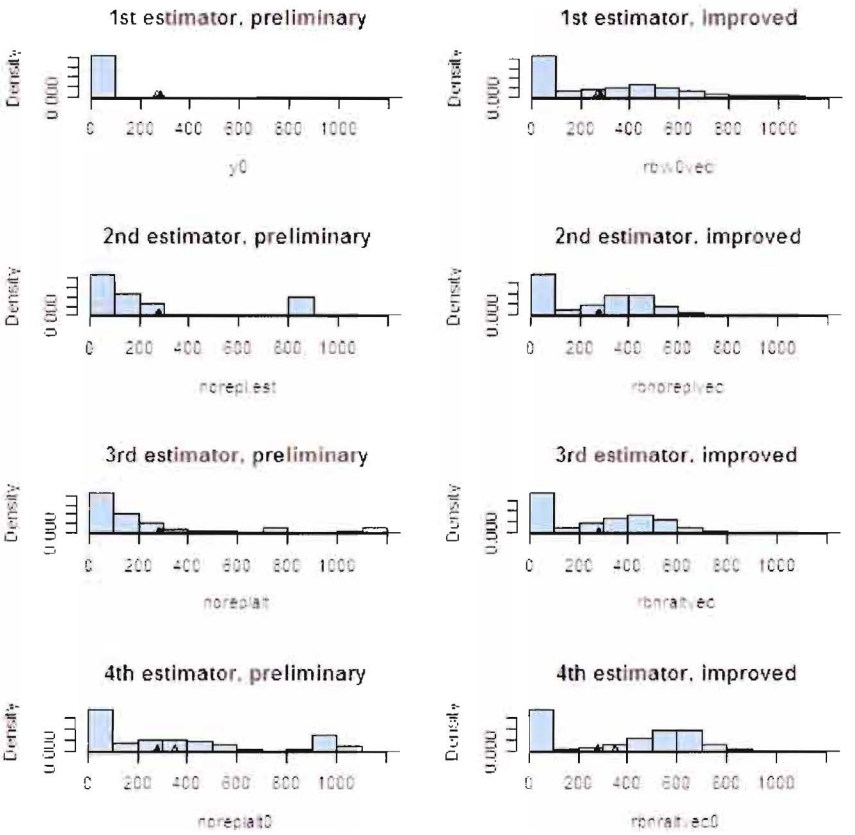


Figure 3.19: Histograms of Strategy 1 estimators for Population 4

Table 3.13 suggests that the second and third estimates perform the best, as opposed to the fourth and first estimators as seen in Population 3 when using Strategy 1. Rao-Blackwellization of the estimators showed significant improvement over the preliminary estimators. The histograms only show frequency values up to 1200, and this is for visual purposes. Extreme values were taken on by the second estimator, and reached values up to 6000. In comparison to the other estimators, a large amount of bias in the fourth estimator is evident, and accounts for most of the MSE.

Strategy 2

The same three dampening vectors that were used in the simulated spatial population will be used in Population 4.

Table 3.14 gives the MSE scores for the preliminary and improved estimators for each vector.

Table 3.14: MSE scores of Strategy 2 estimators for Population 4

		Vector 1	Vector 2	Vector 3
Estimator 1	Preliminary estimate	174431.02	298381.80	545997.06
	Improved Estimate	36093.01	51855.80	70026.42
Estimator 2	Preliminary estimate	84922.88	81422.89	73051.01
	Improved Estimate	20238.76	29723.42	17493.70
Estimator 3	Preliminary estimate	84972.93	76962.99	70261.21
	Improved Estimate	22132.54	25844.17	18337.99
Estimator 4	Preliminary estimate	63469.09	63149.81	63377.66
	Improved Estimate	26827.33	28082.79	27733.91

Bar charts for the MSE scores of the improved estimators from each vector can be seen in Figure 3.20 on the following page.

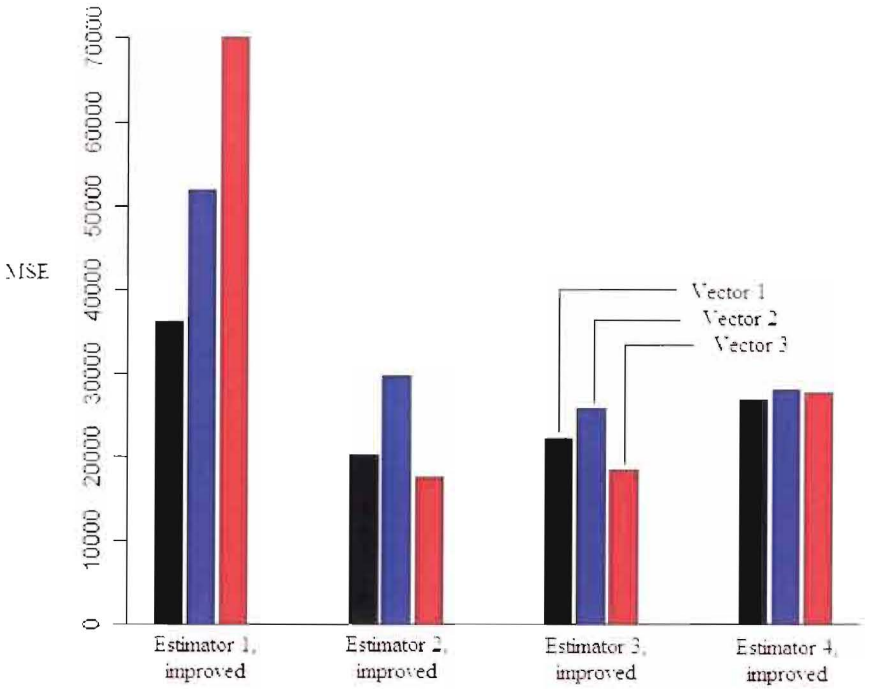


Figure 3.20: Bar charts of MSE scores of Strategy 2 improved estimators for Population 4

It appears, as with the last simulation, that a significant improvement was found in the second and third estimators upon using Vector 3, with little to no change found in the fourth estimator using the new vectors.

The most efficient estimate from the Strategy 1 design came from the second estimate. The most efficient estimate from Strategy 2 came with the second estimate from Vector 3. The relative efficiency of these two estimators was found to be $\frac{17493.70}{20238.76} = 0.864$.

Table 3.15 gives the estimated bias and semi-length of the confidence intervals, as well as the estimated coverage rates of the estimators used with the corresponding vectors for one final sample comprised of six independent samples.

Table 3.15: Bias, CI semi-length, and coverage scores of Strategy 2 estimators for Population 4 upon using six independent samples

Vector 1

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	-4.550	405.111	0.867
	Improved estimate	-0.444	202.008	0.946
Estimator 2	Preliminary estimate	-1.075	258.349	0.876
	Improved estimate	-0.444	146.289	0.945
Estimator 3	Preliminary estimate	4.664	263.670	0.882
	Improved estimate	6.406	152.066	0.937
Estimator 4	Preliminary estimate	59.705	246.272	0.932
	Improved estimate	61.946	154.197	0.767

Vector 2

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	4.648	496.253	0.745
	Improved estimate	0.516	240.842	0.940
Estimator 2	Preliminary estimate	-1.981	245.7117	0.883
	Improved estimate	0.822	144.360	0.945
Estimator 3	Preliminary estimate	6.411	256.739	0.894
	Improved estimate	7.340	151.077	0.935
Estimator 4	Preliminary estimate	65.093	247.661	0.933
	Improved estimate	65.123	158.131	0.769

Vector 3

		Bias	Semi-length	Coverage
Estimator 1	Preliminary estimate	-12.081	554.691	0.515
	Improved estimate	2.121	269.964	0.920
Estimator 2	Preliminary estimate	-0.524	242.470	0.889
	Improved estimate	-3.338	138.202	0.943
Estimator 3	Preliminary estimate	-1.006	248.448	0.890
	Improved estimate	3.255	143.342	0.932
Estimator 4	Preliminary estimate	64.322	245.975	0.927
	Improved estimate	65.253	156.642	0.755

As can be seen in Table 3.15, coverage scores greatly improved with the Rao-Blackwellization of the first, second and third estimators. The significant decrease in the coverage score with the improved fourth estimator may be attributed to the large amount of bias coupled with the small variability in the estimator. Similar to the results found from the simulations from Population 3, significant gains in efficiency were found for estimators two and three when using Vector 3.

Strategy 3

The MSE scores for each estimator when using Strategy 3 is presented in Table 3.16.

Table 3.16: MSE scores of Strategy 3 estimators for Population 4

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary estimate	174962.17	88497.46	85146.47	64157.67
Improved estimate	36482.72	25083.51	25294.03	27578.96

Once again, these estimates have shown no improvement over those found in Strategy 1. Similar to the simulation results with Population 3, when using the changing dampening vectors in conjunction with this strategy, no improvements in the estimates were found over those presented in the previous section.

Chapter 4

Conclusions

Improvements from Rao-Blackwellizing the first three preliminary estimators have proven to give significantly better estimates, while adding some symmetry to the distributions of these estimators to give as good or better coverage rates than those found with their preliminary counterparts. The fourth estimator did not benefit as much from Rao-Blackwellization in comparison to the other three estimators. As expected, the first and second estimators always came out approximately unbiased in the simulations. Though the second estimator had a tendency to take on extreme values in the simulations for both the network and spatial setting populations, the occurrence of these values was always reduced with the ratio estimators. In comparison, the fourth estimator continuously showed much more bias than the third estimator.

The use of Vector 3 in the Strategy 2 design helped to significantly reduce the MSE scores from the second and third estimators. In all cases, using Strategy 2 with Vector 3 has provided at least one estimate that is better than the best estimate that comes with Strategy 1. Coverage rates with these estimators always did as good or better than those found with the Strategy 1 design, while the semi-lengths of the confidence intervals were as small or smaller and the biases remained the same.

Strategy 3 estimators for Population 1 have shown that when setting the parameter $\theta_{H|d}$ to values slightly larger than the true population mean μ, some improvements were made in estimators one and three, while little to no change in MSE was found for the second estimator for all values of $\theta_{H|d}$. However, when setting $\theta_{H|d}$ approximately equal to the population mean, a significant reduction in both the bias and variance was found for estimator four. For Population 2, all four estimators greatly benefitted from setting $\theta_{H|d}$

within close proximity of the population mean.

In both network populations, the MSE score from the fourth estimator was lower than that found with the best estimator from the Strategy 1 design when setting $\theta_{H|d} = \mu$. Although setting $\theta_{H|d}$ equal to the sample mean did not make improvements in the fourth estimator over the best estimate from Strategy 1, significant improvements were still made over some of the estimates from Strategy 1. If one were to select a large number of independent samples, then further improvements may be made by setting $\theta_{H|d}$ equal to the mean of the initial sample means.

Estimators from the Strategy 3 design used in the spatial setting showed no significant improvements over those found from the Strategy 1 design. Moreover, these estimates usually performed worse than the estimates from the Strategy 1 design. While using the changing dampening vectors in conjunction with Strategy 3, it was found that the MSE scores failed to compare with those found in the Strategy 2 design.

Future study on how the initial sample is selected and what size to use would certainly be useful. Thompson (2006a) investigated the same bird population (Population 4), and found that with a final sample of size 20, an initial sample of size 13 or 14 achieved the smallest MSE scores when using the Strategy 1 design. It therefore may be possible to make immediate improvements with the new designs by allowing for different initial sample sizes.

When using Strategy 3, more study on what values of $\theta_{H|d}$ should be used for certain types of populations would certainly be helpful. When uncertain as to which values one should use, research into setting $\theta_{H|d}$ equal to a function of the observed sample values and link variables from the initial sample would also be good.

Investigation on the use of auxiliary information with adaptive web sampling would also be very useful, as new ratio estimators could potentially be developed by exploiting some of the auxiliary information.

In conclusion, the new strategies provide estimates for all four populations which are better than those found in the general design. The new designs are attractive in that they are intuitively appealing and retain many of the features of the general design, such as ease of understanding and application, as well as retaining all the advantages that the previous design has over some of the current link-tracing designs. With the large amount of flexibility available in adaptive web sampling, as well as the results presented in this project from using the new designs, further improvements can be expected to be seen in the future.

Bibliography

Chow, M. and Thompson, S. (2003). Estimation with link-tracing sampling designs- a bayesian approach. *Survey Methodology* **29**, 197–205.

Darrow, W., Potterat, J., Rothenberg, R., Woodhouse, D., Muth, S. and Klovdahl, A. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus* **32**, 143–158.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10**, 53–67.

Frank, O. and Thompson, S. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* **26**, 87–98.

Hastings, W. (1970). Monte-carlo sampling methods using markov chains and their application. *Biometrika* **57**, 97–109.

Hoff, P., Raftery, A. and Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97**, 1090–1098.

Linkletter, C. (2007). *Spatial Process Models for Social Network Analysis*. Ph.D. thesis, Simon Fraser University.

Potterat, J., Woodhouse, D., Rothenberg, R., Muth, S., Darrow, W., Muth, J. and Reynolds, J. (1993). AIDS in Colordao Springs: Is there an epidemic? *AIDS* **7**, 1517–1521.

Smith, D., Conroy, M. and Brakhage, D. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics* **51**, 777–788.

Thompson, S. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association* **85**, 1050–1059.

Thompson, S. (2002). *Sampling*. Wiley, second edition.

Thompson, S. (2006a). Adaptive web sampling. *Biometrics* **62**, 1224–1234.

Thompson, S. (2006b). Targeted random walk designs. *Survey Methodology* **32**, 11–24.

Thompson, S. and Seber, G. (1996). *Adaptive Sampling*. Wiley.