ISPRS Journal of Photogrammetry and Remote Sensing 142 (2018) 292–300

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions

Ronald E. McRoberts[a,*], Stephen V. Stehman[b], Greg C. Liknes[a], Erik Næsset[c], Christophe Sannier[d], Brian F. Walters[a]

[a] Northern Research Station, U.S. Forest Service, Saint Paul, MN, USA
[b] Forest and Natural Resources Management, State University of New York, Syracuse, NY, USA
[c] Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway
[d] Systèmes d'Information à Référence Spatiale (SIRS), Villeneuve d'Ascq, France

ARTICLE INFO

Keywords:
Intepreter error
Bias
Precision
Greenhouse gas inventory
Gain-loss method

ABSTRACT

The gain-loss approach for greenhouse gas inventories requires estimates of areas of human activity and estimates of emissions per unit area for each activity. Stratified sampling and estimation have emerged as a popular and useful statistical approach for estimation of activity areas. With this approach, a map depicting classes of activity is used to stratify the area of interest. For each map class used as a stratum, map units are randomly selected and assessed with respect to an attribute such as forest/non-forest or forest land cover change. Ground observations are generally accepted as the most accurate source of information for these assessments but may be cost-prohibitive to acquire for remote and inaccessible forest regions. In lieu of ground observations, visual interpretations of remotely sensed data such as aerial imagery or satellite imagery are often used with the caveat that the interpretations must be of greater quality than the map data. An unresolved issue pertains to the effects of interpreter error on the bias and precision of the stratified estimators of activity areas.

For a 7500-km^2 study area in north central Minnesota in the United States of America, combinations of forest inventory plot observations, visual interpretations of aerial imagery, and two forest/non-forest maps were used to assess the effects of interpreter error on stratified estimators of proportion forest and corresponding standard errors. The primary objectives related to estimating the bias and precision of the stratified estimators in the presence of interpreter errors, identifying factors and the levels of those factors that affect bias and precision, and facilitating planning to circumvent and/or mitigate the effects of bias. The primary results were that interpreter error induces bias into the stratified estimators of both land cover class proportion and its standard error. Bias increased with greater inequality in stratum weights, smaller map and interpreter accuracies, fewer interpreters and greater correlations among interpreters. Failure to account for interpreter error produced stratified standard errors that under-estimated actual standard errors by factors as great as 2.3. Greater number of interpreters mitigated the effects of interpreter error on proportion forest estimates, and a hybrid variance estimator accounted for the effects on standard errors.

1. Introduction

Two approaches to greenhouse gas emissions accounting are common, the *stock-change* approach and the *gain-loss* approach (IPCC, 2006, Volume 4, Chapter 2, p. 2.10; GFOI, 2016, p. 22). With the stock-change approach, mean annual emissions are estimated as the mean annual difference in carbon stocks between two points in time (IPCC, 2006, Volume 4, Chapter 4, Section 4.2.1.1; GFOI, 2016, Chapter 3). For countries with established forest ground sampling programs such as national forest inventories, the stock-change approach is fairly easy to implement. However, for countries with remote and inaccessible forests, the stock-change approach may be prohibitively expensive. For these countries, the gain-loss approach may be a more feasible alternative. With this approach, emissions are defined to be the net balance of additions to and removals from a carbon pool and are estimated as the product of the areas of "human activity causing emissions", characterized as *activity data*, and the responses of carbon stocks for those activities, characterized as *emission factors* (IPCC, 2006, Volume 1, Chapter 1, Section 1.2; GFOI, 2016, pp. xvii, 22)

Estimation of areas of activities often relies on remote sensing-based

* Corresponding author.
E-mail address: rmcroberts@fs.fed.us (R.E. McRoberts).

https://doi.org/10.1016/j.isprsjprs.2018.06.002
Received 23 February 2018; Received in revised form 27 May 2018; Accepted 2 June 2018
0924-2716/ Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS).

land cover or land cover change maps (Olofsson et al., 2013, 2014; Ban et al., 2015). Of importance, however, estimating areas of these activities by simply adding the areas of map units assigned to activity classes, a practice characterized as *pixel counting*, is a biased procedure because it does not account for map classification errors. Stratified sampling and estimation is a statistically rigorous alternative. With stratified sampling, map classes are used as strata, and within-stratum samples are selected using simple random or systematic sampling designs. Sample unit observations of land cover or land cover change are then used as reference data, and the stratified estimators are used to estimate the areas of activity classes of interest (Olofsson et al., 2013, 2014). In the absence of reference data error, the stratified estimators are unbiased and more precise than simple random sampling estimators (Chen and Wei, 2009).

Reference data in the form of ground observations are often considered optimal, although Foody (2009, 2010) notes that even ground reference data are subject to error. However, regardless of error, acquisition of ground reference data for remote and inaccessible regions may be prohibitively expensive, if not logistically infeasible. For these situations, reference data in the form of visual interpretations of remotely sensed data are often used, albeit with the stipulation that such reference data are of greater quality than the map data with respect to factors such as resolution and accuracy (Mannel et al., 2006; Stehman, 2009; Olofsson et al., 2013; Pengra et al., 2015; Tsendbazar et al. 2015; Boschetti et al., 2016; GFOI, 2016, pp. 125, 139). However, even if visual image interpretations are of greater quality than the map data, they cannot be assumed to be without error. For five trained interpreters of stereo aerial photography, Næsset (1991) reported that interpretations of crown coverage for structurally homogenous Norwegian boreal forests differed substantially among interpreters and among different times of year for the same interpreter. For the same forest conditions, Næsset (1992) reported that interpretations of broad tree species groups by 12 professional, trained interpreters using stereo aerial photography produced only 31–79% agreement with field reference data. For five trained interpreters of videography, Powell et al. (2004) reported interpreter disagreement of almost 30% for five land cover classes in the Brazilian Amazon, two of which were forest-related classes. Thompson et al. (2007) reported errors of 30% when aerial imagery was used to classify boreal forest stands into coniferous, deciduous, and mixed classes in Ontario, Canada. For three trained interpreters, Sun et al. (2017) reported that despite among-interpreter consistency, manual interpretations of Google Earth and other fine resolution imagery were not as reliable as ground measurements for seven land cover classes in Central Asia. In summary, reference data in the form of visual interpretations of remotely sensed data, even by well-trained professional interpreters, are subject to substantial interpreter disagreement and error.

If the reference data are *imperfect* in the sense of being subject to error, then the stratified estimators may be biased, sometimes substantially biased despite only small errors (Foody, 2009, 2010, 2013). Although the effects of imperfect reference data on estimators of class proportions and areas have been at least partially addressed, little has been reported on the effects of imperfect reference data on variance estimators. Compliance with the IPCC good practice guidance for greenhouse gas inventories requires not only avoiding over- and/or under-estimates but also reduction of uncertainties (IPCC, 2006, Volume 1, Chapter 1, Section 1.2; GFOI 2016, p. 15) with the obvious caveat that uncertainties cannot be reduced unless they are first correctly estimated. In particular, correct estimation of uncertainty requires incorporation of the effects of imperfect reference data into variance estimators (Olofsson et al., 2014).

The objectives of the study were fivefold: (1) to assess the effects of imperfect reference data on the bias and precision of stratified estimators of land cover class proportions; (2) to characterize conditions that affect the magnitudes of bias and precision; (3) to develop a variance estimator that incorporates the effects of interpreter error; (4) to illustrate the effects of interpreter error on bias and precision with inventory ground data and visual interpretations of aerial imagery using two forest/non-forest maps; and (5) to facilitate planning for estimation of activity data. Because the ultimate objective is an estimate of the area of a land cover class, and area can be expressed as the product of the class proportion and the total population area which is usually known, the focus of the study was estimation of the class proportion.

2. Data

2.1. Study area

The study area was the 7583 km^2 of Itasca County in north central Minnesota in the United States of America (USA) (Fig. 1). Land cover includes water, wetlands and approximately 80% forest consisting of mixtures of pines (*Pinus* spp.), spruce (*Picea* spp.), and balsam fir (*Abies balsamea* (L.) Mill.) on upland sites and spruce (*Picea* spp.), tamarack (*Larix laricina* (Du Roi) K. Koch), white cedar (*Thuja occidentalis* (L.)), and black ash (*Fraxinus nigra* Marsh.) on lowland sites. Forest stands in the study area are typically naturally regenerated, uneven-aged, and mixed species.

2.2. Forest inventory data

Data were obtained for 310 ground plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the NFI of the USA. The plots were established in permanent field locations using a quasi-systematic sampling design that is regarded as producing an equal probability sample (McRoberts et al., 2010; Mountrakis and Xi, 2013) and were measured between 2014 and 2016. Field crews visually estimate the proportion of each plot that satisfies the FIA definition of forest land: (i) minimum area 0.4 ha (1.0 ac), (ii) minimum tree cover of 10%, (iii) minimum width of 36.58 m (120 ft), and (iv) forest land use. All field crews are well-trained, tested on their ability to assess plot variables, and hence well-qualified to distinguish forest from non-forest based on the FIA definition. A small number of plots in three categories were deleted and considered to be missing at random (Rubin, 1987): (i) plots with mixtures of forest and non-forest cover, (ii) plots with forest use but with no tree cover due to conditions such as recent harvest, and (iii) to the degree possible, plots with non-forest use but with tree cover of which parks and rural residential areas are examples. Plot centers were estimated using global positioning system receivers with sub-meter accuracy. The field crew, plot-level, forest/non-forest observations were used as reference data to produce estimates of proportion forest that served as the standard for comparison for estimates based on visual interpretations of aerial imagery. They also served as the basis for assessing map and interpreter accuracies.

2.3. Percent tree canopy cover datasets

The *Global Forest Change* (GFC) dataset is based on cloud-free, composite, annual growing season Landsat 7 Enhanced Thematic Mapper Plus data (Hansen et al., 2013). For 30-m × 30-m pixels, the dataset includes predictions of maximum percent tree canopy cover in the range of 0–100% for vegetation taller than 5 m for the year 2010. The 2011 *National Land Cover Database* (NLCD) includes percent tree canopy cover values in the range of 0–100% for 30-m × 30-m pixels (Homer et al, 2015). Each of the two datasets was used to construct a forest/non-forest map (Section 3.2) which then facilitated stratified sampling and estimation (Section 3.3).

2.4. Aerial imagery

Aerial imagery was obtained from the Farm Service Agency of the U.S. Department of Agriculture through the National Agriculture

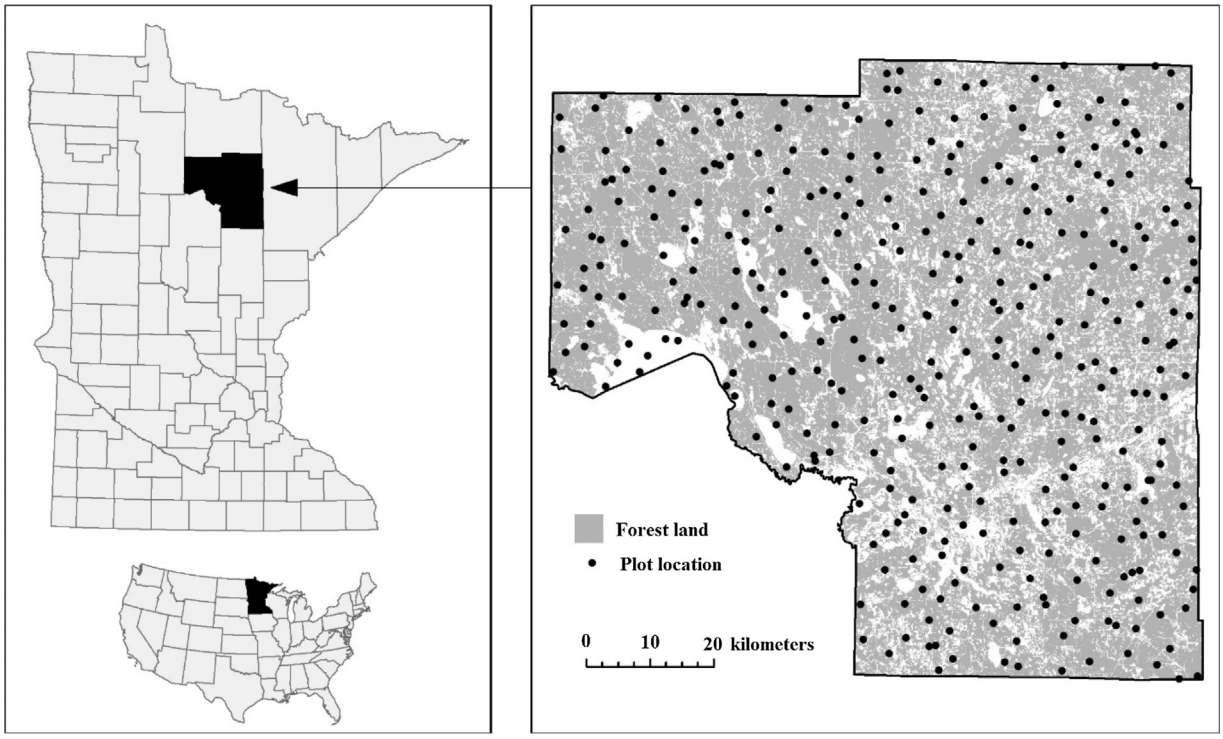

Fig. 1. Study area in Minnesota, USA.

Imagery Program which acquires imagery across the contiguous USA on a 3-year cycle during summer growing seasons. The imagery for this study was acquired in 2010 at 1.0-m resolution in four bands: natural red, natural blue, natural green, and near infrared. Visual interpretations of the imagery were used to produce reference data, albeit possibly imperfect reference data, for estimation of proportion forest area.

3. Methods

3.1. Overview

The GFC and NLCD tree canopy cover datasets were used to construct forest/non-forest maps whose classes served as strata for forest/non-forest stratifications (Section 3.2). Stratified sampling and stratified estimators were used to estimate proportion forest using both inventory plot observations and interpreted aerial imagery as reference data (Section 3.3). Hybrid estimators that incorporated the effects of both sampling variability and interpreter error were used to estimate the standard errors of proportion forest estimates (Section 3.4). Using these basic statistical methods, the effects of interpreter error on the bias and precision of the stratified estimators were assessed using simulated data, the field crew plot-level forest/non-forest observations, and visual interpretations of aerial imagery (Section 3.5).

3.2. Forest/non-forest map construction

Forest/non-forest maps were constructed from the GFC and NLCD datasets by using an optimal threshold value to convert percent tree canopy cover to forest/non-forest. Optimal thresholds were selected by first associating the field crew forest/non-forest observation for each plot with the percent tree canopy cover for each GFC or NLCD map pixel containing the plot center. For these analyses, the plot observations were independent of the data used to construct the maps. For percent tree canopy cover thresholds ranging from 0.01 to 0.99, each FIA plot was assigned to the non-forest class if the corresponding percent canopy cover was less than the threshold and to the forest class if the percent was greater than or equal to the threshold. For each threshold, accuracy was calculated as the proportion of plots for which the plot forest/non-forest class assignments based on percent tree canopy cover were the same as the plot field crew forest/non-forest observations. The threshold with the greatest accuracy was selected as optimal and used to convert percent canopy cover to forest/non-forest. For the GFC data, Sannier et al. (2016, Section 3.1) selected a threshold of 70% for a study area in Gabon; Næsset et al. (2016, Table 5) selected a threshold of 40% for a study area in Tanzania; and McRoberts et al. (2016a, Tables 2) selected a threshold of 95% for a study area in Santa Catarina, Brazil. For the GFC data, the optimal percent canopy cover threshold of 52% for this study produced overall forest/non-forest classification accuracy of 0.88, and for the NLCD data, the optimal threshold of 27% produced overall classification accuracy of 0.81. Each pixel in each dataset was classified as non-forest if the percent tree canopy cover was less than the respective optimal threshold; otherwise, the pixel was classified as forest. For the GFC map, the forest and non-forest map class proportions were 0.733 and 0.267, respectively, and for the NLCD map, the proportions were 0.580 and 0.420, respectively. These map class proportions served as stratum weights for stratified estimation (Section 3.3).

3.3. Probability-based stratified estimators

Estimation of the population proportion, μ, for a particular class among multiple land cover classes can be reduced to a two-class problem characterized by the class of interest, herein designated class 1, and the aggregation of all remaining classes, herein designated class 2. For a sample of reference data, a two-class confusion matrix characterizing the distribution of the sample units with respect to the map and reference classes is the common basis for estimation of proportions (Foody, 2009). When the sampling intensities differ by strata, stratified estimators of the proportion, μ, of the population in class 1 must be used to accommodate the different intensities. The stratified estimators provided by Cochran (1977) can be readily used for this purpose,

Table 1
Confusion matrix and estimators for class 1 proportion.

Map class	Stratum weight* w_h	Reference class		Total	$\hat{p}_h$	$\hat{Var}(\hat{p}_h)$
		1	2			
1	w_1	n_{11}	n_{12}	$n_1 = n_{11} + n_{12}$	$\hat{p}_1 = \frac{n_{11}}{n_1}$	$\hat{Var}(\hat{p}_1) = \frac{\hat{p}_1 \cdot (1-\hat{p}_1)}{n_1 - 1}$
2	w_2	n_{21}	n_{22}	$n_2 = n_{21} + n_{22}$	$\hat{p}_2 = \frac{n_{21}}{n_2}$	$\hat{Var}(\hat{p}_2) = \frac{\hat{p}_2 \cdot (1-\hat{p}_2)}{n_2 - 1}$
Stratified estimators					$\hat{\mu} = \sum_{h=1}^{2} w_h \cdot \hat{p}_h$	$\hat{Var}(\hat{\mu}) = \sum_{h=1}^{2} w_h^2 \cdot \hat{Var}(\hat{p}_h)$

* $w_1 = w$, $w_2 = 1-w$, where w is defined following Eq. (1b).

$$\hat{\mu}_{Str} = \sum_{h=1}^{2} w_h \cdot \hat{p}_h \tag{1a}$$

and

$$\begin{aligned} \hat{Var}_{Str}(\hat{\mu}_{Str}) &= \sum_{h=1}^{2} w_h^2 \cdot \hat{Var}(\hat{p}_h) \\ &= \sum_{h=1}^{2} w_h^2 \cdot \frac{\hat{p}_h \cdot (1-\hat{p}_h)}{n_h - 1} \end{aligned} \tag{1b}$$

where the subscript, h, denotes the map class used as a stratum; w_h is the stratum weight calculated as the proportion of map units in the h^{th} map class (Section 3.2); n_h is the sample size for the h^{th} stratum; and $\hat{p}_h$ is the proportion of sample units in the h^{th} stratum in reference class 1 (Table 1).

Although interpreter errors are known to induce bias into estimators of class proportions (Foody 2009, 2010, 2013), the degree of bias relative to factors such as map accuracy and number of interpreters is mostly unknown. If the conditions were known, then advance planning to mitigate the biasing factors may be possible. A slight reformulation of Bross (1954) and Zimmerman and Liknes (2010) shows that for a single interpreter,

$$E(\hat{\mu}) = \mu \cdot r + (1-\mu) \cdot (1-r), \tag{2a}$$

where r is the map accuracy relative to reference data without error for both classes. Subtracting μ from both sides yields,

$$\begin{aligned} Bias(\hat{\mu}) &= E(\hat{\mu}) - \mu \\ &= (2 \cdot \mu - 1) \cdot (1-r) \end{aligned} \tag{2b}$$

Although informative, the estimators of Eqs. (2a) and (2b) require knowledge of both μ and r, neither of which is known in advance. Further, because $\hat{\mu}$ is a biased estimator of μ, substitution of $\hat{\mu}$ into Eq. (2b) produces a biased estimator of the bias. Thus, this formulation provides little that can be used for advance planning purposes, such as selecting the number of interpreters as a means of minimizing bias.

Alternatively, in the absence of interpreter errors, the expected value of the stratified estimator of proportion forest can also be expressed as,

$$\begin{aligned} E(\hat{\mu}_{Str}) &= w_1 \cdot q_1 + w_2 \cdot (1-q_2) \\ &= \mu \end{aligned} \tag{3}$$

where w_h is the stratum weight and q_h is the map accuracy for the h^{th} stratum. For any particular map, w_1 and $w_2 = 1-w_1$ will be known in advance, and for widely available maps such as the GFC and NLCD maps, at least rudimentary information on map accuracy will likely be available. If the effects on bias and variances of w_1, w_2, q_1, and q_2 are augmented with the effects of interpreter numbers and accuracies, then informed advance planning may be facilitated. Thus the study focused on estimating the latter effects.

3.4. Hybrid inference

Probability-based (design-based) inference assumes a probability sampling design and one and only one possible value with at most negligible uncertainty for each population unit. When only predictions with uncertainty, rather than observations assumed to be without uncertainty, are available for the sample units, two sources of uncertainty must be accommodated when estimating variances, the uncertainty associated with sampling variability and the uncertainty resulting from using sample unit predictions rather than observations. The term *hybrid inference* is used to characterize methods that combine probability-based inferential methods to incorporate the effects of sampling variability and model-based inferential methods to incorporate the effects of uncertainty associated with the sample unit predictions (Fattorini, 2012; Corona et al., 2014; McRoberts et al., 2016b).

For this study, hybrid variance estimation entailed use of a Monte Carlo procedure to incorporate the effects of interpreter error into the variance estimator. The procedure is initiated with values selected for five factors: (i) stratum weights, (ii) within-stratum map accuracies relative to ground classes, (iii) within-stratum interpreter accuracies relative to ground classes, (iv) numbers of interpreters, and (v) pairwise between-interpreter correlations. Operationally, interpreters visually interpret the aerial imagery independently of other interpreters. However, because the same underlying imagery is very likely used by all interpreters, errors arising from differences between plot observation dates and imagery dates are likely to be common to all interpreters. In addition, sample units for which the imagery is easy or difficult to classify for one interpreter are also likely to be easy or difficult to classify for other interpreters. Thus, positive correlations among interpretations by different interpreters should be expected.

For a value for each of the five factors, and a stratified random sample, a Monte Carlo procedure was used to estimate variances that incorporate both sampling variability and interpreter errors. For these analyses, all map and interpreter accuracies were relative to ground classes that are assumed to be without error, whereas reference data may or may not be without error. The Monte Carlo procedure included six steps:

(1) Ground class: for each sample unit in each stratum, a random number was drawn from a uniform [0,1] distribution; if the number was smaller than or equal to the within-stratum map accuracy, the ground class was the map stratum from which the sample unit was drawn; otherwise, the ground class was the complement of the map stratum; although the term *ground truth* is not used for this study, the ground class serves as what is characterized by some other studies as ground truth;
(2) Forest/non-forest interpretation: for each sample unit within each stratum and for each interpreter, a random number was drawn from a multivariate, uniform [0,1] distribution with possibly non-zero, pairwise, between-interpreter correlations (Demirtas, 2004); if the number was smaller than or equal to the within-stratum interpreter accuracy, the interpretation was the same as the ground class;

otherwise, the interpretation was the complement of the ground class;

(3) Reference class: for each sample unit within each stratum, the majority forest/non-forest interpretation among interpreters was selected as the reference class with the provision that for cases of equal numbers of interpretations per class, the reference class was the map stratum; of importance, because of interpreter error, the reference class is not necessarily the same as the ground class;

(4) Stratified estimation: following selection of a reference class for each sample unit, the probability-based, stratified estimators of Eqs. (1a) and (1b) were used to estimate proportion forest and the standard error of the estimate;

(5) Replication: steps (1)–(4) were replicated n_{rep} times;

(6) Variance estimation: the hybrid estimates of proportion forest and their standard errors were calculated as (Rubin, 1987, pp.76–77),

$$\hat{\mu}_{Hyb} = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \hat{\mu}_{Str}(k) \tag{4a}$$

and

$$\hat{Var}_{Hyb}(\hat{\mu}_{Hyb}) = \left(1 + \frac{1}{n_{rep}}\right) \cdot \hat{V}_1 + \hat{V}_2 \tag{4b}$$

where k indexes replications,

$$\hat{V}_1 = \frac{1}{n_{rep}-1} \sum_{k=1}^{n_{rep}} [\hat{\mu}_{Str}(k) - \hat{\mu}_{Hyb}]^2 \tag{4c}$$

is the among-replications variance,

$$\hat{V}_2 = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \hat{Var}_{Str}[\hat{\mu}_{Str}(k)] \tag{4d}$$

is the mean per replication variance, and $\hat{Var}_{Str}[\hat{\mu}_{Str}(k)]$ is calculated using Eq. (1b) regardless of whether the reference data were with or without error.

In Step (5), replication continued until hybrid estimates of both proportion forest and variance stabilized.

3.5. Analyses

3.5.1. Overview of analyses

Three approaches were used to estimate proportion forest and corresponding standard errors. First, the effects of interpreter error on bias and precision were evaluated using a simulation approach applied to a broad range of combinations of the five previously noted factors: forest/non-forest stratum sizes, map accuracy, interpreter accuracy, pairwise interpreter correlation, and number of interpreters (Section 3.5.2). Second, for each of the GFC and NLCD maps, a stratified sample was drawn from the Itasca County study area, and for each sample unit the field crew forest/non-forest observation was used as the reference class (Section 3.5.4). Third, for each map, the majority visual interpretations for the same sample of field plot locations were used as reference data. For all three approaches, the stratified estimators (Section 3.3) were used to estimate proportion forest, and both the stratified and hybrid estimators were used to calculate the standard errors of the proportion forest estimates (Section 3.4).

3.5.2. Simulations

Simulations were conducted to assess the effects of interpreter error on stratified estimates of mean proportion forest and the standard errors of the estimates. For each of the forest and non-forest strata, 75 sample units were used. Separate simulations were conducted for each combination of the following values for the five factors noted in Section 3.2.2: (i) forest stratum weights of 0.25, 0.50, 0.75, and 0.90; (ii) common within-stratum map accuracies of 0.75 and 0.90; (iii) common within-stratum interpreter accuracies of 0.75 and 0.90; (iv) 1, 3, 5, and 7 interpreters; and (v) common pairwise between-interpreter correlations of 0.00, 0.50, and 0.90. Although the data reported in Section 2 were used to select ranges of values for the five factors, the simulations were independent of those data. For the simulation analyses, all map and interpreter accuracies were relative to ground classes assumed to be without error.

3.5.3. Forest/non-forest plot observations

For each of the GFC and NLCD forest/non-forest maps, 75 plots classified by the map as forest and 75 plots classified by the map as non-forest were randomly and independently selected from among the 303 plots that remained after deletions (Section 2.2). Because the FIA plots represented an equal probability sample from the population, stratified random samples drawn from these 303 plots were considered stratified random samples from the entire population. For the GFC map, the forest and non-forest stratum weights were 0.733 and 0.267 (Section 3.2), and the within-stratum forest/non-forest map accuracies were 0.933 and 0.853. For the NLCD map, the forest and non-forest stratum weights were 0.580 and 0.420 (Section 3.2), and the within-stratum forest/non-forest map accuracies were 0.920 and 0.680. For each map, the stratified estimators described in Eqs. (1a) and (1b) and Table 1 were used to estimate proportion forest and the standard error of the estimate using the FIA plot observations as reference data. Because the reference data were field crew forest/non-forest observations assumed to be without error, hybrid estimators were not used for these analyses. All map and interpreter accuracies were relative to ground classes in the form of field crew observations assumed to be without error.

3.5.4. Visual interpretations

For each of the GFC and NLCD maps, stratified estimates of proportion forest were calculated using reference data based on the visual interpretations. For these analyses, the same stratified random samples, the same stratum weights, and the same within-stratum map accuracies as reported in Section 3.5.3 were used. For each sample unit, each of three well-trained and experienced interpreters, independently of each other and independently of field crew assessments, visually interpreted the aerial imagery described in Section 2.4 and selected a forest or non-forest class based on the FIA definition of forest land (Section 2.2). For each sample unit, the majority interpretation among the three interpreters was then used as the reference class, and the stratified estimators described in Eqs. (1a) and (1b) and in Table 1 were used to estimate proportion forest and the standard error of the estimate. In addition, the hybrid estimator (Section 3.4) was used to estimate the standard error of the estimate of proportion forest for each map.

In Step (1) of the Monte Carlo procedure described in Section 3.4, map and interpreter accuracies were calculated using inventory plot observations as ground classes. However, such plot data are often not available; otherwise, they would be used as reference data rather than visual interpretations. Without such ground data the effects of interpreter error cannot be readily assessed or incorporated into the hybrid variance estimator. To compensate for the lack of such data, the majority interpretations determined from Step (3) of the Monte Carlo procedure were used as substitutes for the inventory plot observations when calculating map and interpreter accuracies, and the hybrid variance estimates were calculated. To facilitate these analyses, within-stratum map and interpreter accuracies were calculated relative to majority interpretations rather than inventory plot observations. Of importance, these are the only analyses for which map and interpreter accuracies are not relative to ground classes assumed to be without error.

4. Results and discussion

4.1. Simulations

A primary simulation result was that interpreter error induced bias

Table 2a
Simulation results for pairwise between-interpreter correlation of 0.00.

Forest stratum weight (w_1)	Map accuracy	Expected proportion forest in the absence of interpreter error	Interpreter accuracy	Means of proportion forest estimates and standard errors (number of interpreters)											
				1			3			5			7		
				Mean	Standard error		Mean	Standard error		Mean	Standard error		Mean	Standard error	
					Str	Hyb		Str	Hyb		Str	Hyb		Str	Hyb
0.25	0.75	0.375	0.75	0.438	0.038	0.054	0.414	0.037	0.053	0.400	0.036	0.051	0.393	0.036	0.050
			0.90	0.400	0.036	0.052	0.382	0.035	0.049	0.377	0.034	0.049	0.376	0.034	0.049
	0.90	0.300	0.75	0.400	0.036	0.051	0.361	0.033	0.047	0.341	0.030	0.043	0.328	0.029	0.040
			0.90	0.339	0.030	0.043	0.311	0.026	0.036	0.303	0.024	0.035	0.301	0.024	0.034
0.50	0.75	0.500	0.75	0.500	0.034	0.048	0.500	0.033	0.047	0.500	0.032	0.046	0.500	0.032	0.045
			0.90	0.499	0.032	0.047	0.500	0.031	0.044	0.500	0.031	0.043	0.500	0.031	0.043
	0.90	0.500	0.75	0.499	0.032	0.046	0.499	0.030	0.042	0.500	0.027	0.039	0.501	0.026	0.036
			0.90	0.500	0.027	0.038	0.500	0.023	0.033	0.500	0.022	0.031	0.500	0.021	0.030
0.75	0.75	0.625	0.75	0.563	0.038	0.054	0.585	0.037	0.053	0.599	0.036	0.051	0.608	0.036	0.050
			0.90	0.599	0.036	0.051	0.618	0.035	0.050	0.622	0.034	0.049	0.624	0.034	0.048
	0.90	0.700	0.75	0.600	0.036	0.052	0.636	0.033	0.047	0.659	0.030	0.044	0.671	0.029	0.041
			0.90	0.660	0.030	0.043	0.689	0.026	0.037	0.697	0.024	0.034	0.700	0.024	0.034
0.90	0.75	0.700	0.75	0.600	0.044	0.062	0.638	0.042	0.060	0.658	0.042	0.059	0.671	0.041	0.058
			0.90	0.661	0.041	0.059	0.689	0.040	0.057	0.697	0.039	0.056	0.699	0.039	0.055
	0.90	0.820	0.75	0.660	0.041	0.059	0.720	0.038	0.054	0.755	0.035	0.049	0.775	0.033	0.046
			0.90	0.755	0.035	0.050	0.803	0.029	0.042	0.815	0.028	0.039	0.819	0.027	0.039

Table 2b
Simulation results for pairwise between-interpreter correlation of 0.75.

Forest stratum weight (w_1)	Map accuracy	Expected proportion forest in the absence of interpreter error	Interpreter accuracy	Means of proportion forest estimates and standard errors (number of interpreters)											
				1			3			5			7		
				Mean	Standard error		Mean	Standard error		Mean	Standard error		Mean	Standard error	
					Str	Hyb		Str	Hyb		Str	Hyb		Str	Hyb
0.25	0.75	0.375	0.75	0.437	0.038	0.054	0.433	0.038	0.054	0.432	0.038	0.054	0.432	0.038	0.054
			0.90	0.401	0.036	0.051	0.395	0.036	0.050	0.394	0.036	0.051	0.394	0.036	0.050
	0.90	0.300	0.75	0.400	0.036	0.051	0.394	0.036	0.050	0.392	0.036	0.051	0.390	0.035	0.050
			0.90	0.341	0.030	0.043	0.334	0.029	0.042	0.332	0.029	0.041	0.330	0.029	0.041
0.50	0.75	0.500	0.75	0.500	0.034	0.048	0.501	0.034	0.048	0.501	0.034	0.048	0.500	0.034	0.048
			0.90	0.500	0.032	0.046	0.501	0.032	0.045	0.501	0.032	0.045	0.499	0.032	0.045
	0.90	0.500	0.75	0.500	0.032	0.046	0.501	0.032	0.045	0.500	0.032	0.045	0.500	0.032	0.045
			0.90	0.500	0.027	0.038	0.500	0.026	0.037	0.501	0.026	0.037	0.501	0.026	0.037
0.75	0.75	0.625	0.75	0.563	0.038	0.055	0.567	0.038	0.053	0.567	0.038	0.054	0.569	0.038	0.054
			0.90	0.601	0.036	0.051	0.604	0.036	0.051	0.607	0.036	0.050	0.606	0.036	0.051
	0.90	0.700	0.75	0.601	0.036	0.051	0.606	0.036	0.051	0.608	0.036	0.050	0.610	0.035	0.050
			0.90	0.661	0.030	0.043	0.666	0.029	0.042	0.669	0.029	0.041	0.669	0.029	0.041
0.90	0.75	0.700	0.75	0.601	0.044	0.062	0.607	0.044	0.061	0.609	0.044	0.062	0.610	0.044	0.063
			0.90	0.661	0.041	0.058	0.668	0.041	0.058	0.668	0.041	0.059	0.670	0.041	0.058
	0.90	0.820	0.75	0.661	0.041	0.059	0.670	0.041	0.058	0.673	0.041	0.058	0.676	0.041	0.057
			0.90	0.757	0.035	0.049	0.766	0.034	0.048	0.770	0.033	0.047	0.772	0.033	0.047

into the stratified estimator of proportion forest. For these simulation analyses, bias was assessed by comparing the expected forest proportions from Eq. (3) and the simulation means, both as reported in Tables 2a,b,c. For common within-stratum map and interpreter accuracies and common between-interpreter correlations based on ground classes without error, the bias increased as map and interpreter accuracies decreased, as the numbers of interpreters decreased, and as the between-interpreter correlations increased. In addition, estimator bias increased as the forest stratum weight deviated more from $w = 0.5$ with positive bias for $w > 0.5$, essentially no bias for $w \approx 0.5$, and negative bias for $w < 0.5$. The explanation for the effects of unequal stratum weights is that while similarly distributed within-stratum interpreter errors had similar effects on within-stratum estimates, greater differences in stratum weights weighted these similarly within-stratum estimates more unequally. Multiple studies reported by Foody (2009, 2010, 2013) support these findings. These simulation results should be interpreted in a general sense, because they do not consider different map or interpreter accuracies for different strata, different correlations for different pairs of interpreters or different strata, or correlations between map and interpreter accuracies. Nevertheless, the crucial result was that greater numbers of interpreters were necessary to offset the biasing effects of the other factors.

A second primary result of the simulations was that the stratified estimator of the standard error is also biased in the presence of interpreter error. Ratios of hybrid SEs from Eq. (4b) to means of stratified SEs from Eq. (1b) indicated that the stratified estimator which did not incorporate uncertainty due to interpreter error under-estimated hybrid standard errors by a remarkably consistent factor of approximately 1.4. Comparisons of hybrid standard errors indicated they were smaller for greater map and interpreter accuracies, greater numbers of interpreters,

Table 2c
Simulation results for pairwise between-interpreter correlation of 0.90.

Forest stratum weight (w_1)	Map accuracy	Expected proportion forest in the absence of interpreter error	Interpreter accuracy	Means of proportion forest estimates and standard errors (number of interpreters)											
				1			3			5			7		
				Mean	Standard error		Mean	Standard error		Mean	Standard error		Mean	Standard error	
					Str	Hyb		Str	Hyb		Str	Hyb		Str	Hyb
0.25	0.75	0.375	0.75	0.436	0.038	0.054	0.433	0.038	0.053	0.432	0.038	0.054	0.430	0.038	0.055
			0.90	0.402	0.036	0.051	0.395	0.036	0.050	0.395	0.026	0.051	0.394	0.036	0.051
	0.90	0.300	0.75	0.400	0.036	0.051	0.393	0.036	0.050	0.392	0.036	0.051	0.391	0.036	0.050
			0.90	0.340	0.030	0.043	0.334	0.029	0.041	0.332	0.029	0.041	0.330	0.029	0.041
0.50	0.75	0.500	0.75	0.501	0.034	0.048	0.500	0.034	0.049	0.501	0.034	0.049	0.500	0.034	0.048
			0.90	0.501	0.032	0.045	0.501	0.032	0.046	0.500	0.032	0.045	0.500	0.032	0.045
	0.90	0.500	0.75	0.500	0.032	0.046	0.500	0.032	0.046	0.501	0.032	0.046	0.500	0.032	0.045
			0.90	0.500	0.027	0.038	0.500	0.026	0.037	0.500	0.026	0.037	0.500	0.026	0.037
0.75	0.75	0.625	0.75	0.563	0.038	0.054	0.565	0.038	0.054	0.568	0.038	0.054	0.569	0.028	0.054
			0.90	0.601	0.036	0.052	0.603	0.036	0.050	0.606	0.036	0.050	0.606	0.036	0.050
	0.90	0.700	0.75	0.601	0.036	0.051	0.606	0.036	0.051	0.608	0.036	0.051	0.610	0.035	0.050
			0.90	0.660	0.030	0.043	0.666	0.029	0.042	0.669	0.029	0.041	0.669	0.029	0.041
0.90	0.75	0.700	0.75	0.601	0.044	0.062	0.607	0.044	0.061	0.610	0.044	0.061	0.610	0.043	0.062
			0.90	0.660	0.041	0.059	0.668	0.041	0.059	0.669	0.041	0.059	0.670	0.041	0.058
	0.90	0.820	0.75	0.661	0.041	0.059	0.671	0.041	0.058	0.674	0.041	0.057	0.675	0.041	0.058
			0.90	0.756	0.035	0.049	0.768	0.034	0.048	0.770	0.033	0.048	0.772	0.033	0.047

and smaller pairwise between-interpreter correlations. The crucial result was that unbiased variance estimation requires hybrid inferential methods that incorporate the effects of uncertainty due to both sampling variability and interpreter error.

4.2. Field and interpreter data

4.2.1. Global forest change map

For the GFC map, the forest and non-forest stratum weights were 0.733 and 0.267, and the within-stratum map accuracies relative to the inventory plot observations were 0.933 and 0.853 and relative to the majority interpretations were 0.960 and 0.840. Within-stratum interpreter accuracies relative to the inventory plot observations and the majority interpretations are reported in Table 3. The stratified estimates were $\hat{\mu}_{Str} = 0.723$with $SE_{Str}(\hat{\mu}_{str}) = 0.024$ when using the FIA plot forest/non-forest observations without interpreter error as reference data and $\hat{\mu}_{Str} = 0.738$ with $SE_{Str}(\hat{\mu}_{str}) = 0.016$ when using the majority interpretations with interpreter error as reference data (Table 4). Of importance, the stratified estimator is unbiased when used with reference data without error but biased when using reference data with interpreter error. Although there is no evidence to suggest which of these two stratified standard errors should be smaller or larger, the stratified standard error obtained using the majority interpretations with interpreter error as reference data should be dismissed in favor of the standard error obtained using the hybrid variance estimator when the reference data have error.

As previously (Section 3.5.4), the hybrid standard error from Eq. (4b) was calculated for two scenarios, one using the FIA plot observations assumed to be without error as ground classes, and one using the majority interpretations with possible interpreter error as ground classes where the latter scenario represents operational practice when no ground data are available. The hybrid standard error obtained using the FIA plot observations as the ground classes was $SE_{Hyb}(\hat{\mu}_{Hyb}) = 0.037$ and when using the majority interpretations as substitutes for the ground classes the hybrid standard error was $SE_{Hyb}(\hat{\mu}_{Hyb}) = 0.036$. Using majority interpretations as reference data, the ratios of hybrid to stratified SEs were 2.25–2.31, regardless of whether the FIA plot observation or the majority interpretation was used as ground class. This ratio represents the factor by which the SE is under-estimated by the stratified estimator which does not account for interpreter error.

Table 3
Interpreter accuracies.

Interpreter	GFC*		NLCD*	
	Forest stratum	Non-forest stratum	Forest stratum	Non-forest stratum
Relative to inventory observations				
1	0.975	0.928	0.895	0.968
2	0.870	0.963	0.842	0.978
3	0.913	0.889	0.877	0.892
Relative to majority interpretations				
1	0.960	0.840	0.680	0.960
2	0.787	0.947	0.627	0.960
3	0.880	0.920	0.733	0.933

* NLCD: National Land Cover Database; GFC: Global Forest Change dataset.

4.2.2. National land cover database map

For the NLCD map, the forest and non-forest stratum weights were 0.580 and 0.420, and the within-stratum map accuracies relative to the inventory plot observations were 0.920 and 0.680 and relative to the majority interpretations were 0.960 and 0.680. Within-stratum interpreter accuracies relative to the inventory plot observations and the majority interpretations are reported in Table 3. The stratified estimates were $\hat{\mu}_{Str} = 0.668$ with $SE_{Str}(\hat{\mu}_{Str}) = 0.029$ using the FIA plot forest/non-forest observations as reference data and $\hat{\mu}_{Str} = 0.658$ with $SE_{Str}(\hat{\mu}_{Str}) = 0.025$ using the majority interpretations as reference data (Table 4). The same comments from Section 4.2.1 regarding the relative sizes of these two standard errors, the biasedness of the stratified estimator when using reference data, and the two scenarios for calculating hybrid standard errors pertain to the NLCD analyses. The hybrid standard error obtained using the FIA plot observations as the ground classes was $SE_{Hyb}(\hat{\mu}_{Hyb}) = 0.044$ and using the majority interpretations as ground classes was $SE_{Hyb}(\hat{\mu}_{Hyb}) = 0.042$. Using majority interpretations as reference data, the ratios of the hybrid to the stratified SEs were 1.68–1.76, regardless of whether the FIA plot observations or the majority interpretations were used as ground classes. This ratio represents the factor by which the SE is under-estimated by the stratified estimator which does not account for interpreter error.

Table 4
Estimates based on field and interpreter data.

Map	No interpreter error		With interpreter error			
	$\hat{\mu}_{Str}$	$SE(\hat{\mu}_{str})$	$\hat{\mu}_{Str}$	$SE(\hat{\mu}_{str})$	$SE_{Hyb}(\hat{\mu}_{Hyb})$ (Ground class data)	
					Plot observations	Majority Interpretations
GFC	0.723	0.024	0.738	0.016	0.036	0.037
NLCD	0.668	0.029	0.658	0.025	0.044	0.042

4.3. Summary and discussion

The primary results of the simulation analyses were that for unequal stratum weights, the effects of interpreter error were twofold. First, interpreter errors induced bias into the stratified estimator of proportion forest. The bias is less for greater equality in the stratum weights, greater map and interpreter accuracies, larger numbers of interpreters, and smaller between-interpreter correlations. Of importance for advance planning purposes, under the assumption that all interpreters are well-trained, the only one of these factors that can be controlled is the number of interpreters. In particular, the stratum weights and the map accuracies are a function of the map, and between-interpreter correlations for experienced interpreters using the same aerial imagery should be expected to be large; indeed, small correlations should be a cause for operational concern. Second, interpreter errors induced bias into the stratified estimator of the standard error with the result that stratified standard errors were substantially under-estimated. Ratios of hybrid to stratified SEs were approximately 1.4, were non-negligible, and argue strongly in favor of using the hybrid estimators.

Multiple results obtained from analyses of the inventory plot observations and the visual interpretations merit consideration. First, differences in the stratified proportion forest estimates obtained using the inventory plot observations as reference data and the majority interpretations as reference data were small, 0.015 for the GFC map and 0.010 for the NLCD map. For the GFC map, the small difference can be at least partially attributed to large interpreter accuracies, ranging from 0.870 to 0.975 (Table 3), and relatively small interpreter correlations (Table 5). For the NLCD map, the small difference can be attributed to nearly equal stratum weights and relatively large interpreter accuracies.

Second, the hybrid standard errors based on inventory plot observations as the ground classes were larger than the stratified standard errors by factors of 1.76–2.31. These factors are larger than the factor of 1.40 obtained from the simulations. However, the important consequence is that failure to use the hybrid variance estimator to incorporate interpreter error produced substantial under-estimates of standard errors and, thereby, would lead to non-compliance with the IPCC good practice guidance.

Third, the hybrid standard errors obtained using majority interpretations as substitutes for ground classes were larger than the stratified estimates by factors of 1.68 to 2.25. These factors are similar to the factors obtained using the plot observations as ground classes and suggest that when ground observations are not available, use of the majority interpretations as substitutes is a reasonable alternative.

Several issues merit additional comments. Firstly, the simulations showed that greater between-interpreter correlations, which would be natural for well-trained interpreters, actually produced greater bias in both estimates of proportions and their standard errors. The explanation is that with greater correlations, an error for a particular sample unit by one interpreter is more likely to be associated with similar errors by other interpreters for the same sample unit, whereas with smaller correlations, an error for a particular sample unit is more likely to be offset by correct interpretations by the other interpreters. Secondly, more equal stratum weights tended to produce less bias in estimators of both proportions and standard errors. However, small, fragmented, and interspersed forest and non-forest patches are known to be difficult to classify correctly using remotely sensed data. Therefore, if the total forest and non-forest areas of fragmented and interspersed forest patches are approximately equal, as would be the case for approximately equal stratum weights, then these strata of approximately equal size may be associated with greater interpreter errors.

Finally, alternatives to the simple majority of independent interpretations as reference classes may be considered. For example, prior to operational interpretation, interpreters can calibrate their interpretations with respect to known field conditions and/or to each other (Guyana Forestry Commission, 2012, Appendix 10, Section 5.2). Also, in the absence of unanimous interpretations, interpreters may discuss the specific sample units and agree on a consensus interpretation. In addition, instead of using majority interpretations leading to binary reference observations (0 for non-forest, 1 for forest), continuous reference observations in the form of the proportions of forest interpretations among interpreters for the same sample unit are possible. Confusion matrices can still be used, although variances would be calculated differently.

5. Conclusions

Compliance with the IPCC good practice guidance for greenhouse gas inventories requires the use of unbiased estimators and reduction of uncertainties with the latter guideline pre-supposing rigorous estimation of those uncertainties (IPCC, 2006, Volume 1, Chapter 1, Section 1.2; GFOI 2016, p. 15). Thus, the study focused on issues of bias and rigorous estimation of variances and standard errors. Three conclusions were drawn from the study. First, interpreter error induces bias into the stratified estimator of proportion forest. The bias is greater for greater

Table 5
Interpreter correlations.

		Global Forest Change data						National Land Cover Database					
		Forest stratum (Interpreter)			Non-forest stratum (Interpreter)			Forest stratum (Interpreter)			Non-forest stratum (Interpreter)		
		1	2	3	1	2	3	1	2	3	1	2	3
Interpreter	1	1.000	0.557	0.692	1.000	0.838	0.622	1.000	0.653	0.491	1.000	0.889	0.750
	2	0.557	1.000	0.367	0.838	1.000	0.509	0.653	1.000	0.764	0.889	1.000	0,657
	3	0.692	0.367	1.000	0.622	0.509	1.000	0;491	0.764	1.000	0.750	0.657	1.000

inequality in stratum weights, smaller map and interpreter accuracies, fewer interpreters, and greater correlations among interpreters. Of these factors, only the number of interpreters can be readily controlled. Second, interpreter error induces bias into the stratified estimator of the variance and standard error. Failure to incorporate the effects of interpreter error via the hybrid estimator led to under-estimates of standard errors by factors ranging from 1.4 for the simulation analyses to as great as 2.3 for the inventory data and visual interpretations. Thirdly, in the absence of ground class data without error such as inventory plot observations, use of the majority interpretations as substitutes produced hybrid standard errors that were similar to estimates obtained using the inventory observations. However, the latter conclusion should be the subject of additional research for a greater variety of conditions, particularly greater inequality in stratum weights.

For planning purposes, the expected land class proportion in the absence of interpreter error as expressed by Eq. (3) and the results in the Tables 2a,b,c can be used to guide decisions regarding number of interpreters, the only factor influencing bias that can be readily controlled. In this context, two general recommendations merit consideration. First, at least three experienced interpreters should be used. Use of a single interpreter does not permit substitution of majority interpretations for ground class observations, and use of two interpreters may lead to a large number of interpretation ties when assessing majority interpretations. For extremely unequal stratum weights, the simulation analyses suggest that five or perhaps even seven interpreters may be necessary to mitigate the effects of bias in the stratified estimator of proportion forest. Secondly, hybrid variance estimators are necessary to circumvent the biasing effects of interpreter error on the stratified estimator of the standard error.

Acknowledgements

The authors gratefully acknowledge the assistance of aerial imagery interpreters, Cassandra L. Olson, Brian J. Gasper, and Jeffery S. Wazenegger; coordinator, Daniel J. Kaisershot; and field crew supervisor, James D. Blehm, all of the Forest Inventory and Analysis program, Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota, USA.

References

Ban, Y., Gong, P., Giri, C., 2015. Global land cover mapping using Earth observation satellite data: recent progresses and challenges. ISPRS J. Photogramm. Remote Sens. 103, 1–6.

Boschetti, L., Stehman, S.V., Roy, D.P., 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. Remote Sens. Environ. 186, 465–478.

Bross, I., 1954. Misclassification in 2 x 2 tables. Biometrics 10 (4), 478–486.

Chen, D.M., Wei, H., 2009. The effect of spatial autocorrelation and class proportion on the accuracy measures from different sampling designs. ISPRS J. Photogramm. Remote Sens. 64, 140–150.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley, New York, pp. 428.

Corona, P., Fattorini, L., Franceschi, S., Scrinzi, G., Torresan, C., 2014. Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Can. J. For. Res. 44, 1303–1311.

Demirtas, H., 2004. Pseudo-random number generation in R for commonly used multivariate distributions. J. Modern Appl. Stat. Methods 3 (2) Article 19.

Fattorini, L., 2012. Design-based or model-based inference? The role of hybrid approaches in environmental surveys. Studies in Honor of Claudio Scala, L. Fattorini (Ed.). Department of Economics and Statistics, University of Siena, Siena, Italy. pp. 173–214.

Foody, G.M., 2009. The impact of imperfect ground reference data on the accuracy of land cover change estimation. Int. J. Remote Sens. 30 (12), 3275–3281.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sens. Environ. 114, 2271–2285.

Foody, G.M., 2013. Ground reference data error and the misestimation of the area of land cover change as a function of its abundance. Remote Sensing Letters 4, 8.

GFOI, 2016. Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests: Methods and Guidance from the Global Forest Observations Initiative, Edition 2.0. Food and Agriculture Organization, Rome. 224 p. Available at: < https://www.reddcompass.org/download-the-mgd > (last accessed, July 2017).

GFC (Guyana Forestry Commission), 2012. Guyana REDD+ Monitoring Reporting & Verification System (MRVS) Interim Measures Report, 01 October 2010 – 31 December 2011, Version 3.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-Resolution global maps of 21st-century forest cover change. Science 342, 850–853.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogramm. Eng. Remote Sens. 81 (5), 345–354.

IPCC, 2006. 2006 IPCC guidelines for national greenhouse gas inventories. In: In: Eggleston, H.S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K. (Eds.), Agriculture, Forestry and Other Land Use, vol. 4 Institute for Global Environmental Strategies, Japan. Available at: < http://www.ipcc-nggip.iges.or.jp/public/2006gl/index.html > (last accessed: February 2018).

Mannel, S., Price, M., Hua, D., 2006. A method to obtain large quantities of reference data. Int. J. Remote Sens. 27, 623–627.

McRoberts, R.E., Hansen, M.H., Smith, W.B., 2010. United States of America. In: Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (Eds.), National Forest Inventories, Pathways for Common Reporting. Springer 610p.

McRoberts, R.E., Vibrans, A.C., Sannier, C., Næsset, E., Hansen, M.C., Walters, B.F., Lingner, D.V., 2016a. Methods for evaluating the utilities of local and global maps for increasing the precision of estimates of subtropical forest area. Can. J. For. Res. 46, 924–932.

McRoberts, R.E., Chen, Q., Domke, G.M., Ståhl, G., Saarela, S., Westfall, J.A., 2016b. Hybrid estimators for mean aboveground carbon per unit area. For. Ecol. Manage. 378, 44–56.

Mountrakis, G., Xi, B., 2013. Assessing reference dataset representativeness through confidence metrics based on information density. ISPRS J. Photogramm. Remote Sens. 78, 129–147.

Næsset, E., 1991. The effect of season upon registrations of stand mean height, crown closure and tree species on aerial photos. Commun. Skogforsk 44 (7), 1–28.

Næsset, E., 1992. The effect of scale, type of film and focal length upon interpretation of tree species mixture on aerial photos. Commun. Skogforsk 45 (5), 1–28.

Næsset, E., Ørka, H.O., Solberg, S., Bollandsås, O.M., Hansen, E.H., Mauya, E., Zahabu, E., Malimbwi, R., Chamuya, N., Olsson, H., Gobakken, T., 2016. Mapping and estimating forest area and aboveground biomass in miombo woodlands in Tanzania using data from airborne laser scanning, TanDEM-X, RapidEye, and global forest maps: a comparison of estimated precision. Remote Sens. Environ. 175, 282–300.

Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sens. Environ. 129, 122–131.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.

Pengra, B., Long, J., Dahal, D., Stehman, S.V., Loveland, T.R., 2015. A global reference database from very high resolution commercial satellite data and methodology for application to Landsat derived 30 m continuous field tree cover data. Remote Sens. Environ. 165, 234–248.

Powell, R.L., Matzke, N., de Souza, D., Clark, M., Numata, I., Hess, L.L., Roberts, D.A., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sens. Environ. 90, 221–234.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York, pp. 258.

Sannier, C., McRoberts, R.E., Fichet, L.-V., 2016. Suitability of Global Forest Change data to report forest cover estimates at national level in Gabon. Remote Sens. Environ. 173, 326–338.

Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. Int. J. Remote Sens. 30 (20), 5243–5272.

Sun, B., Chen, X., Zhou, Q., 2017. Analyzing the uncertainties of ground validation for remote sensing land cover mapping in the era of big geographic data. In: Zhou, C., Su, F., Harvey, F., Xu, J. (Eds.), Spatial data handling in big data era. Advances in geographic information science, Springer, Singapore, pp. 31–38.

Thompson, I.D., Maher, S.C., Rouillard, D.P., Fryxell, J.M., Baker, J.A., 2007. Accuracy of forest inventory mapping, some implications for boreal forest management. For. Ecol. Manage. 252, 208–221.

Tsendbazar, N.E., de Bruin, S., Herold, M., 2015. Assessing global land cover reference datasets for different user communities. ISPRS J. Photogramm. Remote Sens. 103, 93–114.

Zimmerman, P.L., Liknes, G.C., 2010. The role of misclassification in estimating proportions and an estimator of misclassification probability. Math. Comput. Forest. Nat. Resour. Sci. 2 (2), 78–85.