ACCURACY OF REMOTELY SENSED CLASSIFICATIONS FOR STRATIFICATION OF FOREST AND NONFOREST LANDS[1]

Raymond L. Czaplewski and Paul L. Patterson[2]

Abstract—We specify accuracy standards for remotely sensed classifications used by FIA to stratify landscapes into two categories: forest and nonforest. Accuracy must be highest when forest area approaches 100 percent of the landscape. If forest area is rare in a landscape, then accuracy in the nonforest stratum must be very high, even at the expense of accuracy in the forest stratum. Accuracy in both strata must be at least 90 percent to achieve appreciable gains in efficiency. We recommend that new remotely sensed data be used to re-stratify landscapes whenever the area in forestland decreases by five percent or more since the previous stratification. Efficiency can increase up to 15 percent with formation of an "indeterminate" stratum, which contains elements that are most likely to be misclassified.

INTRODUCTION

The USDA Forest Service's Forest Inventory and Analysis (FIA) uses a systematic sample of field plots to characterize forest conditions over 300,000,000 hectares of forest and woodland ecosystems in the USA. FIA measures 364,000 1-ha field plots, 120,500 of which are forested. Remote sensing can improve accuracy of FIA statistical estimates. For example, FIA interprets aerial photography for a systematic sample of 9,400,000 plots to improve estimates of forest area and population totals. Stratification with wall-to-wall Landsat satellite data could replace photo-interpretation within the next few years.

We specify accuracy standards for remotely sensed classifications that will be used for stratification into two categories: forest and nonforest. Application of our recommendations requires assumptions, such as the expected prevalence of forestlands in the population. We make generalizations regarding the loss in efficiency caused by change in land use since acquisition of remotely sensed data. These generalizations help determine how old remotely sensed images can become before new imagery is needed for stratification. We also consider formation of a stratum for elements that are most likely to be misclassified. We make recommendations that can help determine a priori the size of this "indeterminate" stratum.

SAMPLE SURVEY ESTIMATORS

Assume a population is subdivided into two sub-populations, such as forest and nonforest. Our goal is estimation of the proportion $P(A_i)$ of each sub-population A_i in the population, where $0<P(A_i)<1$. A simple transformation converts this proportion into a percentage or an area (e.g., number of hectares). Assume every element of the sampled population is composed of one and only one sub-population category, which justifies the binomial distribution. We introduce the "error matrix" for remotely sensed classifications and make the connection to statistical stratification, start with the estimator, give the estimators for simple random sampling and for stratification, and define the "design effect" as a measure of the gain in statistical efficiency with stratification.

Error Matrix for Remote Sensing

The "error matrix" or "confusion matrix" describes accuracy in the remote sensing literature (e.g., Congalton 1991). Let $P(B_j)$ be the proportion of the population in stratum B_j, let $P(A_i \cap B_j)$ denote the proportion of the population that is jointly in sub-population A_i and remotely sensed stratum B_j, and let $P(A_i|B_j)$ denote the proportion of sub-population A_i given that the remotely sensed stratum is B_j, where $P(A_i|B_j)=P(A_i \cap B_j)/P(B_j)$. Figure 1 gives the mathematical notation that we use for the error matrix. We assume remotely sensed classifications are used to define each stratum, and remote sensing measures the size, or area, of each stratum, *i.e.*, $P(B_j)$.

The ideal stratification occurs when each sub-population occurs in one and only one stratum (Cochran 1977). However, remote sensing does not have 100 percent accuracy, and each remotely sensed stratum usually contains both sub-populations. For example, let stratum B_1 be classified as forest with wall-to-wall Landsat data; however, not all sites that are truly forested will be included in this stratum.

The sample of field plots is used to estimate the distribution $P(A_i|B_j)$ of each sub-population A_i within each remotely sensed stratum B_j. Remote sensing improves statistical estimates of each sub-population $P(A_i)$ by introducing ancillary data, namely precise measurement of the size $P(B_j)$ for each stratum.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Project Leader and Mathematical Statistician, USDA Forest Service, Rocky Mountain Research Station, 2150 Centre Ave., Bldg. A, Suite 350, Fort Collins, CO 80526–1891, respectively.

1a. Joint probabilities

$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	$P(A_1) = \sum_{j=1}^{m} P(A_2 \cap B_j)$
$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	$P(A_2) = \sum_{j=1}^{m} P(A_2 \cap B_j)$
$P(B_1) = \sum_{i=1}^{m} P(A_i \cap B_1)$	$P(B_2) = \sum_{i=1}^{m} P(A_i \cap B_2)$	$\sum_{i=1}^{m} \sum_{j=1}^{m} P(A_i \cap B_j) = 1$

1b. Conditional probabilities (accuracy) within strata

$P(A_1 \mid B_1) = \frac{P(A_1 \cap B_1)}{B_1}$	$P(A_1 \mid B_2) = \frac{P(A_1 \cap B_2)}{B_2}$
$P(A_2 \mid B_1) = \frac{P(A_2 \cap B_1)}{B_1}$	$P(A_2 \mid B_2) = \frac{P(A_2 \cap B_2)}{B_2}$
$\sum_{i=1}^{m} P(A_i \mid B_1) = 1$	$\sum_{i=1}^{m} P(A_i \mid B_2) = 1$

Figure 1—The "error matrix", or "confusion matrix", describes classification accuracy with remotely sensed data. The goal is estimation of the prevalence or size of each sub-population, i.e., $P(A_j)$. Post-stratification uses the distribution of sub-population proportions in each stratum, i.e., $P(A_j|B_j)$, and the size of each stratum, i.e., $P(B_j)$, to improve statistical estimates of $P(A_j)$ **.**

Simple Random Sampling

Assuming simple random sampling, Cochran (1977) gives the estimated proportion of sub-population A_j and its variance as:

$$
\begin{aligned}
\hat{P}_{SRS}(A_i) &= \frac{1}{n} \sum_{a=1}^{n} x_a \\
V_{SRS}\left[\hat{P}(A_i)\right] &\equiv \left\{ \begin{array}{l} \text{the variance of the estimate } \hat{P}(A_i) \\ \text{from the simple random sample (SRS)} \end{array} \right\} \\
&= \frac{1}{n} P(A_i)\left[1 - P(A_i)\right] \frac{N - n}{N - 1} \\
&\cong \frac{1}{n} P(A_i)\left[1 - P(A_i)\right]
\end{aligned}
\tag{1}
$$

where

$$
\begin{aligned}
x_a &= \begin{cases} 1 \text{ if sample unit } a \text{ is in sub-population } A_i \\ 0 \text{ otherwise} \end{cases} \\
N &\equiv \{\text{the total number of units in the population}\} \\
n &\equiv \{\text{the sample size, } n \text{ is assumed very small compared to } N\}
\end{aligned}
$$

Stratification

Consider a simple random sample of field plots, each of which is classified into one and only one sub-population. Remotely sensed classifications place each field plot into one and only one stratum, and remote sensing measures the area of each stratum. Cochran (1977) gives the estimated proportion of sub-population A_i in the total population, and its variance:

$$
\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{STR}}\left(A_i\right) &= \sum_{j=1}^{m} \mathrm{P}\left(B_j\right)\hat{\mathrm{P}}\left(A_i \mid B_j\right) \\
V_{\mathrm{STR}}\left[\hat{\mathrm{P}}\left(A_i\right)\right] &\equiv \left\{\begin{matrix}\text{the variance of the estimate } \hat{\mathrm{P}}\left(A_i\right) \\ \text{from the stratified random sample (STR)}\end{matrix}\right\} \\
&= \sum_{j=1}^{m} \frac{\mathrm{P}\left(B_j\right)^2}{n_j}\mathrm{P}\left(A_i \mid B_j\right)\left[1-\mathrm{P}\left(A_i \mid B_j\right)\right]\frac{N_j-n_j}{N_j-1} \\
&\cong \frac{1}{n}\sum_{j=1}^{m}\mathrm{P}\left(B_j\right)\mathrm{P}\left(A_i \mid B_j\right)\left[1-\mathrm{P}\left(A_i \mid B_j\right)\right]
\end{aligned}
\tag{2}
$$

where

$$
\begin{aligned}
\mathrm{P}\left(B_j\right) &= N_j/N \\
N_j &\equiv \left\{\text{the number of units in stratum } j\right\} \\
\mathrm{P}\left(A_i \mid B_j\right) &\equiv \left\{\begin{matrix}\text{the proportion of sub-population } A_i \\ \text{in stratum } B_j\end{matrix}\right\} \\
\hat{\mathrm{P}}\left(A_i \mid B_j\right) &= \frac{1}{n_j}\sum_{a\in n \cap a\in B_j} x_a \cong \frac{1}{\mathrm{P}\left(B_j\right)n}\sum_{a\in n \cap a\in B_j} x_a \\
n_j &\equiv \left\{\begin{matrix}\text{sample size in stratum } j, \\ n_j \text{ much smaller than } N_j\end{matrix}\right\} \cong \mathrm{P}\left(B_j\right)n
\end{aligned}
$$

Equation (2) is sufficient to compute approximations for the variance of a post-stratified estimate. However, the expected variance is often needed for survey planning when the stratum sizes $\mathrm{P}(B_j)$ and conditional probabilities $\mathrm{P}(A_i|B_j)$ are not yet observed. In the following sections, we make realistic assumptions and simplifications that make it easier to anticipate gains from stratification and specify accuracy standards for remote sensing.

Design Effect

The improvement in statistical efficiency with stratification of a simple random or systematic sample is quantified by the ratio of variances, which is designated the "design effect" by Särndal and others (1992). We denote the design effect as k, and it is approximated with equations (1) and (2) as:

$$
k = \frac{V_{\mathrm{STR}}\left[\mathrm{P}\left(A_i\right)\right]}{V_{\mathrm{SRS}}\left[\mathrm{P}\left(A_i\right)\right]} \cong \frac{\sum_{j=1}^{m}\mathrm{P}\left(B_j\right)\mathrm{P}\left(A_i \mid B_j\right)\left[1-\mathrm{P}\left(A_i \mid B_j\right)\right]}{\mathrm{P}\left(A_i\right)\left[1-\mathrm{P}\left(A_i\right)\right]}
\tag{3}
$$

If stratification improves the estimate, then k must be less than 1. Since all variances are positive, $k>0$. In the following sections, we use the design effect to simplify the mathematics and draw broad generalizations.

For two strata ($m = 2$), the design effect k in equation (3) simplifies to:

$$
\begin{aligned}
k &= \frac{V_{str}\left[\mathrm{P}\left(A_1\right)\right]}{V_{srs}\left[\mathrm{P}\left(A_1\right)\right]} \\
&= \frac{\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\left[1-\mathrm{P}\left(A_1 \mid B_1\right)\right]+\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\left[1-\mathrm{P}\left(A_1 \mid B_2\right)\right]}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)} \\
&= \frac{\mathrm{P}\left(B_1\right)\mathrm{P}\left(A_1 \mid B_1\right)\left[\mathrm{P}\left(A_2 \mid B_1\right)\right]+\mathrm{P}\left(B_2\right)\mathrm{P}\left(A_1 \mid B_2\right)\left[\mathrm{P}\left(A_2 \mid B_2\right)\right]}{\mathrm{P}\left(A_1\right)\mathrm{P}\left(A_2\right)}
\end{aligned}
\tag{4}
$$

Now, we express equation (4) in a form that is more useful in deriving the generalizations that follow. First, we use figure 1 to define two useful equalities:

$$\begin{aligned} \mathrm{P}(A_1) &= \begin{cases} \left[\mathrm{P}(A_1 \cap B_1)\right] + \left[\mathrm{P}(A_1 \cap B_2)\right] \\ \left[\mathrm{P}(B_1)\,\mathrm{P}(A_1 \mid B_1)\right] + \left[\mathrm{P}(B_2)\,\mathrm{P}(A_1 \mid B_2)\right] \end{cases} \\ \left[\mathrm{P}(B_2)\,\mathrm{P}(A_1 \mid B_2)\right] &= \mathrm{P}(A_1) - \left[\mathrm{P}(B_1)\,\mathrm{P}(A_1 \mid B_1)\right] \\ \left[\mathrm{P}(B_1)\,\mathrm{P}(A_2 \mid B_1)\right] &= \mathrm{P}(A_2) - \left[\mathrm{P}(B_2)\,\mathrm{P}(A_2 \mid B_2)\right] \end{aligned} \tag{5}$$

Equation (4) can be expressed using the equalities in equation (5) as:

$$\begin{aligned} k\left[\mathrm{P}(A_1)\,\mathrm{P}(A_2)\right] &= \left[\mathrm{P}(B_1)\,\mathrm{P}(A_2 \mid B_1)\right]\mathrm{P}(A_1 \mid B_1) + \left[\mathrm{P}(B_2)\,\mathrm{P}(A_1 \mid B_2)\right]\mathrm{P}(A_2 \mid B_2) \\ &= \begin{Bmatrix} \left[\mathrm{P}(A_2) - \mathrm{P}(B_2)\,\mathrm{P}(A_2 \mid B_2)\right]\mathrm{P}(A_1 \mid B_1) + \\ \left[\mathrm{P}(A_1) - \mathrm{P}(B_1)\,\mathrm{P}(A_1 \mid B_1)\right]\mathrm{P}(A_2 \mid B_2) \end{Bmatrix} \\ &= \begin{Bmatrix} \mathrm{P}(A_2)\,\mathrm{P}(A_1 \mid B_1) - \mathrm{P}(B_2)\,\mathrm{P}(A_2 \mid B_2)\,\mathrm{P}(A_1 \mid B_1) + \\ \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid B_2) - \mathrm{P}(B_1)\,\mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2) \end{Bmatrix} \\ &= \begin{Bmatrix} \mathrm{P}(A_2)\,\mathrm{P}(A_1 \mid B_1) + \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid B_2) \\ - \mathrm{P}(B_2)\,\mathrm{P}(A_2 \mid B_2)\,\mathrm{P}(A_1 \mid B_1) - \mathrm{P}(B_1)\,\mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2) \end{Bmatrix} \\ &= \begin{Bmatrix} \mathrm{P}(A_2)\,\mathrm{P}(A_1 \mid B_1) + \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid B_2) \\ -\left[\mathrm{P}(B_2) + \mathrm{P}(B_1)\right]\mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2) \end{Bmatrix} \end{aligned} \tag{6}$$

Since $\mathrm{P}(B_1) + \mathrm{P}(B_2) = 1$, equation (6) may be rewritten as:

$$\begin{aligned} k &= \frac{\mathrm{P}(A_2)\,\mathrm{P}(A_1 \mid B_1) + \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid B_2) - \mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} \\ &= \frac{\mathrm{P}(A_2)\,\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} + \frac{\mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} - \frac{\mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} \\ &= \frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)} + \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)} - \frac{\mathrm{P}(A_1 \mid B_1)\,\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} \end{aligned} \tag{7}$$

Subtract *k* in equation (7) from 1:

$$\begin{aligned} 1 - k &= 1 - \left\{\left[\frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] + \left[\frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right] - \left[\frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right]\right\} \\ &= 1 - \left[\frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] - \left\{\left[\frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right] - \left[\frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right]\right\} \\ &= \left[1 - \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] - \frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\left[1 - \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \\ &= \left[1 - \frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[1 - \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \end{aligned} \tag{8}$$

Rearranging equation (8):

$$
\begin{aligned}
1 - k &= \left[\frac{\mathrm{P}(A_1)}{\mathrm{P}(A_1)} - \frac{\mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2)}{\mathrm{P}(A_2)} - \frac{\mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \\
&= \left[\frac{\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_2)}\right] \\
&= \frac{\left[\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)\right]\left[\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)\right]}{\mathrm{P}(A_1)\,\mathrm{P}(A_2)} \\
&= \left[-\frac{\mathrm{P}(A_1) - \mathrm{P}(A_1 \mid B_1)}{\mathrm{P}(A_2)}\right]\left[-\frac{\mathrm{P}(A_2) - \mathrm{P}(A_2 \mid B_2)}{\mathrm{P}(A_1)}\right] \\
&= \left[\frac{\mathrm{P}(A_1 \mid B_1) - \mathrm{P}(A_1)}{1 - \mathrm{P}(A_1)}\right]\left[\frac{\mathrm{P}(A_2 \mid B_2) - \mathrm{P}(A_2)}{1 - \mathrm{P}(A_2)}\right]
\end{aligned}
\tag{9}
$$

Note that the final expression in equation (9) is the product of two terms, each of which is independent of each other and the sizes of the strata $\mathrm{P}(B_i)$. This feature greatly simplifies subsequent algebra.

SYMMETRICAL 2x2 ERROR MATRIX

We now derive expressions for the size of the two strata given the sizes for both sub-populations $\mathrm{P}(A_i)$. We assume the relative accuracy is identical in both strata (see below). Under this assumption, we show that both margins of the 2x2 error matrix in figure 1a are identical, i.e., $\mathrm{P}(A_i)= \mathrm{P}(B_i)$, and the off-diagonal joint probabilities are identical, i.e., $\mathrm{P}(A_1 \cap B_2)=\mathrm{P}(A_2 \cap B_1)$. The symmetry under these conditions facilitates derivations in other sections.

Relative Accuracy

Using equation (9), define "relative accuracy" as follows:

$$
\frac{\mathrm{P}(A_i \mid B_i) - \mathrm{P}(A_i)}{1 - \mathrm{P}(A_i)} \equiv \text{relative accuracy in stratum } i
\tag{10}
$$

If the relative accuracies are identical in both strata, then the following proceed from equation (9):

$$
\begin{aligned}
\frac{\mathrm{P}(A_i \mid B_i) - \mathrm{P}(A_i)}{1 - \mathrm{P}(A_i)} &= \sqrt{1 - k} \\
\mathrm{P}(A_i \mid B_i) &= \sqrt{1 - k}\left[1 - \mathrm{P}(A_i)\right] + \mathrm{P}(A_i)
\end{aligned}
\tag{11}
$$

Symmetric Margins

From figure 1b and equation (11), the off-diagonal conditional probabilities are:

$$
\begin{aligned}
\mathrm{P}(A_2 \mid B_1) &= 1 - \mathrm{P}(A_1 \mid B_1) \\
&= 1 - \left\{\sqrt{1 - k}\left[1 - \mathrm{P}(A_1)\right] + \mathrm{P}(A_1)\right\} \\
&= \left[1 - \mathrm{P}(A_1)\right] - \sqrt{1 - k}\left[1 - \mathrm{P}(A_1)\right] \\
&= \left[1 - \sqrt{1 - k}\right]\mathrm{P}(A_2) \\
\mathrm{P}(A_1 \mid B_2) &= 1 - \mathrm{P}(A_2 \mid B_2) \\
&= \left[1 - \sqrt{1 - k}\right]\mathrm{P}(A_1)
\end{aligned}
\tag{12}
$$

By definition, the size of sub-population A_1 equals the sum of sub-population A_1 in each of the two strata B_1 and B_2. Using figure 1a and equations (11) and (12):

$$
\begin{aligned}
\mathrm{P}(A_1) &= \mathrm{P}(A_1 \cap B_1) + \mathrm{P}(A_1 \cap B_2) \\
&= \mathrm{P}(B_1)\left\{\mathrm{P}(A_1 \mid B_1)\right\} + \mathrm{P}(B_2)\left\{\mathrm{P}(A_1 \mid B_2)\right\} \\
&= \mathrm{P}(B_1)\left\{\sqrt{1-k}\,\mathrm{P}(A_2) + \mathrm{P}(A_1)\right\} + \left[1 - \mathrm{P}(B_1)\right]\left\{\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\right\} \\
&= \left\{\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \\ \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} + \left\{\begin{matrix}\left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] - \\ \left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right]\mathrm{P}(B_1)\end{matrix}\right\} \\
&= \left\{\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \\ \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} + \left\{\begin{matrix}\left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] + \\ -\mathrm{P}(A_1)\,\mathrm{P}(B_1) + \sqrt{1-k}\,\mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right\} \\
&= \left[\begin{matrix}\sqrt{1-k}\,\mathrm{P}(A_2)\,\mathrm{P}(B_1) + \sqrt{1-k}\,\mathrm{P}(A_1)\,\mathrm{P}(B_1) \\ + \mathrm{P}(A_1)\,\mathrm{P}(B_1) - \mathrm{P}(A_1)\,\mathrm{P}(B_1)\end{matrix}\right] + \left[\mathrm{P}(A_1) - \sqrt{1-k}\,\mathrm{P}(A_1)\right] \\
&\left.\begin{aligned}
&= \sqrt{1-k}\left[\mathrm{P}(A_2) + \mathrm{P}(A_1)\right]\mathrm{P}(B_1) + \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \sqrt{1-k}\,\mathrm{P}(B_1) + \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)
\end{aligned}\right\} \text{ since } \left[\begin{matrix}\mathrm{P}(A_2) + \\ \mathrm{P}(A_1)\end{matrix}\right] = 1
\end{aligned}
\tag{13}
$$

Solving equation (13) for $\mathrm{P}(B_1)$, we show that both margins of the 2x2 error matrix in figure 1a are identical, i.e., $\mathrm{P}(A_i) = \mathrm{P}(B_i)$:

$$
\begin{aligned}
\sqrt{1-k}\,\mathrm{P}(B_1) &= \mathrm{P}(A_1) - \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \left[1 - 1 + \sqrt{1-k}\right]\mathrm{P}(A_1) \\
\mathrm{P}(B_1) &= \mathrm{P}(A_1) \\
1 - \mathrm{P}(B_1) &= 1 - \mathrm{P}(A_1) \\
\mathrm{P}(B_2) &= \mathrm{P}(A_2)
\end{aligned}
\tag{14}
$$

Symmetric Off-Diagonals

Finally, we show that both off-diagonal joint probabilities in the 2x2 error matrix (fig. 1a) are identical when the relative accuracies in both strata are identical:

$$
\begin{aligned}
&\left.\begin{aligned}
\mathrm{P}(A_1 \cap B_2) &= (A_2)\,\mathrm{P}(A_1 \mid B_2) \\
&= (A_2)\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1) \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\,\mathrm{P}(A_2)
\end{aligned}\right\} \text{for } \mathrm{P}(B_2) = \mathrm{P}(A_2) \\
&\left.\begin{aligned}
\mathrm{P}(A_2 \cap B_1) &= (A_1)\,\mathrm{P}(A_2 \mid B_1) \\
&= (A_1)\left[1 - \sqrt{1-k}\right]\mathrm{P}(A_2) \\
&= \left[1 - \sqrt{1-k}\right]\mathrm{P}(A_1)\,\mathrm{P}(A_2)
\end{aligned}\right\} \text{for } \mathrm{P}(B_1) = \mathrm{P}(A_1) \\
&\mathrm{P}(A_1 \cap B_2) = \mathrm{P}(A_2 \cap B_1)
\end{aligned}
\tag{15}
$$

Symmetric Matrix of Joint Probabilities

Assuming the relative accuracies are identical in both strata, as in equations (11) to (15), the matrix of joint probabilities from figure 1a is symmetric, as given in figure 2.

ACCURACY STANDARDS

Classification accuracy $\mathrm{P}(A_i|B_i)$ in stratum B_i must be greater than the proportion of sub-population $\mathrm{P}(A_i)$ in the population, i.e., $\mathrm{P}(A_i|B_i)>\mathrm{P}(A_i)$; otherwise, the design effect k will be greater than one in equation (9). For example, a 10,000-km^2 geographic area truly contains 7,000-km^2 of forest cover. Stratification will improve precision if, and only if, the remotely sensed forest category has at least 70 percent accuracy [*i.e.*, $0.7<\mathrm{P}(A_1|B_1)<1.0$], and the remotely sensed nonforest category has at least 30 percent accuracy [*i.e.*, $0.3<\mathrm{P}(A_2|B_2)<1.0$]. However, accuracy must be far greater before the gain in precision is substantial, as we now discuss.

Examples of desired accuracy of remotely sensed classifications are given for five different levels of design effect k in table 1. For example, a "substantial" gain in table 1 is defined as a design effect of k=0.5, meaning:

$P(A_1 \cap B_1) = P(A_1)\begin{bmatrix} P(A_1) + \\ P(A_2)\sqrt{1-k} \end{bmatrix}$	$P(A_1 \cap B_2) = \left[1 - \sqrt{1-k}\right] P(A_1) P(A_2) = P(A_2 \cap B_1)$	$P(A_1)$
$P(A_2 \cap B_1) = \left[1 - \sqrt{1-k}\right] P(A_1) P(A_2) = P(A_1 \cap B_2)$	$P(A_2 \cap B_2) = P(A_2)\begin{bmatrix} P(A_2) + \\ P(A_1)\sqrt{1-k} \end{bmatrix}$	$P(A_2)$
$P(B_1) = P(A_1)$	$P(B_2) = P(A_2)$	1

Figure 2—Error matrix for two sub-populations and corresponding strata when the relative accuracy in each stratum is identical (equations (11) to (15)). These assumptions permit generalizations, while maintaining realistic scenarios.

- The estimate with stratification has half the variance of the estimate with simple random sampling;
- Estimates from simple random sampling would require a two-fold increase in the number of field plots to achieve the same variance with stratification; and
- The confidence interval with stratification is approximately 71 percent ($\sqrt{0.5} \times 100\%$) smaller than that with simple random sampling.

Figure 3 illustrates the relative precision of stratified estimates for each level of design effect in table 1. The discrete levels in table 1 simplify mathematical generalizations that follow.

Assume that the relative accuracies are identical in both strata, as in equation (11). Figure 4 displays the classification accuracy $P(A_i|B_i)$ in stratum B_i required to meet various levels of gain in statistical efficiency (k in table 1). From figure 4, classification accuracy in a stratum must be nearly perfect if its corresponding sub-population is very prevalent, *i.e.*, $P(A_i)$»1.0, while the accuracy need not be nearly as great for a rare sub-population, *i.e.*, $P(A_i)\approx 0$. Czaplewski and Patterson (in preparation) show that figure 4 is applicable classification systems having three or more sub-populations under certain assumptions.

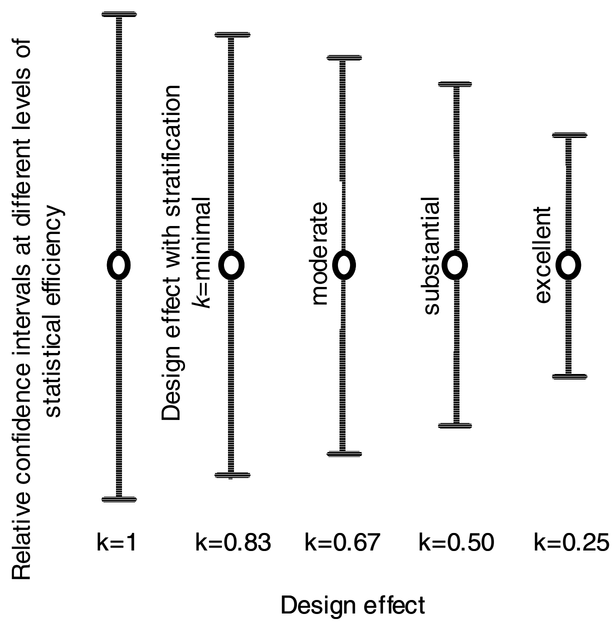

Figure 3—Relative confidence intervals for different levels of design effect *k* in table 1.

Table 1—Levels of gain from stratification used in comparisons

Design effect $k=V_{STR}/V_{SRS}$	Gain in efficiency through post-stratification	Increase in effective number of plots[a] gained through stratification	Relative variance of stratified sampling compared to simple random sampling $100\times V_{STR}/V_{SRS}$	Relative standard error[b] of stratified sampling compared to simple random sampling $100\times \sqrt{V_{STR}/V_{SRS}}$
			- - - - - - - - - - - -*Percent*-	- - - - - - - - - - - -
k=(1/1.0)=1.00	"None"	None	100	100
k=(1/1.2)=0.83	"Minimal"	1.2-fold	83	91
k=(1/1.5)=0.67	"Moderate"	1.5-fold	67	82
k=(1/2.0)=0.50	"Substantial"	2-fold	50	71
k=(1/4.0)=0.25	"Excellent"	4-fold	25	50

[a] The increase in sample size n that would be required to achieve the same variance without stratification.
[b] Approximately proportional to the confidence interval.

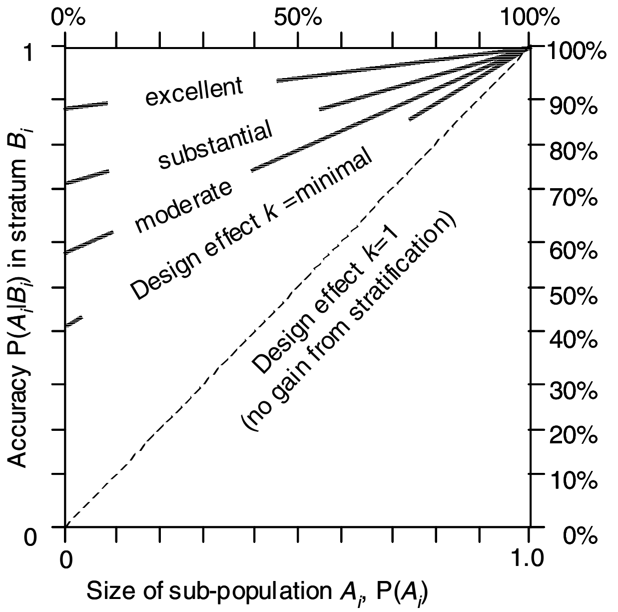

Figure 4—Classification accuracy P($A_i|B_i$) in stratum B_i required for different levels of statistical gain (k in table 1) as a function of sub-population size P(A_i). These results assume that the relative accuracy is identical for all strata as in equation (11).

LOSS OF EFFICIENCY OVER TIME

Landscapes change over time through land management, succession, disturbance and shifting land use. Current field plots may be stratified using remotely sensed data that were acquired many years ago. Some portion of the total "classification error" is caused by changes in the landscape, not by the original accuracy of the remotely sensed classifications. How old can the remotely sensed data become before its value for stratification becomes seriously degraded?

Assume the size of sub-population A_1 at time t decreases by some fraction D of its original size at time 0, where 0<D<1. Since the remotely sensed data were acquired at time t=0, the stratum sizes P(B_i) are the same at times 0 and t. Assume changes in the landscape between times 0 and t are independent of the remotely sensed classification at time 0. The decrease in the size of sub-population A_1 causes a corresponding increase in sub-population A_2. Finally, assume both strata at time 0 have the same relative accuracy. The error matrix in figure 5, which corresponds to figure 2, captures these assumptions. Under these conditions, the design effect k_t at time t equals:

$$k_t = 1 - \left\{ \frac{(1-\Delta)(1-k_0)\left[1-P(A_1)_0\right]}{1-P(A_1)_0 + \Delta P(A_1)_0} \right\} \qquad (16)$$

Equation (16) can be transformed into a more general expression that simultaneously covers all levels of the design effect k_0:

$$\frac{k_t - k_0}{1-k_0} = \frac{\Delta}{1-(1-\Delta)\Delta P(A_1)_0} \qquad (17)$$

Figure 6 is a graphical display of equation (17). When there has been little change in the landscape between time 0 and t ($\Delta \approx 0$), there is little change in design effect ($k_t \approx k_0$) using remotely sensed data acquired at time 0, and there is little loss in statistical efficiency. However, as the net decrease in sub-population size (Δ) becomes larger, the design effect approaches one. This means that the variance with stratification is nearly equal to that under simple random sampling, i.e., the gain in efficiency through stratification is almost entirely lost.

When a sub-population is very common, i.e., P(A_i)≈1, even a small decrease in sub-population size between time 0 and t causes major losses in efficiency. However, if the

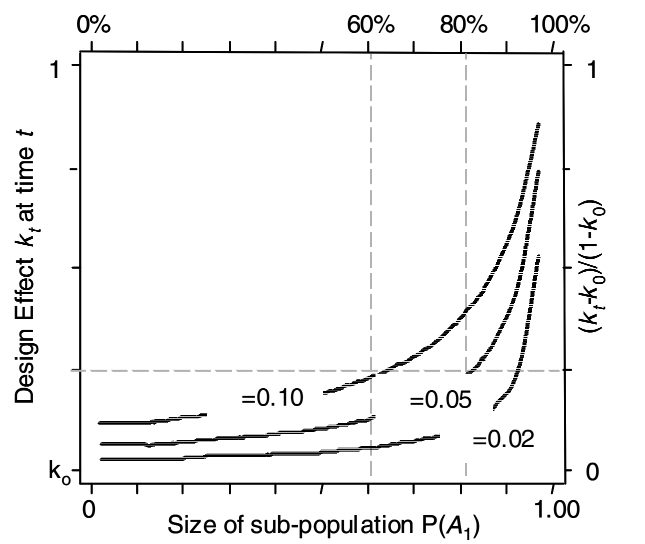

Figure 6—Design effect as function of loss rate for one of the sub-populations. As the rate becomes faster, the design effect approaches 1, meaning that the variance with stratification is no better than the variance with simple random sampling.

$P(A_1 \cap B_1)_t = (1-\Delta)P(A_1 \cap B_1)_0$	$P(A_1 \cap B_2)_t = (1-\Delta)P(A_1 \cap B_2)_0$	$P(A_1)_t = (1-\Delta)P(A_1)_0$
$P(A_2 \cap B_1)_t = \begin{bmatrix} P(A_2 \cap B_1)_0 + \\ \Delta P(A_1 \cap B_1)_0 \end{bmatrix}$	$P(A_2 \cap B_2)_t = \begin{bmatrix} P(A_2 \cap B_2)_0 + \\ \Delta P(A_1 \cap B_2)_0 \end{bmatrix}$	$P(A_2)_t = \begin{bmatrix} P(A_2)_0 + \\ \Delta P(A_1)_0 \end{bmatrix}$
$P(B_1)$	$P(B_2)$	1

Figure 5—Error matrix that includes change (D at time t) in sub-population A_1 after acquisition of the remotely sensed data (time t=0) that are used to specify strata B_1 and B_2.

sub-population represents less than 80 percent of the total population at time 0, and the net decrease at time *t* is less than 5 percent (Δ=0.05), then the decrease in efficiency is less than 25 percent (fig. 6). Most extensive landscapes have less than 80 percent forest.

As example, let sub-population A_1 represent forest land-use, and the stratification is based on ten-year old remotely sensed data. Assume 1 percent per year of the original forest land-use is converted to another land use, such as agriculture or urban. The rate of change over 10 years is $\Delta \approx 0.10$. Assume there are no conversions back to forest land-use. If forest land-uses occupy less than 60 percent of the landscape, then the stratification based on ten-year old remotely sensed data retains 75 percent of its efficiency (fig. 6). Loss of statistical efficiency is most rapid in those landscapes dominated by forest land-uses.

Czaplewski and Patterson (in preparation) extend this model to cases in which changes occur in both sub-populations. They analyze steady state conditions, in which the changes in sub-population A_1 exactly equal the changes in sub-population A_2. They find that statistical efficiency also decreases over time, and the rate of loss in efficiency can be higher in a dynamic stead-state landscape than a landscape that is not at equilibrium.

INDETERMINATE STRATUM

Some population elements (e.g., pixels) are classified with less confidence than other elements with remotely sensed data. For example, the maximum likelihood classifier, which is widely used for image processing, computes the probabilities of a pixel being a member of each remotely sensed category. The pixel is assigned to the category with the highest probability, even if the largest probability is relatively low for some pixels. This often occurs with mixed pixels, or pixels near the boundary of a multivariate cluster. As another example, a binary-tree classifier assigns each element into a single category, but the algorithm estimates the probability of correct classification using its training data. Even with unsupervised classifiers, all multivariate clusters do not have the same proportion of predominate labeling sites. We investigate the opportunity to increase statistical efficiency by creating a new stratum that contains pixels which are classified with less confidence than other pixels. We label this stratum as the “indeterminate stratum.”

The matrix in figure 7 gives one example that is numerically tractable. Let d_{ij} represent the quantity of elements that are removed from sub-population *i* in stratum *j*. We start by moving a small quantity of elements into the indeterminate stratum and increase the quantity until the gain in statistical efficiency is maximized. The first elements removed are those that are most difficult to successfully classify.

The size of each d_{ij} is modeled by functions $f_{i,j}(c_i)$, where c_i starts at 0 and incrementally increases towards 1 until the optimum is realized (fig. 8). These functions have the following conditions:

1. For the matrix in figure 7, $d_{ij}=f_{i,j}(c_i)$ for $0<c_i<1$ in stratum B_i. If the sub-population is correctly classified in stratum B_i, $f_{i,i}(c_i)$ is a linear function of (c_i). If the sub-population is not correctly classified in stratum B_i, $f_{j,i}(c_i)$ is a non-linear function of (c_i) so that we can impose Conditions 3 and 4 that follow. Figure 8 illustrates these two functions.

$$\begin{aligned} d_{i,i} &= f_{i,i}(c_i) \\ &= c_i \, \mathrm{P}(A_i \cap B_i) \\ d_{j,i} &= f_{j,i}(c_i) \\ &= \frac{\mathrm{P}(A_i \cap B_i)}{\alpha}\left(1 - e^{-\alpha c_i}\right) \end{aligned} \tag{18}$$

2. The number of elements correctly classified in a stratum must always be larger than the number of elements that are incorrectly classified, i.e., $\mathrm{P}(A_i|B_i)>0.5$. For large values of the design effect *k* (meaning that the original stratification yields little gain in efficiency over simple random sampling), and a when stratum is very rare, this condition is not always met. Therefore, the following constraint is placed on $\mathrm{P}(A_i)$ in equation (18).

$$\mathrm{P}(A_1) > \frac{1 - 2\sqrt{1-k}}{2\left(1 - \sqrt{1-k}\right)}$$

3. Some classification errors are not removed until virtually all elements in the stratum are shifted into the indeterminate stratum (fig. 8). The last elements to be removed from the stratum ($c \approx 1$) have almost no classification error. The value of a in equation (18) is numerically determined so that the following condition is true for $f_{j,i}(c_i)$, i.e., when the sub-population is not correctly classified in stratum B_i.

$$f_{j,i}(c_i = 1) = \mathrm{P}(A_j \cap B_i)$$

4. The highest proportion of classification errors are removed from each stratum during the first incremental

$\mathrm{P}(A_1 \cap B_1) - d_{1,1}$	$d_{1,1} + d_{1,2}$	$\mathrm{P}(A_1 \cap B_2) - d_{1,2}$	$\mathrm{P}(A_1)$
$\mathrm{P}(A_2 \cap B_1) - d_{2,1}$	$d_{2,1} + d_{2,2}$	$\mathrm{P}(A_1 \cap B_2) - d_{2,2}$	$\mathrm{P}(A_2)$
$\mathrm{P}(A_1) - (d_{1,1} + d_{2,1})$	$d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}$	$\mathrm{P}(A_2) - (d_{1,2} + d_{2,2})$	1

Figure 7—Error matrix in which an “indeterminate” stratum, which includes sites that are most likely to be misclassified. This stratum can increase statistical efficiency by increasing the classification accuracy in the original two strata.

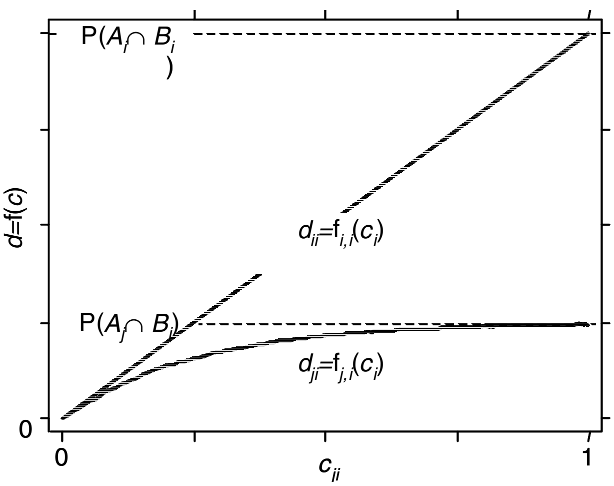

Figure 8—Functions $f_{i,j}(c_i)$ for the size of each d_{ij}, which is used to shift likely classification errors into the "indeterminate" stratum. See figure 7 and equation (18).

shifts of elements into the indeterminate stratum, where $c_i \approx 0$. We assume a 50:50 mixture of the two subpopulations among these first elements. This represents those elements that are most difficult to correctly classify. This constraint is imposed by making the first derivatives of $f_{i,i}(c_i)$ and $f_{j,i}(c_i)$ identical when $c_i=0$ (fig. 8).

$$\left[\frac{\partial}{\partial c_i} f_{i,i}\left(c_i = 0\right)\right] = \left[\frac{\partial}{\partial c_i} f_{j,i}\left(c_i = 0\right)\right]$$

5. As elements are moved into the indeterminate stratum, accuracies increase in both of the original strata. We force their relative accuracies to remain equal so that c_2 can be expressed as a function of c_1. This reduces the number of variables in our evaluation. However, we make an exception to this constraint when $c_1=1$, meaning all of stratum B_1 is moved into the indeterminate stratum.

$$\left[\frac{\dfrac{\mathrm{P}\left(A_1 \cap B_1\right) - d_{1,1}}{\mathrm{P}\left(A_1\right) - d_{1,1} - d_{2,1}} - \mathrm{P}\left(A_1\right)}{1 - \mathrm{P}\left(A_1\right)}\right] = \left[\frac{\dfrac{\mathrm{P}\left(A_2 \cap B_2\right) - d_{2,2}}{\mathrm{P}\left(A_2\right) - d_{1,2} - d_{2,2}} - \mathrm{P}\left(A_2\right)}{1 - \mathrm{P}\left(A_2\right)}\right]$$

$$c_2 = f\left(c_1\right) \quad \text{for } c_1 < 1$$

6. The shift of elements into the indeterminate stratum stops when the design effect in equation (4) reaches its minimum within the interval $0<c_i<1$, i.e., the optimum improvement in statistical efficiency.

$$\left[\frac{\partial}{\partial c_1} k\right] = 0$$

7. We relax Condition 5 (above) when $c_1=1$, meaning the optimum in Condition 6 is not realized as c_1 reaches 1. This situation approximately occurs whenever $\mathrm{P}(A_i)<1.15k-0.67$ in our model from figure 7 and equation (18). In this situation, we merge stratum B_1 with the indeterminate stratum. This returns us to two strata, where the merged stratum contains stratum B_1 plus elements removed from stratum B_2. We use c_2 to increase the quantity of elements shifted from stratum B_2 into this new stratum until we achieve the optimum in Condition 6.

We were unable to find an algebraic solution to this formulation; therefore, we developed a numerical solution. The following describe our results.

We found that statistical efficiency does increase with addition of an indeterminate stratum, at least using the model in figure 7 and equation (18). Let k_{opt} represent the design optimal effect with addition of the indeterminate stratum. Figure 9 shows the proportional improvement in the design effect (k_{opt}/k) relative to the initial prevalence of stratum B_1 for different initial design effects (table 1). The optimal gain in efficiency exceeds 15 percent ($k_{opt}/k<1-0.15$) when classification accuracy is high, *i.e.*, the initial design effect is excellent; however, the gain is less than 5 percent when the initial design effect is marginal.

Given the model in figure 7 and equation (18), optimal size of the indeterminate stratum varies with prevalence of the two strata, as shown in figure 10. The optimal size is under 10 percent of the population when the design effect is excellent (classification accuracy is high), but it can approach 30 percent when the design effect is marginal (fig. 10). The optimal proportion of the indeterminate stratum that originates from each of the original strata in figure 7 is given in equation (19) and illustrated in figure 11:

$$\begin{aligned} \frac{d_{1,1} + d_{2,1}}{d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}} &\approx 0.5 - \left(\frac{e^{4.3k}}{25.6}\right)\left[\mathrm{P}\left(A_1\right) - 0.5\right] \\ \frac{d_{1,2} + d_{2,2}}{d_{1,1} + d_{1,2} + d_{2,1} + d_{2,2}} &\approx 0.5 + \left(\frac{e^{4.3k}}{25.6}\right)\left[\mathrm{P}\left(A_1\right) - 0.5\right] \end{aligned} \qquad (19)$$

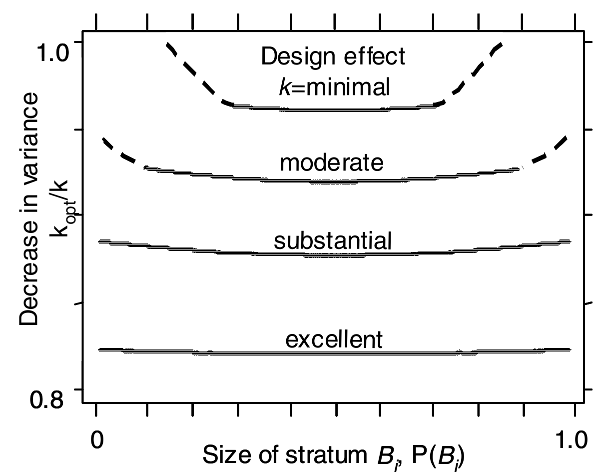

Figure 9—Proportional improvement in the design effect (k_{opt}/k) with addition of an indeterminate stratum for different initial design effects (table 1). The dashed lines indicate when stratum B_1 is merged with the indeterminate stratum to optimize efficiency (Condition 7).

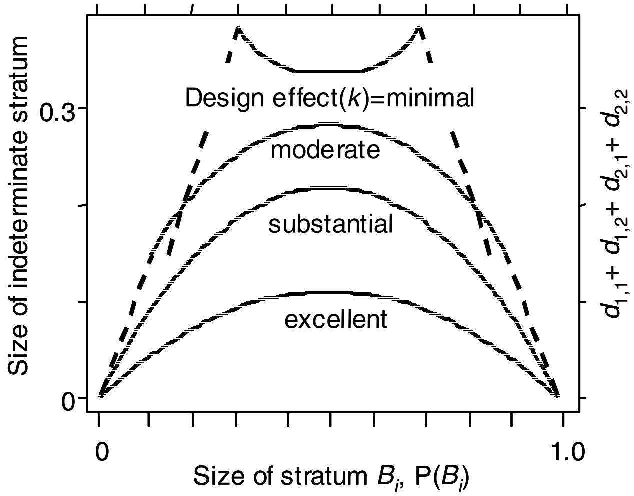

Figure 10—The optimal size of the indeterminate stratum, expressed as a proportion of the total population. Dashed lines indicate when stratum B_1 is merged with the indeterminate stratum to optimize efficiency (Condition 7).

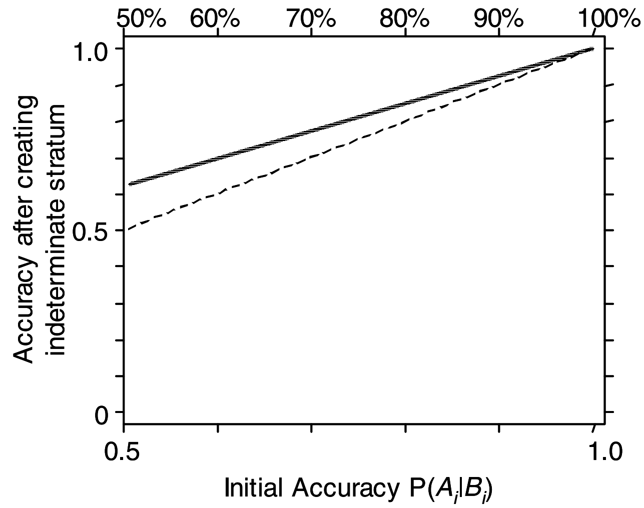

Figure 12—Optimal classification accuracy in stratum B_1 after formation of the indeterminate stratum, which contains elements that are most likely to have classification errors.

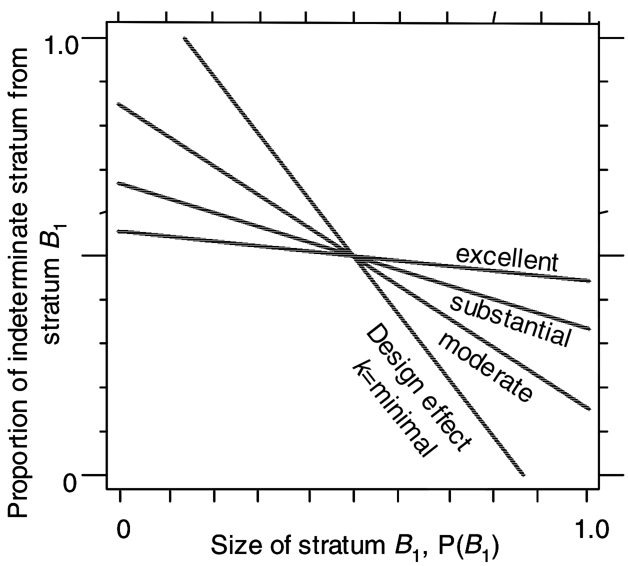

Figure 11—Proportion of indeterminate stratum originally part of stratum B_1, which corresponds to equation (19).

Let $P(A_1|B_1)_{OPT}$ represent the optimal classification accuracy in stratum B_1 after formation of the indeterminate stratum. Figure 12 shows $P(A_1|B_1)_{OPT}$ as a function of the original accuracy in stratum B_1, i.e., $P(A_1|B_1)$. We found that this relationship is approximately $P(A_1|B_1)_{OPT} \approx [0.75\ P(A_1|B_1)+0.25]$ in our model (figure 7 and equation 18), regardless of the initial design effect.

If accuracy is marginal and the stratum size is small (i.e., $P(B_j)<1.15k-0.67$), then the indeterminate stratum should be merged with the rare stratum to increase efficiency. This situation corresponds to the dashed lines in figure 9. However, gains do not exceed 5 percent with the model in figure 7 and equation 18. When the gain in efficiency is optimal, the indeterminate stratum contains 100 percent (c_i=1) of the rare stratum, plus approximately 25 percent ($c_j \approx$ 0.25) of the common stratum. The latter portion contains the most likely classification errors in stratum B_j.

We recommend that the size of the indeterminate stratum be specified before exploring the sample data after they are collected; this avoids "over-fitting" to a given sample. Over-fitting can bias the estimated sampling error, thus producing a variance estimate that is smaller than its true value. In other words, our estimate would not be as precise as we assume, and analyses of these estimates can produce false conclusions. Figures 9 and 10, and equation 19, provide a priori specifications that can help practitioners follow our recommendation.

REFERENCES

Cochran, W.G. 1977. Sampling techniques. New York. John Wiley. 428 p.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment. 37: 35–46.

Czaplewski, R.L.; Patterson, P.L. [In preparation]. Classification accuracy for stratification with remotely sensed data. Submitted to Forest Science.

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