Scientific Research Publishing

Open Journal of Forestry, 2021, 11, 14-36
https://www.scirp.org/journal/ojf
ISSN Online: 2163-0437
ISSN Print: 2163-0429

Calibration of a Confidence Interval for a Classification Accuracy

Steen Magnussen

Natural Resources Canada, Canadian Forest Service, Victoria, British Columbia, Canada
Email: steen.magnussen@canada.ca

How to cite this paper: Magnussen, S. (2021). Calibration of a Confidence Interval for a Classification Accuracy. *Open Journal of Forestry, 11,* 14-36.
https://doi.org/10.4236/ojf.2021.111002

Received: September 30, 2020
Accepted: January 17, 2021
Published: January 20, 2021

Abstract

Coverage of nominal 95% confidence intervals of a proportion estimated from a sample obtained under a complex survey design, or a proportion estimated from a ratio of two random variables, can depart significantly from its target. Effective calibration methods exist for intervals for a proportion derived from a single binary study variable, but not for estimates of thematic classification accuracy. To promote a calibration of confidence intervals within the context of land-cover mapping, this study first illustrates a common problem of under and over-coverage with standard confidence intervals, and then proposes a simple and fast calibration that more often than not will improve coverage. The demonstration is with simulated sampling from a classified map with four classes, and a reference class known for every unit in a population of 160,000 units arranged in a square array. The simulations include four common probability sampling designs for accuracy assessment, and three sample sizes. Statistically significant over- and under-coverage was present in estimates of user's (UA) and producer's accuracy (PA) as well as in estimates of class area proportion. A calibration with Bayes intervals for UA and PA was most efficient with smaller sample sizes and two cluster sampling designs.

Keywords

Overall Accuracy, Producer's Accuracy, User's Accuracy, Area Proportions, Semi-Systematic Sampling, Post-Stratification, Stratified Random Sampling, One-Stage Cluster Sampling, Two-Stage Cluster Sampling

1. Introduction

Accuracy assessment is an important step in any land cover mapping project (Congalton, 2001; Stehman & Foody, 2019). An estimate of accuracy is typically derived from a comparison between the classified map cover class of a unit and a

reference land cover class assigned either after a field visit to the unit, or after a supervised classification by trained interpreters of an image of the same unit rendered in a substantially higher resolution than used for the map production (Khatami, Mountrakis, & Stehman, 2016; McRoberts et al., 2018). Regardless, to be objective, free of bias, and independent of the classification process, it is important that accuracy statistics are derived from a sample of units obtained under a probability sampling design (Stehman, 1999). That is, the sample inclusion probability of each unit to be sampled is known prior to sampling. The sample inclusion probabilities are used as weights to obtain design-consistent estimators of accuracy and their variances (Cochran, 1977).

Simple random or systematic sampling with post-stratification by map class, stratified random sampling, and one- and two-stage cluster sampling are the most popular choices of sampling designs for accuracy assessment (Morales-Barquero, Lyons, Phinn, & Roelfsema, 2019; Olofsson et al., 2014; Wulder, Franklin, White, Linke, & Magnussen, 2006). The sample size for this endeavor is typically a compromise between purely statistical considerations about the anticipated accuracy and a maximum acceptable standard error (Barth & Ståhl, 2011; Dobbertin & Biging, 1996; Stehman, 2012). Pros and cons of sampling designs for accuracy assessments have been detailed (Stehman, 1999). Recommendations with regards to design, sample size, and statistics are followed by many, yet, a review by Morales-Barquero et al. (2019) indicated that the use of probability sampling is still not the norm. Only approximately 30% of the land cover classification studies reviewed published statistics sufficient for reproducing results.

Estimates of accuracy are, as a rule, presented as the proportion of correct classifications or as the proportion of correct classification given either a map or a reference class (Czaplewski, 2003). Alternative metrics used to describe accuracy of fuzzy (soft) classifications (Binaghi, Madella, Montesano, & Rampini, 1997; Ricotta, 2004) and classification of multi-unit objects (Stehman & Wickham, 2011) are beyond the scope of this exposé. The accuracy can be reported by thematic class (map or reference), or as the overall accuracy. Design-specific estimators of accuracy and their variances have been worked out and are now widely available (Breidt & Opsomer, 2017; Czaplewski, 1994; Fattorini, 2015; Olofsson et al., 2014).

It is also considered good practice to provide a confidence interval for a point estimate of accuracy. A common choice is a 95% confidence interval (Olofsson et al., 2014; Stehman & Foody, 2019). The correct interpretation is that in repeated sampling under the same design, the true but unknown accuracy will lie between the limits of an estimated interval 95% of the time (Cochran, 1977). In land-cover accuracy assessments, the standard method for the construction of a confidence interval is to assume a normal or a *t*-distribution of the point estimate with a standard deviation equal to the obtained estimate of the margin of error, i.e. the square root of the estimated variance divided by the sample size (Newcombe, 1998). Such confidence intervals only achieve a nominal coverage (viz. the proportion of intervals that include the true unknown accuracy) asymp-

totically with independent observations (Esty, 1982; Miao & Gastwirt, 2004). In smaller samples or with proportions near the limits of 0 and 1, a coverage based on an assumed distribution can be poor (Agresti & Caffo, 2000; Korn & Graubard, 1998; Wendell & Schmee, 2001). For proportions derived from a univariate binary variable we have many alternative estimators for the upper and lower limits of a confidence interval with improved performance over intervals computed with the standard method (Newcombe, 1998). For complex survey designs, Franco, Little, Louis, & Slud (2019) reported important improvements in coverage of nominal 95% confidence intervals of a proportion when they used Wilson- and Bayes-type intervals, and sample sizes corrected to the nominal sample size divided by the design-effect.

Resampling methods have also been employed in pursuit of improving coverage of a confidence interval for a proportion, but improvements are only realized when the resampling is consistent with the sampling design generating the sample data (Rao & Wu, 1988; Shao, 1996). With an unknown target distribution of quantiles, success is not ascertained with standard bootstrap methods (Antal, 2011; Chambers & Dorfman, 2003; Conti & Marella, 2014). Jackknife based intervals may also suffer from over-coverage (Román-Montoya, Rueda, & Arcos, 2008). Full Bayesian methods have been tried but typically also result in over-coverage and a different inference (e.g. Finley, Banerjee, Ek, & McRoberts, 2008). A pseudo Bayesian empirical likelihood approach proposed by Rao & Wu (2010) appears promising but also requires complex computations.

It is rare, however, to find an accuracy assessment of a land-cover mapping project that acknowledges the potential of poor coverage of confidence intervals for point estimates. The problem of poor coverage is compounded when an accuracy statistic is derived from a joint distribution of two correlated binary variables or estimated from a ratio of two correlated random counts (Chambers & Dorfman, 2003; Fodé & Louis-Paul, 2014; Miao & Gastwirt, 2004). On this background, the objective of this study is to encourage use of a simple and fast method for computing well-calibrated confidence intervals for point estimates of thematic accuracy and area proportions. The encouragement comes in the form of promising results from simulations with a stochastic classification process with four land-cover classes and a Bayes-type confidence interval with a uniform prior. Intervals of this type were recommended for complex surveys by Franco et al. (2019). Their recommendation, however, is restricted to a single univariate binary variable, and may not extend to a set of class-specific bivariate correlated binary variables as encountered in a land-cover mapping project. A re-assessment for application in a land-cover accuracy assessment is called for.

2. Material and Methods

2.1. The Reference Map

A fixed artificial reference map with four classes (A, B, C, and D) was created for 160,000 equal sized square units in an array with 400 rows and 400 columns. The

proportions of units in classes A-D are, respectively, 0.10, 0.20, 0.30, and 0.40. The reference map portrays a mosaic of spatial clusters of different sizes and shapes dominated by a single class. This was achieved by a random draw of a matrix with 400 rows and 400 columns populated with standard normal unit-level values (z) drawn from a Gaussian matrix-distribution (Gupta & Nagar, 1999) with a first-order autoregressive (AR1) covariance structure along rows and columns. The AR1 parameter was set to 0.90 which means that the correlation between z-values separated by one unit along a row or a column is 0.90 and 0.90^r if separated by r units. The z-variate was then converted to class labels. A z-value less than or equal to the 10%-tile of a standard normal distribution was assigned to class A. Values between the 10%-tile and the 30%-tile were assigned to class B, and so on for the 60%- and 100%-tile (viz. infinity). The spatial distribution of the four classes in the first 100 rows and 100 columns is shown separately for each class in **Figure 1**.

2.2. Classification Process

A correct map class was assigned to a unit based on the outcome of random draw from a binomial distribution with a probability provided by a reference class specific spatial latent accuracy distribution. Generalized beta distributions (McDonald & Xu, 1995; Tan, 1969) with a mean target accuracy of 0.88 for reference class A, 0.92 for class B, 0.78 for class C, and 0.75 for class D served as the marginal distributions of the latent classification accuracy (cf. **Figure 2**).

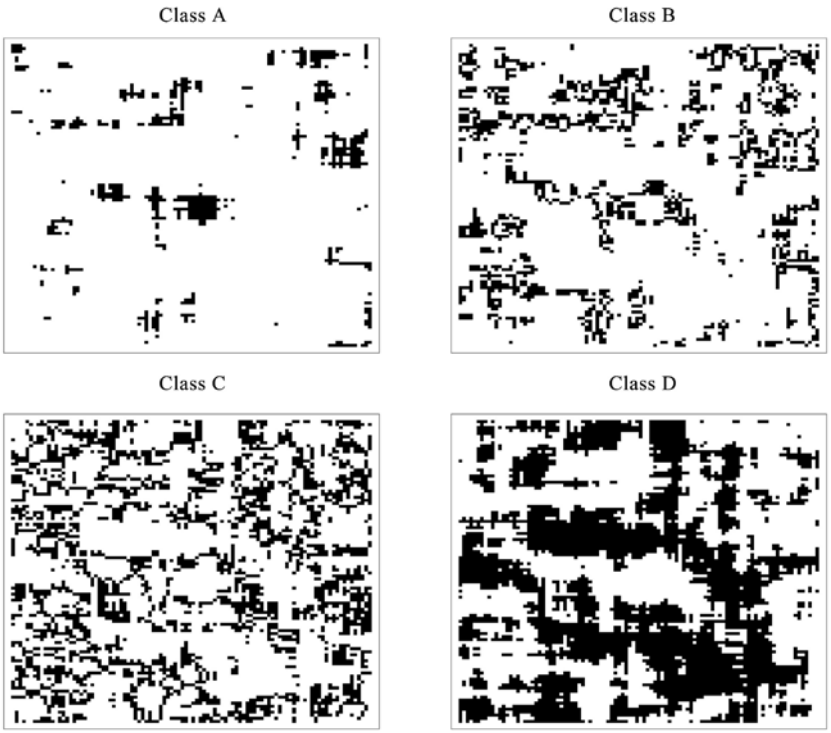

Figure 1. Maps for reference classes A, B, C, and D. A black unit indicate the location of a reference class (A, B, C, or D). Only the area covered by the first 100 rows and first 100 columns are shown.

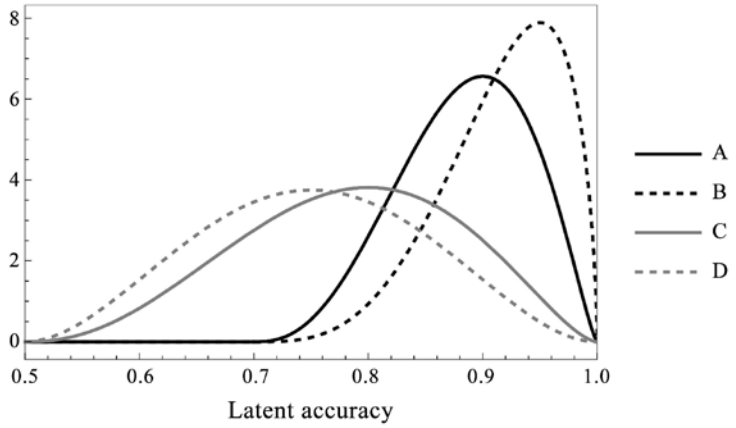

Figure 2. Marginal distributions of the latent accuracy of a classification to a map class (A, B, C, D) given the reference class of a unit.

For reference class A, 95% of the latent accuracies are between 0.77 and 0.98, for class B between 0.80 and 0.99, for class C between 0.60 and 0.95, and between 0.57 and 0.93 for class D. A spatial autocorrelation in latent accuracies was imposed by a third-order moving average process (MA3) along rows and columns in the population. The expected autocorrelation in the binary outcome of a classification (1 for correct, 0 for incorrect) was hereafter 0.58, 0.30, and 0.07 for units separated by a row- or column-distance of one, two, and three units. No autocorrelation was imposed on units separated by four or more row (column) units. A sample of the binary outcome (correct = 1 (white), incorrect = 0 (black)) of the classification is in **Figure 3**. The presence of short-range positive spatial autocorrelation in correct (and incorrect) classification outcome is apparent and deemed more realistic than independence (Khatami, Mountrakis, & Stehman, 2017).

In case of an incorrect classification, the map class was assigned with equal probability to one of the two most similar reference classes. The two most similar reference classes are the class to the right and to the left of a class when the four classes were connected to form a circle as in A-B-C-D-A. Classification results for the units in **Figure 1** are shown in **Figure 4**.

2.3. Sampling Designs

Four commonly employed sampling designs are investigated. They are: semi-systematic sampling (*ssyst*), stratified random sampling (*strat*), one-stage, and two-stage cluster sampling (*clust* and *clust2st*). For each design, three sample sizes with n = 828, 414, and 207 units were employed. The largest sample size was determined as the sample size for a one-stage cluster sampling design with an intra-cluster binary correlation coefficient of 0.4 that in 95 out of 100 trials would generate a relative standard error less than or equal to 0.025 in an estimate of producer's accuracy for a reference class with an area proportion of 0.30 (Fleiss, Levin, & Paik, 2013). The two other sample sizes represent 50% and 25% of the sample size deemed necessary to achieve the above target of precision.

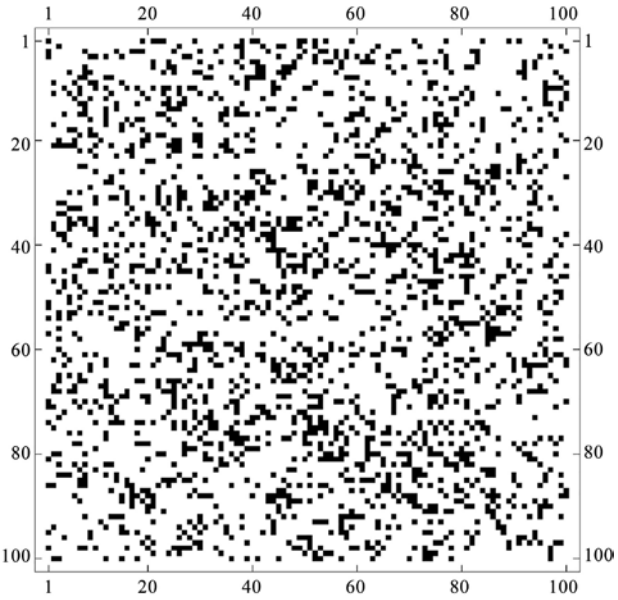

Figure 3. Correct (white) and incorrect (black) classifications in the first 100 rows and 100 columns of the population.

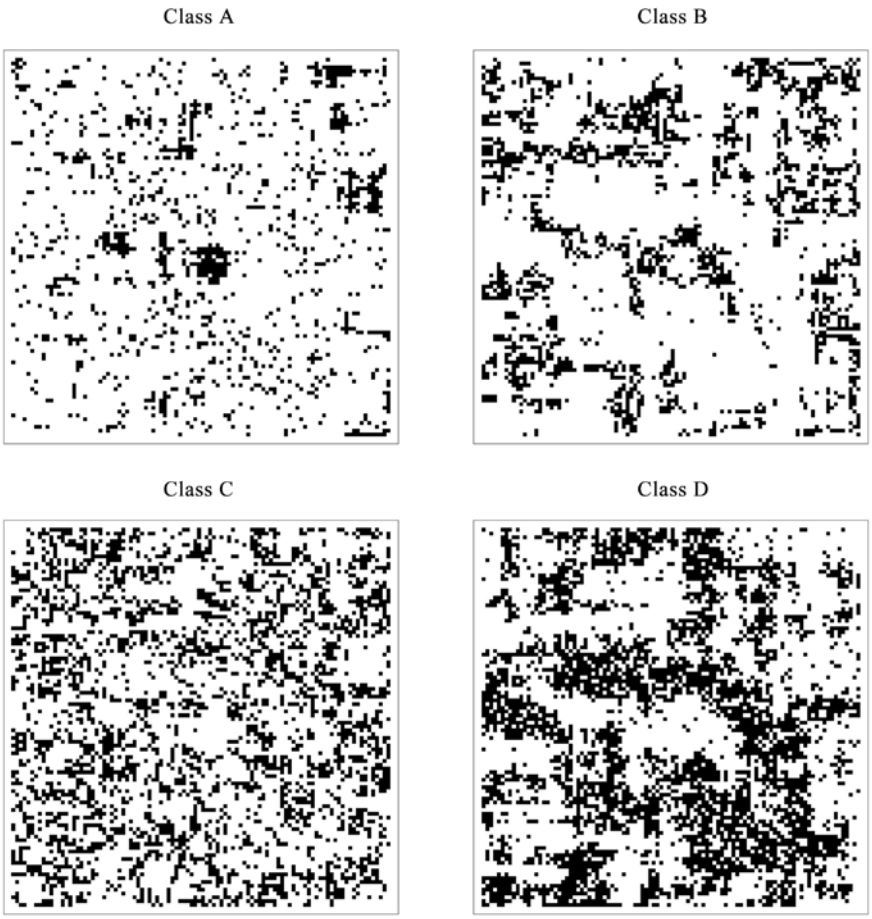

Figure 4. Classification outcome viz. map classes (black) for the first 100 rows and 100 columns in the population. See Figure 1 for a comparison to corresponding reference classes.

The *ssyst* was executed by first subdividing the population into 1600 blocks of 10 × 10 units followed by a random without replacement selection of n of these blocks and a random selection of a single unit within each selected block.

In the stratified random sampling design, the population was stratified to map classes, i.e. to four strata. The known strata sizes in units were 23,512 (14.7%) for map class A, 35,452 (22.2%) for map class B, 46,848 (29.3%) for class C, and 54,190 (33.8%) for class D. Stratum sample sizes were proportional to stratum sizes. Thus, with n = 828 stratum sample sizes were 121, 183, 243, and 281; for n = 414 they were 60, 91, 122, and 141. Finally, for n = 207 they were 39, 45, 61, and 71. Within a stratum, units were sampled at random without replacement.

In the *clust* design the cluster size was 9 units arranged in an array of three rows and three columns. For a design with a sample size of n units, there are $K = n \div 9$ clusters. The K clusters were first aligned with a random set K of the n blocks selected for the *ssyst* design, and then a square cluster was formed as the set of 8 nearest within-block neighbours to a unit selected under *ssyst*.

In the *clust2st* design, the cluster size was 36 units arranged in an array of six rows and six columns. A total of 9 units were sampled within each cluster. Thus, the number of clusters in *clust* (K) and *clust2st* is the same. The selection of clusters followed, apart from issues of scale, the procedure for *clust*. The 9 units from a 6 × 6 cluster were selected as in a systematic design with a sampling interval of 4 units and a random start in one of the columns 1 to 4 in row 1 of a cluster.

2.4. Estimators of Accuracy

The following estimators of accuracy and area proportions are used:

Overall accuracy (OA) is defined as the probability of a correct map classification. It is estimated as the proportion of correct map classifications in the sample of n units. In other words, OA is the chance that a unit selected at random from a map is classified correctly.

Producer's accuracy (PA) is defined as the reference class specific (conditional) probability of a correct classification. It is estimated as the number of correct classifications in a reference class divided by number of sample units in the reference class. For both *ssyst* with a post-stratification, and the *strat* designs, the two counts are weighted sums over the map-classes with weights proportional to map-class strata sizes. For the two cluster designs, PA was computed on a per cluster basis and averaged over the K clusters. PA communicates the accuracy of the classification process applied to units of given reference class.

User's accuracy (UA) is defined as the map class specific (conditional) probability of a correct classification. It is estimated as the number of correct classifications in a map class divided by the number of sample units in the map class. For *ssyst* (with post-stratification) and *strat* UA is computed as the number of correct classifications in a map class divided by either the known sample size in a stratum in case of *strat*, or—in case of ssyst—by the realized number of sample units in the map post stratum class. UA for the two cluster designs was com-

puted on a per cluster basis and averaged over the K clusters. UA provides a user of a map with the probability that the reference class of a unit in the map (the actual state) is the same as the one given in the map.

Reference class area proportions (P_{ref}) for *ssyst* and *strat* were computed as a weighted average of the number of sample units in a reference class divided by the overall sample size n. The weights were again proportional to the size of a map-class stratum. For *clust* and *clust2st* the area proportions were computed for each cluster and averaged over the K clusters.

When an estimate of an accuracy could not be computed, due to a zero sample count in a given class, a correction by adding two successes and two failures as suggested by Agresti & Caffo (2000) was implemented.

Analytical variance estimators are not presented in an effort to reduce the length of this section. All estimators have been detailed elsewhere (Stehman, 1992; Stehman, 1997; Stehman, 2012; Stehman & Czaplewski, 1998; Stehman et al., 2009). For *ssyst* the estimator of variance for post-stratification by map class and simple random sampling within post-strata was employed. In all estimators of variance, accuracy estimates of either 0 or 1 were replaced by, $1/n$, or $n/(n+1)$, respectively.

2.5. Confidence Intervals

Standard nominal 95% confidence intervals for an estimate of accuracy was computed as $\hat{p} \pm t_{0.975,df}\, a\hat{S}E$ where $\hat{p}$ is the sample-based estimate of accuracy, $a\hat{S}E$ is the analytical estimate of the standard error of $\hat{p}$, and $t_{0.975,df}$ is the 0.975 quantile in a student's t-distribution with df degrees of freedom. Here df is the number of sample units in the denominator of the ratio estimator for $\hat{p}$ minus one. Coverage of estimated confidence intervals (CCI95) was computed as the proportion of estimated intervals (cf. sub-section on simulated sampling) for a given accuracy that include the true accuracy p.

In pursuit of improved coverage, the estimated confidence intervals were replaced (calibrated) with Bayes intervals with a uniform prior. We could equally have chosen intervals of the types Wilson, Jeffrey or Clopper Pearson as their performance in this study and in that of Franco et al. (2019) was similar. Specifically, a calibrated interval was equated to the 0.025 and 0.975 quantiles of a beta distribution with parameters $\alpha = \tilde{n}_{correct} + 1$ and $\beta = \tilde{n}_{sample} - \tilde{n}_{correct} + 1$ where $\tilde{n}_{correct}$ and $\tilde{n}_{sample}$ are the adjusted class-specific number of correctly classified units and the class specific number of sample units, i.e. the numerator and denominator used to obtain $\hat{p}$. Both $\tilde{n}_{correct}$ and $\tilde{n}_{sample}$ are derived from the actual counts $n_{correct}$ and n_{sample} after a division by the square root of a sample specific estimate of the design effect $\left(\sqrt{\hat{D}_{eff}} = \sqrt{\hat{var}_{des}(\hat{p}) \div \hat{var}_{srs}(\hat{p})}\right)$ where $\hat{var}_{des}(\hat{p})$ is the design-based estimator of variance of $\hat{p}$, and $\hat{var}_{srs}(\hat{p})$ is the estimated variance under a simple random sampling design (Fuller, 2011). The square root transformation was adopted because it significantly improved the chance that the coverage of a calibrated interval would be improved. The improvement is

ascribed to a reduction of a marked skewness of the distribution of $\hat{D}_{eff}$. Franco et al. also mention a bias in $\hat{D}_{eff}$ due to the fact that $\hat{D}_{eff}$ is a design-biased ratio of two random but correlated variances.

Two examples detail how Bayes intervals differ from the standard intervals. Both cases are for PA in class A and a sample size of 207 units. The first example is with *ssyst*. The average standard 95% interval was from 0.77 to 0.99 and with PA estimated at 0.89 (true PA is 0.88) the coverage over the 2000 replications as 0.88. The average of Bayes intervals is 0.79 to 0.96 (i.e. wider by 0.04) also with a coverage of 0.88. The changes to the lower and upper limit of a standard confidence interval when it is replaces by a Bayes interval is illustrated in Figure 5. Even with changes to both ends, the overall coverage remained unchanged.

The second example is with *clust*. PA was again estimated at 0.89 and the average of the standard intervals was from 0.78 to 0.98 with a coverage of just 0.75. After a calibration with Bayes intervals, the average interval was from 0.69 to 0.96 (i.e. wider by 0.06) and coverage rose to 0.90. Figure 6 illustrates the changes from a calibration to the lower and upper endpoints of a standard interval. While changes are apparent at both ends, they are much larger for the lower limit than for the upper limit.

From here on end, the terms "significant" and "significantly" are used as short forms for "statistically significant at the 5% level or lower of a Type I error".

2.6. Simulated Sampling

Simulated sampling from the fixed population of bivariate units with a reference class and a map class label was executed according to the twelve combinations of four designs and three sample sizes. Following the selection of a sample, the accuracy statistics and their analytical variances were computed and 95% intervals (standard and calibrated) were obtained as outlined. This process was repeated 2000 times. Afterwards, the empirical (Monte-Carlo) estimate of the standard error in an estimate of accuracy (or an area proportion) was obtained as the

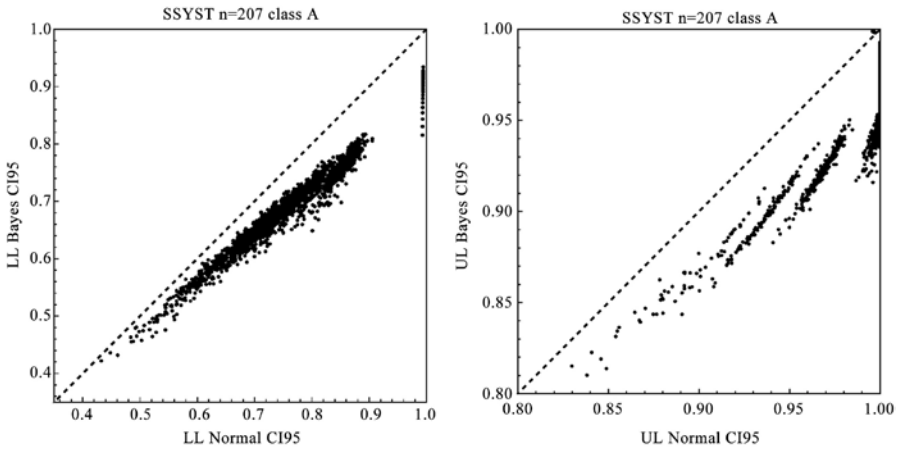

Figure 5. Case one: scatter of lower (left) and upper (right) limit of Bayes intervals plotted against corresponding standard interval limits.

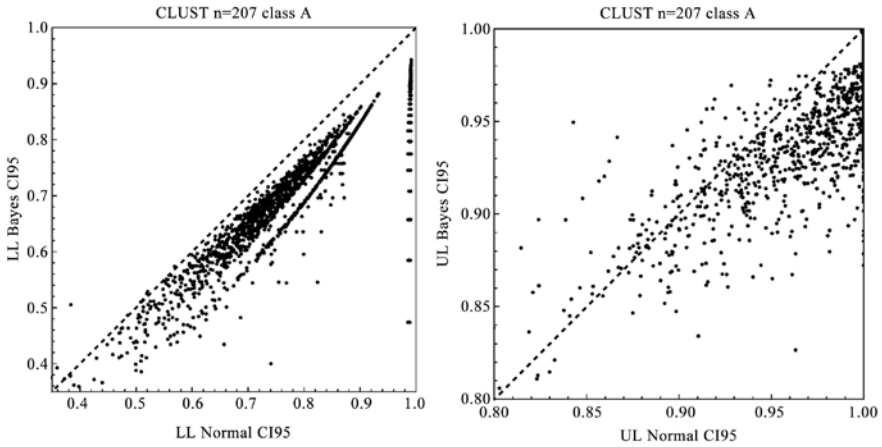

Figure 6. Case two: scatter of lower (left) and upper (right) limit of Bayes intervals plotted against corresponding standard interval limits.

standard deviation of the 2000 available estimates, and the coverage of a confidence interval computed as the proportion of intervals including the true accuracy (proportion). With 2000 replications a coverage below 0.94 or above 0.96 would be statistically significant at the 0.05% level of a Type I error.

3. Results

3.1. Overall Accuracy (OA)

On average over the 2000 replications, the estimate of OA was for all designs and sample sizes within 0.3% from the true value of 0.81 (Table 1). None of the observed differences from the target value were statistically significantly different from 0 at the 5% level.

Empirical standard errors (eSE) with *ssyst* and post-stratification to map classes and with *strat* were within 0.01 of each other. With these two designs, the empirical standard error increased with a decrease in sample size at the expected rate of one over the square root of n. The average of the analytical standard errors (aSE) matched to within 7% the empirical estimates. Empirical standard errors for *clust* and *clust2st* are, as expected, greater than for *ssyst* and *strat*. The average difference was 0.03. In relative terms, the difference decreased from approximately 20% with $n = 828$ to approximately 10% with $n = 207$. The average analytical estimate of standard error under *clust* and *clust2st* matched the empirical estimate to within 10%. Coverage of estimated 95% confidence intervals fluctuated between 0.94 to 0.96 with no statistically significant departure from the nominal level of 0.95. An over-coverage was only encountered with the *ssyst* and *strat* designs, and under-coverage was only seen with the *clust* and *clust2st* designs.

With sample sizes over 100, a calibration of confidence intervals for *OA* computed with standard methods is neither needed nor recommended. For illustration only, Bayes uniform prior confidence intervals would have increased the over-coverage of *ssyst*-based intervals by 0.006, but leave the coverage of *strat*-, *clust*- and *clust2st*-based intervals nearly unchanged.

Table 1. Summary statistics of overall accuracy (OA). The true OA is 0.81.

Design	n[a]	OA[b]	eSE[c]	aSE[d]	CCI95[e]
ssyst	828	0.81	0.014	0.013	0.95
strat	828	0.81	0.014	0.013	0.96
clust	828	0.81	0.016	0.016	0.94
clust2st	828	0.81	0.015	0.015	0.96
ssyst	414	0.81	0.019	0.019	0.96
strat	414	0.81	0.019	0.019	0.96
clust	414	0.81	0.022	0.022	0.95
clust2st	414	0.81	0.032	0.030	0.94
ssyst	207	0.81	0.027	0.027	0.96
strat	207	0.81	0.027	0.026	0.96
clust	207	0.81	0.032	0.032	0.94
clust2st	207	0.81	0.032	0.030	0.94

[a]Sample size in units; [b]Overall accuracy; [c]Empirical standard error; [d]Average analytical standard error; [e]Coverage of standard 95% confidence intervals.

3.2. Producer's Accuracy (PA)

All estimates of PA were nearly unbiased. Estimates of apparent bias fluctuate between −0.3% and 0.0% with *ssyst,* between −0.5% and 0.0% with *strat*, and between −0.5% and +0.5% with *clust* and *clust2st* (Table 2). There was no trend in these estimates across classes, but smaller sample sizes typically generated a higher variability in tabled estimates. For the two cluster sampling designs, the apparent bias in *PA* for class A was statistically significant at the 5% level.

With *ssyst* and *strat*, the average aSE was within ±10% of the eSE; with a larger overestimation limited to the largest sample size and class A. In all other cases, an underestimation of approximately 10% was the norm. Standard errors changed with sample size in agreement with expectations. With *clust* and *clust2st* and the two largest sample sizes, aSE and eSE were typically within 10% of each other, but also 10% to 30% greater than standard errors reported for *ssyst* and *strat*. Larger discrepancies between eSE and aSE were limited to the smallest sample size. In class A and *clust*, the average aSE underestimates the eSE by 14%, and in class B and *clust2st* the mean of aSE was 40% greater than the eSE.

A combination of an apparent bias and aSEs that do not match the eSEs sets the stage for under- or over-coverage of estimated confidence intervals. Table 3 confirms this. A coverage that was either significantly above or below the nominal value of 0.95 occurred in 27 out of 48 combinations of design × class × sample size. Class B had the highest number of significant departures (11 out of 12) which agrees with the fact that coverage of a confidence interval computed by standard method deteriorates as the estimated proportion approaches the

limits of 0 and 1. Class A with the second highest PA also had the second highest number of significant cases of either over or under-coverage (9). In classes C and D, the number of significant departures was 5 and 2, respectively.

Table 2. Summary Statistics of producer's accuracy (*PA*). True *PA*s are 0.88 for class A, 0.92 for class B, 0.78 for class C, and 0.75 for class D.

		Class A			Class B			Class C			Class D		
Design	**n[a]**	**PA**	**eSE[b]**	**aSE[c]**	**PA**	**eSE**	**aSE**	**PA**	**eSE**	**aSE**	**PA**	**eSE**	**aSE**
ssyst	828	0.88	0.033	0.03	0.92	0.020	0.02	0.79	0.022	0.022	0.75	0.018	0.02
strat	828	0.89	0.033	0.03	0.92	0.020	0.02	0.78	0.022	0.022	0.75	0.018	0.02
clust	828	0.89	0.036	0.04	0.92	0.022	0.02	0.78	0.027	0.028	0.75	0.025	0.03
clust2st	828	0.89	0.036	0.04	0.92	0.030	0.02	0.78	0.025	0.026	0.75	0.023	0.02
ssyst	414	0.88	0.047	0.05	0.92	0.028	0.03	0.79	0.032	0.030	0.75	0.026	0.03
strat	414	0.88	0.047	0.05	0.92	0.028	0.03	0.79	0.032	0.031	0.75	0.026	0.03
clust	414	0.88	0.051	0.05	0.92	0.031	0.03	0.78	0.039	0.040	0.75	0.035	0.04
clust2st	414	0.89	0.080	0.08	0.92	0.063	0.05	0.78	0.054	0.054	0.75	0.051	0.05
ssyst	207	0.89	0.065	0.07	0.92	0.040	0.04	0.79	0.045	0.045	0.75	0.037	0.04
strat	207	0.89	0.064	0.07	0.92	0.039	0.04	0.79	0.044	0.045	0.75	0.036	0.04
clust	207	0.89	0.071	0.08	0.92	0.044	0.05	0.78	0.055	0.057	0.75	0.050	0.05
clust2st	207	0.89	0.080	0.08	0.92	0.063	0.05	0.78	0.054	0.054	0.75	0.051	0.05

[a]Sample size in units; [b]Empirical standard error; [c]Average analytical standard error.

Table 3. Coverage of nominal 95% confidence interval for PA estimated with standard methods ($CCI95_{SE}$) and coverage of Bayes uniform prior 95% confidence intervals ($CCI95_{BAY}$). Table entries marked with a star (*) have a coverage significantly different from 0.95 at the 5% level or lower.

		Class A		Class B		Class C		Class D	
Design	**n**	**$CCI95_{SE}$**	**$CCI95_{BAY}$**	**$CCI95_{SE}$**	**$CCI95_{BAY}$**	**$CCI95_{SE}$**	**$CCI95_{BAY}$**	**$CCI95_{SE}$**	**$CCI95_{BAY}$**
ssyst	828	0.93*	0.95	0.94	0.96	0.95	0.98*	0.96	0.98*
strat	828	0.92*	0.96	0.93*	0.95	0.95	0.96	0.95	0.98*
clust	828	0.90*	0.94	0.93*	0.94	0.93*	0.94	0.95	0.94
clust2st	828	0.95	0.96	0.99*	0.97*	0.93*	0.94	0.94	0.95
ssyst	414	0.92*	0.96	0.92*	0.96	0.96	0.97*	0.95	0.98*
strat	414	0.91*	0.96	0.92*	0.97*	0.95	0.96	0.95	0.98*
clust	414	0.88*	0.94	0.89*	0.94	0.93*	0.94	0.93*	0.94
clust2st	414	0.95	0.94	0.98*	0.95	0.95	0.96	0.95	0.95
ssyst	207	0.88*	0.88*	0.88*	0.93*	0.94	0.97*	0.95	0.98*
strat	207	0.90*	0.89*	0.85*	0.93*	0.93*	0.97*	0.95	0.97*
clust	207	0.75*	0.90*	0.86*	0.91*	0.92*	0.94	0.91*	0.93*
clust2st	207	0.95	0.94	0.98*	0.95	0.95	0.96	0.95	0.95

The *clust* design accounted for most (11) of the significant departures, followed by *strat* with 7, *ssyst* with 5, and *clust2st* with 4. The number of significant departures increased weakly with a decrease in sample size from 8 (n = 828) to 10 (n = 207). Somewhat surprisingly, the calibration was least effective with the smallest sample size, and most effective with a sample size of 828 with just 3 significant departures compared to 8 before the calibration. The result may reflect that an increase in sample size also improves the accuracy of an estimate of a design-effect.

Following a calibration to Bayes posterior 95% confidence intervals, the number of significant departures from the nominal coverage dropped from 27 to 19. The calibration was most successful in classes A and B with just 3 and 5 post-calibration departures from the target coverage. For class C, the calibration did not improve coverage, and in class D it worsened the coverage with a rise in the number of significant departures from 2 to 7. A calibration was most effective with cluster sampling in that the number of significant departures dropped from 15 to 4. In *ssyst* the calibration was counterproductive: the number of significant departures from 0.95 increased from 5 to 8. In *strat*, calibration did not achieve any improvements. Even for two cases with an important under-coverage in class A (i.e. with *strat*, and *ssyst* and n = 207), the calibration did not address the under-coverage.

If merely improving the coverage is deemed worthwhile, then the calibration achieved this in 25 cases, yet also worsened it in 14 cases. When under-coverage is considered as more serious than over-coverage, then there is ample support for a calibration. Of the 27 cases with a significant departure from the nominal coverage, 24 reported under-coverage. Post calibration, only 4 cases retained a significant under-coverage. In 9 cases, a calibration changed a non-significant departure from 0.95 into a significant over-coverage, almost exclusively in class D.

3.3. User's Accuracy (UA)

Estimates of UA were nearly unbiased in classes B, C and D, but significantly underestimated by 1% to 3% in class A and the two cluster designs (**Table 4**). Analytical standard errors with the *ssyst* and *strat* designs were similar and within 5% of the analytical counterparts, yet with a clear tendency towards a slight overestimation in case of *ssyst*, and a slight underestimation in case of *strat*. Standard errors with *clust* were, as expected, greater than standard errors with *ssyst* and *strat*. In class A, the difference was approximately 80%, and approximately 50% in classes B, C, and D. The average of aSE was close the eSE in all classes and sample sizes with the exception of class A and a sample size of 207. Here the average aSE is approximately 10% greater than the eSE. A more complex pattern emerges for *clust2st* with a substantial (~30%) overestimation in eSE for class A and an even greater (30% - 60%) underestimation in classes B, C, and D.

Confidence intervals for user's accuracy (UA) often (27 times out of 48) had

either a significant under-coverage (18 cases) or a significant over-coverage (9 cases) (Table 5). Most cases (18) came from *clust* (6), and *clust2st* (12). Although *ssyst* and *strat* only contributed 9 cases, they were all with a significant under-coverage. Settings with the smallest sample size accounted for 15 of the 25 cases with a significant over- or under-coverage.

Table 4. Summary Statistics of user's accuracy (UA). True *UA*s are 0.60 for class A, 0.83 for class B, 0.80 for class C, and 0.89 for class D.

		Class A			Class B			Class C			Class D		
Design	**n**	**UA**	**eSE**	**aSE**	**UA**	**eSE**	**aSE**	**UA**	**eSE**	**aSE**	**UA**	**eSE**	**aSE**
ssyst	828	0.60	0.044	0.042	0.83	0.028	0.027	0.80	0.025	0.025	0.89	0.019	0.018
strat	828	0.60	0.044	0.045	0.83	0.028	0.028	0.80	0.025	0.025	0.89	0.019	0.019
clust	828	0.59	0.077	0.080	0.82	0.034	0.034	0.80	0.034	0.034	0.89	0.024	0.025
clust2st	828	0.59	0.053	0.067	0.82	0.054	0.030	0.80	0.042	0.031	0.89	0.049	0.022
ssyst	414	0.60	0.063	0.061	0.83	0.040	0.039	0.80	0.036	0.036	0.89	0.027	0.026
strat	414	0.60	0.063	0.064	0.83	0.039	0.040	0.80	0.036	0.037	0.89	0.027	0.026
clust	414	0.59	0.108	0.109	0.82	0.048	0.048	0.80	0.048	0.048	0.89	0.035	0.033
clust2st	414	0.58	0.115	0.146	0.82	0.111	0.061	0.80	0.086	0.061	0.89	0.099	0.047
ssyst	207	0.60	0.088	0.087	0.83	0.056	0.055	0.80	0.051	0.051	0.89	0.038	0.037
strat	207	0.60	0.088	0.090	0.83	0.056	0.057	0.80	0.051	0.050	0.89	0.037	0.038
clust	207	0.57	0.150	0.163	0.82	0.070	0.071	0.80	0.069	0.068	0.89	0.051	0.050
clust2st	207	0.58	0.115	0.146	0.82	0.111	0.061	0.80	0.086	0.061	0.89	0.099	0.047

Table 5. Coverage of nominal 95% confidence interval estimated for UA with standard methods and coverage of Bayes uniform prior 95% confidence interval based on the effective sample size n_{eff}. Table entries marked with a star (*) have a coverage that is significantly different from 0.95 at the 5% level or lower.

		Class A		Class B		Class C		Class D	
Design	**n**	**CCI95$_{SE}$**	**CCI95$_{BAY}$**	**CCI95$_{SE}$**	**CCI95$_{BAY}$**	**CCI95$_{SE}$**	**CCI95$_{BAY}$**	**CCI95$_{SE}$**	**CCI95$_{BAY}$**
ssyst	828	0.96	0.96	0.95	0.96	0.94	0.95	0.95	0.96
strat	828	0.93*	0.94	0.95	0.95	0.95	0.95	0.95	0.95
clust	828	0.94	0.95	0.94	0.93*	0.94	0.94	0.93*	0.94
clust2st	828	0.86*	0.88*	1.00*	0.99*	0.99*	0.96	1.00*	0.97*
ssyst	414	0.95	0.96	0.95	0.96	0.94	0.95	0.95	0.96
strat	414	0.93*	0.95	0.95	0.95	0.94	0.95	0.96	0.97*
clust	414	0.93*	0.98*	0.94	0.93*	0.94	0.92*	0.94	0.94
clust2st	414	0.87*	0.94	1.00*	0.98*	0.99*	0.96	1.00*	0.97*
ssyst	207	0.94	0.96	0.93*	0.95	0.93*	0.94	0.92*	0.96
strat	207	0.93*	0.96	0.89*	0.95	0.93*	0.97*	0.89*	0.96
clust	207	0.87*	0.87*	0.91*	0.92*	0.92*	0.91*	0.89*	0.91*
clust2st	207	0.87*	0.94	1.00*	0.98*	0.99*	0.96	1.00*	0.97*

Bayes uniform prior confidence intervals had, overall, better coverage properties. The calibration reduced the number of significant departures (from 0.95) from 17 to 10; with an even split between under- and over-coverage. A calibration was most effective for the *strat* and *clust2st* designs, but nearly counterproductive for the *clust* design with 8 cases of a significant under-coverage compared to 6 cases with standard confidence intervals. The calibration was most effective with the smallest sample size. Here it achieved a reduction from 12 to 7 cases with a significant departure from the target coverage.

On balance, the coverage of Bayes uniform prior confidence intervals was in 26 cases closer to the 0.95 target than a standard confidence interval. Calibration also reduced the number of cases with a significant under-coverage from 18 to 10. It only generated a single case with a significant over-coverage where there was none before.

3.4. Reference Area Proportions (P_{ref})

The average estimate of an area proportion was within 0.002 of the true value. No estimate of apparent bias was significant at the 5% level (Table 6). With *ssyst* (and post-stratification), the average aSE was within 0.001 from eSE. With *strat*, the average aSE was also close to the eSE but with an overestimation of 0.002 in class D and an underestimation of 0.002 in class A. Monte-Carlo simulations, with a multinomial distribution and true class proportions, suggest that with n = 828 or n = 414 a difference between aSE and eSE of 0.002 would be significant at the 5% level, while a difference of 0.003 would be significant with n = 207. Accordingly, a significant underestimation of the empirical standard error was isolated to settings with *clust2st* in classes A-C and sample sizes of 414 and 207 units. With *clust2st* and sample size 828, the estimates of eSE and aSE were at most 0.001 apart.

Coverage with a majority (37 of 48) of confidence intervals for P_{ref} was between 0.94 and 0.96 (non-significant different from 0.95) (Table 7). Of the 11 confidence intervals with a coverage deviating significantly from 0.95, seven exhibited under-coverage and four over-coverage. Most of the cases with a coverage significantly different from 0.95 were with the *strat* and *clust* and sample sizes of 414 and 207 units. No significant departures from 0.95 were found with *ssyst* and only 2 with *clust2st*, both cases had over-coverage.

For P_{ref} a calibration of the standard confidence intervals was counterproductive. With Bayes uniform prior intervals, the number of cases with a significant over- or under-coverage rose from 11 to 39 with over-coverage. Seven cases, with a significant under-coverage in standard intervals had a significant over-coverage with Bayes intervals. In just six cases was the coverage of a Bayes interval closer to 0.95 than with standard interval, in 35 cases it was further away.

4. Discussion

A poor coverage of standard confidence interval for a proportion estimated from

Table 6. Summary Statistics of reference class proportions (P_{ref}). True reference class proportions are 0.10 (A), 0.20 (B), 0.30 (C), and 0.40 (D).

		Class A			Class B			Class C			Class D		
Design	n	P_{ref}	eSE[b]	aSE[c]	P_{ref}	eSE	aSE	P_{ref}	eSE	aSE	P_{ref}	eSE	aSE
ssyst	828	0.10	0.008	0.007	0.20	0.008	0.007	0.30	0.011	0.011	0.40	0.011	0.011
strat	828	0.10	0.008	0.007	0.20	0.008	0.008	0.30	0.011	0.011	0.40	0.010	0.011
clust	828	0.10	0.024	0.024	0.20	0.026	0.026	0.30	0.029	0.028	0.40	0.041	0.041
clust2st	828	0.10	0.022	0.020	0.20	0.023	0.022	0.30	0.024	0.023	0.40	0.037	0.036
ssyst	414	0.10	0.011	0.010	0.20	0.011	0.011	0.30	0.016	0.016	0.40	0.016	0.016
strat	414	0.10	0.012	0.011	0.20	0.011	0.011	0.30	0.016	0.016	0.40	0.014	0.016
clust	414	0.10	0.034	0.033	0.20	0.037	0.037	0.30	0.041	0.039	0.40	0.059	0.057
clust2st	414	0.10	0.051	0.042	0.20	0.052	0.045	0.30	0.053	0.047	0.40	0.077	0.074
ssyst	207	0.10	0.015	0.015	0.20	0.015	0.015	0.30	0.022	0.022	0.40	0.023	0.022
strat	207	0.10	0.017	0.015	0.20	0.015	0.015	0.30	0.022	0.023	0.40	0.020	0.023
clust	207	0.10	0.048	0.047	0.20	0.052	0.054	0.30	0.058	0.058	0.40	0.083	0.083
clust2st	207	0.10	0.051	0.042	0.20	0.052	0.045	0.30	0.053	0.047	0.40	0.077	0.074

Table 7. Coverage of nominal 95% confidence interval estimated for P_{ref} with standard methods and coverage of Bayes uniform prior 95% confidence interval based on the effective sample size n_{eff}. Table entries marked with a star (*) have a coverage that is significantly different from 0.95 at the 5% level or lower.

		Class A		Class B		Class C		Class D	
Design	n	$CCI95_{SE}$	$CCI95_{BAY}$	$CCI95_{SE}$	$CCI95_{BAY}$	$CCI95_{SE}$	$CCI95_{BAY}$	$CCI95_{SE}$	$CCI95_{BAY}$
ssyst	828	0.96	1.00*	0.96	1.00*	0.96	0.99*	0.96	0.99*
strat	828	0.97*	1.00*	0.95	1.00*	0.96	0.99*	0.92*	0.98*
clust	828	0.94	0.98*	0.94	0.98*	0.94	0.96	0.95	0.96
clust2st	828	0.95	0.97*	0.95	0.96	0.96	0.96	0.96	0.96
ssyst	414	0.96	1.00*	0.96	1.00*	0.95	0.99*	0.96	0.99*
strat	414	0.97*	1.00*	0.95	1.00*	0.94	0.99*	0.91*	0.98*
clust	414	0.92*	0.97*	0.94	0.98*	0.96	0.97*	0.95	0.96
clust2st	414	0.96	0.98*	0.96	0.98*	0.97*	0.97*	0.95	0.96
ssyst	207	0.96	1.00*	0.95	1.00*	0.95	0.99*	0.96	0.99*
strat	207	0.96	1.00*	0.95	1.00*	0.94	0.99*	0.91*	0.97*
clust	207	0.87*	0.98*	0.92*	0.97*	0.93*	0.97*	0.94	0.95
clust2st	207	0.96	0.98*	0.96	0.98*	0.97*	0.97*	0.95	0.96

a small sample size or a proportion close to either 0 or 1 is to be expected (Fleiss et al., 2013). Effective methods of calibration have been worked out for settings when either exact methods are feasible, or the distribution of the pivotal statistic is known (Newcombe, 1998). Unfortunately, these methods do not apply to es-

timates of thematic accuracy obtained under some probability sampling designs because the sampling distribution of an estimated proportion is typically analytically intractable. Take, for example, the estimate of OA under stratified random sampling by map-class. First, the number of sample units by reference class in a map-class is distributed as per a multinomial given the known stratum sample size but unknown reference class proportions. Second, the number of correct classifications in a map class now depends not only on the unknown number of sample units by reference but also on the latent class specific accuracies associated with sampled units. With settings as in this study, a normal, a beta, and an inverse Gaussian distribution would fit the sample distribution of OA equally well (based on Akaike's information criterion); a formal test of a normal distribution would have been rejected at the 0.01 level due to a negative skewness of −0.08 (D'Agostino, 1990). A spatially varying latent accuracy (Khatami et al., 2017) with a limited range of autocorrelation made the simulated land-cover classification process realistic (Stehman & Wickham, 2011) but also makes the expected sampling distribution of an accuracy estimate analytically intractable. Our reliance on standard confidence interval with the implicit assumption of a normal distribution is therefore tenuous, in particular with regards to small sample sizes (Neyman, 1934).

Calibration methods for confidence intervals proposed for complex sampling designs do not apply directly to inference about classification accuracy because they are tailored to a binary variable with a dependence structure limited to a positive intra-cluster correlation coefficient (see Franco et al., 2019, and references therein). Hence, a fully effective and reliable calibration method for confidence intervals of an accuracy statistic may not exist for land-cover map projects. The highly variable number and diversity of thematic classes, a multitude of possible spatial covariance structures in both the reference map and the classification process, paired with typically relatively small affordable sample sizes for accuracy assessment (Congalton, 1991; Morales-Barquero et al., 2019; Olofsson, Foody, Stehman, & Woodcock, 2013) suggest that no single calibration will be overall best.

In an application an analyst will, of course, not know if a calibration of confidence intervals is successful or not and may stick with standard intervals or resort to computational-intensive resampling methods (Magnussen & Köhl, 2002; Magnussen, Stehman, Corona, & Wulder, 2004). The performance of a calibration method requires simulated sampling in populations with the map and reference class known for every unit in the population. A simulation study like this one is expedient given available software allowing us to introduce a spatial covariance structure at all levels, and distributions of the latent class specific probabilities of a correct classification (Gupta & Nagar, 1999; Li, 2007; White & Ghosh, 2009). Simulated sampling from two classified images, one of which is treated as a reference map, is even more expedient when available (Andersen, 1998; Khatami et al., 2016).

An across-the-board calibration of all confidence intervals in an accuracy as-

sessment of a land-cover map product is neither warranted nor necessary. With sample sizes of 20 units or more per map class and the four sampling designs used here, there would be no strong impetus to calibrate a confidence interval for an OA. It is an open question whether this applies equally to designs with double sampling (Kalkhan, Reich, & Stohlgren, 1998; Westfall, Lister, Scott, & Weber, 2019), or to designs with a selection of units based on the distribution of remotely sensed auxiliary variables (Grafström, Saarela, & Ene, 2014; Pagliarella et al., 2016). The above recommendation for OA also applies, by and large, to P_{ref} since the sampling distribution is approximately normal for sample sizes that are not too small (say > 20 in most rare class) and an area proportion above 0.05.

For PA and UA, a calibration of confidence intervals is encouraged as the coverage, at least in this study, could—more often than not—be improved. Calibration with Bayes intervals is obviously not a win-win, given a general tendency to widen the confidence interval with a concomitant over-coverage for standard intervals that achieved their nominal coverage. On balance, for risk adverse users/producers, a calibration seems in order.

In sampling with a single binary variable, it is straightforward to compute the design effect for any design (Fuller, 2011) and from there Bayes intervals, as demonstrated recently by Franco et al. (2019)—with the caveat that an estimate of the design effect from a single sample will be biased as it is a ratio of two random variables (variances). For estimates of multi-class thematic accuracies, we do not yet have a statistic for gauging the design efficiency. The proposed (novel) square root transformation of an estimate of the design effect lacks a foundation in theory. It has only intuitive appeal since the width of a confidence interval has a stronger correlation with the square root of a variance that the variance in an assumed distribution for an estimated proportion. A search for more efficient transformations firmer grounded in theory is recommended.

A still popular measure of accuracy, the Kappa coefficient, was left out on purpose as repeated studies have shown that it is not what it pretends to be: a measure of accuracy corrected for chance agreement (Foody, 2020; Lin, Hedayat, Sinha, & Yang, 2002; Pontius & Millones, 2011).

The main objective of this study is to encourage a calibration of confidence intervals for multi-class estimates of accuracy. A single simulation study is a start, but more concerted efforts on the side of statistical theory and methods are still needed.

5. Conclusion

A calibration of confidence intervals for sample based estimates of user's and producer's accuracy can be accomplished with a few additional statistics that are easy to compute. It is recommended to calibrate confidence intervals when the sample size in a specific combination of map and reference class drops below 20. The benefit is confidence intervals that are more likely to achieve the nominal coverage than a non-calibrated interval.

Acknowledgements

Valued contributions and improvements to the original submission were provided by the journal reviewers.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Agresti, A., & Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician, 54,* 280-288. https://doi.org/10.1080/00031305.2000.10474560

Andersen, G. L. (1998). *Classification and Estimation of Forest and Vegetation Variables in Optical High Resolution Satellites: A Review of Methodologies* (pp. 1-19). IR-98-085. Laxenburg, AU. http://pure.iiasa.ac.at/id/eprint/5566/1/IR-98-085.pdf

Antal, E. (2011). A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population. *Journal of the American Statistical Association, 106,* 534-543. https://doi.org/10.1198/jasa.2011.tm09767

Barth, A., & Ståhl, G. (2011). Determining Sample Size in National Forest Inventories by Cost-plus-Loss Analysis: An Exploratory Case Study. *European Journal of Forest Research, 131,* 339-346. https://doi.org/10.1007/s10342-011-0505-5

Binaghi, E., Madella, P., Montesano, M. G., & Rampini, A. (1997). Fuzzy Contextual Classification of Multisource Remote Sensing Images. *IEEE Geoscience and Remote Sensing, 35,* 326-339. https://doi.org/10.1109/36.563272

Breidt, F. J., & Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science, 32,* 190-205. https://doi.org/10.1214/16-STS589

Chambers, R., & Dorfman, A. (2003). *Robust Sample Survey Inference via Bootstrapping and Bias Correction: The Case of the Ratio Estimator*. S3RI Methodology Working Papers, M03/13, S3RI Methodology Working Papers, M03/13. Southampton: Southampton Statistical Sciences Research Institute. https://www.researchgate.net/publication/242367532

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

Congalton, R. G. (1991). A Review of Assessing the Accuracy of Classification of Remotely Sensed Data. *Remote Sensing of Environment, 37,* 35-45. https://doi.org/10.1016/0034-4257(91)90048-B

Congalton, R. G. (2001). Accuracy Assessment and Validation of Remotely Sensed and Other Spatial Information. *International Journal of Wildland Fire, 10,* 321-328. https://doi.org/10.1071/WF01031

Conti, P. L., & Marella, D. (2014). Inference for Quantiles of a Finite Population: Asymptotic versus Resampling Results. *Scandinavian Journal of Statistics, 42,* 545-561. https://doi.org/10.1111/sjos.12122

Czaplewski, R. L. (1994). *Variance Approximations for Assessment of Classification Accuracy*. Research Paper, RM-316, Fort Collins.

Czaplewski, R. L. (2003). Statistical Design and Methodological Considerations for the

Accuracy Assessment of Maps of Forest Conditions. In M. A. Wulder, & J. Franklin (Eds.), *Remote Sensing of Forest Environments: Concepts and Case Studies* (pp. 115-129). Boston, MA: Kluwer Acad. Publisher. https://doi.org/10.1007/978-1-4615-0306-4_5

D'Agostino, R. B. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician, 44,* 316-321. https://doi.org/10.1080/00031305.1990.10475751

Dobbertin, M., & Biging, G. S. (1996). A Simulation Study of the Effect of Scene Autocorrelation, Training Sample Size and Sampling Method on Classification Accuracy. *Canadian Journal of Remote Sensing, 22,* 360-367. https://doi.org/10.1080/07038992.1996.10874660

Esty, W. W. (1982). Confidence Intervals for the Coverage of Low Coverage Samples. *Annals of Statistics, 10,* 190-196. https://doi.org/10.1214/aos/1176345701

Fattorini, L. (2015). Design-Based Methodological Advances to Support National Forest Inventories: A Review of Recent Proposals. *iForest-Biogeosciences and Forestry, 8,* 6-11. http://iforest.sisef.org/contents/?id=ifor1239-007 https://doi.org/10.3832/ifor1239-007

Finley, A. O., Banerjee, S., Ek, A. R., & McRoberts, R. E. (2008). Bayesian Multivariate Process Modeling for Prediction of Forest Attributes. *Journal of Agricultural Biological and Environmental Statistics, 13,* 60-83. https://doi.org/10.1198/108571108X273160

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons.

Fodé, T., & Louis-Paul, R. (2014). Some New Random Effect Models for Correlated Binary Responses. *Dependence Modeling, 2,* 15. https://doi.org/10.2478/demo-2014-0006

Foody, G. M. (2020). Explaining the Unsuitability of the Kappa Coefficient in the Assessment and Comparison of the Accuracy of Thematic Maps Obtained by Image Classification. *Remote Sensing of Environment, 239,* Article ID: 111630. https://doi.org/10.1016/j.rse.2019.111630

Franco, C., Little, R. J. A., Louis, T. A., & Slud, E. V. (2019). Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys. *Journal of Survey Statistics and Methodology, 7,* 334-364. https://doi.org/10.1093/jssam/smy019

Fuller, W. A. (2011). *Sampling Statistics*. New York: Wiley.

Grafström, A., Saarela, S., & Ene, L. T. (2014). Efficient Sampling Strategies for Forest Inventories by Spreading the Sample in Auxiliary Space. *Canadian Journal of Forest Research, 44,* 1156-1164. https://doi.org/10.1139/cjfr-2014-0202

Gupta, A. K., & Nagar, D. K. (1999). *Matrix Variate Distributions*. Boca Raton: Chapman & Hall/CRC.

Kalkhan, M. A., Reich, R. M., & Stohlgren, T. J. (1998). Assessing the Accuracy of Landsat Thematic Mapper Classification Using Double Sampling. *International Journal of Remote Sensing, 19,* 2049-2060. https://doi.org/10.1080/014311698214857

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A Meta-Analysis of Remote Sensing Research on Supervised Pixel-Based Land-Cover Image Classification Processes: General Guidelines for Practitioners and Future Research. *Remote Sensing of Environment, 177,* 89-100. https://doi.org/10.1016/j.rse.2016.02.028

Khatami, R., Mountrakis, G., & Stehman, S. V. (2017). Mapping Per-Pixel Predicted Accuracy of Classified Remote Sensing Images. *Remote Sensing of Environment, 191,* 156-167. https://doi.org/10.1016/j.rse.2017.01.025

Korn, E. L., & Graubard, B. I. (1998). Confidence Intervals for Proportions with Small

Expected Number of Positive Counts Estimated from Survey Data. *Survey Methodology, 24,* 193-201. https://www150.statcan.gc.ca/n1/pub/12-001-x/1998002/article/4356-eng.pdf

Li, W. D. (2007). A Fixed-Path Markov Chain Algorithm for Conditional Simulation of Discrete Spatial Variables. *Mathematical Geology, 39,* 159-176. https://doi.org/10.1007/s11004-006-9071-7

Lin, L., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical Methods in Assessing Agreement: Models, Issues and Tools. *Journal of the American Statistical Association, 97,* 257-270. https://doi.org/10.1198/016214502753479392

Magnussen, S., & Köhl, M. (2002). Polya Posterior Frequency Distributions for Stratified Double Sampling of Categorical Data. *Forest Science, 48,* 569-581.

Magnussen, S., Stehman, S. V., Corona, P., & Wulder, M. A. (2004). A Pòlya-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.

McDonald, J. B., & Xu, Y. J. (1995). A Generalization of the Beta Distribution with Applications. *Journal of Econometrics, 66,* 133-152. https://doi.org/10.1016/0304-4076(94)01612-4

McRoberts, R. E., Stehman, S. V., Liknes, G. C., Næsset, E., Sannier, C., & Walters, B. F. (2018). The Effects of Imperfect Reference Data on Remote Sensing-Assisted Estimators of Land Cover Class Proportions. *ISPRS Journal of Photogrammetry and Remote Sensing, 142,* 292-300. https://doi.org/10.1016/j.isprsjprs.2018.06.002

Miao, W., & Gastwirt, J. L. (2004). The Effect of Dependence on Confidence Intervals for a Population Proportion. *The American Statistician, 58,* 124-130. https://doi.org/10.1198/0003130043303

Morales-Barquero, L., Lyons, M. B., Phinn, S. R., & Roelfsema, C. M. (2019). Trends in Remote Sensing Accuracy Assessment Approaches in the Context of Natural Resources. *Remote Sensing, 11,* 2305. https://doi.org/10.3390/rs11192305

Newcombe, R. G. (1998). Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine, 17,* 857-872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, 97,* 558-606. https://doi.org/10.2307/2342192

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 Sensing of Environment, 129,* 122-131. https://doi.org/10.1016/j.rse.2012.10.031

Olofsson, P., Foody, G., Herold, M., Stehman, S., Woodcock, C., & Wulder, M. (2014). Good Practices for Estimating Area and Assessing Accuracy of Land Change. *Remote Sensing of Environment, 148,* 42-57. https://doi.org/10.1016/j.rse.2014.02.015

Pagliarella, M. C., Sallustio, L., Capobianco, G., Conte, E., Corona, P., Fattorini, L., & Marchetti, M. (2016). From One- to Two-Phase Sampling to Reduce Costs of Remote Sensing-Based Estimation of Land-Cover and Land-Use Proportions and Their Changes. *Remote Sensing of Environment, 184,* 410-417. https://doi.org/10.1016/j.rse.2016.07.027

Pontius, R. G., & Millones, M. (2011). Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment. *International Journal of Remote Sensing, 32,* 4407-4429. https://doi.org/10.1080/01431161.2011.552923

Rao, J. N., & Wu, C. (1988). Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association, 83,* 231-241. https://doi.org/10.1080/01621459.1988.10478591

Rao, J., & Wu, C. (2010). Bayesian Pseudo-Empirical-Likelihood Intervals for Complex Surveys. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72,* 533-544. https://doi.org/10.1111/j.1467-9868.2010.00747.x

Ricotta, C. (2004). Evaluating the Classification Accuracy of Fuzzy Thematic Maps with a Simple Parametric Measure. *International Journal of Remote Sensing, 25,* 2169-2176. https://doi.org/10.1080/01431160310001618130

Román-Montoya, Y., Rueda, M., & Arcos, A. (2008). Confidence Intervals for Quantile Estimation Using Jackknife Techniques. *Computational Statistics, 23,* 573-585. https://doi.org/10.1007/s00180-007-0099-z

Shao, J. (1996). Resampling Methods in Sample Surveys. *Statistics, 27,* 203-254. https://doi.org/10.1080/02331889708802523

Stehman, S. V. (1992). Comparison of Systematic and Random Sampling for Estimating the Accuracy of Maps Generated from Remotely Sensed Data. *Photogrammetric Engineering and Remote Sensing, 58,* 1343-1350. https://www.asprs.org/wp-content/uploads/pers/1992journal/sep/1992_sep_1343-1350.pdf

Stehman, S. V. (1997). Estimating Standard Errors of Accuracy Assessment Statistics under Cluster Sampling. *Remote Sensing of Environment, 60,* 258-269. http://www.sciencedirect.com/science/article/B6V6V-3SWK0SH-4/2/6a55da419a54bd4682a0024b12a8ac14 https://doi.org/10.1016/S0034-4257(96)00176-9

Stehman, S. V. (1999). Basic Probability Sampling Designs for Thematic Map Accuracy Assessment. *International Journal of Remote Sensing, 20,* 2423-2441. https://doi.org/10.1080/014311699212100

Stehman, S. V. (2012). Impact of Sample Size Allocation When Using Stratified Random Sampling to Estimate Accuracy and Area of Land-Cover Change. *Remote Sensing Letters, 3,* 111-120. https://doi.org/10.1080/01431161.2010.541950

Stehman, S. V., & Czaplewski, R. L. (1998). Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sensing of Environment, 64,* 331-344. https://doi.org/10.1016/S0034-4257(98)00010-8

Stehman, S. V., & Foody, G. M. (2019). Key Issues in Rigorous Accuracy Assessment of Land Cover Products. *Remote Sensing of Environment, 231,* Article ID: 111199. https://doi.org/10.1016/j.rse.2019.05.018

Stehman, S. V., & Wickham, J. D. (2011). Pixels, Blocks of Pixels, and Polygons: Choosing a Spatial Unit for Thematic Accuracy Assessment. *Remote Sensing of Environment, 115,* 3044-3055. https://doi.org/10.1016/j.rse.2011.06.007

Stehman, S. V., Wickham, J. D., Fattorini, L., Wade, T. D., Baffetta, F., & Smith, J. H. (2009). Estimating Accuracy of Land-Cover Composition from Two-Stage Cluster Sampling. *Remote Sensing of Environment, 113,* 1236-1249. http://www.sciencedirect.com/science/article/B6V6V-4VY16DK-1/2/9b60c8f1154f77615e7dc0ec2f6e5af6 https://doi.org/10.1016/j.rse.2009.02.011

Tan, W. Y. (1969). Note on the Multivariate and the Generalized Multivariate Beta Distributions. *Journal of the American Statistical Association, 64,* 230-241. https://doi.org/10.1080/01621459.1969.10500966

Wendell, J. P., & Schmee, J. (2001). Likelihood Confidence Intervals for Proportions in

Finite Populations. *The American Statistician, 55,* 55-61. https://doi.org/10.1198/000313001300339941

Westfall, J. A., Lister, A. J., Scott, C. T., & Weber, T. A. (2019). Double Sampling for Post-Stratification in Forest Inventory. *European Journal of Forest Research, 138,* 375-382. https://doi.org/10.1007/s10342-019-01171-9

White, G., & Ghosh, S. K. (2009). A Stochastic Neighborhood Conditional Autoregressive Model for Spatial Data. *Computational Statistics and Data Analysis, 53,* 3033-3046. https://doi.org/10.1016/j.csda.2008.08.010

Wulder, M. A., Franklin, S. E., White, J., 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. https://doi.org/10.1080/01431160500185284