A REVIEW OF THE BC VEGETATION RESOURCES INVENTORY SAMPLING DESIGN AND THE PROPOSED ESTIMATORS.

by

Carolyn G. Taylor
B.Sc. Simon Fraser University 1994

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in the Department
of
Statistics and Mathematics

SIMON FRASER UNIVERSITY
April 1998

APPROVAL

Name: Carolyn G. Taylor
Degree: Master of Science
Title of thesis: A review of the BC vegetation resources inventory sampling design and the proposed estimators.

Examining Committee:
Chair: Dr. Richard Lockhart

_____
Dr. Carl J. Schwarz
Senior Supervisor
Department of Statistics and Mathematics

_____
Dr. Richard Routledge
Professor
Department of Statistics and Mathematics

_____
Dr. Charmaine Dean
Professor
Department of Statistics and Mathematics

_____
Dr. Randy Sitter
External Examiner
Department of Statistics and Mathematics
Simon Fraser University

ABSTRACT

The Vegetation Resources Inventory (VRI) is a new provincial survey that collects data on the type, amount, and location of vegetation in British Columbia. The survey is carried out in inventory units which may be as small as a watershed or as large as a Forest District. In the first step of the survey process, aerial photographs are taken of the entire inventory unit. The photographs are used to divide the inventory unit into homogeneous sub-areas called polygons and to provide photo-interpreted estimates for every polygon in the inventory unit. The next step of the survey process involves selecting a sample of polygons from which ground locations are selected for taking ground measurements. The goal of this project is to investigate the current and alternative methods of sample selection and to answer the two following questions:

1. Will the current metho used to select a sample of polygons for ground measurement affect the integrity of the inventory unit estimates.
2. What estimators should be used to estimate attribute totals, means, and variances at the inventory unit level?

Acknowledgments

This project is dedicated to my husband, Doug, whose commitment to me and my education played an essential part to ward the completion of this work.

Thanks to Carl Schwarz, for being a caring teacher and whose guidance throughout this project and my degree I highly valued. Thanks also to Sam Otukol from the Ministry of Forests for his interest, help, and dedication to this project.

Table of Contents

1 INTRODUCTION

The growing demand for better management of British Columbia's land and natural resources has increased the need for more reliable and comprehensive information. In response to this growing need, a review of the province's resource inventories was carried out in 1991 and revealed that they were badly outdated and lacked the information essential for resource managers to make effective, scientifically-based land use decisions. The Vegetation Resources Inventory (VRI) is a new inventory design that is in the process of being developed to help meet the demands for more and better information. The general purpose of the VRI is to collect data on the type, amount, and location of vegetation in British Columbia. Unlike the previous forest inventories, it covers the entire province and, in addition to estimating timber volumes and land areas, includes information on a full range of vegetation types, as well as data on coarse woody debris, forest health, range, soils, and a host of other characteristics (Linnell Nemac 1997).

An important requirement of the VRI survey design is that it be practical within the constraints of economic and resource management needs. In order to meet the need for convenient survey management and data handling, the survey process is carried out separately within inventory units, which collectively make up the total provincial land base. An inventory unit is typically a large-scale management unit such as a Timber Supply Area or a Tree Farm License but it can also be as small as a watershed or as large as a Forest District. Another important requirement of the VRI survey is that it provide

precise and accurate information at the inventory unit level as well as for smaller areas or sections within the inventory unit called polygons. The proposed method for meeting this requirement is to collect data on the various attributes of interest in two phases.

In the first phase, aerial photographs are taken of the entire inventory unit. These photographs are used to divide the entire inventory unit into non-overlapping, subunits called polygons. The polygons are delineated so that they are relatively homogeneous in terms of the characteristics that are relevant to the inventory and useful for management. It is also at this point that polygon-specific estimates are made for some attributes through photo-interpretation. Table 1 gives a list of the tree attributes that can be estimated in Phase I. The intent of Phase I, therefore, is to use aerial photographs to divide the entire inventory unit into smaller, homogeneous subareas called polygons and to provide initial polygon-specific estimates of some of their attributes.

In the second phase, areas are selected where ground measurements will be taken of various attributes including those listed in Table 1. The areas that are selected for ground measurement are chosen according to a two-step process. The first step involves randomly selecting a sample of polygons from the list of polygons delineated in Phase I. Not only is it unrealistic that the entire polygon be measured for every polygon selected but it is also unnecessary and uneconomical given that the polygons are formed so that they are homogeneous. A subsampling approach is used instead, where a 100 m x 100 m

TABLE 1: Vegetation Resources Inventory Tree Attributes for Estimation

Tree Attribute[1]	*Phase I*	*Phase II*
Tree Sp (leading)	√	√
Tree Sp (2nd)	√	√
Cover pattern	√	√
Species comp (% 1st)	√	√
Species comp (% 2nd)	√	√
Age - species 1 (yrs)	√	√
Age - species 2 (yrs)	√	√
Random age (yrs)		√
Height - species 1 (m)	√	√
Height - species 2 (m)	√	√
Random height (m)		√
Lorey height (m)		√
Crown closure (%)	√	
Gross volume (m3)		√
Net volume (m3)	√	√
Basal area (m2)	√	√
Snag frequency (#/ha)	√	√
Vertical complexity	√	√
Tree layer	√	√
Density (stems/ha)	√	√
CWD - freq (#/ha)		√
CWD - volume (m3)		√
Stump volume (m3)		√
Stand diameter (cm)	√	√
Tree past growth		√
Hidden decay %		√
Pest damage incidence		√
Pest damage severity		√

[1] This table is from Otukol, 1996.

grid is used to select ground locations within each of the selected polygons. At each selected ground location, fixed-area plots, prism plots, and line transects are established for measuring ecological, tree, and range attributes. The measurements taken at a specific ground location are then used to estimate the total per hectare (i.e. m^3/ha for tree volume) for that location. Therefore, the final product of Phase II sampling is an estimated total per hectare for each ground location that is selected for measurement.

If proper sampling and estimation methodology are used in Phase II of the VRI survey, the ground measurements alone, can be used to calculate unbiased estimates of the inventory unit total or mean and its estimated precision for each of the measured attributes. In Phase I, however, the polygon-specific estimates based on aerial photographs are biased as a result of the photo-interpretation process used. For those attributes where both Phase I and Phase II estimates are available, the hope is that the results from Phase II combined with the complete coverage from Phase I will help improve the polygon-specific estimates as well as the inventory unit estimates. Various adjustment methods have been considered by Stauffer (1995) which incorporate using both the Phase I and II estimates in calculating inventory unit estimates.

One of the fundamental requirements of the VRI survey design is that it be well founded in statistical theory. Even though many components of the VRI have been accepted, there still remain design and analysis issues that the BC Ministry of Forests wish to have

resolved. The purpose of this project is to address two specific aspects of the survey process:

1. The selection of polygons in Phase II of the VRI sampling design
2. Estimators of the total, mean and variance for the inventory unit

The motivation for investigating these two areas is to answer the following two questions:

1. Will the current method used to select a sample of polygons for ground measurement affect the integrity of the inventory unit estimates?
2. What estimators should be used to estimate attribute totals, means and variances at the inventory unit level?

The tree data collected from the first operational trial, conducted in 1995 in the Boston Bar area of the Fraser timber supply area, will be used to help address these issues.

2 NOTATION

n - number of polygons selected

N - number of polygons in the inventory unit

z_i - area (ha) for polygon i

Z - total area (ha) for the inventory unit

$$Z = \sum_{i=1}^{N} z_i$$

m_i - number of ground locations measured in polygon i

m - total number of ground locations measured

$$m = \sum_{i=1}^{n} m_i$$

M_i - number of grid locations in polygon i

M - total number of grid locations in the inventory unit

$$M = \sum_{i=1}^{N} M_i$$

X_i - Phase I total for polygon i

X - Phase I total for the inventory unit

$$X = \sum_{i=1}^{N} X_i$$

x_i - Phase I total per hectare for polygon i

$$x_i = \frac{X_i}{z_i}$$

x - Phase I total per hectare for the inventory unit

$$x = \frac{\sum_{i=1}^{N} X_i}{Z}$$

$\hat{y}_{il}$ - Phase II total per hectare for location l in polygon i

$\hat{y}_i$ - Phase II total per hectare for polygon i

$$\hat{y}_i = \frac{\sum_{l=1}^{m_i} \hat{y}_{il}}{m_i}$$

y_i - true total per hectare for polygon i

y - true total per hectare for the inventory unit

$\hat{Y}_i$ - Phase II total for polygon i

$$\hat{Y}_i = z_i \hat{y}_i$$

Y_i - true total for polygon i

Y - true total for the inventory unit

$$Y = \sum_{i=1}^{N} Y_i$$

$\hat{s}_i^2$ - estimated variability in total per hectare within-polygon i

$$\hat{s}_i^2 = \frac{\sum_{l=1}^{m_i} \left(\hat{y}_{il} - \hat{y}_i\right)^2}{m_i - 1}$$

s_i^2 - true variability in total per hectare within-polygon i

$\hat{\sigma}_i^2$ - estimated variance of $\hat{Y}_i$ for polygon *i*

$$\hat{\sigma}_i^2 = z_i^2 \frac{\hat{s}_i^2}{m_i}\left(\frac{M_i - m_i}{M_i}\right)$$

σ_i^2 - true variance of $\hat{Y}_i$ for polygon *i*

$$\sigma_i^2 = z_i^2 \frac{s_i^2}{m_i}\frac{(M_i - m_i)}{M_i}$$

π_i - probability of selecting polygon *i*

π_{ij} - probability of selecting both polygon *i* and polygon *j*

p_i - probability proportional to size for polygon *i*

$$p_i = \frac{z_i}{Z}$$

$v(\bullet)$ - estimated variance of the estimate

Estimator acronyms:

srs - simple random sampling

r - ratio

lreg - linear regression

HT - Horvitz-Thompson

gr - generalized ratio

greg - generalized regression

DR - Des Raj

RHC - Rao-Hartley-Cochran

LMS - Lahiri-Midzuno-Sen

pps - probability proportional to size

Sampling acronyms:

SRS - simple random sampling

PPS - probability proportional to size

PPSWR - probability proportional to size with replacement

PPSWOR - probability proportional to size without replacement

OS - ordered systematic

3 CURRENT METHODS OF SAMPLE SELECTION AND ESTIMATION

This section discusses the method of sample selection used in Phase II of the VRI survey design and the currently recommended estimators for estimating the total, mean and estimated precision for the inventory unit.

3.1 Description

3.1.1 Phase II sample selection

As previously mentioned, Phase II of the VRI survey design selects a sample of ground locations for measurement according to a two-step process. The current method for selecting polygons in the first step of the sampling design is similar to a method that was first introduced by Madow (1949) which Brewer and Hanif (1983) refer to as the ordered systematic (OS) procedure. This method is used to select a sample of polygons by arranging the list of all polygons delineated in Phase I first in an order under which systematic sampling performs well. Systematic sampling will perform well if the polygons are arranged in such a way that will cause the sample to be spread evenly over the range of values for the characteristics being ground-measured. In the Boston Bar operational trial, the polygons were first sorted into vegetated and non-vegetated types. The vegetated polygons were then further sorted by the type of vegetation, being either tree, shrub, herb or cryptogam. The treed polygons were then further sorted by leading species and site index. If the estimated polygon totals from Phase II correlate with these

variables used to sort the list of polygons the precision of the inventory unit estimates will be improved. Once the list is sorted, the selection interval, k, is calculated by dividing the total area of all the polygons in the inventory unit (in hectares), Z, by the desired sample size, m (note that the sample size refers to the number of ground locations and not the number of polygons). If $Z = km$ exactly, k will be an integer, otherwise it will be a decimal. Finally, a systematic sample of polygons is selected from the sorted list by choosing a random starting point between 1 and k hectares and then selecting polygons that contain every k^{th} hectare.

The second step of the Phase II sampling design selects locations for ground measurement within each of the polygons chosen in the first step. A province-wide 20 km x 20 km grid is used to establish a sampling frame for selecting the sampling points. This grid is a permanent grid with the latitudes and longitudes of each grid point stored in the VRI database. For each polygon that is selected for ground measurement, the Forest Inventory GIS system will create a 100 m x 100 m grid that lines up with the 20 km x 20 km provincial grid. Each grid point that falls inside the polygon is numbered and ground locations are randomly selected without replacement from these numbered points. If the area of a selected polygon is less than the selection interval then the current method proposes that one grid location be randomly selected for ground measurement. Polygons larger than the selection interval will have the number of grid locations chosen for ground measurement equal to the number of systematic points that fall within its area.

3.1.2 Recommended estimators in Stauffer (1995)

Stauffer (1995) in his report on the statistical estimation and adjustment for the VRI

recommends a number of estimators to estimate attribute totals and means and their variances for the inventory unit. His report states that one can view the ground points, selected with equal probability, as the sampling units and suggests the simple random sampling (*srs*) estimator and regression estimators (weighted, linear and ratio) as possible choices to use. Another option that he considers is to view the polygons, selected with unequal probability, as the sampling units and then use the probability proportional to size without replacement (PPSWOR) estimator, the generalized ratio (*gr*) estimator or the generalized regression (*greg*) estimator. The following are the equations for these estimators:

SRS Estimator

$$\hat{y}_{srs} = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m_i}\hat{y}_{il}}{m} \qquad \hat{Y}_{srs} = Z\hat{y}_{srs} \tag{1a}$$

$$v(\hat{y}_{srs}) = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m_i}\left(\hat{y}_{il} - \hat{y}_{srs}\right)^2}{m(m-1)}\frac{(M-m)}{M} \qquad v(\hat{Y}_{srs}) = Z^2 v(\hat{y}_{srs}) \tag{1b}$$

Ratio Estimator

$$\hat{y}_r = rx \qquad \hat{Y}_r = Z\hat{y}_r \tag{2a}$$

$$r = \frac{\hat{y}_{srs}}{\hat{x}}$$

$$\hat{x} = \frac{\sum_{i=1}^{n} m_i x_i}{m}$$

$$v(\hat{y}_r) = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m_i}\left(\hat{y}_{il} - r x_i\right)^2}{m(m-1)}\frac{(M-m)}{M} \qquad (2b)$$

$$v(\hat{Y}_r) = Z^2 v(\hat{y}_r)$$

Linear Regression Estimator

$$\hat{y}_{lreg} = a + bx$$
$$\hat{Y}_{lreg} = Z\hat{y}_{lreg} \qquad (3a)$$

$$b = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m_i}\left(x_i - \hat{x}\right)\left(\hat{y}_{il} - \hat{y}_{srs}\right)}{\sum_{i=1}^{n} m_i\left(x_i - \hat{x}\right)^2}$$

$$a = \hat{y}_{srs} - b\hat{x}$$

$$v(\hat{y}_{lreg}) = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m_i}\left(\hat{y}_{il} - (a + bx_i)\right)^2}{m(m-2)}\frac{(M-m)}{M} \qquad (3b)$$

$$v(\hat{Y}_{lreg}) = Z^2 v(\hat{y}_{lreg})$$

PPSWOR or Horvitz-Thompson Estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n}\frac{\hat{Y}_i}{\pi_i} \qquad (4a)$$

$$\pi_i = n\frac{z_i}{Z}$$

$$v_1(\hat{Y}_{HT}) = \sum_{i=1}^{n}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)\hat{Y}_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right)\hat{Y}_i\hat{Y}_l \qquad (4b)$$

$$v_2(\hat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_i}\right)^2 \qquad (4c)$$

Generalized Ratio Estimator

$$\hat{Y}_{gr} = rX \qquad (5a)$$

$$r = \frac{\hat{Y}_{HT}}{\hat{X}}$$

$$\hat{X} = \sum_{i=1}^{n}\frac{X_i}{\pi_i}$$

$$v_1(\hat{Y}_{gr}) = \sum_{i=1}^{n}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)e_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right)e_ie_l \qquad (5b)$$

$$v_2(\hat{Y}_{gr}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_i}\right)^2 \qquad (5c)$$

$$e_i = \hat{Y}_i - rX_i$$

Generalized Regression Estimator

$$\hat{Y}_{greg} = \hat{Y}_{HT} + b(X - \hat{X}) \qquad (6a)$$

$$b = \frac{\sum_{i=1}^{n}\frac{X_i\hat{Y}_i}{\pi_i} - \frac{\hat{X}\hat{Y}_{HT}}{\sum_{i=1}^{n}\frac{1}{\pi_i}}}{\sum_{i=1}^{n}\frac{X_i^2}{\pi_i} - \frac{\hat{X}^2}{\sum_{i=1}^{n}\frac{1}{\pi_i}}}$$

$$v_1(\hat{Y}_{greg}) = \sum_{i=1}^{n}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)e_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right)e_i e_l \qquad (6b)$$

$$v_2(\hat{Y}_{greg}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_i}\right)^2 \qquad (6c)$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{HT} - b\hat{X}}{\sum_{i=1}^{n}\frac{1}{\pi_i}}$$

Equations 4c (Yates and Grundy 1953), 5c and 6c are not the equations that Stauffer (1995) recommends but they are believed to be less often negative than compared to equations 4b, 5b and 6b respectively (Raj 1956).

3.2 Critique

3.2.1 Sources of error and bias

There are four sources or levels of variability in Phase II of the VRI survey: polygon-to-polygon, grid-to-grid, point-to-point on the grid, and measurement error. It is possible to estimate the variability between polygons since more than one polygon is selected. An estimate of the grid-to-grid variability is not available since the chosen grid is a permanent grid that is used in each survey. This is not a problem if the grid-to-grid variability can justifiably be assumed to be small. But one possible effect of the fixed grid location is that it may lead to unknown biases. An estimate of the point-to-point or within-polygon variability is available if some of the polygons have more than one ground location that is measured. Given the nature of the current method for selecting ground locations, an

estimate of the within-polygon variability will not be available in those cases where only one ground location is measured in each of the selected polygons. Furthermore, the formulas for the unequal probability estimators (equations 4-6) are only single-stage estimators which do not incorporate the estimated within-polygon variability even when it is available in those cases where more than one ground location is measured in some of the selected polygons. In the cases where the within-polygon variability is not available the assumption that this source of variability is insignificant compared to the between polygon variability could probably be justified. On the other hand, one important requirement of the VRI survey is that it provide reasonable polygon-specific estimates and therefore, it is important to estimate the within-polygon variability. The last level of variability is the measurement error which is the variability in the total per hectare ($\hat{y}_{il}$) calculated for each ground location measured. The assumption is made that there is no error in the total per hectare since the variability in the estimates calculated from the measurements taken in the prism plots and transects at each of the ground locations is not taken into account. Also, it is explicitly assumed that the total per hectare gives an unbiased estimate of the true total per hectare. This may or may not be the case and is beyond the scope of this report.

3.2.2 Polygon selection

One of the purposes for using the OS procedure in the first step of Phase II sampling is to select a sample of polygons with probability proportional to size or PPS. This, however, will only occur if each polygon in the inventory unit is less than or equal to *k*.

Consider the scenario given in the following table for Boston Bar where the selection interval is

TABLE 2: Phase II sample selection scenarios

	Boston Bar List	Lilloet List	Dawson Creek List
Total Area, Z (ha)	273,793.2	1,124,351.2	2,988,590.2
Sample Size, m	200	500	500
Selection Interval, k	1,368.966	2,248.7024	5,977.1804
Max(z_i) (ha)	993.1	12,762.4	12,569.2

larger than the area of every polygon in the inventory unit (Max(z_i) ≤ Z/m). Every polygon in this case is selected with probability z_i/k, and therefore, every polygon chosen from the Boston Bar list is selected with PPS (Figure 1a). The polygon lists for Lilloet and Dawson Creek both contain polygons that are larger than their corresponding selection intervals for the sample sizes given in Table 2. The probability of selecting each polygon in these lists is shown in Figures 1b and 1c. These figures reveal that many of the polygons will be chosen with certainty and not PPS.

3.2.3 Equal probability selection

It is important to note, however, that it is not the intent of the design to have polygons chosen strictly proportional to size as much as it is to have every hectare being chosen

with equal probability. The reason it is desirable for every hectare to be chosen with equal probability is so that the *srs*, ratio and regression estimators of the total (equations 1a, 2a and 3a respectively) can be used. Also, in the cases where several samples have been taken on different occasions from the same population to satisfy different objectives, the results are more easily combined if each hectare has an equal chance of being selected. In theory, the current method tries to ensure that this is the case:

Let g_{ij} be grid point j in polygon i

P(selecting g_{ij}) = P(selecting polygon i) x P(selecting g_{ij} | polygon i is selected)

$$\text{P(selecting polygon } i) = \begin{cases} \frac{z_i}{k} & \text{if } z_i < k \\ 1 & \text{if } z_i \geq k \end{cases}$$

$$\text{P(selecting } g_{ij} \text{ | polygon } i \text{ is selected)} = \frac{\text{Expected number of times polygon } i \text{ is selected}}{z_i}$$

$$= \begin{cases} 1\left(\frac{1}{z_i}\right) = \frac{1}{z_i} & \text{if } z_i < k \\ \left(\frac{z_i}{k}\right)\left(\frac{1}{z_i}\right) = \frac{1}{k} & \text{if } z_i \geq k \end{cases}$$

$$\text{P(selecting } g_{ij}) = \begin{cases} \left(\frac{z_i}{k}\right)\left(\frac{1}{z_i}\right) = \frac{1}{k} & \text{if } z_i < k \\ 1\left(\frac{1}{k}\right) = \frac{1}{k} & \text{if } z_i \geq k \end{cases}$$

Providing there is a one-to-one correspondence between the area of a polygon and the number of grid points, the above shows that, in theory, every grid point in the inventory unit has the same chance of being selected whether there are polygons with areas greater than k or not in the list. A one-to-one correspondence between the area of a polygon and

the number of grid points is probably not true for every polygon in the inventory unit. For example, long skinny polygons may not contain any grid points while others may contain extra ones if many points fall just inside their boundaries. Also, each location won't have an equal probability of being selected in practice if locations are changed in the field from those originally specified in the sample selection. One way this can occur is when a location turns out to be inaccessible for ground measurement. In such situations, an alternative location is selected in a different polygon with similar characteristics. Samples chosen in this way are no longer probability samples and therefore, may lead to unknown biases if equal probability sampling is assumed. The assumption of equal probability sampling will also not hold if an alternative location is chosen within the same polygon as the inaccessible location. In this situation, if inaccessible locations exist within a polygon then the accessible locations in that particular polygon have an increased chance of being selected which would not be the case for those polygons where inaccessible locations do not exist. For example, suppose one ground location is to be selected in a polygon for measurement. If it turns out to be inaccessible in the field, and another one is selected in its place, then two draws from the locations within the polygon are actually made and not one, thus increasing the probability of selecting each ground location within that particular polygon.

3.2.4 Estimators in Stauffer (1995)

If, in practice, not every ground location is sampled with equal probability then the formulas for the *srs* estimate of the total and mean (equation 1a) for the inventory unit are

not unbiased and the various regression estimators of the total and mean (equations 2a and 3a) are not approximately unbiased. Therefore, the *srs* and regression estimators of the total and mean are not appropriate to use unless, in fact, every ground sample is actually selected with equal probability. In any case, their variance estimators are not appropriate regardless of whether the ground samples are selected with equal probability or not because their variance formulas are derived on the basis that every pair of ground locations in the inventory unit has an equal chance of being selected. This is not true in the case of the OS procedure where the joint probability of inclusion is zero for many pairs of ground locations. As a result, the *srs* variance estimator (equation 1b) and the regression variance estimators (equations 2b and 3b) will be biased. Another problem in using the *srs* and regression estimators is that they are not flexible enough to take into account the situation in which the actual ground locations measured differ from those selected at the planning stage.

In order for the VRI survey to provide good variance estimates of the total and mean for the inventory unit, estimators other than the *srs* and regression estimators need to be considered even in the case where ground locations are sampled with equal probability. Stauffer (1995) suggests unequal probability estimators such as the *HT* estimator (equation 4) and the generalized ratio and regression estimators (equations 5 and 6) as other possible alternatives. The equations that Stauffer (1995) recommends as unbiased variance estimators (equations 4b, 5b and 6b) are, in fact, only unbiased if every pair of polygons have joint inclusion probabilities that are greater than zero. This, as already

alluded to, is not the case for the OS sampling method and, therefore, eliminates these three estimators as possible alternatives for estimating the variance.

Out of the various problems outlined, the most serious problem with the Phase II sampling method is in the use of the OS procedure. Since the OS procedure involves sorting the list of polygons before sample selection, many pairs of polygons and ground locations have joint inclusion probabilities of zero. Consequently, the estimated precision will always be biased and, depending on what actually happens in practice, the estimate of the total and mean for equations 1a, 2a and 3a may also be biased. Therefore, it is important to consider other methods of Phase II sample selection and estimation that may provide not only at least approximately unbiased estimates for the total but also for the variance.

4 ALTERNATIVE METHODS OF SAMPLE SELECTION AND ESTIMATION

This section discusses alternative Phase II methods of sample selection and alternative estimators for estimating the total and estimated precision for the inventory unit.

4.1 Two-stage Design

One advantage of the current Phase II sampling method is that its two-step sample selection avoids the large effort involved in setting up a 100 m x 100 m grid for the entire inventory unit and numbering every single one of its grid locations. Therefore, a similar two-stage design will be recommended where the first stage still involves the selection of polygons with the second stage selecting a random sample of ground locations within each of the selected polygons from the first stage. A positive feature of the current method for selecting the sample of polygons is that it uses the information on polygon area gathered in Phase I to weight the sample selection. If the area of the polygon is strongly correlated with some of the attributes that are being ground-measured then this will help to improve the precision of their inventory unit estimates. Auxiliary information available from the other attributes that are estimated in Phase I could also be used to weight the selection of polygons. Given that there is a variety of characteristics measured at each of the selected ground locations, there most likely is not one estimated attribute in Phase I that is correlated with every characteristic that is ground-measured. Consequently, the measure of size that is chosen is likely to help improve only some of the characteristics' estimates

at the inventory unit level. Polygon area is an obvious choice since it is likely to help improve the inventory unit estimates for a number of attributes including tree volume.

In terms of the second-stage sampling, recall that one drawback of the current method is that it does not allow for any flexibility in terms of the number of ground locations that are measured within each of the selected polygons. This in turn could result in a lack of data for estimating within-polygon variability. The second-stage, therefore, will be changed so that it allows for any number of ground locations to be measured in each of the selected polygons in order that sufficient information can be obtained for estimating the within-polygon variability. In order to incorporate properly the estimates of the within-polygon variability, two-stage estimators will be used to calculate the estimates of total and its variance for the inventory unit.

Recall, that in reality, the entire hectare at each of the selected ground locations is not measured and therefore, the total per hectare for each ground location that is measured ($\hat{y}_{il}$) is not known exactly but estimated from measurements taken in prism plots and transects. It will be assumed that the true total per hectare is known for each of the selected ground locations and therefore, the measurement error due to sampling within the hectare will be ignored. Also, since a fixed grid is used, the grid-to-grid variability can never be estimated. Consequently, the two-stage estimators will capture the between and within-polygon variability but will ignore the measurement error and the grid-to-grid variability.

4.2 Description

This section outlines four alternative methods for selecting the first-stage sample of polygons. The two-stage inventory unit estimators incorporating only the ground measurements are given for each of these alternative methods as well as for the previously described OS procedure. The form of the two-stage generalized ratio (*gr*) and generalized regression (*greg*) estimator is also given for each of the five different sampling methods. These adjustment estimators use the estimates of an attribute from both Phase I and Phase II to calculate a total and its estimated precision for the inventory unit. The difference between these two estimators is in the models they use to fit the data. The *gr* estimator fits a relationship between y_i and x_i that is linear with no intercept ($E(y_i) = \beta x_i$) and lets the variance of y_i increase proportionately to x_i ($V(y_i) = \sigma^2 x_i$). The *greg* estimator fits a relationship between y_i and x_i that is linear with an intercept ($E(y_i) = \beta_1 + \beta_2 x_i$) and lets the variance of y_i be constant ($V(y_i) = \sigma^2$). The general form of the two-stage *gr* and *greg* estimate of the total, its approximate variance and estimated precision is outlined in SÖrndal, Swensson and Wretman (1992, pages 313 and 311 respectively).

There has been a considerable amount of work done on unequal probability without replacement sampling evident by the review given by Brewer and Hanif (1983) which looks at 50 such designs. Much of the research has focused on producing strategies that satisfy the following requirements (SÖrndal 1996):

1. Strictly PPS, i.e. $\pi_i = Cz_i$ for all $i = 1, ..., N$, for some positive constant C

2. Fixed sample size
3. Exact design unbiasedness for the estimator of population total and its variance
4. Simple variance calculation

There is a problem, however, in having all four requirements met in that the first three usually conflict with the fourth one. If the first requirement is to be met in the VRI survey, then there is a limit to the number of polygons that can be selected for ground measurement given by the following:

$$n \leq \frac{Z}{\max(z_i)} \qquad i = 1,...,N$$

In order for this restriction to hold, for instance, in the case of the Lilloet list, *n* cannot be greater than 89 which may not be adequate for the kind of precision required. Given that it may not always be practical for *n* to satisfy the above restriction, designs that relax the strictly PPS requirement are considered. There are only three such procedures given in Brewer and Hanif (1983) and they are the Des Raj (1956), Rao-Hartley-Cochran (1962), and Lahiri-Midzuno-Sen (Lahiri 1951, Midzuno 1952, Sen 1953) procedures. A fourth method will also be considered which selects a sample of polygons with replacement and with PPS (PPSWR). Even though sampling without replacement is often preferred over sampling with replacement the difference in the efficiency of the estimators is only important if the sample size is large relative to the population size. This is due to the fact that when the population size is much larger than the sample size there is only a limited chance that a unit is selected more than once. Another reason sampling with replacement

is considered is that it allows for very simple expressions for the estimated variance in multistage sampling.

4.2.1 Des Raj

After each draw, the selected polygon is removed from the list and then the probabilities of selection are recalculated so that the next polygon is drawn with probability proportional to its area. For instance, the probability associated with the first polygon that is drawn is its area, z_i, divided by Z. The probability associated with the second polygon that is drawn is its area, z_j, divided by Z-z_i and so on.

The Raj (1956) and Murthy (1957) estimates of the total are two possible estimators that may be used with this method of sampling. Both Pathak (1967) and Bayless and Rao (1969 and 1970) compared these two estimators and found that the gains in using Murthy's over Raj's estimate and variance estimator is small when n is relatively small compared to N. Given this and the fact that Murthy's variance estimator is substantially more complicated to use, Murthy's estimators will not be considered further. The estimates of the total and variance are given by equations 7a and 7b respectively (Raj 1956):

$$\hat{Y}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\hat{t}_i \qquad (7a)$$

$$\hat{t}_1 = \frac{\hat{Y}_1}{p_1}$$

$$\hat{t}_2 = \hat{Y}_1 + \frac{\hat{Y}_2}{p_2}(1 - p_1)$$

...

$$\hat{t}_n = \sum_{i=1}^{n-1} \hat{Y}_i + \frac{\hat{Y}_n}{p_n}(1 - \sum_{i=1}^{n-1} p_i)$$

$$v(\hat{Y}_{DR}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{t}_i - \hat{Y}_{DR})^2 + \frac{1}{n} \sum_{i=1}^{n} \hat{\upsilon}_i \qquad (7b)$$

$$\hat{\upsilon}_1 = \frac{\hat{s}_1^2}{p_1}$$

$$\hat{\upsilon}_2 = \hat{s}_1^2 + \frac{\hat{s}_2^2}{p_2}(1 - p_1)$$

...

$$\hat{\upsilon}_n = \sum_{i=1}^{n-1} \hat{s}_i^2 + \frac{\hat{s}_n^2}{p_n}(1 - \sum_{i=1}^{n-1} p_i)$$

The following is the generalized ratio estimate of the total and its estimated variance:

$$\hat{Y}_{gr} = rX \qquad (8a)$$

$$r = \frac{\hat{Y}_{DR}}{\hat{X}}$$

$$\hat{X} = \frac{1}{n} \sum_{i=1}^{n} w_i$$

$$w_1 = \frac{X_1}{p_1}$$

$$w_2 = X_1 + \frac{X_2}{p_2}(1 - p_1)$$

...

$$w_n = \sum_{i=1}^{n-1} X_i + \frac{X_n}{p_n}(1 - \sum_{i=1}^{n-1} p_i)$$

$$v(\hat{Y}_{gr}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(d_i - \frac{1}{n} \sum_{i=1}^{n} d_i \right)^2 + \frac{1}{n} \sum_{i=1}^{n} \hat{\upsilon}_i \qquad (8b)$$

$$d_1 = \frac{e_1}{p_1}$$

$$d_2 = e_1 + \frac{e_2}{p_2}(1 - p_1)$$

...

$$d_n = \sum_{i=1}^{n-1} e_i + \frac{e_n}{p_n}(1 - \sum_{i=1}^{n-1} p_i)$$

$$e_i = \hat{Y}_i - rX_i$$

The following is the generalized regression estimate of the total and its estimated variance:

$$\hat{Y}_{greg} = \hat{Y}_{DR} + b(X - \hat{X}) \qquad (9a)$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} z_i - \frac{\hat{X}\hat{Y}_{DR}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i}}}{\frac{1}{n}\sum_{i=1}^{n} h_i - \frac{\hat{X}^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i}}}$$

$$\pi_1 = p_1$$

$$\pi_2 = \frac{p_2}{1 - p_1}$$

...

$$\pi_n = \frac{p_n}{1 - \sum_{i=1}^{n-1} p_i}$$

$$z_1 = \frac{X_1\hat{Y}_1}{p_1} \qquad h_1 = \frac{X_1^2}{p_1}$$

$$z_2 = X_1\hat{Y}_1 + \frac{X_2\hat{Y}_2}{p_2}(1 - p_1) \qquad h_2 = X_1^2 + \frac{X_2^2}{p_2}(1 - p_1)$$

... ...

$$z_n = \sum_{i=1}^{n-1} X_i\hat{Y}_i + \frac{X_n\hat{Y}_n}{p_n}(1 - \sum_{i=1}^{n-1} p_i) \qquad h_n = \sum_{i=1}^{n-1} X_i^2 + \frac{X_n^2}{p_n}(1 - \sum_{i=1}^{n-1} p_i)$$

$$v(\hat{Y}_{greg}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(d_i - \frac{1}{n}\sum_{i=1}^{n} d_i\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{\upsilon}_i \qquad (9b)$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{DR} - b\hat{X}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i}}$$

4.2.2 Rao-Hartley-Cochran

The list of N polygons is divided into n groups at random with N_i polygons in the i^{th} group, $i = 1,...,n$. One polygon from each of the groups is then selected with PPS. This is done by assigning the g^{th} polygon, $g = 1,...,N_i$ in the i^{th} group, the probability of selection z_{ig}/Z_i where Z_i is the total area of the polygons in the i^{th} group. The estimates of the total and variance are given by equations 10a and 10b respectively (Rao et al. 1962):

$$\hat{Y}_{RHC} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i} \tag{10a}$$

$$\pi_i = \frac{z_{ig}}{Z_i}$$

$$Z_i = \sum_{g=1}^{N_i} z_{ig}$$

z_{ig} - area of the selected polygon in the i^{th} group

$$v(\hat{Y}_{RHC}) = \left(\frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \right) \sum_{i=1}^{n} \frac{Z_i}{Z} \left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{RHC} \right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{10b}$$

It is best, in the sense of minimizing the actual precision, if the number of polygons within each group is chosen to be as equal as possible because this will cause the multiplier in the equation for the true variance (which is the same as that in equation 10b) to be minimized. If N/n is an integer, the choice of $N_i=N/n$ will minimize the multiplier. If $N=nR+q$, with R integral and $q<n$, the best choice is to make q groups of size $(R+1)$ and the remaining $(n-q)$ of size R. The following is the generalized ratio estimate of the total and its estimated variance:

$$\hat{Y}_{gr} = rX \tag{11a}$$

$$r = \frac{\hat{Y}_{RHC}}{\hat{X}}$$

$$\hat{X} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$$

$$v(\hat{Y}_{gr}) = \left(\frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \right) \sum_{i=1}^{n} \frac{Z_i}{Z} \left(\frac{e_i}{p_i} - \sum_{i=1}^{n} \frac{e_i}{\pi_i} \right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{11b}$$

$$e_i = \hat{Y}_i - rX_i$$

The following is the generalized regression estimate of the total and its estimated variance:

$$\hat{Y}_{greg} = \hat{Y}_{RHC} + b(X - \hat{X}) \tag{12a}$$

$$b = \frac{\sum_{i=1}^{n} \frac{X_i \hat{Y}_i}{\pi_i} - \left(\frac{\hat{X} \hat{Y}_{RHC}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \right)}{\sum_{i=1}^{n} \frac{X_i^2}{\pi_i} - \frac{\hat{X}^2}{\sum_{i=1}^{n} \frac{1}{\pi_i}}}$$

$$v(\hat{Y}_{greg}) = \left(\frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \right) \sum_{i=1}^{n} \frac{Z_i}{Z} \left(\frac{e_i}{p_i} - \sum_{i=1}^{n} \frac{e_i}{\pi_i} \right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{12b}$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{RHC} - b\hat{X}}{\sum_{i=1}^{n} \frac{1}{\pi_i}}$$

4.2.3 Lahiri-Midzuno-Sen

This method selects a sample of *n* units with probability proportional to its total or aggregate measure of size. This is done by choosing the first polygon with PPS in the usual manner and then drawing a simple random sample of *n*-1 polygons without replacement from the remaining *N*-1 polygons. The estimate of the total is given by equation 13a (Raj 1954).

$$\hat{Y}_{LMS} = Z \frac{\sum_{i=1}^{n} \hat{Y}_i}{\sum_{i=1}^{n} z_i} \tag{13a}$$

One variance estimator suggested by Raj (1954) is given by equation 13b and another possible variance estimator suggested by Rao and Vijayan (1977) is given by equation 13c.

$$v_1(\hat{Y}_{LMS}) = \hat{Y}_{LMS}^2 - \frac{Z}{\sum_{i=1}^{n} z_i}\left[\sum_{i=1}^{n} \hat{Y}_i^2 + 2\left(\frac{N-1}{n-1}\right)\sum_{i=1}^{n}\sum_{j>i}^{n} \hat{Y}_i\hat{Y}_j\right] + Z\frac{\sum_{i=1}^{n} \hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \tag{13b}$$

$$v_2(\hat{Y}_{LMS}) = \frac{Z}{\sum_{i=1}^{n} z_i}\left[\left(\frac{N-1}{n-1}\right) - \frac{Z}{\sum_{i=1}^{n} z_i}\right]\sum_{i=1}^{n}\sum_{j>i}^{n} z_i z_j \left(\frac{\hat{Y}_i}{z_i} - \frac{\hat{Y}_j}{z_j}\right)^2 + Z\frac{\sum_{i=1}^{n} \hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \tag{13c}$$

The following is the generalized ratio estimate of the total and its estimated variance:

$$\hat{Y}_{gr} = rX \tag{14a}$$

$$r = \frac{\hat{Y}_{LMS}}{\hat{X}}$$

$$\hat{X} = Z\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} z_i}$$

$$v_1(\hat{Y}_{gr}) = Z\frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n} z_i} - \frac{Z}{\sum_{i=1}^{n} z_i}\left[\sum_{i=1}^{n} e_i^2 + 2\left(\frac{N-1}{n-1}\right)\sum_{i=1}^{n}\sum_{j>i}^{n} e_i e_j\right] + Z\frac{\sum_{i=1}^{n}\hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \tag{14b}$$

$$v_2(\hat{Y}_{gr}) = \frac{Z}{\sum_{i=1}^{n} z_i}\left[\left(\frac{N-1}{n-1}\right) - \frac{Z}{\sum_{i=1}^{n} z_i}\right]\sum_{i=1}^{n}\sum_{j>i}^{n} z_i z_j\left(\frac{e_i}{z_i} - \frac{e_j}{z_j}\right)^2 + Z\frac{\sum_{i=1}^{n}\hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \tag{14c}$$

$$e_i = \hat{Y}_i - rX_i$$

The following is the generalized regression estimate of the total and its estimated variance:

$$\hat{Y}_{greg} = \hat{Y}_{LMS} + b(X - \hat{X}) \tag{15a}$$

$$b = \frac{Z\frac{\sum_{i=1}^{n} X_i\hat{Y}_i}{\sum_{i=1}^{n} z_i} - \frac{\hat{X}\hat{Y}_{LMS}}{\frac{Z}{\sum_{i=1}^{n} z_i}}}{Z\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} z_i} - \frac{\hat{X}^2}{\frac{Z}{\sum_{i=1}^{n} z_i}}}$$

$$v_1(\hat{Y}_{greg}) = Z\frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n} z_i} - \frac{Z}{\sum_{i=1}^{n} z_i}\left[\sum_{i=1}^{n} e_i^2 + 2\left(\frac{N-1}{n-1}\right)\sum_{i=1}^{n}\sum_{j>i}^{n} e_i e_j\right] + Z\frac{\sum_{i=1}^{n}\hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \quad (15b)$$

$$v_2(\hat{Y}_{greg}) = \frac{Z}{\sum_{i=1}^{n} z_i}\left[\left(\frac{N-1}{n-1}\right) - \frac{Z}{\sum_{i=1}^{n} z_i}\right]\sum_{i=1}^{n}\sum_{j>i}^{n} z_i z_j\left(\frac{e_i}{z_i} - \frac{e_j}{z_j}\right)^2 + Z\frac{\sum_{i=1}^{n}\hat{\sigma}_i^2}{\sum_{i=1}^{n} z_i} \quad (15c)$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{LMS} - b\hat{X}}{\frac{Z}{\sum_{i=1}^{n} z_i}}$$

4.2.4 Probability proportional to size with replacement

One polygon is selected at each of n draws with probability proportional to size (p_i). The selection probability, p_i, associated with each polygon remains the same throughout the draws since polygons are not removed from the list after they have been selected. If a polygon is selected more than once, the whole set of ground locations is replaced and a new independent drawing of m_i ground locations is made. The estimates of the total and variance are given by equations 16a and 16b respectively (Cochran 1977):

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{p_i} \quad (16a)$$

$$v(\hat{Y}_{pps}) = \frac{\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{pps}\right)^2}{n(n-1)} \quad (16b)$$

If a polygon is selected more than once, a different second-stage sample of ground locations is selected each time and its corresponding $\hat{Y}_i$ calculated. Therefore, there is a different $\hat{Y}_i$ value calculated for every repeated selection of a polygon and each of the different $\hat{Y}_i$ values for that polygon are included in calculating the above estimates. The following is the generalized ratio estimate of the total and its estimated variance:

$$\hat{Y}_{gr} = rX \tag{17a}$$

$$r = \frac{\hat{Y}_{pps}}{\hat{X}}$$

$$\hat{X} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{p_i}$$

$$v(\hat{Y}_{gr}) = \frac{\sum_{i=1}^{n}\left(\frac{e_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{p_i}\right)^2}{n(n-1)} \tag{17b}$$

$$e_i = \hat{Y}_i - rX_i$$

The following is the generalized regression estimate of the total and its estimated variance:

$$\hat{Y}_{greg} = \hat{Y}_{pps} + b(X - \hat{X}) \tag{18a}$$

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{X_i\hat{Y}_i}{p_i} - \frac{\hat{X}\hat{Y}_{pps}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}}}{\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^2}{p_i} - \frac{\hat{X}^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}}}$$

$$v(\hat{Y}_{greg}) = \frac{\sum_{i=1}^{n}\left(\frac{e_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{p_i}\right)^2}{n(n-1)} \tag{18b}$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{pps} - b\hat{X}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}}$$

4.2.5 Ordered Systematic

If the previously discussed OS procedure is used to select the first-stage sample of polygons then the following two-stage formulas are recommended to estimate the total and the variance (Raj 1956):

$$\hat{Y}_{HT} = \sum_{i=1}^{n}\frac{\hat{Y}_i}{\pi_i} \tag{19a}$$

$$\pi_i = \begin{cases} \frac{z_i}{k} & \text{if } z_i < k \\ 1 & \text{if } z_i \geq k \end{cases}$$

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{n}\frac{\hat{\sigma}_i^2}{\pi_i} \tag{19b}$$

The following is the generalized ratio estimate of the total and its estimated variance:

$$\hat{Y}_{gr} = rX \tag{20a}$$

$$r = \frac{\hat{Y}_{HT}}{\hat{X}}$$

$$\hat{X} = \sum_{i=1}^{n}\frac{X_i}{\pi_i}$$

$$v(\hat{Y}_{gr}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2 + \sum_{i=1}^{n}\frac{\hat{\sigma}_i^2}{\pi_i} \tag{20b}$$

$$e_i = \hat{Y}_i - rX_i$$

The following is the generalized regression estimate of the total and its estimated variance:

$$\hat{Y}_{greg} = \hat{Y}_{HT} + b(X - \hat{X}) \tag{21a}$$

$$b = \frac{\sum_{i=1}^{n} \frac{X_i \hat{Y}_i}{\pi_i} - \frac{\hat{X}\hat{Y}_{HT}}{\sum_{i=1}^{n} \frac{1}{\pi_i}}}{\sum_{i=1}^{n} \frac{X_i^2}{\pi_i} - \frac{\hat{X}^2}{\sum_{i=1}^{n} \frac{1}{\pi_i}}}$$

$$v(\hat{Y}_{greg}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{21b}$$

$$e_i = \hat{Y}_i - (a + bX_i)$$

$$a = \frac{\hat{Y}_{HT} - b\hat{X}}{\sum_{i=1}^{n} \frac{1}{\pi_i}}$$

4.3 Comparison

The main advantage of using any one of the four alternative procedures over the OS procedure to select a sample of polygons is that they all provide an unbiased estimate of both the total and the estimated precision. Since using any one of the four alternative procedures is hardly more difficult compared to the OS method and given that their estimators of the total and estimated precision are both unbiased and straightforward to calculate, is enough justification to consider using one of them instead to select the sample of polygons. Also, the *gr* and the *greg* estimator derived for any of the above alternative sampling methods will provide approximately unbiased estimates of the total and its estimated precision unlike the adjustment estimators (equations 5 and 6) outlined in

Stauffer (1995). The *gr* and *greg* estimators may help to substantially reduce the standard error of the estimates of the total by incorporating the auxiliary information available from Phase I. If the model is appropriate in describing the relationship between the Phase I and Phase II values then the adjustment estimators will have a considerably smaller variance compared to the two-stage estimates that use Phase II data only. A positive benefit of the *gr* and *greg* estimators is that even if the relationship between the Phase I and Phase II values is not well described by the model being used, it still guarantees approximate unbiasedness since these estimators are model assisted and not model dependent (SÖndal et al. 1992).

5 SIMULATION STUDY

This section discusses the various simulations that were done to explore the current method of Phase II sample selection and its effect on the various estimators given in Stauffer (1995) as well as the alternative methods of polygon selection and the behaviour of their corresponding estimators on estimating the total and its variance for the inventory unit.

5.1 Population

The tree data collected from the first operational trial, conducted in 1995 in the Boston Bar area of the Fraser timber supply area, is used to construct a population of values for the simulations.

5.1.1 Polygons

Photo-interpreted estimates of total tree volume and polygon areas are available for 21,965 polygons that were delineated in Phase I of the trial. Of the 21,965 polygons in the Boston Bar list, 20,809 are vegetated, 1127 are non-vegetated and 29 are unknown. The 29 that are unknown have Phase I estimates of tree volume that are zero and therefore, are characterized as being non-vegetated which gives a total of 1156 polygons that are characterized as non-vegetated. These non-vegetated polygons are each given a total tree volume (Y_i) of zero. Each of the vegetated polygons in the inventory unit is

given a value of tree volume per hectare (y_i) by applying their photo-interpreted or Phase I tree volume per hectare (x_i) to the following equation:

$$y_i = 11.47679 + 0.78302 x_i$$

This equation is based on the relationship between the Phase I and Phase II estimates that were available for 143 polygons in the Boston Bar operational trial. Each y_i, determined for each of vegetated polygons from the above relationship, is multiplied by the polygon area to get the total tree volume (Y_i). Summary statistics of total tree volume for the 21,965 polygons in the population are given in the following table:

TABLE 3: Polygon Summary Statistics

	Polygon Area (ha)	**Polygon Tree Volume (m^3)**
Mean	12.5	1858.1
Median	7.4	274.3
Minimum	0.1	0
Maximum	993.1	120,155.7
Std. Dev.	21.6	4275.7
Total	273,793.2	40,812,522

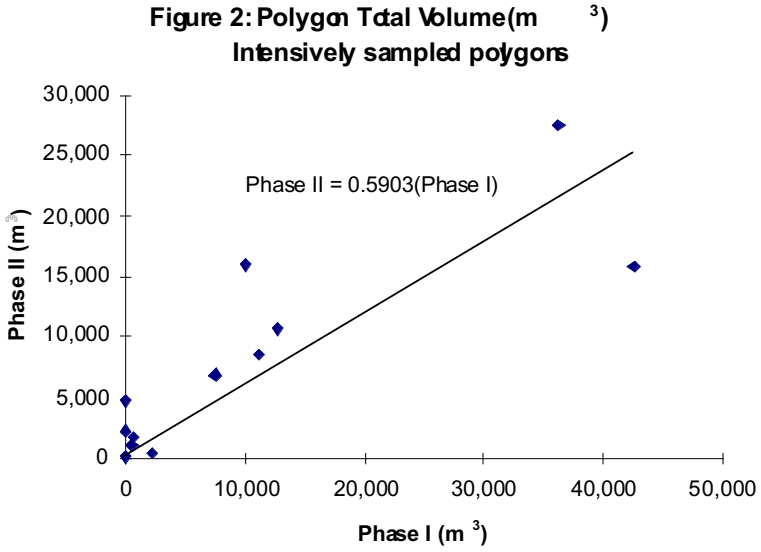
5.1.2 Photo-interpreted values

In order to investigate the behaviour of the adjustment estimators, Phase I estimates of total tree volume for each polygon (X_i) need to be generated. Instead of using the Phase I and Phase II estimates available for all 143 polygons in the Boston Bar trial, this relationship is determined from the Phase I and Phase II estimates for those twelve

polygons that were intensively sampled only. The reason is because many anomalies exist due to subsampling and the fact that only one ground location is measured in each of the other 131 polygons. For instance, in a number of cases the Phase I estimate was zero since the polygon had very few trees. Yet the ground location that was measured happened to fall in one of the places which had the trees. This resulted in a large discrepancy between the Phase I and Phase II estimates. Discrepancies between Phase I and II also occurred when the ground location happened to fall on a bare spot. These discrepancies don't describe the actual differences between reality and the photo estimates but are an artifact of the subsampling. Therefore, a closer or more precise description of the differences between reality and the Phase I estimates may be obtained by looking at the Phase I and II estimates from polygons that have more than one ground location that is measured (Figure 2). Based on the line that is fit to the data in Figure 2, the following equation is used to generate Phase I estimates based on the population values (Y_i):

$$X_i = 1.694Y_i$$

The method used to generate the values for Phase I is not expected to give a perfect representation of the real population but is done to provide a reasonable set of values. By fitting a linear regression line to the data in Figure 2, the assumption is made that there is no variability in the independent variable which, in this case, is photo-interpretation. In other words, if an interpreter was given the same set of photos to interpret twice or three times, the assumption implies that they would come up with exactly the same estimates each time. This assumption is not realistic especially if more than one interpreter is used



in Phase I. No information, however, is available on the magnitude of interpreter variability. Using a linear regression line also assumes that the photo-interpretation process is consistent meaning that if an interpreter's judgment is off, it is off consistently in the same direction and with the same magnitude. It is known, however, that this may not be the case, especially for polygons at the lower tree volume end of the scale. For instance, if there are only a few clusters of trees in a polygon, the interpreter will give an overall estimate of zero tree volume for that polygon. This implies the interpreter will give estimates of zero for polygons with no trees as well as for those with trees up until they cover a certain proportion of the polygon. The effect of the assumptions of consistency and zero variability in the photo-interpretation process on the simulation results is that the adjustment estimators will be more precise than what would actually occur in reality. The results, however, will at least give an indication of the upper bound on the improvement that can be made by using the adjustment estimators.

5.1.3 Second-stage variability

Second-stage variability must also be generated since subsamples of ground locations are taken within each of the polygons selected in the first stage. The intensively sampled polygons from the Boston Bar trial are again used to provide the information for generating this second-stage variability. Figure 3a shows the relationship between the variability of the volume per hectare for the sampled locations within the polygon and its average volume per hectare. The variance appears to increase linearly with the average for the majority of the polygons. Figure 3b is the same as 3a except that the two most influential points have been removed. A log-normal distribution is used to generate the values for the ground samples ($\hat{y}_{il}$) by using y_i as the mean and $169.06y_i$ as the variance of the log-normal distribution. The method used in the simulations to select ground samples does not mimic a fixed grid since different ground-measured values are randomly generated for the selected polygons in each simulation run. This will not have a significant effect on the results since the grid-to-grid variation is expected to be small relative to the other sources of variation. Note that the error in the value at each sampled grid location ($\hat{y}_{il}$) that would normally exist in reality is ignored as well.

5.2 Simulation 1

One set of simulations is conducted where the first-stage sample draws 200 polygons from the inventory unit. In the second stage, only one ground location is measured within each of the polygons selected in the first stage. This process is repeated 1000 times for

both the current and alternative methods of polygon selection and estimation. If polygons are selected which have an area of one hectare or less, it is assumed that the total volume

Figure 3a: Intensively Sampled Polygons
Net.Mer Volume per Hectare

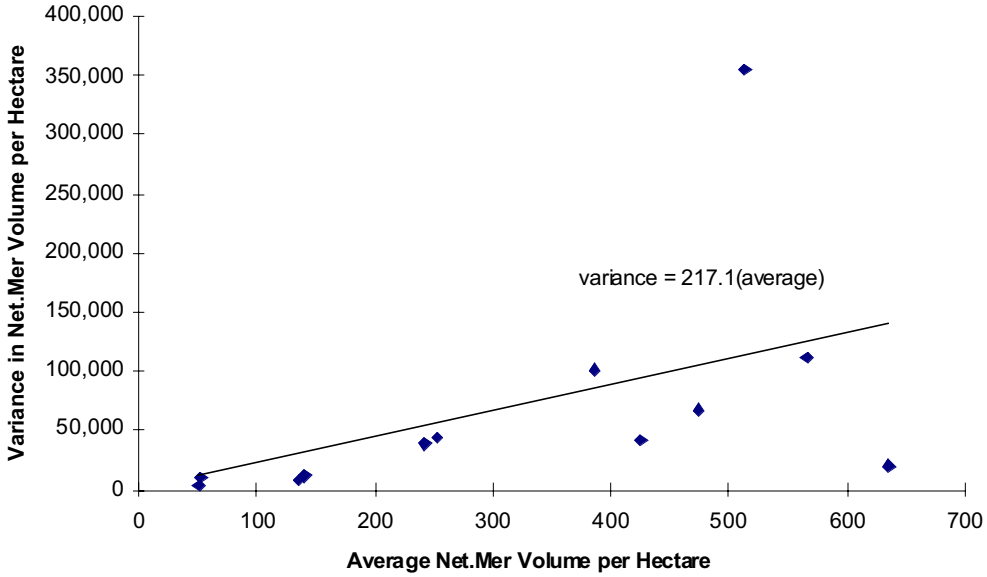
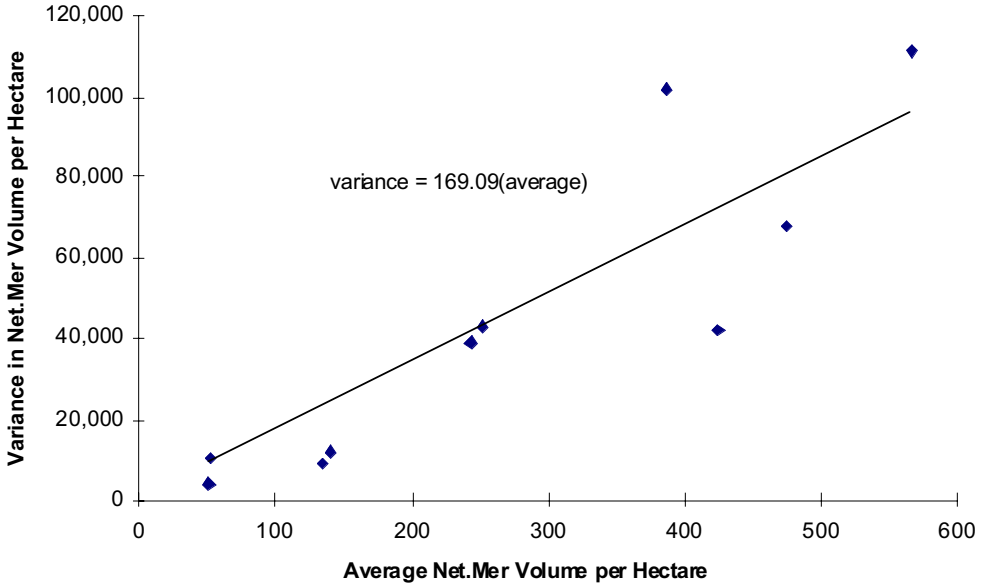


Figure 3b: Two Influencial Points Removed



for the entire polygon is known without error. The results for this set of simulations are given in Tables 4-6. The *Relative Bias* given in Tables 4a, 5a and 6a is calculated as follows:

$$Relative\ Bias = \left(\frac{\bar{\hat{Y}} - Y}{Y} \right) 100\%$$

The *MSE* given in Tables 4b, 5b and 6b is an estimate of the true variance of the total and is calculated as follows (*s* refers to the simulation number):

$$MSE = \frac{\sum_{s=1}^{1000} \left(\hat{Y}_s - \bar{\hat{Y}} \right)^2}{999}$$

Using the above calculation of the *MSE* to estimate the true variability is only valid if the estimators are unbiased. The *Relative Bias* given in Tables 4b, 5b and 6b is calculated as follows:

$$Relative\ Bias = \left(\frac{\overline{v(\hat{Y})} - MSE}{MSE} \right) 100\%$$

The average of the estimated variances, $\overline{v(\hat{Y})}$, calculated for each method, includes all 1000 estimates.

Table 4a shows that each of the unadjusted estimates of total volume give unbiased estimates of the true total. The *srs* and *HT* estimates of the total and *MSE* are exactly the same which is to be expected in this case as explained by Stauffer (1995). Table 4b shows that the *MSE* for the current method of selecting a sample of polygons is smaller compared to the other methods. This seems to indicate that sorting the list of polygons

TABLE 4a: Unadjusted estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*srs* (1a)	40,896,353	0.21%
Current	*HT* (4a)	40,896,353	0.21%
DR	*DR* (7a)	41,021,335	0.51%
RHC	*RHC* (10a)	40,810,273	-0.006%
LMS	*LMS* (13a)	40,823,045	0.03%
PPSWR	*pps* (16a)	40,878,473	0.16%

TABLE 4b: Unadjusted estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*srs* (1b)	$1.63\text{x}10^{13}$	$2.16\text{x}10^{13}$	32%	0
Current	*HT* (4c)	$1.63\text{x}10^{13}$	$-4.06\text{x}10^{15}$	-25,000%	1000
DR	*DR* (7b)	$2.02\text{x}10^{13}$	$2.13\text{x}10^{13}$	5.5%	0
RHC	*RHC* (10b)	$2.26\text{x}10^{13}$	$2.16\text{x}10^{13}$	-4.5%	0
LMS	LMS_1 (13b)	$6.38\text{x}10^{13}$	$6.12\text{x}10^{13}$	-4.0%	377
LMS	LMS_2 (13c)	$6.38\text{x}10^{13}$	$7.01\text{x}10^{13}$	9.9%	493
PPSWR	*pps* (16b)	$2.32\text{x}10^{13}$	$2.20\text{x}10^{13}$	-5.3%	0

TABLE 5a: Regression estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*lreg* (3a)	40,870,332	0.14%
Current	*greg* (6a)	40,865,294	0.13%
DR	*greg* (9a)	40,943,980	0.32%
RHC	*greg* (12a)	40,785,859	-0.07%
LMS	*greg* (15a)	40,903,978	0.22%
PPSWR	*greg* (18a)	40,812,953	0.001%

TABLE 5b: Regression estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*lreg* (3b)	$9.32\text{x}10^{12}$	$9.21\text{x}10^{12}$	-1.2%	0
Current	*greg* (6c)	$9.40\text{x}10^{12}$	$-1.77\text{x}10^{15}$	-19,000%	1000
DR	*greg* (9b)	$9.27\text{x}10^{12}$	$9.32\text{x}10^{12}$	0.54%	0
RHC	*greg* (12b)	$1.04\text{x}10^{13}$	$9.39\text{x}10^{12}$	-10%	0
LMS	$greg_1$ (15b)	$2.14\text{x}10^{13}$	$-3.01\text{x}10^{13}$	-240%	269
LMS	$greg_2$ (15c)	$2.14\text{x}10^{13}$	$2.88\text{x}10^{14}$	1200%	493
PPSWR	*greg* (18b)	$9.66\text{x}10^{12}$	$9.73\text{x}10^{12}$	0.73%	0

TABLE 6a: Ratio estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*r* (2a)	40,870,728	0.14%
Current	*gr* (5a)	40,870,728	0.14%
DR	*gr* (8a)	40,941,298	0.32%
RHC	*gr* (11a)	40,785,402	-0.07%
LMS	*gr* (14a)	40,904,581	0.22%
PPSWR	*gr* (17a)	40,819,243	0.02%

TABLE 6b: Ratio estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*r* (2b)	$9.30\text{x}10^{12}$	$9.20\text{x}10^{12}$	-1.1%	0
Current	*gr* (5c)	$9.30\text{x}10^{12}$	$-1.74\text{x}10^{15}$	-18,800%	1000
DR	*gr* (8b)	$9.25\text{x}10^{12}$	$9.11\text{x}10^{12}$	-1.4%	0
RHC	*gr* (11b)	$1.04\text{x}10^{13}$	$9.16\text{x}10^{12}$	-12%	0
LMS	gr_1 (14b)	$2.14\text{x}10^{13}$	$1.99\text{x}10^{13}$	-6.9%	0
LMS	gr_2 (14c)	$2.14\text{x}10^{13}$	$2.40\text{x}10^{13}$	12%	493
PPSWR	*gr* (17b)	$9.62\text{x}10^{12}$	$9.51\text{x}10^{12}$	-1.1%	0

before sample selection helped to reduce the variability in the estimates of the total. On the other hand, it was mentioned previously that the *srs* and *HT* estimates of the variance are both biased which is evident by the results in Table 4b. The *srs* estimate of the variance overestimates the true variability by over 30% on average and every *HT* estimate of the variance is negative. The fact that the *HT* estimates of the variance are always negative implies that the $\pi_i\pi_j - \pi_{ij}$ term is negative for most pairs of polygons that get selected in the sample. The reason this occurs is because once one polygon is selected, the conditional probabilities of selecting many of those polygons in the sequence using the OS procedure is greater than π_j. Negative variance estimates are also a problem for both of the *LMS* variance estimators. It is expected that the remaining alternative variance estimates in Table 4b would underestimate the true variability since second-stage variability is not taken into account in this simulation. The results show that ignoring the second-stage variability has little effect because the estimated variances are still fairly close to the estimated *MSE*'s.

Tables 5 and 6 compare the regression and ratio estimators respectively for each of the different sampling methods. Similar to the unadjusted estimates of the total, both the regression and ratio estimates of the total are approximately unbiased for all sampling methods. The ratio (*r*) and generalized ratio (*gr*) estimates of the total and *MSE* for the current sampling method in Table 6a and 6b are exactly the same which is to be expected in this case as explained by Stauffer (1995). The estimated *MSE*'s for each of the

adjustment estimators are consistently lower compared to their unadjusted equivalents in Table 4b. This is to be expected given that auxiliary information from Phase I is incorporated in these estimators. The model which actually describes the relationship between the Y_i and X_i values used in the simulations is the ratio model. The regression estimator, however, is based on a different model. Even though the regression model is not the most appropriate model in this case, the regression estimates of the total remain unbiased and the variability in the estimates of the total is still reduced. This demonstrates the fact that the adjustment estimators of the total remain approximately unbiased even if the model that is used is not the best model. The only effect of using a less appropriate model is that it won't reduce the variability as much as if the best model is used. This is evident in the simulation results where the estimated *MSE*'s for the ratio estimators are consistently lower than those for the regression estimators.

Recall that the estimated variances of the linear regression (*lreg*) and ratio (*r*) estimator are biased for the current sampling method. The results in Tables 5b and 6b reveal that the biases are small unlike their unadjusted equivalent, the *srs* estimator. The generalized regression (*greg*) and generalized ratio (*gr*) estimates of the variance for the current and LMS sampling method are negative just like their unadjusted equivalents. It is expected that the other *greg* and *gr* variance estimates would underestimate the true variability because second-stage variability is zero since only one ground sample is selected. Except for the *RHC* adjustment estimators, it appears that excluding second-stage variability had

little effect on the *DR* and *pps* adjustment estimators because their estimated variances are pretty close to their estimated *MSE*'s.

5.3 Simulation 2

This second set of simulations is different from the first set in that 100 polygons are selected instead of 200 and two ground samples are measured in each instead of one. This process is repeated 5000 times for both the current and alternative methods of polygon selection and estimation. If polygons are selected which have an area of two hectares or less, it is assumed that the total volume for the entire polygon is known without error. The results for this set of simulations are given in Tables 7-9. The method column refers to the method used to select a first-stage sample of polygons. The *MSE* given in Tables 7b, 8b and 9b is an estimate of the true variance of the total and is calculated as follows (*s* refers to the simulation number):

$$MSE = \frac{\sum_{s=1}^{5000}\left(\hat{Y}_s - \bar{\hat{Y}}\right)^2}{4999}$$

Since the number of polygons that are selected is half of the number that was selected in the first simulation, the *MSE* is now higher than in Tables 4b, 5b and 6b. This occurs because the between polygon variability is the main source of variability in the inventory unit estimates and therefore, sampling fewer polygons will increase the *MSE*.

Table 7a shows that each of the unadjusted estimates of total volume give unbiased estimates of the true total. Since the same number of subsamples is taken in every polygon that is selected from the Boston Bar list, every ground location still has an equal

chance of being selected when the OS method of selecting a first-stage sample of polygons is used. This, however, is only true if the list of polygons is like the Boston Bar list where the area of each polygon in the inventory unit is smaller than the selection

TABLE 7a: Unadjusted estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*srs* (1a)	40,829,683	0.04%
Current	*HT* (4a)	40,829,683	0.04%
Current	*HT* (19a)	40,829,683	0.04%
DR	*DR* (7a)	40,842,912	0.07%
RHC	*RHC* (10a)	40,681,014	-0.32%
LMS	*LMS* (13a)	40,792,146	-0.05%
PPSWR	*pps* (16a)	40,899,362	0.21%

TABLE 7b: Unadjusted estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*srs* (1b)	2.45×10^{13}	2.17×10^{13}	-11%	0
Current	*HT* (4c)	2.45×10^{13}	-4.74×10^{13}	-293%	5000
Current	*HT* (19b)	2.45×10^{13}	-4.73×10^{13}	-293%	5000
DR	*DR* (7b)	3.47×10^{13}	3.37×10^{13}	-2.7%	0
RHC	*RHC* (10b)	3.39×10^{13}	3.40×10^{13}	0.5%	0
LMS	LMS_1 (13b)	1.06×10^{14}	1.03×10^{14}	-2.6%	1692
LMS	LMS_2 (13c)	1.06×10^{14}	1.01×10^{14}	-5.2%	2434

PPSWR	*pps* (16b)	3.38×10^{13}	3.41×10^{13}	0.7%	0

TABLE 8a: Regression estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*lreg* (3a)	40,806,065	-0.016%
Current	*greg* (6a)	40,802,332	-0.025%
Current	*greg* (21a)	40,802,332	-0.025%
DR	*greg* (9a)	40,872,258	0.15%
RHC	*greg* (12a)	40,806,013	-0.016%
LMS	*greg* (15a)	40,866,076	0.13%
PPSWR	*greg* (18a)	40,807,615	-0.012%

TABLE 8b: Regression estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*lreg* (3b)	$9.34\text{x}10^{12}$	$9.22\text{x}10^{12}$	-1.3%	0
Current	*greg* (6c)	$9.41\text{x}10^{12}$	$-1.41\text{x}10^{13}$	-250%	5000
Current	*greg* (21b)	$9.41\text{x}10^{12}$	$-1.40\text{x}10^{13}$	-249%	5000
DR	*greg* (9b)	$9.67\text{x}10^{12}$	$9.43\text{x}10^{12}$	-2.5%	0
RHC	*greg* (12b)	$9.83\text{x}10^{12}$	$9.58\text{x}10^{12}$	-2.5%	0
LMS	$greg_1$ (15b)	$2.26\text{x}10^{13}$	$-3.89\text{x}10^{13}$	-270%	1316
LMS	$greg_2$ (15c)	$2.26\text{x}10^{13}$	$3.45\text{x}10^{14}$	1400%	2434

PPSWR	*greg* (18b)	9.53×10^{12}	9.44×10^{12}	-0.9%	0

TABLE 9a: Ratio estimators of total volume

Method	**Estimator (Equation)**	$\bar{\hat{Y}}$	***Relative Bias***
Current	*r* (2a)	40,801,203	-0.03%
Current	*gr* (5a)	40,801,203	-0.03%
Current	*gr* (20a)	40,801,203	-0.03%
DR	*gr* (8a)	40,867,920	0.14%
RHC	*gr* (11a)	40,797,242	-0.04%
LMS	*gr* (14a)	40,864,621	0.13%
PPSWR	*gr* (17a)	40,795,846	-0.04%

TABLE 9b: Ratio estimators of the variance

Method	**Estimator (Equation)**	***MSE***	$\overline{v(\hat{Y})}$	***Relative Bias***	**No. negative estimates**
Current	*r* (2b)	$9.27\text{x}10^{12}$	$9.21\text{x}10^{12}$	-0.7%	0
Current	*gr* (5c)	$9.27\text{x}10^{12}$	$-1.34\text{x}10^{13}$	-245%	5000
Current	*gr* (20b)	$9.27\text{x}10^{12}$	$-1.33\text{x}10^{13}$	-244%	5000
DR	*gr* (8b)	$9.41\text{x}10^{12}$	$9.06\text{x}10^{12}$	-3.7%	0
RHC	*gr* (11b)	$9.50\text{x}10^{12}$	$9.23\text{x}10^{12}$	-2.8%	0
LMS	gr_1 (14b)	$2.24\text{x}10^{13}$	$1.98\text{x}10^{13}$	-12%	0
LMS	gr_2 (14c)	$2.24\text{x}10^{13}$	$2.06\text{x}10^{13}$	-8%	2434

PPSWR	*gr* (17b)	9.33×10^{12}	9.10×10^{12}	-2.5%	0

interval. If the number of subsamples taken in each polygon varies in these types of lists, the *srs* estimate of the total will be biased. Table 7b shows that the same problem of negative variance estimates still exist for the *HT* and *LMS* estimators and that the *srs* estimate of the variance is biased in that it underestimates the *MSE* by 11% on average. This *srs* variance is expected to underestimate the *MSE* because it assumes that all points are independent whereas multiple points within the same polygon are likely to be correlated.

Tables 8 and 9 compare the regression and ratio estimators respectively for each of the different sampling methods. Similar to the unadjusted estimates of the total, both the regression and ratio estimates of the total are approximately unbiased for all sampling methods. Similar to the first simulation, the estimated variances of the linear regression (*lreg*) and ratio (*r*) estimator for the current sampling method have very small bias unlike their unadjusted equivalent, the *srs* estimator. The generalized regression (*greg*) and generalized ratio (*gr*) estimates of the variance for the current and LMS sampling method are negative just like their unadjusted equivalents. Recall that the *greg* and *gr* variance estimators for the RHC, DR and PPSWR methods are approximately unbiased. The simulation results in Tables 8b and 9b show that they consistently underestimate the *MSE* by only a small amount on average. These three methods are compared further in Table 10 which gives the average of the estimated standard errors, the variability in the variance estimates and the percent coverage of the 95% confidence intervals. The standard deviation of the 5000 estimated variances is calculated as follows:

$$SD(v(\hat{Y})) = \sqrt{\frac{\sum_{s=1}^{5000}\left(v(\hat{Y}_s) - \overline{v(\hat{Y})}\right)^2}{4999}}$$

TABLE 10: Comparing the variability of the DR, RHC and PPSWR estimators

Method	**Estimator**	***Average SE***	$SD(v(\hat{Y}))$	**95% C.I. coverage**
DR	*DR*	5,760,278	8.80×10^{12}	93.7%
DR	*greg*	2.996,921	4.67×10^{12}	92.1%
DR	*gr*	2,940,151	4.39×10^{12}	91.8%
RHC	*RHC*	5,784,230	9.54×10^{12}	93.4%
RHC	*greg*	3,011,096	6.37×10^{12}	92.3%
RHC	*gr*	2,958,573	6.03×10^{12}	92.6%
PPSWR	*pps*	5,791,435	8.60×10^{12}	93.8%
PPSWR	*greg*	3,000,861	4.57×10^{12}	92.1%
PPSWR	*gr*	2,947,143	4.34×10^{12}	92.2%

The table shows that if the adjusted estimators are used, the average reduction in the *SE* from the unadjusted estimators is just under 50% for all three of the methods. Note that this is an upper bound on the improvement that is possible when using the adjustment estimators because of the assumptions that are made about the photo-interpretation process in the simulations. Table 10 also shows that the estimated precision is not significantly different between each of the three methods.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Preferred method of sample selection and estimation

The first question posed in the introduction that needs to be answered is, "Will the current method used to select a sample of polygons for ground measurement affect the integrity of the inventory unit estimates?". If the OS procedure is used to select a sample of polygons then the *srs*, *r* and *lreg* estimators (equations 1a, 2a and 3a) suggested by Stauffer (1995) are unbiased estimates of the mean and total for the inventory unit but only if every ground location has an equal chance of being selected. The *HT*, *gr* and *greg* estimators (equations 4a, 5a and 6a) suggested by Stauffer (1995) are unbiased estimates of the total regardless of whether the ground locations are selected with equal probability or not. The problem with the OS method for selecting a sample of polygons is that when it sorts the polygons before selection it causes many pairs of polygons to have a joint selection probability of zero. Because of this fact, no unbiased estimate of the variance can be calculated when the OS method of selecting polygons is used. Both sets of simulations show that under the current method, the *srs*, *HT*, *greg* and *gr* estimators of the variance recommended by Stauffer (1995) all perform poorly. Even though the *lreg* and *r* estimators of the variance are biased, the simulations show that the bias is relatively small.

Even though the *lreg* and *r* estimators suggested by Stauffer (1995) appear to give good estimates of the variance, one drawback of using these estimators is that they aren't

flexible in that they can only be used if every measured ground location has an equal chance of being selected. In the case of selecting a sample of 200 polygons from the Boston Bar list, this implies that only one ground location can be measured in each selected polygon. This is an unrealistic constraint since it is desirable to measure some polygons more than once in order to estimate the second-stage variability. This, in fact, was done in the Boston Bar operational trial where twelve polygons had more than one ground location that was measured. The *lreg* and *r* estimators (equations 2 and 3) suggested by Stauffer (1995) won't correctly incorporate the results from the intensively sampled polygons or the second-stage variability.

Since an important feature of the VRI survey design is that it is flexible, it is recommended that estimators which allow the ground samples to be selected with unequal probability to be used in the survey. This way if changes are made in the field and if additional locations are measured to determine within-polygon variability, the validity of the inventory unit estimates won't be affected. Stauffer (1995) recommended the *HT*, *gr* and *greg* estimators for the OS method (equations 4-6) for such cases but are inappropriate given that they produce negative variance estimates. The second simulation revealed that the DR, RHC and PPSWR methods of selecting a sample of polygons are all capable of producing unbiased estimates of the total and its estimated precision. Since neither one appeared to significantly out perform the others in terms of precision, the PPSWR method is being recommended as the method to select the sample of polygons since it is the easiest method to use in terms of both sample selection and estimation.

The second question posed in the introduction that needs to be answered is, "What estimators should be used to estimate attribute totals, means and variances at the inventory unit level?". The simulation results show that the adjustment estimators can only help to improve the precision of the results and therefore, the two-stage *gr* and *greg* estimators for the PPSWR method (equations 17 and 18) are recommended as the ones to be used to estimate the inventory unit estimates. Both estimators should be calculated and the one that gives the smallest estimated variance is the one that should be used. One thing that needs to be kept in mind when using the adjustment estimators, however, is that if confidence intervals are calculated, the actual level of confidence is reduced because the *gr* and *greg* variance estimates slightly underestimate the true variability. But the results of the simulations (Table 10) showed that the level of confidence is reduced only slightly.

One final recommendation is in the area of subsampling and estimating the within polygon variability. Given that the between polygon variability is the main source of variability in the inventory unit estimates, it is better to allocate the majority of the effort to sampling more polygons and less points within the polygons. This, in fact, was done for the Boston Bar operational trial, where the majority of the polygons had only one ground location that was measured. Even if only one location is measured within most of the polygons, information on the within polygon variability can be obtained from those polygons where more than one ground location is measured. For those polygons where more than one ground location is measured, an estimate of the within-polygon variability

($\hat{s}_i^2$) can be calculated. Figures 3a and 3b plot the $\hat{s}_i^2$ values for each of the intensively sampled polygons in the Boston Bar trial against their estimated average volume per hectare ($\hat{y}_i$). These graphs indicate that the within-polygon variability ($\hat{s}_i^2$) increases with its estimated average volume per hectare ($\hat{y}_i$). The linear equation fit to the relationship between the $\hat{s}_i^2$ and $\hat{y}_i$ values for the intensively sampled polygons can be used to predict the within-polygon variability for those polygons where only an estimate of the average volume per hectare is available.

6.2 Future areas for research

The following is a list of areas that should be considered for future investigation:

1. Incorporating the photo-interpreter's measure of confidence for their Phase I estimates into the adjustment estimators.
2. Pre-stratifying the polygons by the variables that were used to sort the list to help reduce variability.
3. Incorporating the Phase I information on the percentage of the polygon covered by tree, shrub, herb and cryptogam into the Phase II estimates.
4. Incorporating the variability in the total per hectare at each measured ground location into the overall estimator.
5. A study which estimates grid-to-grid variation to verify that this additional source of variation is small.

6. Explore through simulation the effect of selecting polygons with certainty and allowing the number of ground-measured points in selected polygons to vary.

7 LIST OF REFERENCES

Bayless, D.L. and Rao, J.N.K. (1970). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). *Journal of the American Statistical Association*, 65, 1645-1667.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, 33, 133-140.

Linnell Nemec, A.F. (1997). *Vegetation Resources Inventory Development and Design*. Draft prepared for the BC Ministry of Forests, Resources Inventory Branch, Victoria, B.C.

Madow, W.G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20, 333-354.

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, 18, 379-390.

Otukol, S. (1996). *Types of Attributes and Proposed Analyses for the BC Vegetation Resources Inventory*. BC Ministry of Forests, Resources Inventory Branch, Victoria, B.C.

Pathak, P.K. (1967). Asymptotic efficiency of Des Raj's strategy - I. *Sankhya, Series A*, 29, 283-298.

Raj, D. (1954). Ratio estimation in sampling with equal and unequal probabilities. *Journal of the Indian Society of Agricultural Statistics*, 6, 127-138.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, 51, 269-284.

Rao, J.N.K. and Bayless, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.

Rao, J.N.K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Journal of the American Statistical Association*, 72, 579-584.

SÖndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÖndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Stauffer, H.B. (1995). *The Statistical Estimation and Adjustment Process for the British Columbia Vegetation Resource Inventory*. Report prepared for the BC Ministry of Forests, Resources Inventory Branch, Victoria, B.C.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, 15, 253-261.