Quantifying Uncertainty for Estimates Derived from Error Matrices in Land Cover Mapping Applications: The Case for a Bayesian Approach

Jordan Phillipson[1], Gordon Blair[1], Peter Henrys[2]

[1] *School of Computing and Communications, Lancaster University, Lancaster*
[2] *Centre for Ecology and Hydrology, Lancaster*
Email:j.phillipson@lancaster.ac.uk

Abstract. The use of land cover mappings built using remotely sensed imagery data has become increasingly popular in recent years. However, these mappings are ultimately only models. Consequently, it is vital for one to be able to assess and verify the quality of a mapping and quantify uncertainty for any estimates that are derived from them in a reliable manner.

For this, the use of validation sets and error matrices is a long standard practice in land cover mapping applications. In this paper, we review current state of the art methods for quantifying uncertainty for estimates obtained from error matrices in a land cover mapping context. Specifically, we review methods based on their transparency, generalisability, suitability when stratified sampling and suitability in low count situations. This is done with the use of a third-party case study to act as a motivating and demonstrative example throughout the paper.

The main finding of this paper is there is a major issue of transparency for methods that quantify uncertainty in terms of confidence intervals (frequentist methods). This is primarily because of the difficulty of analysing nominal coverages in common situations. Effectively, this leaves one without the necessary tools to know when a frequentist method is reliable in all but a few niche situations. The paper then discusses how a Bayesian approach may be better suited as a default method for uncertainty quantification when judged by our criteria.

Key words: Uncertainty quantification, map assessment, Bayesian, land cover maps.

1 Introduction

National and global scale land cover mappings based on remotely sensed imagery have been shown to be directly useful in many environmental science applications including: carbon emission monitoring [1]–[4], forest monitoring [5], [6], modelling of soil properties [7], land change detection [8]–[10], climate dynamics [11]–[14], natural hazard assessment [15], [16], agriculture, water/wetland monitoring [17], [18] and biodiversity studies [19], [20]. Because of this, along with the increasing availability of satellite imagery data, national and global scale land cover mappings have attracted significant attention from researchers in the environmental sciences in recent decades.

Satellite imagery alone though is generally not enough to build reliable and meaningful land cover maps. One must also collect reference samples (sometimes referred to as ground truth samples) to both train (when using supervised learning techniques) and validate maps.

When estimating standard performance metrics and area estimates in land cover mapping (e.g. user, producer and overall accuracies and area estimates), a popular method of estimating these quantities is with the use of a post-hoc validation set. This is done by comparing the ground-truth values of these validation samples with their respective predicted values and inferring estimates based on the forms of agreements and disagreements between these values. Since these estimations usually only require the number of different types of agreements and disagreements, it is often convenient to tabulate these results. When this is the case, the subsequent tabulated results are often presented as an error (or confusion) matrix. As a validation set is itself only a sample, such estimations are inevitably going to have uncertainties associated with them. In order for policy makers, stakeholders and other users to have the appropriate level of confidence in such estimations, it is vital that any quantification of these uncertainties are justified.

A major advantage of a using a post-hoc validation sample for estimating these quantities (and subsequently quantifying the associated uncertainties) is that it does not place any requirements on the methods used to create the strata. This means that one is free to build mappings with machine learning techniques (such as Random Forests, Support Vector Machines and Artificial Neural Networks [21]) without needing to be concerned that many of these methods can be black box in nature. Another advantage is that one has much more freedom when collecting training samples. This is because one is not restricted to the specific stochastic structures of sampling, which are necessary when inferring uncertainties directly from a model. This is especially important when dealing with machine learning techniques, as we often have to rely on cheaper, less structured methods, of collecting training data (e.g. polygon sampling, using found data, etc.). Thirdly, it is possible to apply this method with nothing more than the results from an error matrix. This especially useful when analysing historical or third party maps.

The current recommended approach of uncertainty quantification from error matrices is to take a frequentist approach and rely on asymptotic normality estimates to provide confidence intervals [22], [23] . The drawback of this approach is that it is not appropriate when relevant entries of an error matrix are not sufficiently large. Furthermore, because relevant events may be rare (e.g. instances of incorrect labelling between two contrast classes) additional sampling of validation data is not always a practical solution to this problem. Whilst there are methods for dealing with low entry counts in simple situations [24]–[26] , complications arise when one needs to correct estimates for disproportionate sampling across the strata. The main consequences of these complications is that the resultant confidence intervals are of little practical use, either due to their excessively cautious nature, or by the fact the fundamental statement that is implicitly made by confidence intervals (i.e. nominal coverage) cannot be reliably verified.

The goal of this paper is to review existing methods for quantifying uncertainty with the aim of providing an approach that can deal with these aforementioned complica-

tions. In what follows, we firstly review the current recommended practice for uncertainty quantification under a frequentist perspective based on the following criteria: transparency, generalisability, suitability when stratified sampling and suitability in low count situations (see section 2 for further details).
We then make a case that a Bayesian approach is more suited as a default for method uncertainty quantification when judged by these criteria.

2 Terminology and formulating the problem.

We begin by supposing that we have k mutually exclusive strata and that, within in each stratum, instances can be classified as belonging to one of c discrete values. Typically, in land cover mapping applications these instances are single pixels or small clusters of pixels, each approximately of equal size. For the sake of convince we assume that these instances are always at the single pixel level and hence refer to instances as pixels.

Let $\boldsymbol{p}_i \in [0,1]^c, i = 1,..,k$ denote the proportion vector for population i where $(\boldsymbol{p}_i)_j$ is the proportion of pixels that are within strata i that belong to class j where $j = 1,..,c$. We define a global quantity as any quantity that can be expressed as a function of $\boldsymbol{p} := (\boldsymbol{p}_1', \dots, \boldsymbol{p}_k')'$. I.e. a global quantity is any quantity that can be expressed in the form $g(\boldsymbol{p})$. Examples of global quantities in land cover mapping applications are performance metrics such as user, producer, and overall accuracies along with large scale measurements such as the total areas. In practice not all entries of $\boldsymbol{p}$ will be needed in the calculation of g. Here we will write "relevant $\boldsymbol{p}$" as a short hand to "all entries $\boldsymbol{p}$ that are necessary in the calculation of g".

Next suppose that for each of the k strata, we draw a random sample of pixels (with replacement) of size $n_1,\dots n_k$ respectively and let $\boldsymbol{x}_i$ denote the response vector for strata i with $(\boldsymbol{x}_i)_j$ indicating the number of the n_i pixels drawn from strata i that belong to class j.

Within this notation, the aim of this paper is to review current methods of quantifying uncertainty for estimates of $g(\boldsymbol{p})$ made with $\boldsymbol{n} := (n_1,\dots n_k)'$ and $\boldsymbol{x} := (\boldsymbol{x}_1', \dots, \boldsymbol{x}_k')'$.

Note that estimates obtained from an error matrix are a special case of this whereby $k = c$ and $\boldsymbol{x}$ is a vector representation of said error matrix. The evaluation of methods discussed in this paper will be based on the following four criteria.

Transparency – the extent to which one can explicitly state, justify and analyse any assumptions or choices necessary within the method. This is an important criterion as this will influence how likely end users will have confidence in the results of methods.

Generalisability – the suitability and ease of applying a method when considering a wide variety of global quantities or when estimates for global quantities are part of a modelling chain. Essentially, this criterion is included to assess how flexible a given method is to choices of g. This important land cover mapping applications as g is regularly a non-trivial function of the components of $\boldsymbol{p}$ (e.g ratios, weighted sums etc.). In addition, global quantities are regularly used inputs in other models. Hence, it is common that one may wish to propagate the uncertainty for an estimate of g into another quantity.

Suitability for stratified sampling - how appropriate the method is in situations when a stratified random sampling has taken place. Stratified random sampling has

been common practice when collecting test samples as it allows for a more efficient reduction in uncertainty under the currently recommended approach [27]. Hence, it is important that a method of uncertainty quantification can also handle the case of stratified random sampling in order to similar advantage of these practices.

Suitability in low count situations - how appropriate the method is in situations when relevant entries (or combinations of entries) in $\boldsymbol{x}$ are close to, or exactly, zero (around 5 or less). Note, that low sample sizes can cause low count situations but these are not the same thing. For example, a sample of 25 success and 25 failures is not a low count situation but a sample of 499 successes and 1 failure would be a low count situation. It is important that a method of uncertainty quantification can handle low count situations as there serval naturally occurring factors that make them quite frequent. Such factors include a demand for higher resolutions (e.g. thematic, temporal), the relatively high cost of test sampling (reducing the total sample size) and situations when a single class dominates a stratum (making the alternative classes in said strum rare). The latter factor here is interconnected with the efficiency gains that can arise from stratified random sampling. This is because stratified random sampling is most effective when one can create strata in which a single class of pixel heavily dominates each stratum. However, such a stratification is likely to induce a low count situation. This can lead to a peculiar situation in where one can be a victim of one's own success when quantifying uncertainty with a method that cannot handle low count situations.

3 A motivating example: Georgian deforestation

To motivate the work, we consider an example case study of estimating the total deforestation with the use of a land cover change map of Georgia [28] .
This case study was chosen as it provides an example in which stratified sampling has taken place and one is in a low count situation for some of the entries of the error matrix. The general problem of monitoring deforestation plays an important role in estimating carbon emissions and is now required as part of recent EU policy [29].

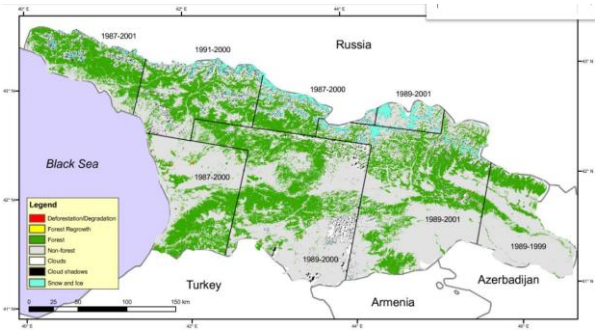

Fig 1 Change map for Georgia, from circa 1990 to 2000 as presented in [28]

Table 1. Error table for the map presented in Fig.1. 1 = forest-to-non-forest; 2 = stable forest; 3 = stable non-forest. W_i denotes the total area of the predicted classes in hectares.

		Reference				
		(1)	*(2)*	*(3)*	n_i	W_i
Prediction	*(1)*	51	23	13	87	22,044
	(2)	0	416	15	431	2,694,787
	(3)	1	20	410	431	4,071,576
	Total	52	459	438	949	6,788,387

For the sake of brevity, we shall only consider providing uncertainty quantification for estimates of the total area of deforestation along with the user accuracy and producer accuracy for the deforestation class. In terms of our notation, we define the total area ($\mathcal{A}_1$) , user accuracy ($\mathcal{U}_1$) and producer accuracy $\mathcal{P}_1$ for the forest-to-non-forest class as

$$\mathcal{A}_1 = \sum_{i=1}^{k} W_i(\boldsymbol{p}_i)_1 \,, \quad \mathcal{U}_1 \coloneqq (\boldsymbol{p}_1)_1 \,, \quad \mathcal{P}_1 = \frac{W_1\,(\boldsymbol{p}_1)_1}{\sum_{i=1}^{k} W_i(\boldsymbol{p}_i)_1} = \frac{W_1\mathcal{U}_1}{\mathcal{A}_1}$$

which we need to estimate from $\boldsymbol{x} \coloneqq (\boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{x}_3')'$ with

$$\boldsymbol{x}_1 = (51\,, 23, 13\,)', \qquad \boldsymbol{x}_2 = (0, 416, 15)', \qquad \boldsymbol{x}_3 = (1{,}20{,}410)'$$

We chose these accuracy quantities as the they are standard practice in many land cover mapping applications and will allow us to demonstrate how different methods behave when assessing them against our chosen criteria. The user accuracy for the forest-to-non-forest class is intended to act as simple base case. The total area of deforestation is a quantity in this case in which we are in a low count situation and must account for stratified sampling for a relatively simple function (i.e. a weighted sum). The producer accuracy has the same qualities as total area but considers a slightly more complex case of g that involves a ratio of two unknown values. We also make a note that only $(\boldsymbol{p}_1)_1$ is relevant to $\mathcal{U}_1$ and $(\boldsymbol{p}_i)_1, i = 1{,}2{,}3$ are relevant to $\mathcal{A}_1$ and $\mathcal{P}_1$.

4 Methods of uncertainty quantification

One way of quantifying uncertainty is to take a frequentist approach and use measures of uncertainty such as confidence intervals. Here the unknown value of $g(\boldsymbol{p})$ is assumed fixed and confidence intervals are probabilistic statements made in relation to the test sample, to which $\boldsymbol{x}$ is one realisation of this process.

It is here that we introduce the concept of nominal coverage. Suppose we have a method of generating confidence intervals for $g(\boldsymbol{p})$ and we repeat a sampling process a large number of times to generate a large number of test samples. Next suppose one was to apply said method to each of these test samples to generate a large number of confidence intervals. The nominal coverage for $g(\boldsymbol{p})$ for a method under this sampling process is then the proportion of these confidence intervals containing the unobserved

true value of $g(\boldsymbol{p})$. For a method that quantifies uncertainty in terms of confidence intervals the validity of said method in particular situations is determined by how closely the stated level of coverage relates to its nominal coverage. For example, a method that creates a confidence interval at the $100(1-\alpha)\%$ level is valid in a given scenario if it is reasonable to believe that the nominal coverage is approximately $1-\alpha$. For the sake of simplicity, this paper will only focus on equal tailed intervals but much of the analysis will extend to the case when tails are not equal.

The use of confidence intervals is currently the recommended practice within the land cover mapping community [22], [23]. We place methods of creating confidence intervals in three categories, exact, heuristic and asymptotic.

Exact methods are methods that rely on using the exact distribution of the sampling processes (in relation to $g(\boldsymbol{p})$) to generate confidence intervals that have rational gauntness regarding nominal coverage. An example of this is Clopper-Pearson intervals [30].

Heuristic methods are methods that rely on approximations of sampling distributions or make slight amendments to exact methods. Typically, heuristic methods are in response to specific weakness of exact methods or when exact methods are not easily be derivable. An example of a heuristic approaches would be Agresti–Coull intervals [31] or using credible intervals from Bayesian methods with uninformative priors.

Asymptotic methods are methods that rely on asymptotic theory to generate confidence intervals. Whilst they could be considered specific cases of heuristic methods, we have chosen to separate them as they act differently when judge by our four criteria (see section 5). The current recommended practice, that assumes a normal distribution based on asymptotic properties of the central limit theorem and bootstrapping methods [32] are examples of asymptotic methods.

An alternative approach to uncertainty quantification seen in land cover mapping applications is to express uncertainties in the form of probability density functions through **Bayesian inference** [33], [34]. Here allow for the uncertainty of relevant $\boldsymbol{p}$ to be represented as a probability distribution given the observed data and a predetermined prior distribution. From this, we can then quantify uncertainty for $g(\boldsymbol{p})$, either through direct inference or through simulation based methods.

Because frequentists and Bayesian methods take different perspectives on probability, it does not make sense to judge a Bayesian approach through the assessment of nominal coverage. In a frequentist setting, a confidence interval is a statement related to the behaviour of a large number of (hypothetical) samples. The uncertainty is on the sampling process, not on the parameter itself. Whereas a measure of uncertainty such as a credible interval (often described as a parallel to confidence intervals in a Bayesian setting) is a measure for the spread of the posterior distribution of model unknowns including parameters. This distribution is a rational quantification of uncertainty based on an observed sample and prior knowledge (or belief). Technically speaking, providing that we believe the prior placed on relevant $\boldsymbol{p}$ to be suitable, the resultant posterior distribution for $g(\boldsymbol{p})$ is valid. A potential difference in results due to set of priors deemed suitable is consistent here. An intuitive interpretation of this is that if two or more actors have different beliefs before observing sample, their beliefs after seeing the sample may also be different if their prior beliefs were sufficiently strong.

Hence when a method takes a Bayesian approach to uncertainty, we shall judge its suitably based on how sensitive the posterior distributions are to **similar** choices of prior distributions.

5 Analysis of Methods

We begin by applying several methods of uncertainty quantification on our Georgian deforestation example. For each method, we calculate an equal tailed confidence (or credible) interval at the 95% level. For the frequentist methods we apply the currently recommended normal approximation method as well as naïve bootstrapping (both asymptotic), a method based on using bounds of multiple Clopper-Pearson intervals for $(\boldsymbol{p}_i)_1$ created at the $100\left(1-\sqrt[3]{0.95}\right)\%$ level (exact method) and the 95% credible intervals from the Bayesian methods (heuristic). For the Bayesian methods, we use a set of uninformative priors for each $(\boldsymbol{p}_i)_1$ (Jeffery $(\boldsymbol{p}_i)_1 \sim Beta(0.5,0.5)$, uniform $(\boldsymbol{p}_i)_1 \sim Beta(1,1)$, (close to) improper $(\boldsymbol{p}_i)_1 \sim Beta(0.01,0.01)$).

Table 2. Limits for the confidence and credible intervals under various methods at (equal tailed, 95% level) for the Gregorian deforestation example.

Method	Forest-to-non-forest					
	User Accuracy		Area (hectares)		Producer Accuracy	
	Lower	Upper	Lower	Upper	Lower	Upper
Normal Approximation	0.4935	0.6975	3734	41002	0.0986	1.0573
Bootstrap (Naive)	0.4944	0.6966	11201	42767	0.2881	1.0000
Clopper–Pearson (+)	0.4862	0.6983	10183	109361	0.0734	6.0035
Bayes (Jeffery)	0.4918	0.6931	14738	61445	0.2075	0.8620
Bayes (Uniform)	0.4913	0.6916	17830	73925	0.1721	0.7174
Bayes (Improper)	0.4923	0.6946	12488	48051	0.2669	0.9804

Form table 2 we can see that all methods largely agree for the user accuracy of the deforestation class. This is largely expected as all methods will eventually converge to normality and the user accuracy is a relatively simple case. What is particular striking however, is the substantial differences we see for the intervals around the area of deforestation. This is problematic as the choice of method here could potentially have a serious impact on decision making.

In a frequentist setting, one must be able to confirm it is likely that the stated level of confidence is at least close to nominal coverage. However, analysing nominal coverage in general is difficult and will depend on many factors including the sample size $\boldsymbol{n}$, the level of confidence, g and the value of unknown $\boldsymbol{p}$. In practice, the dependence on unknown $\boldsymbol{p}$ will generally mean that one will never be able to give the exact nominal coverage. Rather, we may need to consider multiple plausible values of relevant entries of $\boldsymbol{p}$ based on $\boldsymbol{x}$ to build a case of representative coverage.

As an example, let us consider a confidence interval at the 95% level for $(\boldsymbol{p}_1)_3$ generated with the normal approximation method and naive bootstrapping based on a sample size in the presented example (431). In Fig 2 we can see that both methods suffer from considerable under-coverage if $(\boldsymbol{p}_1)_3$ is close to 0. A low count situation observed in $(\boldsymbol{x}_1)_3$ in the present example is evidence that $(\boldsymbol{p}_1)_3$ may be close to 0.

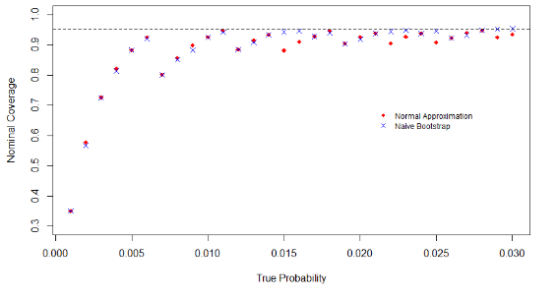

Fig. 2. Coverage plot for the normal approximation method and naïve bootstrapping based on a sample size of 431 (the same as stratum (3) in the Georgian example).

This may call in to question the validity of both these methods when quantifying uncertainty for the total area of deforestation as this quantity replies heavily on $(\boldsymbol{p}_1)_3$ (especially with W_3 being so large). However, a more robust analysis is likely unviable since the relative area relies on three values of $\boldsymbol{p}$, and so one would need a 3-dimisional equivalent of Fig.2.

For the other frequentists methods, whilst the Clopper–Pearson (+) intervals are guaranteed to provide sufficient coverage, the unknown over-coverage may be so high that this is also misleading.

For each of the different types of frequentist methods, there is a trade-off between how well they satisfy of each of our criteria. A more systematic analysis of how each type of method meets our criteria is presented in the appendix in Table A1. The major findings are that, most suffer from transparency issues in more complex situations due to the difficulties in coverage assessment. No type of method is likely to be suitable across all four criteria. Of course, not all criteria are relevant in all situations and so some methods may be suitable in individual cases. The problem is that no method is consistent enough across all four criteria to be a good default approach.

In practice, this can mean having to choose between many methods when taking a frequentist approach to uncertainty. This type of approach would rely on expertise in said methods and suitable diagnostic tests (which may not even be available in situations that are more complex).

In comparison, one is more likely to better satisfy our criteria when taking a Bayesian approach to uncertainty quantification (see Table A1 for further details). This is mainly due to three important advantages

The first advantage is that **issues related to coverage are avoided** as Bayesian analysis is an entirely different form of uncertainty quantification.

The second advantage is that **sensitivity to prior choice can be assessed post-hoc**. This means we can effectively "wait and see" if prior sensitivity is going to be an issue at all. In comparison, one must have assurances related to nominal coverage for every new situation with frequentist methods.

Thirdly, **the problem of prior sensitivity is not as detrimental as the problems we face in frequentist settings** due to coverage assessment. This is because the validity of the results is determined how reasonable we believe the priors to be. In practice, a set of standard prior choices is often agreed in advanced by communities (e.g. a set of

uninformative priors). We would argue that this is an easier task than say having communities agree which frequentist methods to use in which particular situations.

In the context of the Georgian deforestation example, consider the Bayesian results for the area of deforestation. The results in this case differ by a relatively considerable amount. Whilst this is not ideal, these results are robust and informative for decision makers. This is different to the frequentist setting we have considerably different confidence intervals that are, potentially, misleading due to their level of miss-coverage.

One may be tempted to make a similar statement with the frequentist approaches. The problem with this is that we need to assume that all the individual methods are appropriate to begin with (otherwise, they may act as disinformation). In order to do this, one must assess the coverage for all methods in a given situation, which as we have discussed already, is often very difficult.

6 Discussion and future work.

So far we have discussed the advantages of a Bayesian approach to quantifying uncertainty from an error matrix produced by single sample and map. However, a Bayesian approach to uncertainty quantification has the potential to offer many more advantages that are not available in frequentist approaches. Whilst an in-depth exploration of these advantages goes beyond the scope of this paper, they do offer some insights in to where future work may lead in terms of uncertainty quantification from a Bayesian perspective. Such work could include:

A formal means of including prior information. The inclusion of a prior distribution means that we can formally include information in to our uncertainty quantification before observing our sample. This information may come from historical maps or biased samples (e.g. citizen science data). This could substantially the reduce the sizes of test samples needed to reduce uncertainty to satisfactory levels , especially investigating the prevalence of rare classes (e.g. when monitoring land-use change). Note that prior information cannot be formally incorporated in frequentist approaches, as statements such as confidence intervals are related to the sample itself under fixed parameter values.

Predicting the impact of additional test samples on uncertainty reduction. Suppose we wish to reduce the uncertainty further by collecting a further test sample. When predicting the effects of further test sampling, their impact on the degree of uncertainty is often governed by relevant $\boldsymbol{p}$, which we can estimate based on an initial test sample. When taking a frequentist approach to uncertainty quantitation, it is difficult to propagate uncertainty in these initial estimates.

However, when the uncertainty for relevant $\boldsymbol{p}$ is represented as a probability distribution, one can propagate this forward. The practical advantage this gives is that one can have a more reliable means of assessing the trade-offs between the cost of additional samples and their likely impact on uncertainty. In addition, it allows us to compare how different distributions across strata may effect uncertainty. This would be a key step in any work assessing the efficiency of different sampling strategies.

7 Conclusion.

When making estimates from mappings built with machine learning techniques, one must often rely on error matrices obtained from test sampling to quantify uncertainty for these estimates. The current recommended approach in this setting is a frequentist one that assumes asymptotic normality of estimates. This is often unsuitable when estimating the prevalence of rare classes or when strata are homogenous. Alternative methods may exist for simple cases, but they do not extend to more advanced situations that are relied upon in land cover mapping applications. Furthermore, the assessment of any frequentist method itself is near impossible in more complex situations because of the difficulties in analysing nominal coverage.

In comparison, Bayesian inference can offer an approach to uncertainty quantification that is better suited for land cover mapping applications. It is for these reasons that we recommend that future work related to uncertainty quantification from error matrices should be focused on the development and refinement of Bayesian approaches rather than looking towards more advanced frequentist methods.

References

[1] R. Birdsey *et al.*, "Approaches to monitoring changes in carbon stocks for REDD+," *Carbon Manag.*, vol. 4, no. 5, pp. 519–537, 2013.

[2] R. S. DeFries, R. A. Houghton, M. C. Hansen, C. B. Field, D. Skole, and J. Townshend, "Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 1990s," *Proc. Natl. Acad. Sci.*, 2002.

[3] R. B. Myneni *et al.*, "A large carbon sink in the woody biomass of Northern forests," *Proc. Natl. Acad. Sci.*, 2001.

[4] C. R. Schwalm *et al.*, "Reduction in carbon uptake during turn of the century drought in western North America," *Nature Geoscience*. 2012.

[5] G. P. Asner, E. N. Broadbent, P. J. C. Oliveira, M. Keller, D. E. Knapp, and J. N. M. Silva, "Condition and fate of logged forests in the Brazilian Amazon," *Proc. Natl. Acad. Sci.*, 2006.

[6] P. V. Potapov *et al.*, "Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive," *Remote Sens. Environ.*, 2015.

[7] W. Shi, J. Liu, Z. Du, A. Stein, and T. Yue, "Surface modelling of soil properties based on land use information," *Geoderma*, 2011.

[8] L. Giustarini, R. Hostache, P. Matgen, G. J. P. Schumann, P. D. Bates, and D. C. Mason, "A change detection approach to flood mapping in Urban areas using TerraSAR-X," *IEEE Trans. Geosci. Remote Sens.*, 2013.

[9] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS Journal of Photogrammetry and Remote Sensing*. 2013.

[10] R. R. Rindfuss, S. J. Walsh, B. L. Turner, J. Fox, and V. Mishra, "Developing a science of land change: Challenges and methodological issues," *Proc. Natl. Acad. Sci.*, 2004.

[11] K. M. Keegan, M. R. Albert, J. R. McConnell, and I. Baker, "Climate change and forest fires synergistically drive widespread melt events of the Greenland Ice Sheet," *Proc. Natl. Acad. Sci.*, 2014.

[12] Y. Knyazikhin *et al.*, "Hyperspectral remote sensing of foliar nitrogen content," *Proc. Natl. Acad. Sci.*, 2013.

[13] S. K. McMenamin, E. A. Hadly, and C. K. Wright, "Climatic change and wetland desiccation cause amphibian decline in Yellowstone National Park," *Proc. Natl. Acad. Sci.*, 2008.

[14] T. H. Syed, J. S. Famiglietti, D. P. Chambers, J. K. Willis, and K. Hilburn, "Satellite-based global-ocean mass balance estimates of interannual variability and emerging trends in continental freshwater discharge," *Proc. Natl. Acad. Sci.*, 2010.

[15] Y. Fialko, D. Sandwell, M. Simons, and P. Rosen, "Three-dimensional deformation caused by the Bam, Iran, earthquake and the origin of shallow slip deficit," *Nature*, 2005.

[16] R. Khatami and G. Mountrakis, "Implications of classification of methodological decisions in flooding

analysis from Hurricane Katrina," *Remote Sens.*, 2012.
[17] C. Alcantara, T. Kuemmerle, A. V. Prishchepov, and V. C. Radeloff, "Mapping abandoned agriculture with multi-temporal MODIS satellite data," *Remote Sens. Environ.*, 2012.
[18] M. C. Anderson, R. G. Allen, A. Morse, and W. P. Kustas, "Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources," *Remote Sens. Environ.*, 2012.
[19] G. P. Asner *et al.*, "Large-scale impacts of herbivores on the structural diversity of African savannas," *Proc. Natl. Acad. Sci.*, 2009.
[20] C. D. Mendenhall, C. H. Sekercioglu, F. O. Brenes, P. R. Ehrlich, and G. C. Daily, "Predictive model for sustaining biodiversity in tropical countryside," *Proc. Natl. Acad. Sci.*, 2011.
[21] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sens. Environ.*, 2016.
[22] P. Olofsson, G. M. Foody, S. V. Stehman, and C. E. Woodcock, "Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation," *Remote Sens. Environ.*, vol. 129, pp. 122–131, 2013.
[23] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," *Remote Sens. Environ.*, vol. 148, pp. 42–57, 2014.
[24] A. DasGupta, T. T. Cai, and L. D. Brown, "Interval Estimation for a Binomial Proportion," *Stat. Sci.*, vol. 16, no. 2, pp. 101–133, 2001.
[25] S. Wallis, "Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods," *J. Quant. Linguist.*, 2013.
[26] A. Agresti and B. A. Coull, "Approximate Is Better than " Exact " for Interval Estimation of Binomial Proportions Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : http://www.jstor.org/stable/2685469 Approximate is Better than " Ex," *Am. Stat.*, vol. 52, no. 2, pp. 119–126, 1998.
[27] J. E. Wagner and S. V. Stehman, "Optimizing sample size allocation to strata for estimating area and map accuracy," *Remote Sens. Environ.*, 2015.
[28] P. Olofsson *et al.*, "Implications of land use change on the national terrestrial carbon budget of Georgia.," *Carbon Balance Manag.*, vol. 5, p. 4, 2010.
[29] O. T. E. U. Council, "Regulation (EU) No 2018/841 of 30 May 2018 on the inclusion of greenhouse gas emissions and removals from land use, land use change and forestry in the 2030 climate and energy framework, and amending Regulation (EU) No 525/2013 and Decision No 529/2013/EU," vol. 2018, no. October 2003, pp. 1–25, 2018.
[30] C. J. CLOPPER and E. S. PEARSON, "the Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.
[31] A. Agresti and B. A. Coull, "Approximate is better than 'Exact' for interval estimation of binomial proportions," *Am. Stat.*, 1998.
[32] A. C. (Anthony C. Davison and D. V. Hinkley, *Bootstrap methods and their application*. Cambridge : Cambridge University Press, 1997.
[33] R. Denham, K. Mengersen, and C. Witte, "Bayesian analysis of thematic map accuracy data," *Remote Sens. Environ.*, 2009.
[34] A. O. Finley, S. Banerjee, and R. E. McRoberts, "A Bayesian approach to multi-source forest area estimation," *Environ. Ecol. Stat.*, 2008.

Appendix

Table A1. Analysis of the method types based on the four chosen criteria.

Method Type	Transparency	Generalisability	Suitability for Stratified Sampling	Suitability in Low Count Situations
Frequentists (General)	(+) Coverage can be empirically verified in simple cases. (-) Exact coverage is impossible due to the discrete nature of $\boldsymbol{x}$. (-) Assurances regarding coverage must be obtained beforehand and is dependent on many parameters (some of which are unknown).	(+) Any method can easily be extended when is when g with only one relevant on a single entry of $\boldsymbol{p}$. (-) Difficult to verify coverage empirically when g requires multiple entries of $\boldsymbol{p}$.	(-) Issues arise when analysing nominal coverage empirically when one needs to correct for stratified sampling. (g requires multiple entries of $\boldsymbol{p}$)	(-) Low count situations, are an indicator of extreme miss-coverage for many popular methods.
Frequentists (Exact)	(+) Sufficient coverage can by rationally guaranteed in specific cases. (-) Over-coverage may be extreme in low sample sizes.	(-) Difficult to apply for general g (without relying on overly conservative methods).	(-) Over-coverage becomes more severe as the number of strata increases.	(+) Guaranteed to provide sufficient coverage in low count situations. (-) Over-coverage may be extreme in low count situations.
Frequentists (Heuristic)	(-) Issues analysing nominal coverage empirically when stratified sampling or g requires multiple entries of $\boldsymbol{p}$.	(+) Easy to apply, even for more complex forms of g. (-/+) Theoretically has the potential to be computationally expensive. Mitigatable with aggregation properties and rarely an issue in many applications.	(+) Application can be easily extended when stratified sampling. (+) Theoretical results extend easily when stratified sampling.	(+) Near appropriate coverage is verifiable in simple low count situations.
Frequentists (Asymptotic)	(-) Verifying the assumptions based on asymptotic results is difficult for complex g (-) difficult to know how large sample sizes should be in general.	(+) Theoretical results extend for complex g. (+) Some methods (e.g. bootstrapping) can be calculated with simulation based methods, lowering the mathematical expertise required (useful when as g becomes more complex)	(+) Many asymptotic based methods can be easily applied when accounting for stratified sampling.(e.g. normal approximation, bootstrapping). (+) Theoretical results extend for the case of stratified sampling.	(-) Common methods are verifiably unsuitable in low count situations in simple cases. (e.g. normal approximation, naive bootstrapping).
Bayesian (Simple method)	(+) Credible intervals are a measure of spread of a posterior; hence, one does not need to show sufficient coverage. (+) Differences due to prior choice can be analysed post-hoc.	(+) Easy to apply, even for more complex forms of g. (+) Posterior distributions can be generated through simulation based methods. (-/+) Can become computationally expensive but this mitigatable with conjugate priors and aggregation properties.	(+) Application can be easily extended when stratified sampling. (-) Increasing the number of strata increases the number of subjective prior choices one has to make.	(-/+) Results can be sensitive to prior choice in low count situations. Uncertainty quantitation is still viable (if slightly weaker) in prior sensitive situations.