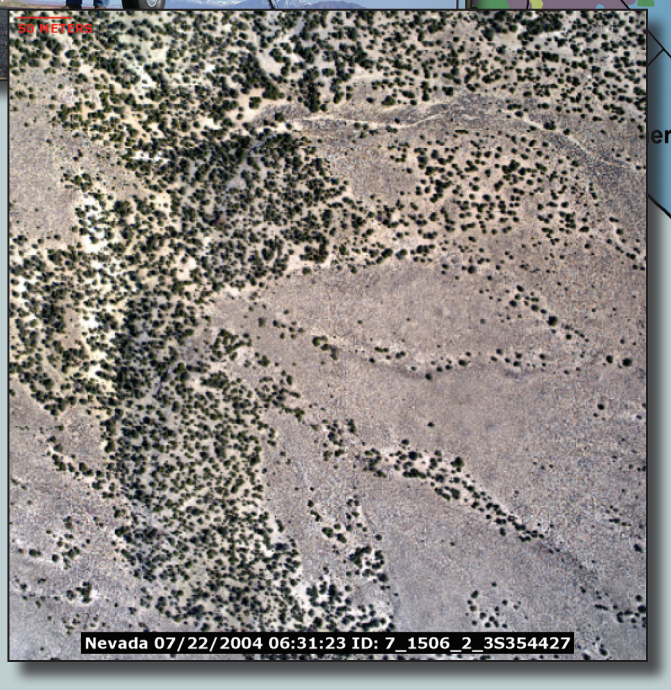
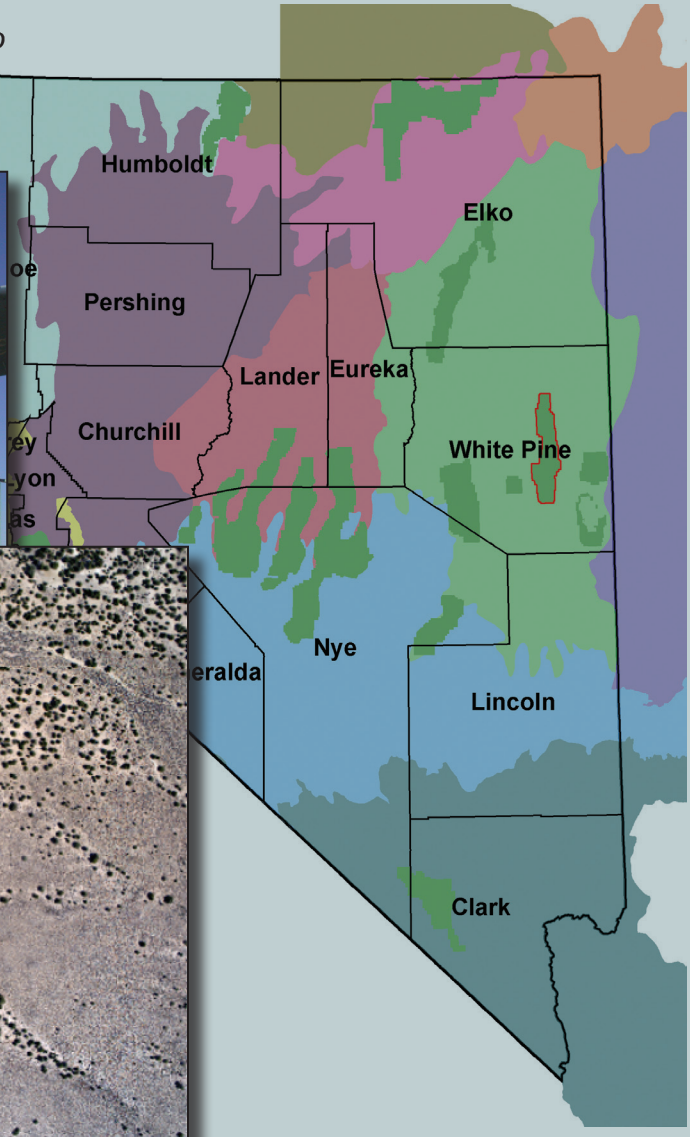
Photo-Based Estimators for the Nevada Photo-Based Inventory

Paul L. Patterson

$$I_{R,\bar{y}_D} = \sum_{h=1}^{H} \int_{R_h} \bar{y}_D(s)ds = \sum_{h=1}^{H} I_{R_h,\bar{y}_D}$$





USDA

United States Department of Agriculture

Forest Service

Rocky Mountain Research Station

Research Paper RMRS-RP-92

July 2012

$$\hat{\hat{I}}_{R_h,\bar{y}_D} = \frac{\|R_h\|}{n_h} \sum_{i=1}^{n_h} \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y(s_{hij})$$

Patterson, Paul L. 2012. **Photo-based estimators for the Nevada photo-based inventory**. Res. Pap. RMRS-RP-92. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 14 p.

Abstract

The U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis Program conducted the Nevada Photo-Based Inventory Pilot in an effort to improve precision in estimates of forest parameters, reduce field data collection costs on margin lands that are covered by slow growing woodland species, and address the potential of strategic-level inventory on lands not traditionally sampled. One part of the project involved the use of large-scale aerial photography instead of traditional field plot visits to produce three types of estimates: (1) area by a variety of forest and nonforest types; (2) percent cover of object types in the landscape; and (3) percent cover of object types within forest or nonforest type. In the context of the infinite sampling paradigm and using the support region construct, we construct the estimators used in the project, derive the variance of the estimator, and give an unbiased estimate of the variance. All estimates are constructed under the assumptions of a stratified sample of an infinite population with an independent simple random sample of each strata and a simple random sample of points in the support region around each sample point in the stratified sample.

Keywords: infinite sampling, support region, inclusion field, photo-based inventory, large-scale aerial photography (LSP)

Authors

Paul L. Patterson is a Statistician for the U.S. Forest Service, Rocky Mountain Research Station, Interior-West Forest Inventory and Monitoring Program, Ogden, UT.

Acknowledgments

The author thanks all the reviewers for their comments that greatly improved the quality of the manuscript, and Tracey Frescino for programming the estimators.

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and number.

Publishing Services

Telephone	(970) 498-1392
FAX	(970) 498-1122
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rmrs
Mailing Address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

Photo-Based Estimators for the Nevada Photo-Based Inventory

Paul L. Patterson

Introduction

The U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis (FIA) Program is a national program that conducts an annual forest inventory on a permanent grid of plots across the United States (Bechtold and Patterson 2005). In an effort to improve precision in estimates of forest parameters, reduce field data collection costs on margin lands that are covered by slow growing woodland species, and address the potential of strategic-level inventory on lands not traditionally sampled by FIA, the Interior-West region of FIA (IW-FIA) conducted the Nevada Photo-Based Inventory Pilot (NPIP) project (Frescino and others 2009). One part of the project involved the use of large-scale aerial photography instead of traditional field plot visits to produce three types of estimates: (1) area by a variety of forest and nonforest types; (2) percent cover of object types in the landscape (for example, tree cover, cover of specific trees species or species grouping, bare ground, and human structures); and (3) percent cover of object types within forest or nonforest type (for example, percent cover of pinyon within pinyon juniper woodland group and percent cover of bare ground within the pinyon juniper woodland group).

The IW-FIA sample is organized into 10 panels, with 1 panel measured each year. Each panel has the same geographic distribution properties as the entire sample, only at one-tenth of the sampling intensity, so each panel can be used to construct estimates under the same methods that pertain to the overall FIA design. For the NPIP project, the State of Nevada was pre-stratified using a pixel-based, 250-m resolution map of predicted timberland, woodland, and nonforested areas (Frescino and others 2009). Within the forested stratum (timberland and woodland classes), all IW-FIA plot locations were selected, and within the nonforested stratum, one panel was selected. For this subset of the IW- FIA plot locations, high-resolution photographs were obtained of an area containing the FIA plot location and a 250-m radius photo-plot was installed on the photograph with the photo-plot center collocated with FIA plot center. For a sample of points within the photo-plot, a photo interpreter assessed two properties for each point: first, the characteristics of the vegetation cover and second, the object the point fell on. The photo-based estimates were constructed using these data.

Historically, FIA derives the properties of estimators based on the finite sampling paradigm, in other words, under the assumption of a finite population. This approach has the theoretical difficulties of specifying what the population unit is and whether area is subdivided into distinct, non-overlapping population units. A different approach is to construct an estimator and derive its properties using the infinite sampling paradigm where the probability sample is a set of points from a continuous population (Cordy 1993). From an infinite sampling perspective, the NPIP sample is a set of clusters of sample points (the clusters are the photo-interpreted points within each photo-plot) with the clusters centered on the FIA plot centers. From this perspective, determining

the probability of selection for a population element is difficult, and this difficulty is compounded for photo-plots that straddle a stratum boundary. Instead, one can consider the photo-plot as a support region for a measurement assigned to the FIA plot center, where a support region is the region over which the measurement is calculated (for example, the proportion of bare ground on a photo-plot is the measurement that is assigned to the photo-plot center and is calculated on the support region of the photo-plot). Then the sample is the FIA plot centers and the photo-interpreted points within each photo-plot are a separate point sample of the support region. An advantage of using a support region is that the independent samples of each stratum are maintained; the disadvantage is the value assigned to the FIA plot center is an average of the attribute of interest over the support region. Stevens and Urquhart (2000) introduced conditions on the support region so that using the average of attribute of interest over the support region produces the same results as using the value of the attribute of interest at the FIA plot center. Cordy (1993) introduced an extension of the Horvitz-Thompson estimator to sampling from an infinite universe.

The purpose of this paper is to: (1) use Steven and Urquhart's results on support regions along with Cordy's (1993) definitions and results to construct an unbiased estimate of the total and unbiased estimated variance of the estimated total; (2) construct an unbiased estimate of the covariance between two estimates; and (3) apply these results to construct estimates and estimated variances for the NPIP study. All estimates are constructed under the assumptions of a stratified sample of an infinite population with an independent simple random sample of each strata and a simple random sample of points in the support region around each sample point in the stratified sample.

The remainder of the paper is two sections: the first contains the construction of the estimator, the derivation of the covariance between two estimates, and a derivation of an unbiased estimator of the covariance; and the second section applies these results to the NPIP study, with a numeric example provided.

Estimator and Estimated Variance

This section is organized into three sub-sections. The first presents the necessary background material of sampling from infinite population and support regions. After presenting Cordy's (1993) extend of the Horvitz-Thompson estimator in the second sub-section, we construct the unbiased estimator. In the third sub-section, the covariance of two estimates is derived, and an estimator of covariance is presented and shown to be an unbiased estimator of covariance. Restricting to a single estimate yields a formula for an unbiased estimator of the variance of the estimator.

Sampling From Infinite Populations and Support Regions

In infinite sampling, the probability sample is a set of points from a continuous universe. Let R be the region that will be sampled; the population characteristics of interest are $I_y = \int_R y(s)ds$, where $y(s)$ is the value of the attribute of interest at point s. For example, if we are interested in percent cover of bare ground over R, let $y(s)$ be equal to 1 if s falls on bare ground, and 0 if s does not fall on bare ground. Then, the percent bare ground is $I_y/||R|| * 100$, where $||R||$ is the area of R. However, we are interested in an estimate of I_y, in instances where an attribute is not measured at a point but is averaged over a neighborhood of support of the point (the neighborhood will also be referred to as a plot). Let $\bar{y}_D(s)$ denote the average value of the attribute over the

neighborhood of support and $D(s)$ denote the plot. The quantity being estimated is $\int_R \bar{y}_D(s)ds$ instead of $\int_R y(s)ds$. Stevens and Urquhart (2000) derive a sufficient condition for $\int_R \bar{y}_D(s)ds$ to equal $\int_R y(s)ds$. The condition is stated in terms of area of $D(s)$ and the area of the inclusion field $D^{-1}(t)$, where the inclusion field is the set of all point s with $t \in D(s)$. The sufficient condition is that the area of plot and the area of inclusion field are equal for all points, in other words,

$$\|D(s)\| = \|D^{-1}(t)\| = \|D\| \qquad [1]$$

If the plot is a circle, then the condition is satisfied for all "internal" points in the region, R, in other words, a point for which the distance to the nearest edge of R is greater than the radius of the circle. For points near the edge, part of the plot falls outside of R. Near the edge, the plot needs to be deformed. Stevens and Urquhart (2000) give examples of plot designs that are circles for interior plots and deformed for points near the edge so that equation [1] holds. An example is shown in Figure 1 (from Stevens and Urquhart 2000). We will assume that a plot design has been adopted that satisfies equation [1].

Construction of Two-Step Estimator

We wish to construct a strategy for estimating $\int_R \bar{y}_D(s)ds$, where $\bar{y}_D(s)$ is the average of y over the support region s centered at D; where a strategy is a combination of a sampling design and an estimator, in other words, an equation (Sarndal 1992). The sample design for the proposed estimator is a two-step process: the first step is a stratified sample of the points in R with sample values $\bar{y}_D(\cdot)$. Then, a sample of the support region around each point in the stratified sample is drawn to estimate $\bar{y}_D$. This is not two-stage sampling with the support regions being the clusters; the support regions

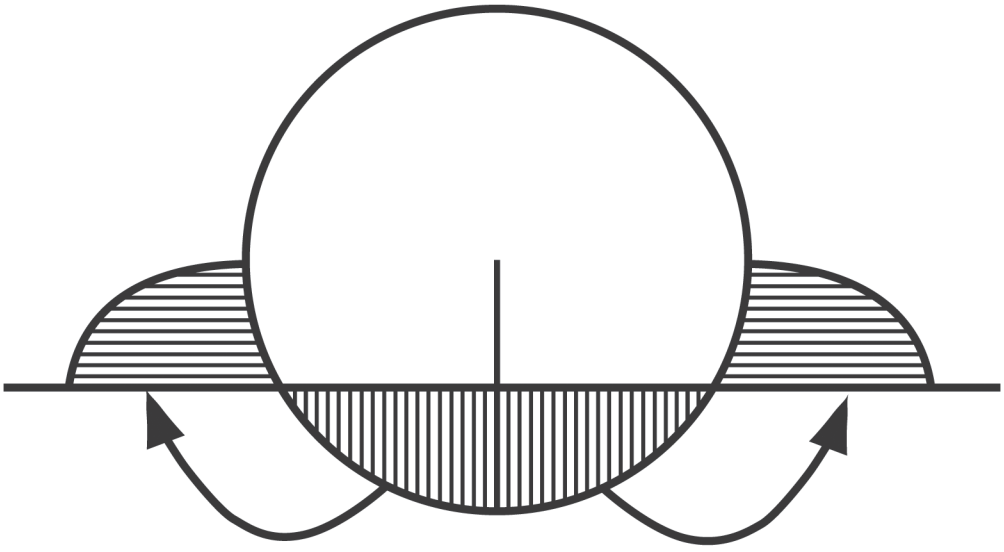


Figure 1—Adapted from Stevens and Urquhart's (2000) figure 12. The part of the plot indicated by the vertical lines is outside the boundary of the region. This portion of the plot is repositioned inside the region—the sections with horizontal lines—so that the condition is preserved.

can overlap, which violates the assumption of disjoint clusters. The process of sampling the support region is mentioned in Stevens and Urquhart (2000). The equation for the estimator is constructed using Cordy's (1993) extension of the Horwitz-Thompson estimator. The specifics of the construction follow.

Cordy (1993) extended the Horwitz-Thompson estimator to continuous domains. We summarize his development and concurrently develop notation. For the region R, we assume z is a Lebesgue integrable function on R and there is a probabilistic sample design defined on R. If S is a sample drawn based on a sample design, then an estimator of $I_z = \int_R z(s)ds$ is defined by

$$\hat{I}_z(S) = \sum_{s \in S} z(s)/\pi(s)$$

where $\pi(s)$ is the inclusion density function; the inclusion density function is the infinite sampling version of first order inclusion probability in finite sampling (Sarndal 1992). Cordy (1993) showed that $\hat{I}_z(S)$ is an unbiased estimator of I_z.

At times, the region will be explicitly included in the notation. For the population characteristic, the region will be denoted using a subscript, in other words, $I_{R,y}$; for the estimator, the region is designated by $\hat{I}_{R,y}(S) = \sum_{s \in S(R)} y(s)/\pi_R(s)$, where π_R is the inclusion density function for the sample design on R, $S(R)$ denotes a sample of the region R, and to simplify the notation, the sample S inside the parenthesis refers to the sample of the region in subscript, that is, $S(R)$.

From this point on, we focus on estimating $\int_R \bar{y}_D(s)ds$, where $\bar{y}_D(s) = \frac{1}{\|D\|}\int_{D(s)} y(t)dt$. We use stratified sampling and the associated estimator. That is, if the region R has been partitioned into H subregions, $R_1, R_2, ..., R_H$, the linearity of the integral implies

$$I_{R,\bar{y}_D} = \int_R \bar{y}_D(s)ds = \sum_{h=1}^{H} \int_{R_h} \bar{y}_D(s)ds = \sum_{h=1}^{H} I_{R_h,\bar{y}_D}$$

If an independent sample $S(R_h)$ is drawn from each R_h, $h = 1, H$, the stratified estimator of $I_{R,\bar{y}_D}$ is given by

$$\hat{I}_{R,\bar{y}_D}(S) = \sum_{h=1}^{H} \hat{I}_{R_h,\bar{y}_D}(S)$$

where, to simplify notation, the sample S inside the parentheses refers to the sample of the region in the subscript, that is, $S(R)$ or $S(R_h)$ in this equation.

When using stratified sampling, the deformation of the plots that is required so that equation [1] holds is relative to the boundary of the entire region and not relative to the boundary of each stratum. For example, if there are two strata, R_1 and R_2, and $s \in R_1$, then D_s can contain points in R_2. This may appear to violate the constraint in stratified sampling that each population element is in one, and only one, stratum. For infinite domains, the population element is the center point, not the plot; the information on the plot is collapsed to the value that is assigned to the center point of the plot, and the center point of the plot is the population element.

We restrict our attention to $I_{R,\bar{y}_D}$ and the estimator

$$\hat{I}_{R_h,\bar{y}_D}(S) = \sum_{s \in S(R_h)} \bar{y}_D(s)/\pi_h(s) \qquad [2]$$

where π_h is the plot-level inclusion density function on R_h.

Instead of calculating $\bar{y}_D(s)=\frac{1}{\|D\|}\int_{D(s)} y(t)dt$, we estimate $I_{D(s),y_D} = \int_{D(s)} y(t)/\|D\|\, dt$ by using an infinite domain Horvitz-Thompson estimator based on a simple random sample of $D(s)$. So

$$\hat{\bar{y}}_D(s) = \hat{I}_{D(s),y_D}(S) = \sum_{t\in S(D(s))} \frac{y(t)}{\|D\|\pi_D(t)} \quad [3]$$

where s represents a sample of $D(s)$ and π_D is the inclusion density function on $D(s)$.

Substituting equation [3] into equation [2], the final form of the estimator is

$$\hat{\hat{I}}_{R_h,\bar{y}_D} = \hat{I}_{R_h,\hat{I}_{D(s),y_D}(S)}(S) = \sum_{s\in S(R_h)} \frac{1}{\pi_h(s)}\left[\sum_{t\in S(D(s))} \frac{y(t)}{\|D\|\pi_D(t)}\right] \quad [4]$$

where, by the convention adopted earlier, the symbol S in the middle expression represents either a sample of subscript R_h or a sample of the subscript $D(s)$.

The variance of the infinite domain Horvitz-Thompson estimator contains an extension of the finite sampling second order inclusion probability. In the next result, which follows directly from example 1 in Cordy (1993), we evaluate inclusion density function and the pairwise inclusion density function for the stratified sampling design on the region R and simple random sample design on support region $D(s)$.

Lemma 1:

(a) For stratified sampling with simple random sample of size n_h within each of the strata, (R_h), the inclusion density function and the pairwise inclusion density function is given by

$$\pi_h(s) = \frac{n_h}{\|R_h\|} \text{ and } \pi_h(s,s') = \frac{n_h(n_h-1)}{\|R_h\|^2}$$

(b) Using simple random sampling for the i^{th} support region in the h^{th} stratum with a sample size of m_{hi}, the inclusion density function and the pairwise inclusion density function is given by

$$\pi_D(t) = \frac{m_{hi}}{\|D\|} \text{ and } \pi_D(t,t') = \frac{m_{hi}(m_{hi}-1)}{\|D\|^2}$$

Denote the sample elements of R_h by s_{hi}, $i = 1, \ldots, n_h$, and denote the sample elements of the simple random sample of $D(s_{hi})$ by s_{hij}, $j = 1, \ldots, m_{hi}$, . Using this notation and lemma 1, equations [2] though [4] can be rewritten as

$$\hat{I}_{R_h,\bar{y}_D}(S) = \frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h} \bar{y}_D(s_{hi}) \quad [5]$$

where S indicates a sample of R_h.

$$\hat{\bar{y}}_D(s_{hi}) = \hat{I}_{D(s_{hi}),y_D}(S) = \frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}} y(s_{hij}) \quad [6]$$

where S indicates a sample of $D(s_{hi})$. Combining, we get

$$\hat{\hat{I}}_{R_h,\bar{y}_D} = \hat{I}_{R_h,\hat{I}_{D(s),y_D}} = \frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D} = \frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h}\frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}} y(s_{hij}) \quad [7]$$

where, to simplify the notation, we have dropped the (S) in the two most left-hand symbols. It is clear from the construction process that $\hat{\hat{I}}_{R_h,\bar{y}_D}$ is an unbiased estimator of $I_{R_h,\bar{y}_D}$.

Variance and Estimated Variance of the Two-Step Estimator

Cordy (1993) derived the variance of $\hat{I}_y(S)$ and gave an unbiased estimator of Var $(\hat{I}_y(S))$; the following proposition extends this to the covariance between estimates of two functions. The proof is a straightforward extension of the proof in Theorem 2 in Cordy (1993).

Proposition 1: Suppose the functions y and z are bounded, $\pi(s) > 0$ for each $s \in R$, and $\int_R (1/\pi(s))ds < \infty$.

(a) Then Cov $(\hat{I}_y(S), \hat{I}_z(S))$ exists and is given by

$$\text{Cov}\left(\hat{I}_y(S),\ \hat{I}_z(S)\right) = \int_R \frac{y(s)z(s)}{\pi(s)}ds + \int_R\int_R y(s)z(s')\left(\frac{\pi(s,s') - \pi(s)\pi(s')}{\pi(s)\pi(s')}\right)ds\,ds'$$

(b) If, in addition, $\pi(s,s') > 0$ for all $s,s' \in R$, then

$$\hat{C}\left(\hat{I}_y(S),\ \hat{I}_z(S)\right) = \sum_{s\in S}\left(\frac{y(s)z(s)}{\pi(s)^2}\right) + \sum_{s\in S}\sum_{\substack{s'\in S\\ s\neq s'}} y(s)z(s')\left(\frac{\pi(s,s') - \pi(s)\pi(s')}{\pi(s,s')\pi(s)\pi(s')}\right)$$

is an unbiased estimator of Cov $(\hat{I}_y(S), \hat{I}_z(S))$.

The following equation is used in the proof of the next proposition and is a straightforward calculation using Lemma 1. For stratified sampling with simple random sample of size n_h within each of the strata, (R_h),

$$\frac{\pi_h(s,s') - \pi_h(s)\pi_h(s')}{\pi_h(s,s')\pi_h(s)\pi_h(s')} = -\frac{\|R_h\|^2}{n_h^2(n_h-1)} \quad [8]$$

We extend the use of the subscripts h and D to indicate conditioning with respect to sample design within the stratum and within the plot, respectively.

Proposition 2: Let y and z be bounded functions on the region R_h, and let $D(s)$ be a plot design that satisfies equation [1] on a region R containing R_h. If a simple random sample of size n_h, $\{s_{hi}\}_{i=1}^{n_h}$, is drawn from R_h, and for each plot $D(s_{hi})$ a simple random sample is drawn, then for the estimator $\hat{\hat{I}}_{R_h,\bar{y}_D}$ defined in equations [4] and [7]

(a) the

$$\mathrm{Cov}\left(\hat{\hat{I}}_{R_h,\bar{y}_D},\hat{\hat{I}}_{R_h,\bar{z}_D}\right)=\mathrm{Cov}_h\left[\hat{I}_{R_h,I_{D(s),y_D}},\hat{I}_{R_h,I_{D(s),z_D}}\right]+\frac{\|R_h\|}{n_h}I_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}$$

(b) and

$$\hat{C}\left(\hat{\hat{I}}_{R_h,\bar{y}_D},\hat{\hat{I}}_{R_h,\bar{z}_D}\right)=\frac{\|R_h\|^2}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hi}),z_D}-\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\sum_{j=1}^{n_h}\hat{I}_{D(s_{hj}),z_D}\right]$$

is an unbiased estimator of $\mathrm{Cov}\left(\hat{\hat{I}}_{R_h,\bar{y}_D},\hat{\hat{I}}_{R_h,\bar{z}_D}\right)$

Proof: (a) We use the standard result

$$\mathrm{Cov}(\hat{I}_y,\hat{I}_z)=\mathrm{Cov}_h\left[E_D(\hat{I}_y),E_D(\hat{I}_z)\right]+E_h\left[\mathrm{Cov}_D(\hat{I}_y,\hat{I}_z)\right] \quad [9]$$

We start by calculating the $E_h\mathrm{Cov}_D$ term on the right-hand side of equation [9].

$$\mathrm{Cov}_D\left(\hat{I}_{R_h,\hat{I}_{D(s),y_D}},\hat{I}_{R_h,\hat{I}_{D(s),z_D}}\right)=\mathrm{Cov}_D\left(\frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D},\frac{\|R_h\|}{n_h}\sum_{j=1}^{n_h}\hat{I}_{D(s_{hj}),z_D}\right)$$

$$=\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}\mathrm{Cov}_D\left(\hat{I}_{D(s_{hi}),y_D},\hat{I}_{D(s_{hi}),z_D}\right)$$

where the second equality follows since the photo-plots are sampled independently. The last expression is the product of $\|R_h\|/n_h$ and $\hat{I}_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}$ since $\mathrm{Cov}_D\left(\hat{I}_{D(\cdot),y_D},\hat{I}_{D(\cdot),z_D}\right)$ is a function defined on R_h and $\hat{I}_{R_h,y}(S)=\frac{\|R_h\|}{n_h}\sum_{s\in S(R_h)}y(s)$. Taking the expectation with respect to the sample design on R_h yields

$$E_h\left[\mathrm{Cov}_D\left(\hat{I}_{R_h,\hat{I}_{D(s),y_D}},\hat{I}_{R_h,\hat{I}_{D(s),z_D}}\right)\right]=E_h\left[\frac{\|R_h\|}{n_h}\hat{I}_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}\right] \quad [10]$$

$$=\frac{\|R_h\|}{n_h}I_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}$$

We now turn our attention to calculating $\mathrm{Cov}_h\left[E_D(\cdot),E_D(\cdot)\right]$ of equation [9]. Note that

$$E_D\left(\hat{I}_{R_h,\hat{I}_{D(s),y_D}}\right)=E_D\left(\frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\right)$$

$$=\frac{\|R_h\|}{n_h}\sum_{i=1}^{n_h}I_{D(s_{hi}),y_D}$$

$$=\hat{I}_{R_h,I_{D(s),y_D}}$$

Thus

$$\mathrm{Cov}_h\left[E_D\left(\hat{I}_{R_h,\hat{I}_{D(s),y_D}}\right),E_D\left(\hat{I}_{R_h,\hat{I}_{D(s),z_D}}\right)\right]=\mathrm{Cov}_h\left(\hat{I}_{R_h,I_{D(s),y_D}},\hat{I}_{R_h,I_{D(s),z_D}}\right) \quad [11]$$

Adding equations [10] and [11] completes the proof of part (a).

(b) To calculate the expectation, we use $E = E_h E_D$. Using the identity $\sum_i a_i \sum_j b_j = \sum_i a_i b_i + \sum_i \sum_{j,j\neq i} a_i b_j$, we first express $\hat{C}\left(\hat{\hat{I}}_{R_h,\bar{y}_D},\hat{\hat{I}}_{R_h,\bar{z}_D}\right)$ in the equivalent form

$$\frac{\|R_h\|^2}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hi}),z_D}-\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\sum_{j=1}^{n_h}\hat{I}_{D(s_{hj}),z_D}\right]=$$

$$\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hi}),z_D}-\frac{\|R_h\|^2}{{n_h}^2(n_h-1)}\sum_{i=1}^{n_h}\sum_{\substack{j=1\\ j\neq i}}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hj}),z_D} \quad [12]$$

The expectation of the first term of expression [12] is

$$E_h E_D\left[\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hi}),z_D}\right]=$$

$$E_h\left[\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}\left[\mathrm{Cov}_D\left(\hat{I}_{D(s_{hi}),y_D},\hat{I}_{D(s_{hi}),z_D}\right)+E_D\left(\hat{I}_{D(s_{hi}),y_D}\right)E_D\left(\hat{I}_{D(s_{hi}),z_D}\right)\right]\right]$$

where E_D was distributed across the sum and we used the identity $Cov_D(X,Y) = E_D(\mathrm{XY}) - E_D(X)E_D(Y)$. The first term in this expression is the product of $\|R_h\|/n_h$ and $\hat{I}_{R_h,Cov_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}$, so we can rewrite the above expression in the equivalent form

$$\frac{\|R_h\|}{n_h}E_h\left[\hat{I}_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}\right]+E_h\left[\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}E_D\left(\hat{I}_{D(s_{hi}),y_D}\right)E_D\left(\hat{I}_{D(s_{hi}),z_D}\right)\right]$$

Since $\hat{I}$ is an unbiased estimator, this expression is equal to

$$\frac{\|R_h\|}{n_h}I_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}+E_h\left[\left(\frac{\|R_h\|}{n_h}\right)^2\sum_{i=1}^{n_h}I_{D(s_{hi}),y_D}I_{D(s_{hi}),z_D}\right] \quad [13]$$

Before continuing with this calculation, we calculate the expectation of the second term of expression [12].

$$E_h E_D\left[-\frac{\|R_h\|^2}{{n_h}^2(n_h-1)}\sum_{i=1}^{n_h}\sum_{\substack{j=1\\ j\neq i}}^{n_h}\hat{I}_{D(s_{hi}),y_D}\hat{I}_{D(s_{hj}),z_D}\right]$$

Distributing the expectation, E_D, across the sum in combination with the independent sampling on the plots implies the previous expression is equal to

$$E_h\left[-\frac{\|R_h\|^2}{{n_h}^2(n_h-1)}\sum_{i=1}^{n_h}\sum_{\substack{j=1\\ j\neq i}}^{n_h}E_D\left(\hat{I}_{D(s_{hi}),y_D}\right)E_D\left(\hat{I}_{D(s_{hi}),z_D}\right)\right]$$

Since $\hat{I}$ is an unbiased estimator, the previous expression is equal to

$$E_h\left[-\frac{\|R_h\|^2}{{n_h}^2(n_h-1)}\sum_{i=1}^{n_h}\sum_{\substack{j=1\\ j\neq i}}^{n_h}I_{D(s_{hi}),y_D}I_{D(s_{hi}),z_D}\right] \qquad [14]$$

By part (b) of Proposition 1 in conjunction with equation [10], the addition of the second term of expression [13] and expression [14] is equal to $E_h\left[\hat{C}_h\left(\hat{I}_{R_h,I_{D(s),y_D}},\hat{I}_{R_h,I_{D(s),z_D}}\right)\right]$. By Proposition 1, $\hat{\mathrm{C}}_h$ is an unbiased estimator of Cov_h, hence the addition of expressions [13] and [14] is equal to

$$\mathrm{Cov}_h\left[\hat{I}_{R_h,I_{D(s),y_D}},\hat{I}_{R_h,I_{D(s),z_D}}\right]+\frac{\|R_h\|}{n_h}I_{R_h,\mathrm{Cov}_D\left(\hat{I}_{D(s),y_D},\hat{I}_{D(s),z_D}\right)}$$

and the proof is complete.

In Proposition 2, the case of y and z being the same function gives the equation for the variance and estimated variance of $\hat{\hat{I}}_{R_h,\bar{y}_D}$. For those familiar with two-stage sampling for finite populations, there is a similarity between the variance and estimated variance in Proposition 1 and the equations for variance and estimated variance for two-stage sampling for finite populations (Cochran 1977: Theorems 10.1 and 10.2). However, there is a striking dissimilarity. If we view the plots as clusters, then as with finite sampling, the the variance contains a between cluster component, $\mathrm{Var}_h\left[\hat{I}_{R_h,I_{D(s),y_D}}\right]$, and a within cluster component, $I_{R_h,\mathrm{Var}_D\left(\hat{I}_{D(s),y_D}\right)}$; what is dissimilar is the estimated variance in Proposition 2 does not contain a within cluster component. In both this setting and the two-stage sampling for finite populations, the between cluster component of the estimated variance contains information about the within cluster component of the variance. The difference is, in this setting, the expectation of the between cluster component of estimated variance contains both an unbiased estimate of the between cluster component of variance and the within cluster component of the variance. In two-stage sampling, the expectation of estimate for the between cluster component of the variance contains an unbiased estimate of the between cluster component of the variance and a biased estimate of the within cluster component of the variance and hence needs to be "adjusted" with a within cluster component to produce an unbiased estimate of the variance. This remark is from an intuitive point of view; formally, the plots are not clusters since they

can straddle strata boundaries and a primary sampling unit (cluster) can be in one, and only one, stratum. In the construction of $\hat{\hat{I}}_{R_h,\bar{y}_D}$, the information on the plot is collapsed to the primary sampling unit, which is the center of the plot; the sample of the plot is not a subsample but a separate sample from the sample of the region.

The following restatement of Proposition 2 will be useful in the next section.

Proposition 2.A. Using the same assumptions as Proposition 2, let $\bar{\hat{I}}_{D,y_D}$ *denote* $\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{I}_{D(s_{hi}),y_D}$,

where $\hat{I}_{D(s_{hi}),y_D}(S) = \frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}} y(s_{hij})$, *with* s_{hij}, $j = 1, \ldots, m_{hi}$ *the simple random sample of* $D(s_{hi})$,

then

(a) $\hat{C}\left(\hat{\hat{I}}_{h,\bar{y}_D}, \hat{\hat{I}}_{h,\bar{z}_D}\right) = \frac{\|R_h\|^2}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}\left(\hat{I}_{D(s_{hi}),y_D} - \bar{\hat{I}}_{D,y_D}\right)\left(\hat{I}_{D(s_{hi}),z_D} - \bar{\hat{I}}_{D,z_D}\right)\right]$

(b) $\hat{V}\left(\hat{\hat{I}}_{h,\bar{y}_D}\right) = \frac{\|R_h\|^2}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}\left(\hat{I}_{D(s_{hi}),y_D} - \bar{\hat{I}}_{D,y_D}\right)^2\right]$

Application to Nevada Photo-Based Inventory

A complete description of the sample design for the NPIP is contained in Frescino and others (2009). We present the details that are germane to the development of the estimators and their estimated variances.

The FIA sample design has been described as a quasi-systematic sample in that plot locations were randomly selected within hexagons, which are a systematic tiling of the United States (Bechtold and Patterson 2005). For the NPIP project, the State of Nevada was pre-stratified into predicted timberland/woodland forest and nonforested areas (Frescino and others 2009). Within the forested stratum, all IW-FIA plot locations were selected, and within the nonforested stratum, one panel was selected. For the purpose of estimation, FIA treats the FIA quasi-systematic sample as a simple random sample (Bechtold and Patterson (2005: p. 25). We treat the NPIP sample as a pre-stratified sample, with a simple random sample of each stratum.

For each 250-m radius photo-plot that was co-located at the FIA plot location, a systematic grid of 49 points was located within each photo-plot. Cochran (1977) showed that systematic sampling has a smaller variance than simple random sampling when units within systematic samples have greater variability than the population as a whole. We assume the systematic grid that was located within each photo-plot does not coincide with any systematic land feature within the photo-plot; hence, treating the systematic sample as a simple random sample produces a conservative estimate of the variance.

The other assumption that was used in the development of the estimators in the "Estimator and Estimated Variance" section is the plot design satisfied equation [1]. All of the photo-plots were sufficient distance from the estimation unit boundaries that they did not cross the estimation unit boundaries; thus, they satisfy equation [1].

For the photo-plot portion of NPIP, three classes of population characteristics are of interest: (1) area of land classified as being in a condition, for example, proportion of land in pinyon-juniper forest type, or proportion of land that is privately owned and in pinyon-juniper forest type; (2) the percent cover of the land by an object type, for example, percent cover of land by bare ground, or the percent cover of land by pinyon trees; and (3) the percentage of a condition that has an attribute, for example, percentage

of bare ground in the pinyon-juniper forest type, or the percentage of pinyon-juniper forest type covered by juniper trees. Estimates of these population characteristics use the following three functions:

$$y_c(s) = \begin{cases} 1, & \text{if } s \text{ is in condition } c \\ 0, & \text{otherwise} \end{cases}$$

$$y_o(s) = \begin{cases} 1, & \text{if } s \text{ falls on object } o \\ 0, & \text{otherwise} \end{cases}$$

$$y_{o|c}(s) = \begin{cases} 1, & \text{if } s \text{ is in domain } c \text{ and falls on object } o \\ 0, & \text{otherwise} \end{cases}, \text{ note } y_{o|c}(s) = y_c(s)y_o(s)$$

All the population characteristics can be stated in terms of proportions; for example, the area of land classified as being in condition c, I_{R,y_c}, equal to $\|R\|P_c$, where P_c is the proportion of area classified as being in condition c. The rest of this paper is devoted to estimating the proportion of land by characteristic or type (classes 1 and 2) and the proportion of land within a condition that has an attribute (class 3).

Proportion of Condition or Cover

When estimating the proportion of a region R that is classified as being in condition c or the proportion of R that is covered by object o, the function y_c is used for conditions, while the function y_o is used for objects. The equations are stated in terms of the proportion of land classified as being in condition c. To obtain the proportion of land covered by object o, simply substitute y_o for y_c in the equations.

Under the assumption that the photo-plot design satisfies equation [1], then the proportion of land in region R classified as being in condition c is $P_c = \frac{1}{\|R\|} I_{R,\bar{y}_{cD}}$, which we can estimate using

$$\hat{P}_c = \frac{1}{\|R\|}\hat{I}_{R,\bar{y}_{cD}} = \frac{1}{\|R\|}\sum_{h=1}^{H}\hat{\hat{I}}_{R_h,\bar{y}_{cD}} \qquad [15]$$

The functional notation used in "Estimator and Estimated Variance" section was useful for the derivation and proofs, but the notation is not standard to FIA. We introduce notation that is more familiar to FIA users.

Let s_{hij} be the jth point of the ith photo-plot in stratum h. If we let

$$y_{chij} = y_c(s_{hij}), \quad p_{chi} = \frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}} y_{chij} \text{ and } \bar{p}_{ch} = \frac{1}{n_h}\sum_{i=1}^{n_h} p_{chi} \qquad [16]$$

then from equations [6] and [7],

$$\hat{I}_{D(s_{hi}),y_{cD}}(S) = p_{chi} \text{ and } \hat{\hat{I}}_{R_h,\bar{y}_{cD}} = \|R_h\|\bar{p}_{ch} \qquad [17]$$

Combining equations [15] and [17] yields

$$\hat{P}_c = \sum_{h=1}^{H} W_h \bar{p}_{ch}, \text{ where } W_h = \|R_h\|/\|R\| \qquad [18]$$

An unbiased estimated variance of $\hat{P}_c$ follows from first noting that since the samples of the strata are independent, equation [17] implies $\hat{V}(\hat{P}_c)$ is equal to $\|R\|^{-2}\sum_{h=1}^{H}\hat{V}\left(\hat{\bar{I}}_{R_h\bar{y}_{cD}}\right)$, which, combining with Proposition 2.A and $\bar{\hat{I}}_{D,y_D}=\bar{p}_{ch}$, implies

$$\hat{V}(\hat{P}_c)=\sum_{h=1}^{H}W_h^2\frac{1}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}(p_{chi}-\bar{p}_{ch})^2\right] \qquad [19]$$

with the usual calculational form

$$\hat{V}(\hat{P}_c)=\sum_{h=1}^{H}W_h^2\frac{1}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}p_{chi}^2-\frac{1}{n_h}\left(\sum_{i=1}^{n_h}p_{chi}\right)^2\right] \qquad [20]$$

Percent Cover Within Condition

We denote the proportion of the condition *c* covered by attribute *o* by $P_{o\subset c}$. Since spatial distribution of condition *c* is unknown, $P_{o\subset c}$ is the ratio of the proportion that is both classified as condition *c* and covered by object *o* to the proportion that is classified as condition *c*, that is $P_{o\subset c}=\mathrm{P}_{o|c}/\mathrm{P}_c$. An estimate of $P_{o\subset c}$ is given by

$$\hat{P}_{o\subset c}=\frac{\hat{P}_{o|c}}{\hat{P}_c} \qquad [21]$$

where

$$\hat{P}_{o|c}=\sum_{h=1}^{H}W_h\bar{p}_{o|ch},\ \text{with}\ \bar{p}_{o|ch}=\frac{1}{n_h}\sum_{i=1}^{n_h}p_{o|chi}\ \text{ and }\ p_{o|chi}=\frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}}y_{o|chij} \qquad [22]$$

An approximate variance is calculated using Taylor linearization (see Särndal and others 1992: p. 172-175 and bottom p. 177 through p. 178).

$$Var(\hat{P}_{o\subset c})=\frac{1}{P_c^2}\left[Var(\hat{P}_{o|c})+P_{o\subset c}^2Var(\hat{P}_c)-2P_{o\subset c}Cov(\hat{P}_{o|c},\hat{P}_c)\right]$$

An estimate of the approximate variance is obtained by estimating all of the unknown quantities, that is,

$$\hat{V}(\hat{P}_{o\subset c})=\frac{1}{\hat{P}_c^2}\left[\hat{V}(\hat{P}_{o|c})+\hat{P}_{o\subset c}^2\hat{V}(\hat{P}_c)-2\hat{P}_{o\subset c}\hat{C}(\hat{P}_{o|c},\hat{P}_c)\right] \qquad [23]$$

where, by Proposition 2.A and equation [22]

$$\hat{C}(\hat{P}_{o|c},\hat{P}_c)=\sum_{h=1}^{H}W_h^2\frac{1}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}(p_{o|chi}-\bar{p}_{o|ch})(p_{chi}-\bar{p}_{ch})\right] \qquad [24]$$

Which, for computational proposes, can be expressed as

$$\hat{C}(\hat{P}_{o|c},\hat{P}_c)=\sum_{h=1}^{H}W_h^2\frac{1}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}p_{o|chi}p_{chi}-\frac{1}{n_h}\sum_{i=1}^{n_h}p_{o|chi}\sum_{j=1}^{n_h}p_{chi}\right] \qquad [25]$$

Example

To illustrate the use of these estimators, a couple of results, using the NPIP data are presented. To estimate the percentage of Nevada covered by pinyon-juniper forest type, use equation [18] (multiplied by 100) with $y_c(s)$ equal to 1 if the point is in the pinyon-juniper forest type and 0 otherwise. The estimate is 13.5 percent with a standard error of 0.55, with the standard error calculated by multiplying the square root of equation [20] by 100. The estimates of percent of land cover by forest type are standard for FIA; what is not standard is coverage by object type. For instance, the percentage of land covered by the various sage species (denoted sage complex) is 12.3, with a standard error of 0.49 (see Frescino and others [2009] for the definition of sage complex). The estimate was calculated using equations [18] and [20] with the function $y_o(s)$ equal to 1 if the point s falls on a member of the sage complex and 0 otherwise. Similarly, the percentage of land covered by other shrub types is 19.7, with a standard error of 0.5, and the percentage of land covered by soil or rock is 45.1, with a standard error of 0.66.

The above percentages are for the entire State of Nevada; what can also be estimated is the percentage of the pinyon-juniper forest type that is covered by sage complex, other shrubs, and soil or rock. These are estimated using equation [21] and the standard error estimated using the square root of equation [23]. First, estimate the percentage of land covered by sage complex within the pinyon-juniper forest type; this estimate is calculated using equation [15] and the function $y_{o|c}(s)$, which is 1 if the point s falls on a sage complex and is in pinyon-juniper forest type and is 0 otherwise. The percentage is 1.1; the ratio of this estimate and the estimate of pinyon-juniper forest type cover yields the estimate of 8.2 percent of pinyon-juniper forest type is covered by the sage complex, with a standard error of 0.61. Similarly, the percentage of the pinyon-juniper forest type covered by other shrubs is 7.4, with a standard error of 0.43; while the percentage of pinyon-juniper forest type covered by soil or rock is 44.9, with a standard error of 0.95. As would be expected, the percentage of soil or rock cover within forest types that typically have a high density of tree cover is significantly less; for example, 8.5 percent of aspen forest type is covered by soil or rock, with a standard error of 2.

Conclusion

An unbiased estimator was derived for the total of a population attribute from an infinite population. The assumptions are: (1) a stratified sample of an infinite population with an independent simple random sample of each stratum; and (2) a simple random sample of points in the support region around each sample point in the stratified sample. Readers who are mostly interested in calculating estimates should reference equation [7]. The forms of an unbiased estimated variance and an unbiased estimated covariance for use in calculations are presented in Proposition 2.A.

These estimators were applied to the situation of using high-resolution photographs to estimate for a region the (1) area of land classified as being in a condition; (2) the percent cover of the land by an object type; and (3) the percentage of a condition that is covered by an object type. For items 1 and 2, the form estimator used in calculations is given in equation [18] and the estimated variance is presented in equation [20]. For item 3, a ratio estimator is used (equation [21]), and an estimated variance is presented in equations [23] and [24].

References

Bechtold, W.A.; Patterson, P.L., editors. 2005. The enhanced Forest Inventory and Analysis program–national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station.

Cochran, W.G. 1977. Sampling Techniques. 3rd ed. John Wiley, New York. 428 p.

Cordy, C.B. 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Statistics & Probability Letters. 18: 353-362.

Frescino, Tracey S.; Moisen, G.G.; Megown, K.A.; Nelson, V.J.; Freeman, E.A.; Patterson, P.L.; Finco, M.; Brewer, K; Menlove, J. 2009. Nevada photo-based inventory pilot (NPIP) photo sampling procedures. Gen. Tech. Rep. RMRS-GTR-222. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 30 p.

Särndal, C.; Swensson, B.; Wretman, J. 1992. Model assisted survey sampling. Springer-Verlag, New York. 694 p.

Stevens, D.L.; Urquhart, S. 2000. Response designs and support regions in sampling continuous domains. Enviornmetrics. 11: 13-41.

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of the National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals. Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Station Headquarters
Rocky Mountain Research Station
240 W Prospect Road
Fort Collins, CO 80526
(970) 498-1100

Research Locations

Flagstaff, Arizona	Reno, Nevada
Fort Collins, Colorado	Albuquerque, New Mexico
Boise, Idaho	Rapid City, South Dakota
Moscow, Idaho	Logan, Utah
Bozeman, Montana	Ogden, Utah
Missoula, Montana	Provo, Utah

www.fs.fed.us/rmrs

To learn more about RMRS publications or search our online titles:

www.fs.fed.us/rm/publications

www.treesearch.fs.fed.us