# Estimating accuracy of land-cover composition from two-stage cluster sampling

Stephen V. Stehman [a,*], James D. Wickham [b], Lorenzo Fattorini [c], Timothy D. Wade [b], Federica Baffetta [d], Jonathan H. Smith [e]

[a] SUNY College of Environmental Science and Forestry, 320 Bray Hall, 1 Forestry Drive, Syracuse, New York, 13210, USA
[b] United States Environmental Protection Agency, E243-05, Research Triangle Park, North Carolina, USA
[c] Dipartimento di Metodi Quantitativi, Università di Siena, Siena, Italy
[d] Dipartimento di Statistica "G. Parenti", Università di Firenze, Firenze, Italy
[e] United States Geological Survey, Geographic Analysis and Monitoring Program, 12201 Sunrise Valley Dr., 519 USGS National Center, Reston, Virginia, USA

ARTICLE INFO

*Article history:*
Received 14 March 2008
Received in revised form 10 February 2009
Accepted 14 February 2009

*Keywords:*
Sampling design
Horvitz–Thompson estimation
Multiple objectives

ABSTRACT

Land-cover maps are often used to compute land-cover composition (i.e., the proportion or percent of area covered by each class), for each unit in a spatial partition of the region mapped. We derive design-based estimators of mean deviation (*MD*), mean absolute deviation (*MAD*), root mean square error (*RMSE*), and correlation (*CORR*) to quantify accuracy of land-cover composition for a general two-stage cluster sampling design, and for the special case of simple random sampling without replacement (*SRSWOR*) at each stage. The bias of the estimators for the two-stage *SRSWOR* design is evaluated via a simulation study. The estimators of *RMSE* and *CORR* have small bias except when sample size is small and the land-cover class is rare. The estimator of *MAD* is biased for both rare and common land-cover classes except when sample size is large. A general recommendation is that rare land-cover classes require large sample sizes to ensure that the accuracy estimators have small bias.



## 1. Introduction

A common use of land-cover maps is to calculate the area or proportion of area of each land-cover class within each unit of a spatial partition of the region mapped. The area classified as forest, agriculture, and developed land within each 10 km by 10 km block forming a partition of the region mapped is an example of land-cover composition data that may be obtained from a land-cover map. Applications of land-cover composition data span a variety of spatial units, including watersheds (e.g., Jones et al., 2001a; Stanfield et al., 2002), subbasins or subwatersheds (Wimberly & Ohmann 2004; Jennings et al., 2004), buffer areas around a sampling station (Comeleo et al., 1996; Driscoll & Donovan 2004), townships (Glennon & Porter 1999), 7.5 km by 7.5 km blocks (Riitters et al., 2004), 400, 800, and 1600 m radii circular plots (Bakker et al., 2002), or 5 km by 5 km blocks (Jones et al., 2001b; Wickham et al., 2002).

Quantifying the accuracy of land-cover composition is an important component of validation of a land-cover map. A direct assessment of land-cover composition accuracy (see also Pontius's (2000) "quantification error") requires comparing the map-derived area or percent cover for each land-cover class to the true area or percent cover (i.e. reference data) on a per-spatial unit basis. The error matrix commonly used to summarize the results of an accuracy assessment (cf. Foody 2002) allows for computing non-site-specific accuracy for each class, i.e. the difference between the proportion of area mapped as that class and the proportion of area that is truly that class. The non-site-specific accuracy obtained from an error matrix can be regarded as an assessment of land-cover composition accuracy of a single spatial unit, where that one unit is the entire region mapped.

In many applications land-cover composition is derived for each of the spatial units making up a partition of the region, so our objective extends to assessing land-cover composition accuracy for such a population of spatial units or blocks. This assessment would require data such as shown in Table 1. The reference land-cover composition is the area or percent of area of each land-cover class as determined from the true ground condition (or the best determination of that ground condition) and the map land-cover composition is the area or percent of area of each land-cover class as determined from the map. Hollister et al. (2004) is one of the few examples in which this type of accuracy assessment has been implemented. They took advantage of complete coverage reference data available for Massachusetts and Rhode Island to conduct an accuracy assessment of land-cover composition provided by the 1992 National Land Cover Data, NLCD 1992 (Vogelmann et al., 2001). For each land-cover class, Hollister et al. (2004) compared the NLCD 1992 area to their reference data area and estimated agreement from a sample of circular units of different sizes ranging from 0.1 $km^2$ to 200 $km^2$.

Assessing the accuracy of land-cover composition should be recognized as distinct from reporting non-site-specific accuracy for spatial domains or subregions of the region mapped. For example,

* Corresponding author.
*E-mail address:* svstehma@syr.edu (S.V. Stehman).

**Table 1**
Example data for an assessment of land-cover composition accuracy.

| Composition by map class (ha) | | | | Composition by reference class (ha) | | | | Difference, map − reference (ha) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Urban | Forest | Ag | Wetl | Urban | Forest | Ag | Wetl | Urban | Forest | Ag | Wetl |
| 1.26 | 181.26 | 699.30 | 16.38 | 65.25 | 197.28 | 624.69 | 6.93 | −63.99 | −16.02 | 74.61 | 9.45 |
| 0.63 | 608.22 | 266.31 | 3.15 | 80.55 | 609.30 | 191.52 | 0.81 | −79.92 | −1.08 | 74.79 | 2.34 |
| 4.14 | 427.23 | 416.16 | 19.62 | 42.57 | 296.10 | 359.37 | 160.20 | −38.43 | 131.13 | 56.79 | −140.58 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.81 | 436.41 | 367.83 | 89.28 | 73.80 | 472.23 | 304.38 | 11.61 | −72.99 | −35.82 | 63.45 | 77.67 |
| 93.78 | 412.83 | 390.15 | 2.16 | 290.79 | 432.00 | 168.30 | 0.00 | −197.01 | −19.17 | 221.85 | 2.16 |
| 8.91 | 685.17 | 175.59 | 4.14 | 58.05 | 710.46 | 121.14 | 0.99 | −49.14 | −25.29 | 54.45 | 3.15 |
| 1.98 | 247.23 | 624.15 | 0.00 | 48.78 | 278.90 | 514.17 | 0.00 | −46.80 | −31.68 | 109.98 | 0.00 |
| 3.60 | 529.02 | 363.87 | 2.25 | 28.35 | 590.04 | 270.54 | 0.00 | −24.75 | −61.02 | 93.33 | 2.25 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 292.68 | 601.92 | 2.79 | 67.86 | 272.97 | 515.34 | 15.93 | −67.86 | 19.71 | 86.58 | −13.14 |
| 2.16 | 615.06 | 267.93 | 2.25 | 31.05 | 691.20 | 161.73 | 1.35 | −28.89 | −76.14 | 106.20 | 0.90 |

The area (ha) covered by each of four land-cover classes (Urban, Forest, Ag = Agriculture, and Wetl = Wetland) is shown for $K = 12$ support units. Each spatial unit (or block) is 3 km by 3 km (900 ha) and contains 10,000 30 m by 30 m pixels. Because not all land-cover classes are included in the table, the total area of the four classes shown for a spatial unit may be less than 900 ha.

Blackard et al. (2008) reported the accuracy of mapped area of forest on a per-state basis for their forest cover map of the United States. In this application each state in the U.S. is treated as a spatial domain and accuracy reported by domain. This is a different feature from accuracy of land-cover composition. If the composition objective had been targeted, the states would be viewed as forming a spatial partition of the U.S. and the evaluation would assess how accurately forest area was quantified by the map for this collection or "population" of states (e.g. the data structure of Table 1 where each row of the table would be a state and the column representing forest area would be of interest). Wu et al. (2008) is another example in which non-site-specific accuracy is reported for various spatial domains. In this study, accuracy of the area of cropland is reported on a per-region and per-province basis, but this is different from evaluating land-cover composition accuracy for a population of regions or a population of provinces.

The data structure shown in Table 1 helps to illustrate the nature of the basic sampling design challenge confronting the assessment of composition accuracy. The sampling protocol must be able to provide enough sample pixels within a spatial unit to adequately characterize the reference land-cover composition of that spatial unit. For example, if only 2 or 3 sample pixels fall within a spatial unit, the estimate of the reference land-cover composition of that unit will be poor unless the unit is dominated by a single class. It is not necessary that every single spatial unit be characterized well (i.e. a sample of these units should suffice), but the sampling protocol must to some degree concentrate or group sample pixels within several of the spatial units of the size targeted for assessment. The basic sampling designs often considered for accuracy assessment, simple random, stratified random (stratifying by the map land-cover class), and systematic, will not likely achieve the desired concentration of sample pixels unless the sample size for the accuracy assessment is large. By construction, cluster sampling has the desired feature that sample pixels are grouped within larger spatial units. Cluster sampling generates a sample of units of different sizes, for example a 10 km by 10 km block applicable to the assessment of composition accuracy, and a pixel to assess per-pixel accuracy.

Although the importance of evaluating accuracy of spatially aggregated land-cover data has been recognized (Strahler et al., 2006, p. 37), sampling design and analysis protocols specifically targeting the assessment of land-cover composition accuracy have not been prescribed. Most accuracy assessments of large-area land-cover maps (cf. Bossard et al., 2000, Couturier et al., 2007, Edwards et al., 1998; Fuller et al., 1994; Han et al., 2004; Heiskanen 2008; Kuemmerle et al., 2006; Latifovic & Olthof 2004; Mayaux et al., 2006; Miettinen et al., 2008; Mücher et al., 2000; Reese et al., 2002; Scepan 1999; Stehman et al., 2003; Wickham et al., 2004a; Wulder et al., 2006) have not addressed the objective of quantifying land-cover composition accuracy, but instead have focused on a per-pixel, error matrix based description of accuracy. Because the accuracy assessment objectives (Stehman et al., 2008) specified for the NLCD 2001 (Homer et al., 2007) include evaluating accuracy of land-cover composition for the Level I classification, it is necessary to develop a sampling design and estimation strategy for conducting an assessment of land-cover composition accuracy.

In this article, we address the problem of assessing the accuracy of land-cover composition depicted by an end product land-cover map. The methodology we present is applicable regardless of the procedure used to create the map. The objective is to quantify how well the map represents the area or proportion of area each land-cover class covers within spatial units forming a partition of the region mapped. As is typical of most accuracy assessment problems, a sampling approach is necessary because it is not practical to collect reference land-cover data for the entire region mapped. Cluster sampling is recommended to satisfy simultaneously the reference data requirements for the composition accuracy objective and to accommodate the traditional per-pixel accuracy objectives (e.g. estimation of the error matrix and associated accuracy parameters). The parameters we use to quantify land-cover composition accuracy are mean deviation, mean absolute deviation, root mean square error, and correlation. We derive estimators of these accuracy parameters for a general two-stage cluster sampling design, and then present the estimators for a simple special case design in which simple random sampling without replacement (SRSWOR) is implemented at each stage. We conduct a simulation study based on hypothetical populations of land-cover classification error to investigate the bias and precision of the composition accuracy estimators for the two-stage SRSWOR design.

## 2. Components of the accuracy assessment protocol

### 2.1. Reference land cover

Reference land cover is the area or the percent of area each land-cover class covers within a spatial unit of the chosen size and shape (Table 1). If a per-pixel accuracy assessment is also desired, the reference land-cover class of each pixel must also be determined. The specific details of the "response design" (Stehman & Czaplewski 1998) used to collect the reference data and to determine the reference land-cover class will be dependent on the features of the particular map being evaluated (class legend, pixel size, source of information for reference data such as ground visit or aerial photography, and practical considerations). The reference land-cover labels determined at the pixel level could then be used to obtain the area or percent area for each land-cover class for the larger spatial units targeted by the composition accuracy objective. If the spatial unit for the composition accuracy objective is large, it may be prohibitively expensive to obtain the

reference land-cover data for the entire unit, and it may be necessary to estimate the reference area or percent of area of each class from a sample of the spatial unit.

### 2.2. Sampling design

To focus on the fundamental sampling design and estimation issues, we will limit attention to the case in which accuracy of land-cover composition is assessed for only one size spatial assessment unit (i.e. cluster) larger than a pixel. As stated previously, cluster sampling is well-suited for this purpose of collecting per-pixel and per-cluster reference data within a unified design. In the terminology of cluster sampling, the spatial unit selected for the composition accuracy objective is the primary sampling unit (PSU), and the pixel is the secondary sampling unit (SSU). In one-stage cluster sampling, all SSUs within each sampled PSU are observed. Hollister et al.'s (2004) sampling design is an example of one-stage cluster sampling. In two-stage cluster sampling, each PSU selected in the first-stage of the sample is subsampled (i.e., a subsample of pixels is selected from within each sampled spatial unit).

One-stage cluster sampling is perhaps the simplest design adequate to the task of providing the reference data needed for assessing composition accuracy. However, practical considerations limit the utility of this design. In particular, the cost of obtaining the reference data for one-stage cluster sampling may constrain the size of the spatial unit for which composition accuracy is assessed. For example, suppose that the land-cover map consists of 30 m pixels and the accuracy assessment is designed to target a 3 km by 3 km unit (PSU). Each PSU would then contain 10,000 pixels, and a one-stage cluster sample of just 20 PSUs would require interpreting the reference land cover for a sample of 200,000 pixels. Two-stage cluster sampling becomes a practical necessity if the spatial unit desired for composition accuracy is too large to obtain reference data for all pixels in the PSU. Different combinations of sampling designs can be implemented at the two stages, with simple random sampling without replacement (SRSWOR) implemented at each stage being one of the simplest options. In the first stage, a simple random sample of PSUs is selected without replacement from all PSUs available. Then within each sampled PSU, a simple random sample of pixels is selected without replacement from all pixels present in the PSU. This constitutes the second stage of the sampling design.

Stratifying by map land-cover class is often implemented to gain control over the sample size allocated to rare land-cover classes, so a design combining stratification with cluster sampling has great appeal. For a per-pixel accuracy assessment, it is straightforward to assign each pixel to a stratum based on the map land-cover class of the pixel. Assigning clusters to strata for the purpose of increasing the sample size allocated to rare land-cover classes is complicated because a cluster may be comprised of pixels from several different land-cover classes. Mayaux et al. (2006) implemented a stratification of Landsat-scene clusters based on diversity of the land cover within the scene and whether the scene contained a specified percent of three designated critical land-cover classes. This combination of stratification with cluster sampling does not directly control the sample size allocated to each individual land-cover class. Stehman et al. (2003) and Wickham et al. (2004a,b) approached the problem differently by first selecting a sample of clusters (e.g. 6 km by 6 km blocks), and then assigning the pixels within the sampled clusters to strata determined by the map land-cover class of each pixel. This design, based on stratification of the pixels rather than stratification of the clusters, allows for specifying a fixed sample size selected for each land-cover class. Wulder et al. (2006) also proposed a two-stage design in which the pixels within a first-stage sample of PSUs (145 km by 111 km map sheets) are stratified by map land-cover class. A subsample of pixels is then selected within each PSU, so the stratified sampling is implemented within each individual sample PSU rather than implemented for the collection of sample PSUs as in Stehman et al. (2003).

In these three examples, combining stratification with cluster sampling introduces additional complexity to the analysis. However, to establish the groundwork for understanding these more complicated designs and analyses, we focus primarily on deriving estimators and illustrating the properties of these estimators for the simple design in which SRSWOR is implemented at both stages of the two-stage cluster sampling design.

### 2.3. Parameters to describe accuracy of land-cover composition

Land-cover composition accuracy can be quantified on a per-class basis by the parameters mean deviation (*MD*), mean absolute deviation (*MAD*), root mean square error (*RMSE*), and correlation (*CORR*). We focus on these parameters because they are relatively easy to compute and familiar to users. *MD* is equivalent to the "non-site-specific" accuracy for each class for the entire map, and allows users to assess for each land-cover class if the per-unit land-cover area derived from the map has a tendency to be higher or lower than the actual (reference) percent cover. However, *MD* can be close to 0 even if the composition data are poor. For example, blocks having large positive errors (map area higher than reference area) are "cancelled out" by blocks having large negative errors (map area lower than reference area) resulting in *MD* close to 0. *MAD* and *RMSE* are both based on simple averages — *MAD* is the average of the absolute deviations between the reference map areas and *RMSE* is the square root of the average of the squared deviations. A difference between *MAD* and *RMSE* is that *RMSE* places greater weight on outliers (large deviations). *CORR* quantifies the strength of the linear association between the map and reference percent cover data for each class. However, an important disadvantage of *CORR* is that it can still be high if the percent cover mapped is consistently too high or consistently too low relative to the reference percent cover, so *CORR* should not be used alone to describe composition accuracy. Together, the suite of these accuracy parameters provides a comprehensive description of composition accuracy. Willmott (1982); Pontius and Cheuk (2006); Ji and Gallo (2006), and Pontius et al. (2008) offer additional options for parameters that potentially could be used to describe accuracy of land-cover composition.

The statistical population in an assessment of land-cover composition accuracy is the region of interest (i.e. the mapped area) partitioned into spatial units of the target size(s). Let $X_i$ denote the area mapped as land-cover class $c$ in PSU $i$, and $Y_i$ denote the area of class $c$ determined from the ground condition or reference data (see Appendix A for a complete list of notation). Although $X_i$ and $Y_i$ depend on class $c$, to simplify notation, a subscript $c$ will not be used. In the statistical developments that follow, $X_i$ and $Y_i$ are treated as fixed quantities, not random variables. This is in keeping with the tenets of design-based inference (Särndal et al. 1992, Sec. 2.7; Stehman, 2000) subsequently used to characterize the expected value, bias, and precision of the estimators of composition accuracy. In practice, accuracy assessments are impacted by problems such as labeling error of the reference data or inability to exactly spatially co-register the reference sample and map locations (Powell et al., 2004). These additional sources of error are not accounted for in the design-based framework, and a fully comprehensive analysis of the error structure would require application of measurement error models (e.g. Särndal et al. 1992, Chapter 16).

In the parameter definitions that follow, $\sum_{U_1}$ indicates summation over all $K$ PSUs in the universe $U_1$ of such PSUs:

$$MD = \frac{1}{K}\sum_{U_1}(X_i - Y_i) \tag{1}$$

$$MAD = \frac{1}{K}\sum_{U_1}|X_i - Y_i| \tag{2}$$

$$MSE = \frac{1}{K}\sum_{U_1}(X_i - Y_i)^2 \quad (3)$$

$$CORR = \frac{\sum_{U_1} X_i Y_i - \frac{1}{K}\left(\sum_{U_1} X_i\right)\left(\sum_{U_1} Y_i\right)}{\sqrt{\sum_{U_1} X_i^2 - \frac{1}{K}\left(\sum_{U_1} X_i\right)^2}\sqrt{\sum_{U_1} Y_i^2 - \frac{1}{K}\left(\sum_{U_1} Y_i\right)^2}} \quad (4)$$

*RMSE* is obtained by taking the square root of *MSE* = mean square error. The accuracy parameters will take on different values depending on the spatial unit chosen.

A simple numerical illustration of the application of these accuracy parameters is created by viewing the Table 1 data as representing a hypothetical region of $K = 12$, 3 km by 3 km PSUs, where each PSU contains 10,000 pixels, and each pixel is 30 m by 30 m. The accuracy parameters (Table 2) calculated for this hypothetical region illustrate the reporting format of results for describing land-cover composition accuracy. To convert accuracy parameters expressed in terms of area to parameters expressed in terms of percent of area, *MD*, *MAD*, and *RMSE* are divided by the area of the spatial unit and multiplied by 100%. For the Table 2 example, the area-based parameters are divided by 900 ha, the area of each PSU, and then multiplied by 100%.

**Table 2**
Reporting format for composition accuracy.

| | Urban (1.1%) | | Forest (41.1%) | | Agriculture (38.6%) | | Wetland (1.3%) | |
|---|---|---|---|---|---|---|---|---|
| | Area (ha) | Area% | Area (ha) | Area% | Area (ha) | Area% | Area (ha) | Area% |
| *MD* | −55.8 | −6.2% | −9.6 | −1.1% | 78.5 | 8.7% | −4.6 | −0.5% |
| *MAD* | 55.8 | 6.2 | 34.8 | 3.9 | 78.5 | 8.7 | 21.0 | 2.3 |
| *RMSE* | 74.6 | 8.3 | 50.5 | 5.6 | 95.8 | 10.6 | 46.6 | 5.2 |
| *CORR* | 0.94 | | 0.98 | | 0.97 | | 0.16 | |

The parameters are computed treating the Table 1 data as a population of $K = 12$, 3 km by 3 km spatial units. The percent of the region covered by each of the four land-cover classes is shown in parentheses. The accuracy parameters are *MD* = mean deviation, *MAD* = mean absolute deviation, *RMSE* = root mean square error, and *CORR* = correlation.

## 3. Estimation under simple random sampling without replacement (SRSWOR)

The derivations of sample-based estimators of *MD*, *MAD*, *RMSE*, and *CORR* are presented in Appendix B for general one-stage and two-stage cluster sampling allowing for any design at each stage as long as the inclusion probabilities are known for the sampled PSUs and SSUs. An inclusion probability is defined as the *a priori* (i.e. prior to selecting the sample) probability of a sampling unit, either a PSU or an SSU, being included in the sample (Särndal et al. 1992, Sec. 2.4). The following formulas are applicable to the special case of SRSWOR implemented at both stages, where all PSUs have the same area and the sample size of pixels selected from each PSU at the second stage is the same for all PSUs (i.e. $N_i = N$ and $n_i = n$). In the notation used (see Appendix A), lower case letters denote the sample-based estimators of the accuracy parameters and the subscript 1 or 2 indicates whether the estimator is for one-stage or two-stage cluster sampling; e.g., $mad_1$ is the one-stage sampling estimator of *MAD*.

In one-stage cluster sampling, the PSU total area of class $c$ for the map and reference classification, $X_i$ and $Y_i$, are known for all sampled PSUs. Suppose a sample ($S_1$) of $k$ PSUs has been selected from the $K$ PSUs by SRSWOR leading to the first-stage inclusion probabilities for each PSU of $k/K$ (see Appendix A). In one-stage cluster sampling, the accuracy estimates are obtained by substituting the sample PSU values into the accuracy parameter formula. Then following from Appendix B Eqs. (B1)–(B4), the one-stage estimators are

$$md_1 = \frac{1}{k}\sum_{S_1}(X_i - Y_i) \quad (5)$$

$$mad_1 = \frac{1}{k}\sum_{S_1}|X_i - Y_i| \quad (6)$$

$$mse_1 = \frac{1}{k}\sum_{S_1}(X_i - Y_i)^2 \quad (7)$$

$$corr_1 = \frac{\sum_{S_1} X_i Y_i - \frac{1}{k}\left(\sum_{S_1} X_i\right)\left(\sum_{S_1} Y_i\right)}{\sqrt{\sum_{S_1} X_i^2 - \frac{1}{k}\left(\sum_{S_1} X_i\right)^2}\sqrt{\sum_{S_1} Y_i^2 - \frac{1}{k}\left(\sum_{S_1} Y_i\right)^2}} \quad (8)$$

where $\sum_{S_1}$ indicates summation over all $k$ PSUs in the first-stage sample, $S_1$.

For two-stage cluster sampling, the true (reference) area of a PSU covered by class $c$, $Y_i$, is not known and must be estimated from the second-stage sample within each PSU. We introduce the notation $r_i$ to denote the sample area of reference class $c$ in PSU $i$ and $m_i$ to denote the sample area of map class $c$ in PSU $i$, where these sample areas are derived from the second-stage sample of pixels from PSU $i$. If $N_i = N$, $n_i = n$, and SRSWOR is implemented at the second stage, then the second-stage inclusion probability of each pixel is $n/N$ and the pairwise inclusion probability of each pair of pixels is $n(n-1)/N(N-1)$ (Appendix A). From Appendix B expressions (B5) and (B6), the Horvitz–Thompson estimators of $X_i$ and $Y_i$ simplify to

$$\hat{X}_i = \frac{N}{n}m_i \quad (9)$$

and

$$\hat{Y}_i = \frac{N}{n}r_i \quad (10)$$

and, following from Appendix B expressions (B7) and (B8), the two-stage sampling estimators of *MD* and *MAD* can be expressed simply in terms of the sample areas $r_i$ and $m_i$,

$$md_2 = \frac{N}{nk}\sum_{S_1}(r_i - m_i) \tag{11}$$

$$mad_2 = \frac{N}{nk}\sum_{S_1}|r_i - m_i| \tag{12}$$

The two-stage sampling estimators of *RMSE* and *CORR* cannot be simplified to expressions in terms of the sample areas $r_i$ and $m_i$, but instead require the data for each individual pixel in the second-stage sample. To this end, we define $x_u = 1$ if pixel $u$ is mapped as class $c$, and $x_u = 0$ if pixel $u$ is mapped to a class other than $c$, and we define $y_u = 1$ if pixel $u$ is reference class $c$, and $y_u = 0$ if pixel $u$ is a reference class other than $c$. Each pixel $u$ is contained within a PSU, but to simplify notation, only the $u$ subscript will be used and the subscript $i$ indicating the PSU to which the SSU belongs will not be shown. The estimators of *RMSE* and *CORR* follow from the general formulas (B12) and (B13) in Appendix B:

$$mse_2 = \frac{N}{nk}\sum_{S_1}\left\{\sum_{S_{2i}}(x_u - y_u)^2 + \frac{N-1}{n-1}\sum_{S_{2i}}\sum_{u\neq v}(x_u x_v + y_u y_v - 2x_u y_v)\right\} \tag{13}$$

$$corr_2 = \frac{\sum_{S_1}\left\{\sum_{S_{2i}} x_u y_u + \frac{N-1}{n-1}\sum_{S_{2i}}\sum_{u\neq v} x_u y_v\right\} - \frac{N}{nk}\left(\sum_{S_1} m_i\right)\left(\sum_{S_1} r_i\right)}{\sqrt{\sum_{S_1}\left\{\sum_{S_{2i}} x_u^2 + \frac{N-1}{n-1}\sum_{S_{2i}}\sum_{u\neq v} x_u x_v\right\} - \frac{N}{nk}\left(\sum_{S_1} m_i\right)^2}\sqrt{\sum_{S_1}\left\{\sum_{S_{2i}} y_u^2 + \frac{N-1}{n-1}\sum_{S_{2i}}\sum_{u\neq v} y_u y_v\right\} - \frac{N}{nk}\left(\sum_{S_1} r_i\right)^2}} \tag{14}$$

where $\sum_{S_{2i}}$ indicates summation over all $n$ pixels in the second-stage sample $S_{2i}$ selected from PSU $i$, and $\sum_{S_{2i}}\sum_{u\neq v}$ indicates summation over all sample pairs of pixels $u$ and $v$ in the second-stage sample selected from PSU $i$.

It is possible that for some samples, $mse_2$ is negative, and in such cases $rmse_2$ cannot be estimated. Because $mse_2$ is computationally complex and to avoid the problem of negative $mse_2$ estimates, we also evaluated a simpler estimator of *MSE* constructed by substituting $\hat{X}_i$ (Eq. (9)) and $\hat{Y}_i$ (Eq. (10)) directly into the one-stage estimator of *MSE* (Eq. (7)). We will denote the square root of this estimator by $rmse_2^*$. By construction, $rmse_2^*$ is guaranteed to be positive. For the estimator $corr_2$, the terms inside the square roots in the denominator are not guaranteed to be positive, nor is $corr_2$ necessarily constrained to be between −1 and 1 as should be the case for an estimator of *CORR*. We constructed an alternative, simpler estimator of *CORR* (denoted by $corr_2^*$) by substituting $\hat{X}_i$ and $\hat{Y}_i$ directly into the one-stage estimator (Eq. (8)).

## 4. Empirical assessment of two-stage estimators

### 4.1. Methods

The bias and variance of the estimators derived for the two-stage sampling design with SRSWOR at both stages were empirically evaluated via a simulation study using populations created from the NLCD 1992 (Vogelmann et al., 2001) and NLCD 2001 (Homer et al., 2007). The NLCD products provide complete coverage land-cover composition data for the test regions selected (Fig. 1). In our simulation study, NLCD 1992 is regarded as the map to be evaluated, and NLCD 2001 is treated as the reference classification. The change between the NLCD products creates a reasonably realistic quantity and spatial distribution of known classification error for the empirical evaluation of properties of the proposed estimators of land-cover composition accuracy (Table 3). The test regions were partitioned into either 3 km by 3 km blocks or 6 km by 6 km blocks to assess whether the accuracy estimator properties depended on PSU size. We report results for a representative set of land-cover classes chosen to span a range from rare classes to common classes.

The estimators of composition accuracy are evaluated using the criterion of relative bias, which is the bias of the estimator divided by the true parameter value. For example, the relative bias of $mad_2$ is $100\% \times \{E(mad_2) - \text{MAD}\}/\text{MAD}$, where $E(mad_2)$ is the expected value of $mad_2$. The expected values of the estimators are approximated by simulating 5000 two-stage samples for each combination of first-stage ($k$) and second-stage ($n$) sample sizes and computing the average of the 5000 estimates of each accuracy parameter. The first-stage sample sizes evaluated are $k = 10$, 25, 50, and 100, and the second-stage sample sizes evaluated are $n = 10$, 25, 50, 100 and 250 (we omitted $n = 10$ with $k = 10$ because we considered the total sample size too small).

### 4.2. Bias of the accuracy estimators

By construction $md_2$ is an unbiased estimator of *MD* (see Appendix B), so simulation results evaluating $md_2$ are not reported. Because $mad_2$ is not an unbiased estimator of *MAD* for the two-stage design the considerable positive relative bias of $mad_2$ (Fig. 2) is not surprising. The pattern of relative bias of $mad_2$ is little affected by the choice of first-stage cluster sample size $k$, but the relative bias improves as the second-stage sample size $n$ increases. The high relative bias of $mad_2$ is present even for the more common classes (forest and the two agriculture examples), where a sample size of $n = 100$ is required to reduce the magnitude of the relative bias to 15% or smaller. The best performance of $mad_2$ occurs for the urban examples where relative bias is below 15% for all three urban cases. For the rare class water, relative bias of $mad_2$ is never below 15%.

The two-stage sampling estimator of *MSE* is not guaranteed to be positive, and in those cases where $mse_2$ is negative, it would not be possible to use $rmse_2$ to estimate *RMSE*. Negative $mse_2$ values do not commonly occur in the three urban populations as fewer than 5% of the samples result in negative estimates (Fig. 3). For the three common class examples (first column of Fig. 3), the proportion of samples resulting in negative $mse_2$ is high when $n = 10$ (except for $k = 250$), and the proportion of negative estimates can exceed 5% even for $n = 25$ when $k = 10$ or $k = 25$. For these three examples

**Fig. 1.** Location of the test region for the nine example populations evaluated in the simulation study. The large shaded region delineates the area that was partitioned into 6 km by 6 km blocks for the populations Agri, Urban, and Water (Table 1). Regions A and B delineate the areas that were partitioned into 3 km by 3 km blocks for the populations UrbanA, AgriA, ForestB, UrbanB, WaterB, and WetlandB.

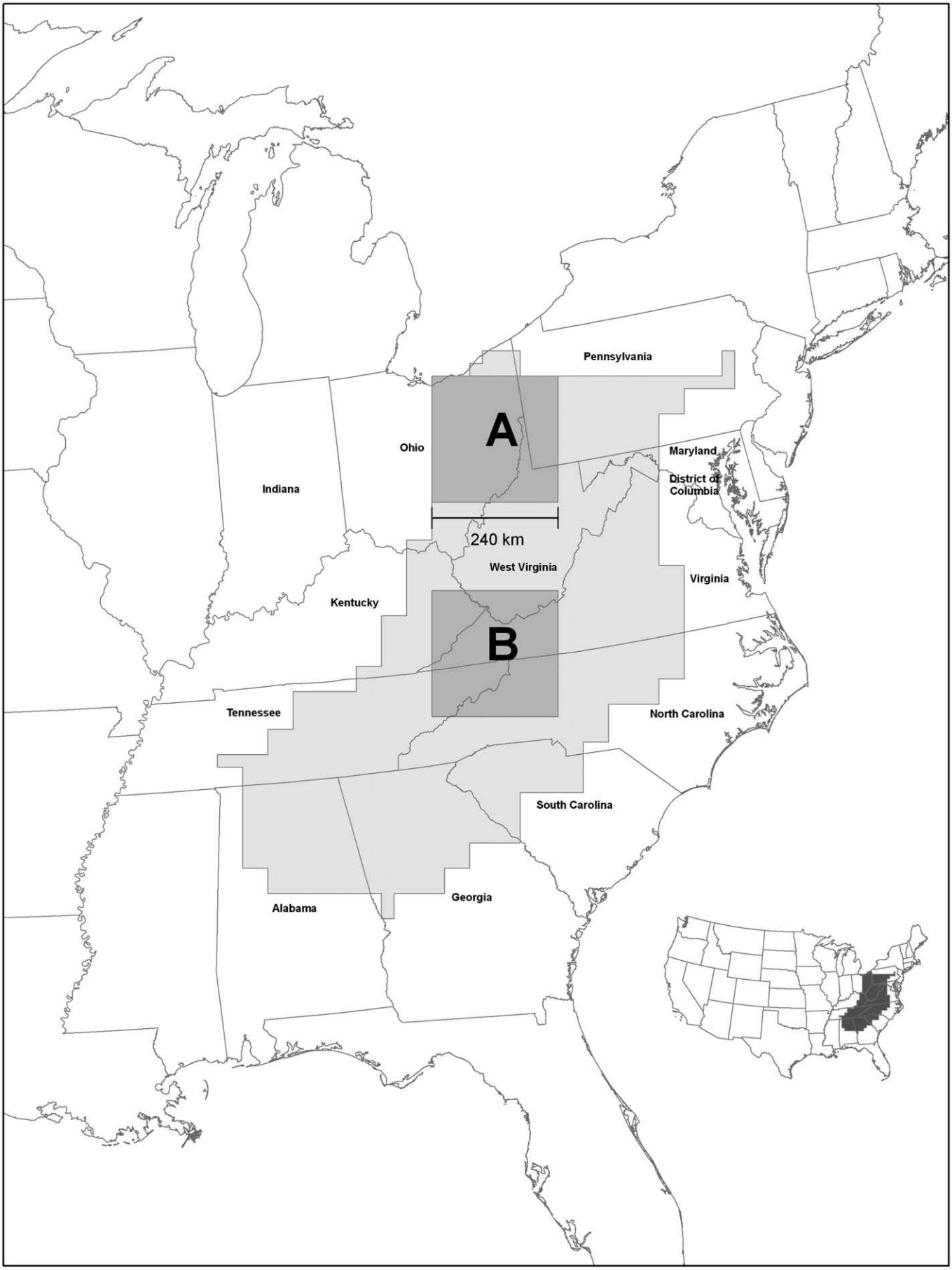
Pennsylvania
A
Ohio
Maryland
District of Columbia
Indiana
240 km
West Virginia
Virginia
Kentucky
B
Tennessee
North Carolina
South Carolina
Georgia
Alabama

**Table 3**
Composition accuracy parameters for populations used in the simulation study.

| | NLCD 1992 | | | | | |
|---|---|---|---|---|---|---|
| | % of Map | | | | | User's |
| Class | Area | *MD%* | *MAD%* | *RMSE%* | *CORR* | *Accuracy* |
| Agri (6 km) | 21.81 | 1.53 | 4.53 | 6.31 | 0.95 | 0.68 |
| AgriA (3 km) | 15.73 | −2.17 | 4.83 | 7.17 | 0.94 | 0.71 |
| ForestB (3 km) | 52.23 | 0.38 | 4.72 | 6.76 | 0.96 | 0.85 |
| Urban (6 km) | 3.16 | −6.81 | 6.82 | 9.01 | 0.93 | 0.80 |
| UrbanA (3 km) | 2.28 | −5.52 | 5.56 | 7.13 | 0.91 | 0.71 |
| UrbanB (3 km) | 5.72 | −9.48 | 9.48 | 12.36 | 0.94 | 0.91 |
| Water (6 km) | 1.52 | −0.02 | 0.23 | 0.49 | 0.99 | 0.83 |
| WaterB (3 km) | 1.87 | 0.05 | 0.38 | 0.88 | 0.99 | 0.80 |
| WetlandB (3 km) | 1.12 | 0.32 | 0.90 | 2.29 | 0.63 | 0.19 |

The populations are listed in order corresponding to columns 1 through 3 of Figs. 2–7. The three populations of $K=14{,}320$, 6 km by 6 km blocks are from the full region delineated in Fig. 1. The letters A and B attached to a land-cover class indicate the subregions (Fig. 1) where the populations of $K=6400$, 3 km by 3 km blocks are located.

representing common land-cover classes, the proportion of negative estimates declines as $k$ and $n$ increase. For the three rarest classes, a hump-shaped pattern describes the relationship between the proportion of negative estimates and sample size. The proportion of negative $mse_2$ estimates is actually lower for smaller $n$ and $k$ than it is for the moderate $n$ and $k$ values. Beyond the moderate sample sizes, the proportion of negative estimates decreases as $n$ and $k$ increase. The reason this humpback pattern emerges if the class is very rare is that when $k$ and $n$ are small, it is likely that the rare class does not appear in any of the sample blocks (i.e. $x_u$ and $y_u$ will be 0 for all sampled pixels) and $mse_2$ will be 0. As $k$ and $n$ increase to more moderately sized samples the likelihood of selecting some pixels of the rare class in either the map or reference classification is higher, and $mse_2$ will not be 0. However, because of the instability of the estimator $mse_2$, negative $mse_2$ estimates begin to occur. As $k$ and $n$ get larger still, $mse_2$ becomes more stable and the proportion of samples with negative $mse_2$ estimates decreases. For the rare classes, once the peak proportion of negative $mse_2$ estimates has been reached, the shape of the pattern of negative estimates with increasing $k$ and $n$ mimics the pattern observed for the three common classes. For the three rarest classes, the proportion of negative $mse_2$ estimates is still high even for the largest sample size of $k=250$.

The relative bias of $rmse_2$ is calculated conditional on the sample yielding a non-negative estimate of $mse_2$. In practice, if $mse_2$ is negative, no estimate of $rmse_2$ is available, so the relative bias shown in Fig. 4 represents performance of $rmse_2$ conditional on a sample estimate being reported. The relative bias of $rmse_2$ is generally negative indicating that $rmse_2$ tends to underestimate *RMSE*, and the magnitude of the relative bias of $rmse_2$ decreases as $k$ and $n$ increase (Fig. 4). The few cases where relative bias is positive occur for small sample size combinations of $k$ and $n$. Except for the rarest classes (water and wetland), the relative bias is generally smaller than 10% in absolute value. For the three rare class examples, first- and second-stage sample sizes must be at least $k=50$ and $n=25$ to reach a relative bias of −10%. The improvement in relative bias as $k$ increases is much greater for the rarest classes than for the more common classes. This suggests that if accuracy of rare land-cover classes is important, the first-stage sample size $k$ will need to be large.

In contrast to $rmse_2$, the relative bias of $rmse_2^*$ is generally positive (Fig. 5). But the relative bias of $rmse_2^*$ can be very high, and it is often much higher in absolute value than the relative bias of $rmse_2$. The high relative bias of $rmse_2^*$ is not mitigated by increasing $k$, but rather depends primarily on $n$, a behavior similar to that observed for $mad_2$. For the three common class examples, $rmse_2^*$ does not attain a relative bias below 10% until $n=100$, whereas $rmse_2$ achieves an absolute relative bias smaller than 10% for all combinations of $k$ and $n$. For the three urban examples, $rmse_2^*$ has a relative bias below 10% when $n$ reaches 25 or 50, whereas the relative bias of $rmse_2$ is below 10% in magnitude for all combinations of $n$ and $k$ evaluated. For the rarest classes, $rmse_2^*$ avoids the problem of negative estimates, but does so at the price of very high relative bias (50% and sometimes much higher) except when $k$ approaches 250. On the basis of relative bias, $rmse_2$ is far superior to $rmse_2^*$. A possible solution to the problem of negative $mse_2$ estimates would be to use $rmse_2^*$ in place of $rmse_2$ when $mse_2$ is negative. However, it would still be important to recognize that when $k$ or $n$ is small, no reliable estimate of *RMSE* is available for the very rare classes.

For the three common class examples, $corr_2$ shows practically no bias for any combination of sample sizes $n$ and $k$ (Fig. 6). For the three urban cases and the three rare land-cover class examples, the bias of $corr_2$ is negative (*CORR* is underestimated). In these cases, the bias decreases as $n$ and $k$ increase. For the three urban examples, bias is generally small except when $k=10$ and for a few cases when $k=25$ and $n$ is small. For the three rarest class examples, underestimation of *CORR* can be considerable for combinations of small $k$ and $n$. Approximate guidelines to ensure a small bias of $corr_2$ for the water and wetland populations are $k=250$ (any $n$), $k=100$ with $n>10$, $k=50$ with $n=50$ or higher, or $k=25$ with $n=100$ or higher.

A difficulty with the two-stage estimator of *CORR* is that if the class is rare and the sample size is small, few if any pixels of the rare class may be selected by the sample, and $corr_2$ is 0. This appears to be the cause of the extreme underestimation of *CORR* for the three rarest land-cover classes. The estimator $corr_2$ also is susceptible to the problem that for some samples, one or both of the expressions inside the two square root terms in the denominator (Eq. (14)) may be negative, and $corr_2$ cannot be computed. As was the case with $mse_2$, such samples are much more prevalent when $n$ or $k$ is small and rarely occur when $k$ and $n$ are larger. Still another problem with $corr_2$ is that it may exceed 1, which is not a possible value for *CORR*. Again, this problem occurs when $k$ and $n$ are small, and disappears as $k$ and $n$ increase. For the bias results reported, if negative estimates prevented computation of $corr_2$, it was set to 0, and if $corr_2$ was greater than 1, it was reset to 1. In general, the relative bias of $corr_2^*$ did not improve upon the relative bias of $corr_2$ for small $k$ and $n$ and performed far worse than $corr_2$ for larger $k$ and $n$ (results not shown). As is true for estimating *RMSE*, reliable estimates of *CORR* for rare classes require relatively large sample sizes $k$ and $n$.

### 4.3. Precision of the accuracy estimators

In addition to evaluating bias, it is also of interest to examine the variance (i.e. precision) of the accuracy estimators as a function of $k$ and $n$. We present results only for the relative standard error of $rmse_2$, defined as the standard error of $rmse_2$ divided by *RMSE*. Relative standard error scales variability relative to the magnitude of the parameter being estimated.

As expected, the relative standard error of $rmse_2$ decreases as the sample size $k$ and $n$ increase (Fig. 7). The relationship of relative standard error with $k$ and $n$ depends on whether the land-cover class is common or rare. For the common classes (agriculture, forest, and urban), the marginal improvement in relative standard error is negligible once $n$ reaches 50. For the three rarest class examples, the relative standard error curves flatten out around $n=100$, indicating that a larger second-stage sample size is needed to estimate *RMSE* with acceptable precision. From two-stage sampling theory, the relative standard error will decrease as a function of the square root of $k$, and the results displayed in Fig. 7 confirm this relationship.

The relative standard errors for $rmse_2$ are high for practically all combinations of $k$ and $n$ and all nine cases evaluated. For illustration, suppose a relative standard error of 20% is considered acceptable (e.g. if *RMSE* is 10.0, a standard error of 2.0 for $rmse_2$ would be acceptable). For the three rarest classes, a 20% relative standard error is achieved only for the wetland class and for a sample size of $k=250$ and $n>50$. For the other six populations (the non-rare classes), the 20% relative standard error goal would generally be achieved by sample sizes of $k=50$ with $n=50$ or by

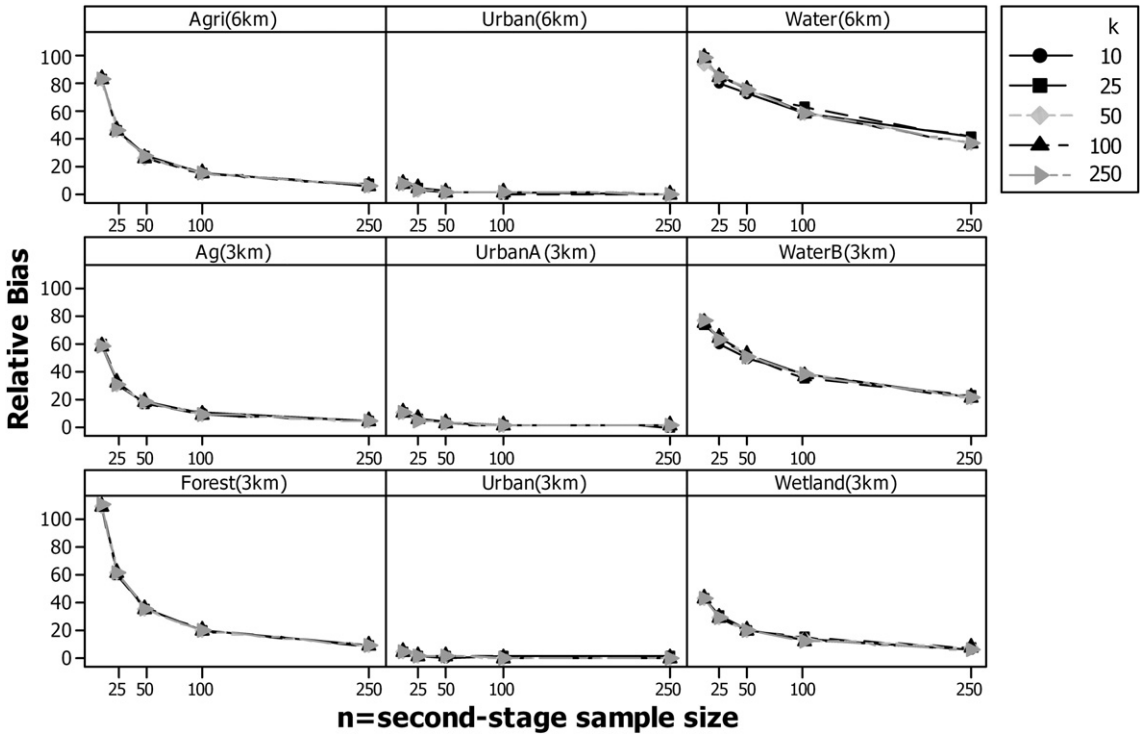


**Fig. 2.** Relative bias of $mad_2$ (estimator of the mean absolute deviation, *MAD*). The nine cases evaluated are organized in the 3×3 arrangement with the 6 km by 6 km sample block results in the top row, and the columns ordered from left to right by decreasing percent cover of the land-cover classes. The first-stage sample size is *k*.

$k = 100$ with $n = 25$. A relative standard error of 20% is not a very strict requirement for precision, so these results suggest that large sample sizes are needed to obtain precise estimates of *RMSE*, and even larger sample sizes will be needed for the very rare classes. Sampling designs other than SRSWOR might require smaller sample sizes from the very rare classes to achieve the 20% relative standard error, and stratified sampling would be an obvious first consideration as an alternative design.

## 5. Discussion

A common application of land-cover maps is to compute land-cover composition for each spatial unit in a partition of the region mapped. To assess accuracy of land-cover composition, two-stage cluster sampling with SRSWOR at each stage is a relatively simple, practical design. The estimators of the composition accuracy

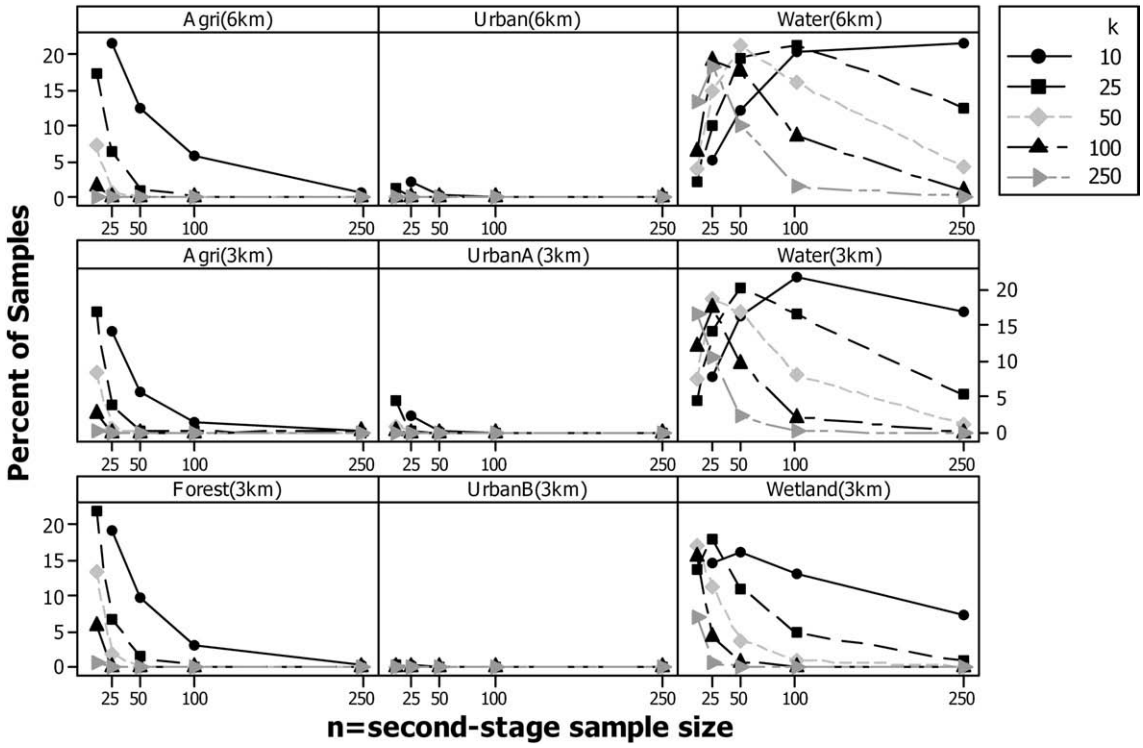


**Fig. 3.** Percent of samples with negative $mse_2$ estimates. The first-stage sample size is *k*.

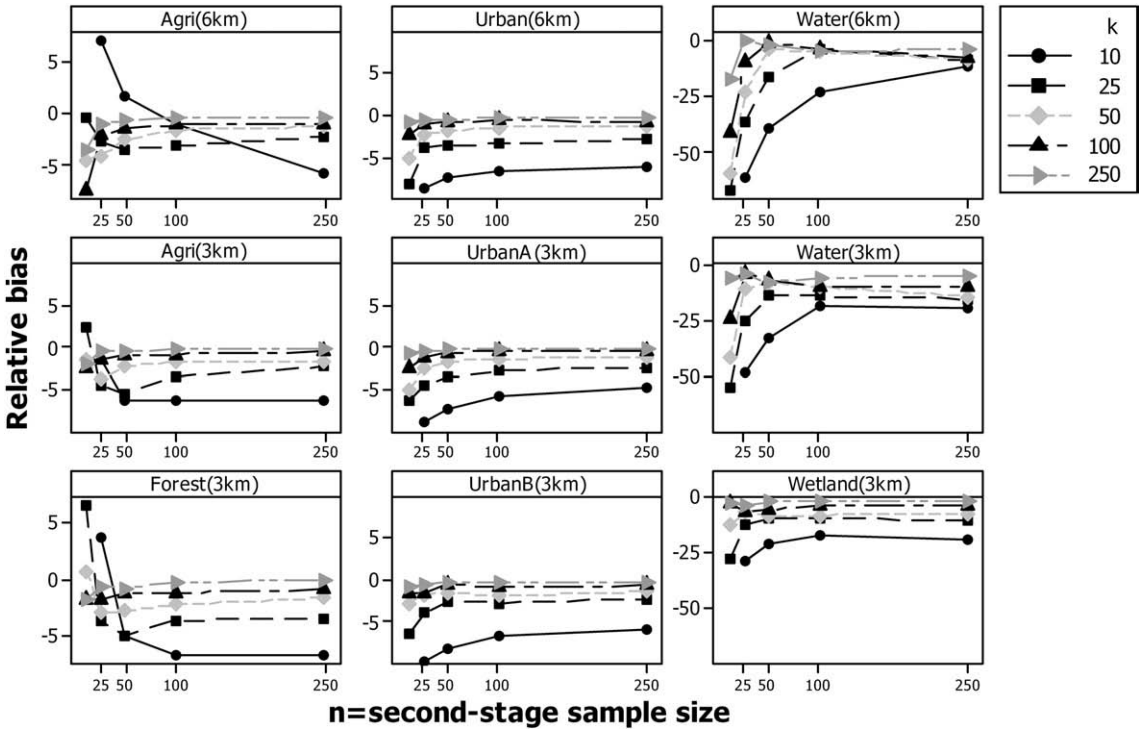


**Fig. 4.** Relative bias of $rmse_2$ (estimator of the root mean square error, *RMSE*).

parameters root mean square error (*RMSE*) and correlation (*CORR*) are demonstrated to have small relative biases as long as the sample size of clusters ($k$) and sample size of pixels per cluster ($n$) are not too small. In general, treating data from a two-stage cluster sample as if they were from a one-stage cluster sample leads to badly biased estimates. This bias is particularly problematic for the estimator of *MAD* because the only option available is to use the one-stage formula. The high relative bias observed for the estimators $rmse_2^*$ and $corr_2^*$ when applied to the two-stage sampling design serves as an important caution. Although intuitively it seems reasonable to construct estimators of these accuracy parameters by substituting the estimated PSU values (e.g. $\hat{X}_i$ and $\hat{Y}_i$ of Eqs. (9) and (10)) into the formula for each parameter, this approach is not appropriate. Subsampling within the PSUs (i.e. two-stage sampling) significantly alters the estimation problem. These estimation considerations likely apply to other representations of accuracy that could be applied to

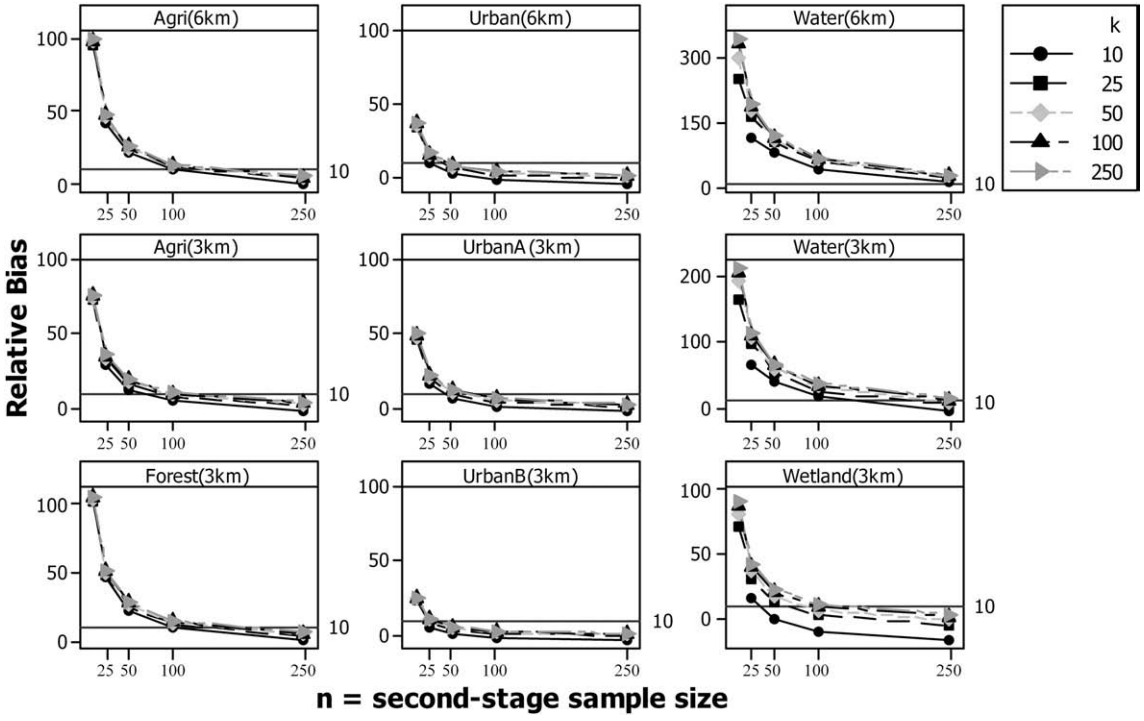


**Fig. 5.** Relative bias of $rmse_2^*$(alternative estimator of the root mean square error, *RMSE*).

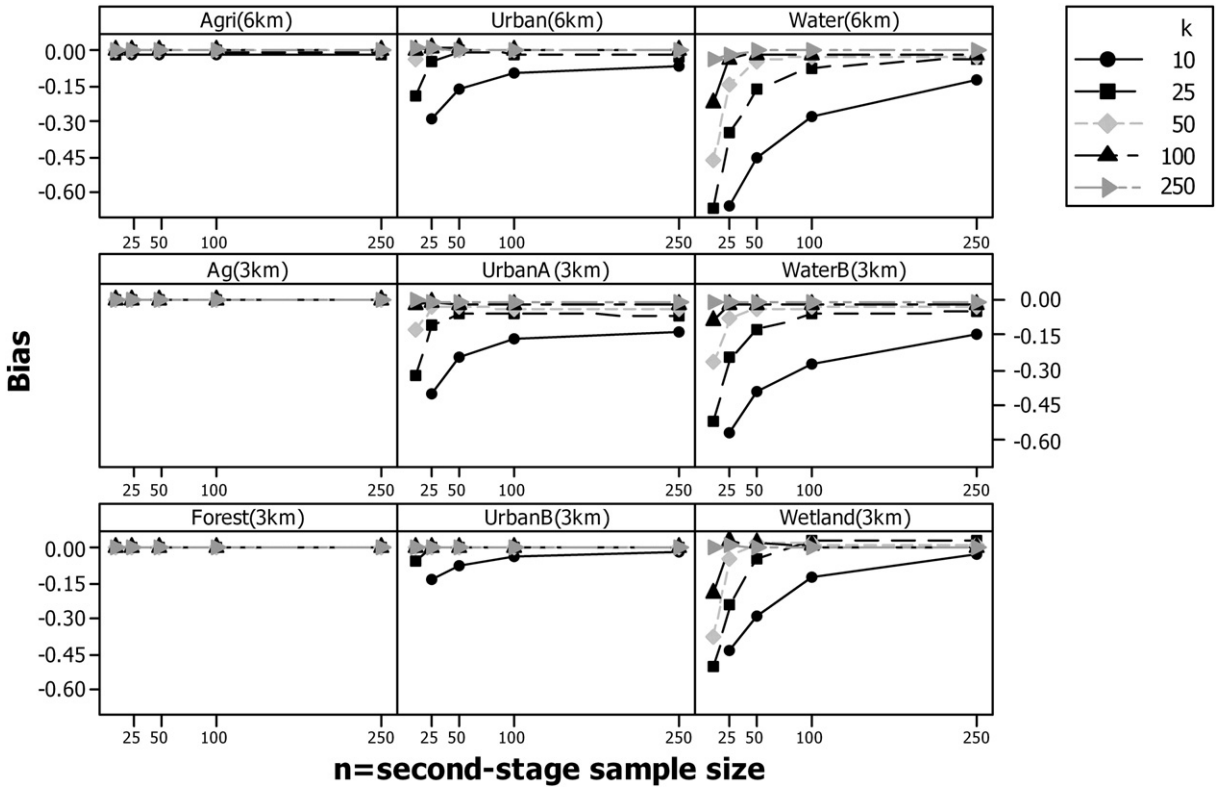


**Fig. 6.** Bias of $corr_2$ (estimator of the correlation, *CORR*). Because *CORR* is constrained to be between −1 and 1, the scaling provided by relative bias is less justified so bias is displayed instead of relative bias.

land-cover composition data, for example, the parameters reviewed by Ji and Gallo (2006) and those proposed by Pontius and Cheuk (2006).

The methodology presented for estimating land-cover composition accuracy is applicable to other attributes or quantities. For example, if the mapped attribute is a particular directional change in

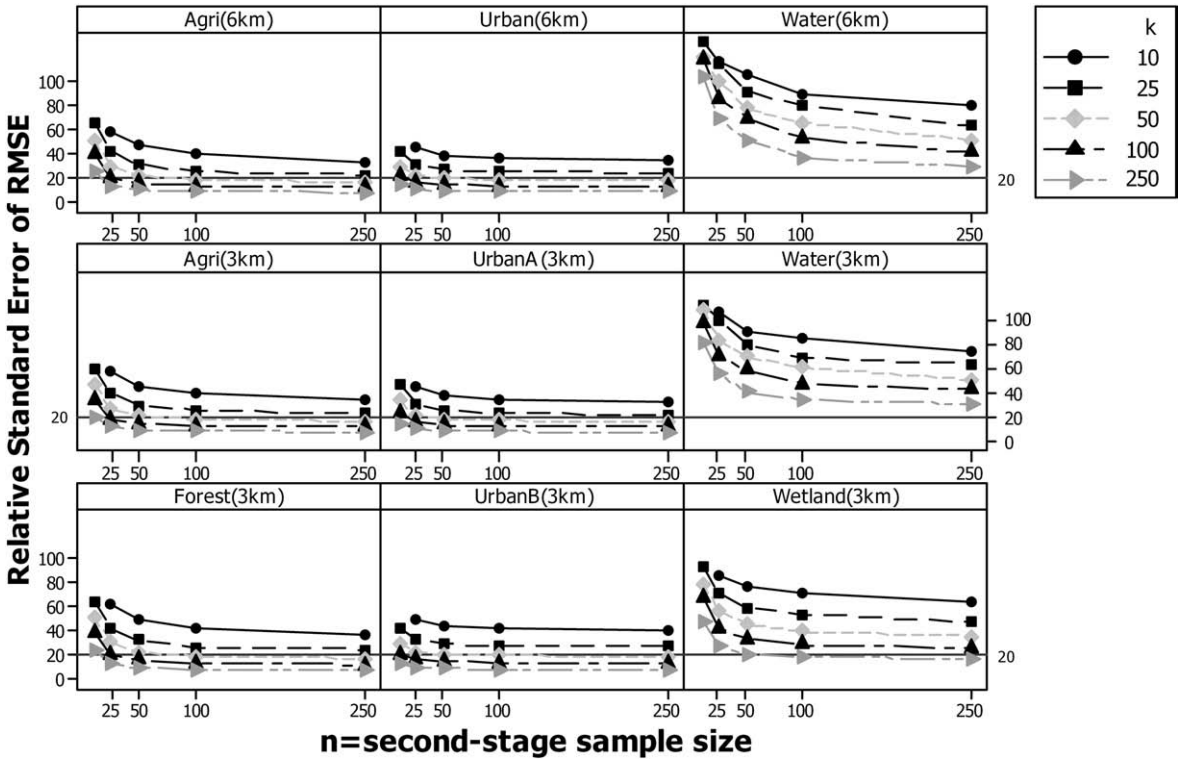


**Fig. 7.** Relative standard error of $rmse_2$ (estimator of *RMSE*). Relative standard error is the standard error of $rmse_2$ divided by *RMSE*.

land-cover class (e.g. forest to developed) and each pixel is mapped as having exhibited this directional change or not, an objective may be to assess how accurately gross change is portrayed by the map when the per-pixel directional data are aggregated to some spatial unit (e.g. $X_i =$ the area of gross change from forest to developed for PSU $i$). This objective can be addressed using the methodology described by defining $y_u = 1$ if pixel $u$ changed from forest to developed according to the reference classification ($y_u = 0$ otherwise) and defining $x_u = 1$ if pixel $u$ changed from forest to developed according to the map classification ($x_u = 0$ otherwise). As another example, suppose a quantity such as percent impervious surface or percent tree canopy cover is associated with each pixel. It would be possible to assess the accuracy of such quantities when the per-pixel data are aggregated to some spatial unit (e.g. accuracy of percent impervious surface of 10 km by 10 km blocks). In this case, the values associated with pixel $u$, $x_u$ and $y_u$, would be the percent impervious surface according to the map and reference data, respectively.

We have not broached the topic of estimating standard errors of the accuracy estimators. Because of the complexity of the estimators $rmse_2$ and $corr_2$, it is likely that the analytic expressions for their standard errors will be extremely complicated to derive and compute. A practical solution may be to implement a computer intensive estimation procedure such as balanced repeated replication (Wolter 1985, Chapter 3), a Polya urn approach (Magnussen et al., 2004), or jackknife (Berger & Skinner 2005). Developing a practical approach to estimating standard errors will be important if these accuracy parameters are adopted for common use.

We have also only briefly touched on the sampling design issues related to estimating composition accuracy. One-stage cluster sampling avoids much of the complexity of the estimation procedure required by two-stage sampling. But this advantage of one-stage cluster sampling is counter-balanced by the practical problems related to the expense and time required to collect complete coverage reference data for each sampled PSU, and also whether practically relevant PSU sizes can be used with one-stage cluster sampling. In practice, accuracy assessment sampling designs typically employ stratification by map land-cover class to ensure that rare classes are represented in the sample. For example, the two-stage sampling designs implemented for the NLCD accuracy assessments (Stehman et al., 2003, 2008) both employed stratification. We have not evaluated precision of the estimators of composition accuracy for these more complex stratified designs. Further investigation of the precision of different design options and different allocations of $k$ and $n$ are needed, and these investigations should include evaluating the precision of the usual error matrix based accuracy estimators. Lastly, it is likely that in practice, assessing composition accuracy for several sizes of spatial units would be of interest. Extending the assessment to more than one size spatial unit would further complicate the sampling design and estimation procedures and raise the question of whether it would be practical to obtain large enough sample sizes to adequately meet the objectives of the assessment for several sizes of spatial units.

## 6. Conclusions

Assessing accuracy of land-cover composition extends the richness of information extracted from an accuracy assessment. To assess accuracy of land-cover composition, it is necessary to construct the sampling design to specifically target this objective. Two-stage cluster sampling provides the capacity to concentrate the sample pixels within a sample of the larger spatial units to provide the necessary sample size per spatial unit to assess land-cover composition accuracy. The estimators of the parameters quantifying land-cover composition accuracy for two-stage cluster sampling are complex, and the naïve approach of constructing estimators as simple sample-based analogs of the accuracy parameters (e.g. $rmse_2^*$ and $corr_2^*$) can produce badly biased estimates. Although we have demonstrated that a strategy for estimating composition accuracy can be implemented using two-stage cluster sampling with SRSWOR at both stages, many practical questions remain to be resolved to create a more operationally practical, efficient strategy for this type of assessment. Two of the critical issues to resolve are determining how to allocate the sample among PSUs and SSUs (i.e. choosing $k$ and $n$) to optimize precision of the accuracy estimators, and how best to combine stratified and cluster sampling to achieve simultaneously precise estimates of composition accuracy as well as precise estimates of class-specific accuracy for the per-pixel assessment.

## Appendix A. Notation

*Land-cover composition accuracy parameters (capital letters denote a parameter, and lower case letters denote an estimator of that parameter)*

| | |
|---|---|
| *MD* | Mean deviation ($md_1$, $md_2$ are one- and two-stage sampling estimators of *MD*) |
| *MAD* | Mean absolute deviation ($mad_1$, $mad_2$ are one- and two-stage sampling estimators of *MAD*) |
| *MSE* | Mean square error ($mse_1$, $mse_2$ are one- and two-stage sampling estimators of *MSE*) |
| *RMSE* | Root mean square error ($rmse_1$, $rmse_2$ are one- and two-stage sampling estimators of *RMSE*) |
| *CORR* | Correlation ($corr_1$, $corr_2$ are one- and two-stage sampling estimators of *CORR*) |

*Universe and sample properties*

| | |
|---|---|
| $U_1$ | Universe of all primary sampling units (PSUs) comprising a partition of the entire region |
| $U_{2i}$ | Universe of secondary sampling units (SSUs) within PSU i (i.e. all pixels within PSU $i$) |
| $S_1$ | First-stage sample of PSUs from $U_1$ |
| $S_{2i}$ | Second-stage sample of SSUs (pixels) selected from PSU $i$ in $S_1$ |
| $K$ | Number of PSUs in $U_1$ |
| $k$ | Number of PSUs selected in the first-stage sample $S_1$ |
| $N_i$ | Number of SSUs in PSU $i$ |
| $N$ | Number of SSUs in PSU $i$ when $N_i$ is the same for all PSUs in $U_1$ |
| $n_i$ | Number of SSUs sampled in the second-stage sample $S_{2i}$ from PSU $i$ |
| $n$ | Number of SSUs sampled in the second-stage sample from PSU $i$ when $n_i$ is the same for all PSUs |

*Subscript notation*

1) $i$ denotes an observation or quantity obtained for PSU $i$
2) $u$ denotes an observation for an SSU, which is a pixel in the examples presented; each pixel is contained within a PSU, but to simplify notation, only the $u$ subscript will be used and the $i$ subscript indicating the PSU to which the SSU belongs will not be shown.
3) Symbols with no subscript are quantities representing the whole universe or region

*Observed quantities and estimators*

| | |
|---|---|
| $y_u$ | Observation determined from the reference data for SSU $u$ in PSU $i$ |
| $Y_i = \sum_{U_{2i}} y_u$ | Total of $y_u$ for PSU $i$ |
| $Y = \sum_{U_1} Y_i$ | Total area of class $c$ in $U_1$ as determined by the reference data |
| $x_u$ | Observation determined from the map or classification for SSU $u$ in a PSU |

$X_i = \sum_{U_{2i}} x_u$ Total of $x_u$ for PSU $i$

$X = \sum_{U_1} X_i$ Total area mapped as class $c$ in $U_1$

$r_i$ Area of second-stage sample with reference label of land-cover class $c$ in PSU $i$

$m_i$ Area of second-stage sample with map label of land-cover class $c$ in PSU $i$

*Inclusion probabilities*

$\pi_{1i}$ Inclusion probability of PSU $i$ (first-stage sample inclusion probability)

$\pi_{2u|i}$ Inclusion probability of SSU $u$ conditional on PSU $i$ being selected in first-stage sample

$\pi_{2uv|i}$ Pairwise inclusion probability of SSUs $u$ and $v$ conditional on PSU $i$ being selected in first stage (i.e., the probability that both pixel $u$ and pixel $v$ are included in the second-stage sample given that PSU $i$ is selected at the first stage)

*Inclusion probabilities for simple random sampling without replacement at both stages*

$\pi_{1i}$ $k/K$

$\pi_{2u|i}$ $n_i/N_i$

$\pi_{2uv|i}$ $n_i(n_i-1)/N_i(N_i-1)$ for any pair of SSUs $u$ and $v$ within the same PSU

## Appendix B. Estimation theory

The general estimators developed in Appendix B can be applied to any two-stage sampling design for which the inclusion probabilities $\pi_{1i}$, $\pi_{2u|i}$, and $\pi_{2uv|i}$ are known for the sampled units. The Horvitz–Thompson (HT) estimator (Lohr 1999, Sec. 6.4.1) provides a general tool for obtaining unbiased estimators of population totals. For any population and any sampling scheme, the HT estimator of a population total is the sum of each sample observation divided by its corresponding inclusion probability. The HT estimator is applicable to any sampling design for which the inclusion probabilities are known for the sampled units.

To estimate the composition accuracy parameters, suppose that a sample ($S_1$) of $k$ PSUs has been selected from the $K$ PSUs in $U_1$ according to a sampling design with inclusion probability $\pi_{1i}$ for PSU $i$. Then the one-stage HT estimators for the population parameters (1)–(4) are

$$md_1 = \frac{1}{K}\sum_{S_1}\frac{X_i - Y_i}{\pi_{1i}} \tag{B1}$$

$$mad_1 = \frac{1}{K}\sum_{S_1}\frac{|X_i - Y_i|}{\pi_{1i}} \tag{B2}$$

$$mse_1 = \frac{1}{K}\sum_{S_1}\frac{(X_i - Y_i)^2}{\pi_{1i}} \tag{B3}$$

$$corr_1 = \frac{\sum_{S_1}\frac{X_iY_i}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{X_i}{\pi_{1i}}\right)\left(\sum_{S_1}\frac{Y_i}{\pi_{1i}}\right)}{\sqrt{\sum_{S_1}\frac{X_i^2}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{X_i}{\pi_{1i}}\right)^2}\sqrt{\sum_{S_1}\frac{Y_i^2}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{Y_i}{\pi_{1i}}\right)^2}} \tag{B4}$$

and the one-stage estimator of *RMSE*, $rmse_1$, is obtained as the square root of $mse_1$. Because (B1)–(B3) are HT estimators of totals, they constitute unbiased estimators of their population counterparts. Although *MSE* is estimated unbiasedly by (B3), and all the totals involved in *CORR* are estimated unbiasedly by (B4), neither $rmse_1$ nor $corr_1$ constitute unbiased estimators of their population counterparts because they are not linear functions of these quantities.

In a two- stage cluster design, a sample of pixels is drawn from each of the PSUs selected in the first-stage sample, and the PSU totals $X_i$ and $Y_i$ are estimated from the second-stage sample. Accordingly, the second-stage sample HT estimators of $X_i$ and $Y_i$ are

$$\hat{X}_i = \sum_{S_{2i}}\frac{x_u}{\pi_{2u|i}} \tag{B5}$$

and

$$\hat{Y}_i = \sum_{S_{2i}}\frac{y_u}{\pi_{2u|i}} \tag{B6}$$

where the summation is over all pixels selected in the second-stage sample from PSU $i$. Although $X_i$ is a known quantity for each PSU (because the map land-cover class is available for the entire PSU), the use of estimated area $\hat{X}_i$ typically improves precision of the estimators because of the positive correlation between $\hat{X}_i$ and $\hat{Y}_i$. For example, $\hat{X}_i - \hat{Y}_i$ is typically less variable over the set of all possible samples than $X_i - \hat{Y}_i$.

Because $\hat{Y}_i$ and $\hat{X}_i$ are unbiased estimators, from the properties of expectation, inserting these estimators into Eq. (B1) suffices to obtain an unbiased two-stage estimator of *MD*,

$$md_2 = \frac{1}{K}\sum_{S_1}\frac{\hat{X}_i - \hat{Y}_i}{\pi_{1i}} \tag{B7}$$

Estimating *MAD* from the two-stage sampling design is problematic. Although an unbiased one-stage estimator of *MAD* exists (i.e., $mad_1$), we are not aware of a way to accommodate the absolute value of the differences to achieve an unbiased two-stage estimator. Thus, the simple substitution approach replacing $X_i$ and $Y_i$ with $\hat{X}_i$ and $\hat{Y}_i$ in Eq. (B2) is employed as the only available option, leading to the two-stage estimator of *MAD* given by

$$mad_2 = \frac{1}{K}\sum_{S_1}\frac{|\hat{X}_i - \hat{Y}_i|}{\pi_{1i}} \tag{B8}$$

Deriving an unbiased estimator of *MSE* for the two-stage design requires re-writing the squared PSU totals $Y_i$ and $X_i$ and their products $X_iY_i$ as sums of totals as follows:

$$X_i^2 = \left(\sum_{U_{2i}} x_u\right)^2 = \sum_{U_{2i}} x_u^2 + \sum_{U_{2i}}\sum_{u\neq v} x_u x_v = Q_{Xi} \tag{B9a}$$

$$Y_i^2 = \left(\sum_{U_{2i}} y_u\right)^2 = \sum_{U_{2i}} y_u^2 + \sum_{U_{2i}}\sum_{u\neq v} y_u y_v = Q_{Yi} \tag{B9b}$$

$$X_iY_i = \left(\sum_{U_{2i}} x_u\right)\left(\sum_{U_{2i}} y_u\right) = \sum_{U_{2i}} x_u y_u + \sum_{U_{2i}}\sum_{u\neq v} x_u y_v = P_{XYi} \tag{B9c}$$

where $\sum_{U_{2i}}\sum_{u\neq v}$ denotes summation over all pairs of pixels within PSU $i$ (excluding pairing a pixel with itself). The corresponding HT estimators of these totals are

$$\hat{Q}_{Xi} = \sum_{S_{2i}}\frac{x_u^2}{\pi_{2u|i}} + \sum_{S_{2i}}\sum_{u\neq v}\frac{x_u x_v}{\pi_{2uv|i}} \tag{B10a}$$

$$\hat{Q}_{Yi} = \sum_{S_{2i}}\frac{y_u^2}{\pi_{2u|i}} + \sum_{S_{2i}}\sum_{u\neq v}\frac{y_u y_v}{\pi_{2uv|i}} \tag{B10b}$$

$$\hat{P}_{XYi} = \sum_{S_{2i}}\frac{x_u y_u}{\pi_{2u|i}} + \sum_{S_{2i}}\sum_{u\neq v}\frac{x_u y_v}{\pi_{2uv|i}} \tag{B10c}$$

where $\sum_{S_{2i}}\sum_{u\neq v}$ denotes summation over all pairs of sample pixels within first-stage sample PSU $i$ (excluding pairing a pixel with itself). To estimate the second term of the previous expressions (i.e. the double summation terms), the pairwise inclusion probabilities must be used because pairs of SSUs (pixels) are involved. Then, rewriting the one-stage unbiased estimator of *MSE* (Eq. (B3)) as

$$mse_1 = \frac{1}{K}\sum_{S_1}\frac{(X_i - Y_i)^2}{\pi_{1i}} = \frac{1}{K}\sum_{S_1}\frac{X_i^2}{\pi_{1i}} + \frac{1}{K}\sum_{S_1}\frac{Y_i^2}{\pi_{1i}} - \frac{2}{K}\sum_{S_1}\frac{X_i Y_i}{\pi_{1i}} \quad \text{(B11)}$$

and substituting estimators (B10a)–(B10c) into Eq. (B11), an unbiased two-stage sampling estimator of *MSE* is

$$mse_2 = \frac{1}{K}\sum_{S_1}\frac{\hat{Q}_{Xi}}{\pi_{1i}} + \frac{1}{K}\sum_{S_1}\frac{\hat{Q}_{Yi}}{\pi_{1i}} - \frac{2}{K}\sum_{S_1}\frac{\hat{P}_{XYi}}{\pi_{1i}} \quad \text{(B12)}$$

Although *MSE* is estimated unbiasedly by $mse_2$, $rmse_2 = \sqrt{mse_2}$ is not an unbiased estimator of *RMSE*.

Finally, as to the two-stage estimation of *CORR*, since even the one-stage estimator $corr_1$ is biased, no two-stage unbiased estimator is possible. However, it seems natural to use the second-stage unbiased estimators of the quantities involved in $corr_1$ to obtain the following two-stage estimator

$$corr_2 = \frac{\sum_{S_1}\frac{\hat{P}_{XYi}}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{X}_i}{\pi_{1i}}\right)\left(\sum_{S_1}\frac{\hat{Y}_i}{\pi_{1i}}\right)}{\sqrt{\sum_{S_1}\frac{\hat{Q}_{Xi}}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{X}_i}{\pi_{1i}}\right)^2}\sqrt{\sum_{S_1}\frac{\hat{Q}_{Yi}}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{Y}_i}{\pi_{1i}}\right)^2}} \quad \text{(B13)}$$

After most of the simulation results had been completed, one of us (LF) derived an alternative estimator of *CORR*,

$$corr_{2a} = \frac{\sum_{S_1}\frac{\hat{P}_{XYi}}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{P}_{XYi}}{\pi_{1i}^2} + \sum_{S_1}\sum_{i\neq j}\frac{\hat{X}_i\hat{Y}_j}{\pi_{1i}\pi_{1j}}\right)}{\sqrt{\sum_{S_1}\frac{\hat{Q}_{Xi}^2}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{Q}_{Xi}^2}{\pi_{1i}^2} + \sum_{S_1}\sum_{i\neq j}\frac{\hat{X}_i\hat{X}_j}{\pi_{1i}\pi_{1j}}\right)}\sqrt{\sum_{S_1}\frac{\hat{Q}_{Yi}^2}{\pi_{1i}} - \frac{1}{K}\left(\sum_{S_1}\frac{\hat{Q}_{Yi}^2}{\pi_{1i}^2} + \sum_{S_1}\sum_{i\neq j}\frac{\hat{Y}_i\hat{Y}_j}{\pi_{1i}\pi_{1j}}\right)}} \quad \text{(B14)}$$

where $\sum_{S_1}\sum_{i\neq j}$ denotes summation over all pairs of PSUs excluding a PSU paired with itself. All of the quantities involved in expression (B14) constitute unbiased estimators of their population counterparts. Limited simulation results indicate that relative to $corr_2$, $corr_{2a}$ is slightly less susceptible but not completely immune to the problems of negative estimates of the terms inside the square roots and producing estimates greater than 1, and that the relative bias of $corr_{2a}$ is slightly smaller than $corr_2$.

## References


Bakker, K. K., Naugle, D. E., & Higgins, K. F. (2002). Incorporating landscape attributes into models for migratory grassland bird conservation. *Conservation Biology*, *16*, 1638–1646.

Berger, Y. G., & Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, *67*, 79–89.

Blackard, J. A., Finco, M. V., Helmer, E. H., Holden, G. R., Hoppus, M. L., Jacobs, D. M., et al. (2008). Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*, 1658–1677.

Bossard, M., Feranec, J., & Otohel, J. (2000). CORINE land cover technical guide — Addendum 2000. *Technical Report 40* Copenhagen, Denmark: European Environment Agency.

Comeleo, R. L., Paul, J. F., August, P. V., Copeland, J. L., Baker, C., Hale, S. S., et al. (1996). Relationships between watershed stressors and sediment contamination in Chesapeake Bay estuaries. *Landscape Ecology*, *11*, 307–319.

Couturier, S., Mas, J. F., Vega, A., & Tapia, V. (2007). Accuracy assessment of land cover maps in sub-tropical countries: A sampling design for the Mexican National Forest Inventory. *OnLine Journal of Earth Sciences*, *1*, 127–135.

Driscoll, M. J. L., & Donovan, T. M. (2004). Landscape context moderates edge effects: Nesting success of wood thrushes in central New York. *Conservation Biology*, *18*, 1330–1338.

Edwards, T. C., Jr., Moisen, G. G., & Cutler, D. R. (1998). Assessing map accuracy in an ecoregion-scale cover-map. *Remote Sensing of Environment*, *63*, 73–83.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185–201.

Fuller, R. M., Groom, G. B., & Jones, A. R. (1994). The land cover map of Great Britain: An automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering & Remote Sensing*, *60*, 553–562.

Glennon, M. J., & Porter, W. F. (1999). Using satellite imagery to assess landscape-scale habitat for wild turkeys. *Wildlife Society Bulletin*, *27*, 646–653.

Han, K. S., Champeaux, J. L., & Roujean, J. L. (2004). A land cover classification product over France at 1 km resolution using SPOT4/VEGETATION data. *Remote Sensing of Environment*, *92*, 52–66.

Heiskanen, J. (2008). Evaluation of global land cover data sets over the tundra–taiga transition zone in northernmost Finland. *International Journal of Remote Sensing*, *29*, 3727–3751.

Hollister, J. W., Gonzalez, M. L., Paul, J. F., August, P. V., & Copeland, J. L. (2004). Assessing the accuracy of National Land-Cover Dataset area estimates at multiple spatial extents. *Photogrammetric Engineering & Remote Sensing*, *70*, 405–414.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., et al. (2007). Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, *73*, 337–341.

Jennings, D. B., Jarnagin, S. T., & Ebert, D. W. (2004). A modeling approach for estimating watershed impervious surface area from National Land Cover Data 92. *Photogrammetric Engineering & Remote Sensing*, *70*, 1295–1307.

Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering & Remote Sensing*, *72*, 823–833.

Jones, K. B., Neale, A. C., Nash, M. S., Van Remortel, D. V., Wickham, J. D., Riitters, K. H., et al. (2001). Predicting nutrient and sediment loadings to streams from landscape metrics: A multiple watershed study from the United States Mid-Atlantic region. *Landscape Ecology*, *16*, 301–312.

Jones, K. B., Neale, A. C., Wade, T. G., Wickham, J. D., Cross, C. L., Edmonds, C. M., et al. (2001). The consequences of landscape change on ecological resources: an assessment of the United States Mid-Atlantic Region, 1973–1993. *Ecosystem Health*, *7*, 229–242.

Kuemmerle, T., Radeloff, V. C., Perzanowski, K., & Hostert, P. (2006). Cross-border comparison of land cover and landscape pattern in Eastern Europe using a hybrid classification technique. *Remote Sensing of Environment*, *103*, 449–464.

Latifovic, R., & Olthof, I. (2004). Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data. *Remote Sensing of Environment*, *90*, 153–165.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole Publishing Co.

Magnussen, S., Stehman, S. V., Corona, P., & Wulder, M. A. (2004). A Polya-urn resampling scheme for estimating precision and confidence intervals under one-stage cluster sampling: Application to map classification accuracy and cover-type frequencies. *Forest Science*, *50*, 810–822.

Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., et al. (2006). Validation of the Global Land Cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, *44*, 1728–1739.

Miettinen, J., Wong, C. M., & Liew, S. C. (2008). New 500 m spatial resolution land cover map of the western insular Southeast Asia region. *International Journal of Remote Sensing*, *29*, 6075–6081.

Mücher, C. A., Steinnocher, K. T., Kressler, F. P., & Heunks, C. (2000). Land cover characterization and change detection for environmental monitoring of pan-Europe. *International Journal of Remote Sensing*, *21*, 1159–1181.

Pontius, R. G., Jr. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, *66*, 1011–1016.

Pontius, R. G., Jr., & Cheuk, M. L. (2006). A generalized cross-tabulation matrix to compare soft-classified maps at multiple spatial resolutions. *International Journal of Geographical Information Science*, *20*, 1–30.

Pontius, R. G., Jr., Thontteh, O., & Chen, H. (2008). Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics*, *15*, 111–142.

Powell, R. L., Matzke, N., de Souza, C., Jr., Clark, M., Numata, I., Hess, L. L., et al. (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sensing of Environment*, *90*, 221–234.

Reese, H. M., Lillesand, T. M., Nagel, D. E., Stewart, J. S., Goldmann, R. A., Simmons, T. E., et al. (2002). Statewide land cover derived from multiseasonal Landsat TM data: A retrospective of the WISCLAND project. *Remote Sensing of Environment*, *82*, 224–237.

Riitters, K. H., Wickham, J. D., & Coulston, J. W. (2004). A preliminary assessment of Montreal process indicators of forest fragmentation for the United States. *Environmental Monitoring and Assessment*, *91*, 257–276.

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

Scepan, J. (1999). Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, *65*, 1051–1060.

Stanfield, B. J., Bliss, J. C., & Spies, T. A. (2002). Land ownership and landscape structure: A spatial analysis of sixty-six Oregon (USA) Coast Range watersheds. *Landscape Ecology*, *17*, 685−697.

Stehman, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, *72*, 35−45.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, *64*, 331−344.

Stehman, S. V., Wickham, J. D., Smith, J. H., & Yang, L. (2003). Thematic accuracy of the 1992 National Land-Cover Data (NLCD) for the Eastern United States: Statistical methodology and regional results. *Remote Sensing of Environment*, *86*, 500−516.

Stehman, S. V., Wickham, J. D., Wade, T. G., & Smith, J. H. (2008). Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, *74*, 1561−1571.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, EUR 22156 EN – DG.* Luxembourg: Office for Official Publications of the European Communities.

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., & Van Driel, N. (2001). Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. *Photogrammetric Engineering & Remote Sensing*, *67*, 650−662.

Wickham, J. D., ONeill, R. V., Riitters, K. H., Smith, E. R., Wade, T. G., & Jones, K. B. (2002). Geographic targeting of increases in nutrient export due to future urbanization. *Ecological Applications*, *12*, 93−106.

Wickham, J. D., Stehman, S. V., Smith, J. H., & Yang, L. (2004). Thematic accuracy of the 1992 National Land-cover Data for the western United States. *Remote Sensing of Environment*, *91*, 452−468.

Wickham, J. D., Stehman, S. V., Smith, J. H., Wade, T. G., & Yang, L. (2004). A priori evaluation of two-stage cluster sampling for accuracy assessment of large-area land-cover maps. *International Journal of Remote Sensing*, *25*, 1235−1252.

Willmott, C. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, *63*, 1309−1313.

Wimberly, M. C., & Ohmann, J. L. (2004). A multi-scale assessment of human and environmental constraints on forest land cover change on the Oregon (USA) coast range. *Landscape Ecology*, *19*, 631−646.

Wolter, K. M. (1985). *Variance Estimation.* New York: Springer-Verlag.

Wu, W., Shibasaki, R., Yang, P., Ongaro, L., Zhou, Q., & Tang, H. (2008). Validation and comparison of 1 km global land cover products in China. *International Journal of Remote Sensing*, *29*, 3769−3785.

Wulder, M. A., Franklin, S. E., White, J. C., Linke, J., & Magnussen, S. (2006). An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. *International Journal of Remote Sensing*, *27*, 663−683.