Key Issues in Rigorous Accuracy Assessment of Land Cover Products

Stephen V. Stehman[a] and Giles M. Foody[b]

[a] Department of Forest and Natural Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, NY 13210, United States (svstehma@esf.edu)

[b] School of Geography, University of Nottingham, Sir Clive Granger Building, University Park, Nottingham, NG7 2RD, United Kingdom (giles.foody@nottingham.ac.uk)

Corresponding Author: Stephen V. Stehman (svstehma@esf.edu)

Abstract

Accuracy assessment and land cover mapping have been inexorably linked throughout the first 50 years of publication of *Remote Sensing of Environment*. The earliest developers of land-cover maps recognized the importance of evaluating the quality of their maps, and the methods and reporting format of these early accuracy assessments included features that would be familiar to practitioners today. Specifically, practitioners have consistently recognized the importance of obtaining high quality reference data to which the map is compared, the need for sampling to collect these reference data, and the role of an error matrix and accuracy measures derived from the error matrix to summarize the accuracy information. Over the past half century these techniques have undergone refinements to place accuracy assessment on a more scientifically credible footing. We describe the current status of accuracy assessment that has emerged from nearly 50 years of practice and identify opportunities for future advances. The article is organized by the three major components of accuracy assessment, the sampling design, response design, and analysis, focusing on good practice methodology that contributes to a rigorous, informative, and honest assessment. The long history of research and applications underlying the current practice of accuracy assessment has advanced the field to a mature state. However, documentation of accuracy assessment methods needs to be improved to enhance reproducibility and transparency, and improved methods are required to address new challenges created by advanced technology that has expanded the capacity to map land cover extensively in space and intensively in time.

1. Introduction

Land cover and land cover change are fundamental environmental variables of broad impact. Land cover is an essential climate variable (Hollmann et al. 2013) that impacts numerous environmental processes and patterns ranging from influencing albedo and hence climate to zoogeographic distributions and hence patterns of biodiversity. The effects of land cover may be relatively direct, such as by influencing albedo, or indirect, such as by impacting environmental processes that determine aspects of ecosystem services valuations which is why land cover can be used as a surrogate variable in these valuations (Costanza et al. 1997). Land cover also greatly influences human use of the land (Turner et al. 2007) and can greatly impact health, wealth and well-being. Changes in land cover can be particularly important. Taking into consideration the impacts on just one essential variable, water, a land-cover conversion such as deforestation might impact processes such as interception, infiltration, evapotranspiration, ground water recharge and flood frequency with consequent impacts on variables such as erosion and sedimentation. Indeed, it has been recognized for many years that land cover change is one of the most important environmental variables impacting major societal concerns such as climate change and biodiversity conservation (Vitousek et al. 1997; Chapin et al. 2000; Foley et al. 2005). Given the importance of land cover and land cover change, there is considerable need for high quality and up-to-date information on them. Because the remotely sensed response typically measured is a function of Earth surface properties, remote sensing has considerable potential as a source of information on land cover and land cover change at a range of spatial and temporal scales.

From a wide range of methods (Lu and Weng 2007; Mather and Tso 2016), information on land cover is commonly extracted from remotely sensed data via a classification analysis and often presented in the form of a thematic map. Similarly, comparisons of mapped representations for different time periods typically provide the basis for monitoring land cover change. Consequently, maps depicting land cover are central to numerous scientific and practical applications. Maps provide a generalised view of the world. As such, while a high quality map depicting the theme of interest can be used successfully to provide desired information it must be recognized that it can be imperfect and require careful interpretation. A fundamental feature of significance to the user of the land cover data is its quality. Quality has many dimensions (e.g., positional accuracy and image segmentation accuracy) but the focus of this article is limited to the critical issue of thematic accuracy. The latter is simple in concept, relating to how closely the map shows reality in terms of the land cover mosaic it represents, but has been surprisingly poorly addressed despite its significance (Olofsson et al. 2013).

Land cover maps can be easily generated from remotely sensed imagery. Using standard image classifiers available in popular software it is trivially easy to generate a land cover map. But the map produced will be a function of the data and classifier used (Foody and Mathur 2006; Pal and Foody 2012; Foody et al. 2016). Different training sets, classification algorithms, and remotely sensed images could be used to produce very dissimilar representations of the same geographic region. Regional to global maps of land cover, for example, provide different representations of the same phenomena which typically disagree on the label for much of the region mapped (Herold et al. 2008; Tchuenté et al. 2011). Which map to use as the closest representation of reality requires an ability to rigorously assess map accuracy; otherwise, each map remains nothing more than a pretty picture representing simply one possible untested hypothesised land cover scenario (Strahler et al. 2006; McRoberts 2011). Good scientific practice demands that the quality of the measurements made by remote sensing be evaluated and integrated into studies. High quality measurement is often seen as a hallmark of science. Without it there is the danger of mis-interpretation and mis-understanding. The assessment must, however, be rigorous and defensible. Poor or weak accuracy assessments offer a veneer of scientific respectability but are a façade that can impede understanding and successful practical application.

Accuracy assessment has a long history in remote sensing as Congalton and Green (1999, Chapter 2) trace the origins back to approximately 1975. The subject has developed greatly and there is a large literature that explores the many dimensions and issues of importance. This article does not aim to revisit the whole subject but is focused on appropriate methodology for a useful and credible accuracy assessment in typical scenarios. For the most part we assume that the map and reference labels represent a hard classification and that accuracy assessment is conducted independently of classifier training. The focus is explicitly on the correct use of design-based inference for the estimation of accuracy expressed on an overall and per-class basis. In this, the accuracy assessment is founded on the comparison of the predicted map class label with the actual label observed on the ground for a sample of selected sites in the mapped area. Although the need for accuracy assessment has long been recognized by the remote sensing community and there has been considerable research on the topic, many land cover maps produced have not been subjected to rigorous evaluation (Olofsson et al. 2013) limiting their scientific and practical value. An accuracy assessment does more than simply describe the quality of a map, it provides a means to enhance its usefulness. An inaccurate map that is accompanied by a rigorous accuracy assessment may, for example, provide highly valuable information.

Good practices have been promoted to help the production of credible accuracy assessments (Olofsson et al. 2014). These good practice recommendations also extend to area estimation, cautioning

against the common practice of using "pixel counting" for area estimation (i.e., summing the number of pixels allocated to a class in the map and multiplying by pixel areal extent). Instead the good practice guidelines advocate estimation of area based on the reference classification and using information contained in the error matrix to improve precision of the estimator (McRoberts 2011; Olofsson et al. 2013, 2014). Although we emphasize elements of good practice, it is also necessary to point out aspects of bad practice such as neglecting to use estimation formulas that correspond to the sampling design, normalizing the error matrix (which equalises user's and producer's accuracy even though they may be meaningfully different and can be a large source of bias) or using the kappa coefficient (which unnecessarily adjusts for chance agreement in a manner in which chance agreement is mis-estimated).

The organization of the article is as follows. Section 1 provides an overview of the historical development of accuracy assessment (Sec. 1.1) and articulates general criteria to guide planning and implementation of accuracy assessment of land cover maps (Sec. 1.2). We then proceed to review the current state of good practice for the three main components of accuracy assessment (Stehman and Czaplewski 1998), sampling design (Sec. 2), response design (Sec. 3), and analysis (Sec. 4). Recognizing the critical role of statistical inference in accuracy assessment, we describe the commonly used design-based inference approach in Section 5 along with alternative options of model-based and Bayesian inference. Practitioners have long been aware of the difficulties of obtaining gold standard (i.e., perfect) reference classifications motivating the discussion in Section 6 of the problems and impacts of imperfect reference data. Section 7 is devoted to future needs and directions of accuracy assessment, and conclusions stated in Section 8 complete the article.

1.1. Historical context

The *Scopus* database was used to search for relevant articles published in the 50-year history of *Remote Sensing of Environment* (*RSE*). Specifically, two searches were undertaken. The first search extracted articles with the term 'land cover' in the title, keywords or abstract, and the second search identified articles via the search phrase 'land cover AND accuracy' to indicate the importance of accuracy assessments in these land cover studies. Through the first 20 years of *RSE* only 6 articles met the latter search phrase and comprised <1% of all articles published in *RSE* (Table 1). In the most recent 15 years of *RSE* (2004-2018), the number of articles including both search terms 'land cover' and 'accuracy' has expanded to 331 comprising about 6.5% of all articles published in *RSE* during that 15 year period. The percent of articles in *RSE* with 'land cover' has increased up until 1999 when it reached a plateau of 15% of all articles. From 1994 to 2008 approximately 40% of all land cover articles also had

'accuracy' as a keyword and this percent increased to about 50% for the past decade, 2009-2018, providing some evidence that accuracy has become more prominently emphasized in land cover studies.

Table 1. Publications in *Remote Sensing of Environment* (*RSE*) in five-year intervals resulting from two *Scopus* searches, one for 'land cover' and another for 'land cover' AND 'accuracy' in titles, keywords, and abstracts (searches conducted 25 January 2019).

Time Interval	Total #*RSE*	Land Cover (% of All)	Accuracy AND Land Cover (% of Land Cover)
1969-1973	64	0 (0)	0
1974-1978	128	0 (0)	0
1979-1983	206	9 (4)	1 (11)
1984-1988	307	6 (2)	3 (50)
1989-1993	434	7 (2)	2 (29)
1994-1998	565	55 (10)	22 (40)
1999-2003	755	117 (15)	45 (38)
2004-2008	1263	185 (15)	72 (39)
2009-2013	1503	227 (15)	110 (48)
2014-2018	2187	286 (13)	149 (52)
Total	7412	892 (12)	404 (45)

The earliest five publications in *RSE* identified by the search phrase of 'land cover AND accuracy' foreshadowed methods and challenges that have persisted throughout the history of accuracy assessment of land cover. The earliest article identified from the Scopus search was Bryan (1975) in which individuals with no experience in the use of side looking airborne radar (SLAR) imagery were asked to interpret the land cover class of 32 test case objects in Melbourne, FL, United States of America (USA). Accuracy was reported simply as the proportion of these objects classified correctly. Walsh (1980) used Landsat imagery to map 12 land-cover classes in the vicinity of Crater Lake, Oregon, USA, and implemented stratified random sampling of 5x5 pixel sampling units to report user's accuracy (without accompanying standard error). Jensen et al. (1980) used MSS to map vegetation types in a 4000 ha non-tidal wetland in South Carolina, USA. The accuracy assessment consisted of field sampling along two transects (not randomly located) with a 1-m length of transect used as the assessment unit.

Jensen et al. (1980) summarized the accuracy information via an error matrix and reported omission and commission error rates (without standard errors). Gordon (1980) used Landsat imagery to map 21 classes within a small area of Ohio, USA (<100 km^2) for two dates, 1973 and 1975, representing an early land-cover change application. User's accuracies for each date were estimated based on photo-interpreted reference classifications for the entire study area (no sampling conducted). Toll (1985) compared accuracy of land cover classified by MSS versus TM imagery for an 800 km^2 test site near Washington, DC, USA. Stratified systematic sampling was implemented with equal allocation of 75 pixels (500 m x 500 m) from each of the 7 classes, and the comparison of MSS versus TM was reported using confidence intervals for overall accuracy of each classifier. In addition to the examples identified from the algorithmic Scopus search, we found accuracy assessments of crop classifications reported in the very fir st two volumes of *RSE*. Haralick et al. (1970) reported error matrices along with overall and producer's accuracies for a classification of crop types, and Anuta and MacDonald (1971) presented an accuracy assessment of classifications of crops in the Imperial Valley of California.

The initial forays into accuracy assessment of land cover included common current practices such as implementing stratified sampling and reporting results in the form of an error matrix accompanied by user's and producer's accuracies. We also observe in these early articles the diversity of objectives of accuracy assessment, including basic description of map quality (see Secs. 4.2 and 4.4 of this review article), accuracy of land-cover change (Sec. 4.5), and comparison of maps (Sec. 4.6). These articles are also illustrative of present day problems such as: 1) absence of standard errors or confidence intervals quantifying the uncertainty of accuracy and area estimates obtained (Sec. 4.4); 2) lack of clarity in the description of the sampling design so that it would be impossible to reproduce the sample selection protocol (Sec. 2.2); 3) absence of a probability sampling design (Sec. 2.1); 4) misplaced concern regarding independence, as reflected in Toll's (1985) statement that non-contiguous pixels were sampled "to increase independence" (Sec. 2.4); and 5) not employing stratified estimation formulas for a stratified sampling design (Sec. 4.2). One major difference of current practice relative to these pioneer land-cover studies is that present-day applications typically map vastly larger areas.

The late 1970s and early 1980s represented an era of emerging methodological developments in accuracy assessment as articles were published in *RSE* and other outlets. Sampling design and sample size questions were commonly addressed (e.g., Hord and Brooner 1976; Fitzpatrick-Lins 1981) with Hay (1979) proposing use of stratified sampling with a minimum sample size of 50 per class, a practice still often followed today. Another focus of these early articles was the question of whether maps were of acceptable quality, leading to an emphasis on the decision-making framework of statistical hypothesis

testing as well as frequent reference to 85% accuracy as a benchmark defining acceptable accuracy (Anderson 1971; Aronoff 1982ab; Fitzpatrick-Lins 1981) (see Section 7.5 for a critique of this 85% benchmark). Among these early articles, Card (1982) warrants special acclamation for recognizing the relevance of estimating proportion of area based on the reference classification, an idea reinvigorated by McRoberts (2011) and Olofsson et al. (2013, 2014). Not surprisingly, categorical data analyses developed in the statistical literature for two-way contingency tables (Bishop et al. 1975) were proposed for use in accuracy assessment (Congalton et al. 1983). These tables represent sample counts for data cross-classified by two categorical variables and thus bear a strong resemblance to an error matrix. Two unfortunate remnants of these contingency table analyses that still persist are the normalized error matrix and kappa coefficient, both practices that are now recommended to be avoided (see Sections 4.1, 4.8, and 7.5).

Congalton (1991) and Congalton and Green (1999) synthesized much of the methodology developed up to 1990 and merit considerable credit for putting accuracy assessment "on the map" of the remote sensing community. A series of subsequent review articles further elucidated the progression of accuracy assessment from the early 1990s to the present (Janssen and van der Wel 1994; Smits et al. 1999; Foody 2002; Strahler et al. 2006; Wulder et al. 2006a; Olofsson et al. 2014). Methodological developments and applications of accuracy assessment have grown exponentially since the early 1990s (Table 1). Advances in technology related to satellite imagery and classification methods have surely fuelled this growth in land cover studies and accuracy assessment (Wulder et al. 2018). Land cover analyses have evolved from rather limited studies of small geographic regions at a single point in time to present day regional, national, and global studies at multiple time periods, often with increasingly smaller spatial resolutions and increasingly greater temporal densities. In this article, we discuss the status of accuracy assessment of land cover highlighting advances in concepts and methods developed primarily since Foody's (2002) review, and focusing on issues relevant to land cover studies covering large areas. We extend and apply the good practice recommendations of Strahler et al. (2006) and Olofsson et al. (2014) to critique current practice. In particular, we discuss how commonly used protocols satisfy or fall short of good practice criteria (defined in the next section) specified to ensure a rigorous and credible assessment of accuracy and area of land cover and land cover change. We also take a glimpse into the future and posit new directions for needed developments to address emerging challenges for accuracy assessment.

1.2. Good practice criteria to guide accuracy assessment methods and reporting

If accuracy assessment is worth doing, then it most certainly is worth doing well. A rigorous accuracy assessment should be viewed as an essential part of a mapping project (Strahler et al. 2006). We thus begin by articulating operational criteria that serve to guide the practice of accuracy assessment toward scientifically credible and informative methods and results. In practice, accuracy assessments should be designed and implemented to achieve these criteria. Criteria 1 through 4 address fundamental requirements for producing a map relevant, scientifically credible accuracy assessment that includes aspects of quality control and quantification of uncertainty. Criteria 5 and 6 are proposed to address the emerging concerns of transparency (openness) and reproducibility of scientific research (Munafò et al. 2017). These criteria are ideals to be aspired to and will be satisfied to a matter of degree in any given application. Although the need for robust accuracy assessments is often espoused, robustness is not included as a good practice criterion because it lacks specific meaning relative to land cover issues (and also because "robust" has a longstanding statistical definition referring to resistance to impacts of outliers and violations of assumptions). The six good practice criteria are:

1. Map relevant - accuracy assessments should address the reality of the landscape mapped which is accomplished when the accuracy estimates and error matrix reflect the proportional areal representation of the study region in terms of both the map and reference classifications (Sec. 4.1). In effect, the map relevant criterion imposes the requirement that the accuracy estimates have fidelity to the region of interest covered by the map. For example, suppose that a map is produced with a binary legend of change and no change, and it is known that the mapped region has experienced approximately 3-5% change. If the error matrix shows roughly equal proportions of area of change and no change (i.e., each class occupies ~50% of the mapped region), the assessment is clearly not map relevant. This criterion is most often violated by use of normalized error matrices (Sec. 4.1) or reporting an error matrix based on sample counts when the sampling design is stratified with equal sample size allocated to each stratum (Sec. 4.4).

2. Statistically rigorous - the primary requirements for a statistically rigorous assessment are to implement a probability sampling design (assuming design-based inference will be used), to employ consistent estimators (i.e., use estimation formulas that correspond to the sampling design implemented), and to quantify the variability of the accuracy and area estimates by reporting standard errors or confidence intervals.

3. Quality assured – protocols should be in place to monitor and evaluate the quality of the reference data and results. An example of quality assurance would be to evaluate consistency of reference class labels obtained by different interpreters by providing two or more interpreters with a randomly chosen

set of sample locations and having the interpreters independently label the locations while blind to the map category. These consistency data can be used to provide feedback to enhance agreement among interpreters and also to address the reliability criterion.

4. Reliable – a repetition of the accuracy assessment protocol should produce approximately similar results. Reliability is tantamount to controlling those features of the assessment protocol that are subject to variability. The two most prominent such features are associated with variability attributable to sampling and variability attributable to interpretation of the reference class labels. Reliability addresses the question, if we had obtained a different sample and had used different interpreters (all trained in the same manner and with equal expertise), would the estimates of accuracy and area be approximately the same? A reliable assessment would be reflected in small standard errors for the accuracy and area estimates and small variability among interpreters when assigning reference class labels. Reporting standard errors or confidence intervals provides an important assessment of reliability because these results quantify how repeatable the estimates would be over different random outcomes of the sample selection.

5. Transparent – provide all relevant details, positive or negative, that would inform readers about the quality of the results (Castilla 2016). For example, any problems with missing data due to cloudy imagery or inability to access sample locations should be reported, and any modifications to standard sampling design and analysis protocols should also be documented. Synonyms for the transparency criterion would include “openness” and “honesty”. The reality is that any accuracy assessment will be confronted with practical difficulties that will interfere with perfect implementation. The transparency criterion simply mandates an honest depiction of the problems encountered and steps taken in response to these problems. Transparency is particularly critical for the response design, and additional details are provided at the end of Section 3.

6. Reproducible – the methodology of the accuracy assessment should be reproducible given the documentation provided for the sampling design, response design, and analysis. The reproducibility criterion places a premium on lucid description of the key details of each protocol. Achieving the reproducibility criterion is critically dependent on the transparency criterion because if documentation of full details of implementation is lacking the assessment protocols cannot be reproduced. A fully reproducible assessment would include any programming code and identification of software used to produce the results as well as the sample data used in the analyses. Although it may not be possible in all applications, providing the sample data would allow users to obtain estimates for subregions of particular interest (e.g., a state or province within a national assessment) as well as to conduct

secondary analyses such as exploring how accuracy is associated with possible causes of misclassification. Additional details necessary to satisfy the reproducibility criterion are presented in each of the sampling design, response design, and analysis sections described below.

2. Sampling design

The sampling design is the protocol for selecting the subset of assessment units for which the reference classification is obtained and then compared to the map classification. Numerous sampling designs have been used for accuracy assessment (Stehman and Czaplewski 1998; Stehman 1999a; Brus et al. 2011) with the most commonly used designs being simple random, stratified random, systematic, and cluster. Stehman (2009a) reviewed the advantages and disadvantages of these basic designs relative to accuracy assessment objectives, practical considerations, and desirable design criteria. If design-based inference is to be invoked, the most important criterion the sampling design must satisfy is that it is a probability sampling design; in the absence of probability sampling, it is necessary to implement model-based inference or Bayesian inference (Sec. 5).

2.1. Probability versus representative sampling

Although the importance of sampling for collecting high quality reference data was recognized in the early days of accuracy assessment (Congalton 1991), it is only in the past 20 or so years that sampling has been placed in the rigorous statistical framework provided by probability sampling and design-based inference (Stehman and Czaplewski 1998). The statistical rigor of accuracy assessment results produced from a sample require that the sampling design satisfy the conditions that define a probability sample. These conditions are specified in terms of the inclusion probability of element u, denoted π_u. If a pixel is the spatial unit of the accuracy assessment, the inclusion probability is the probability that pixel u will be one of the pixels in the sample. The two conditions required to satisfy the definition of probability sampling are that π_u is known for all units selected in the sample, and $\pi_u>0$ for all units in the population (i.e., all pixels in the region of interest). The basic sampling designs, simple random, stratified random, systematic, and cluster, meet these conditions. Common exceptions to probability sampling include purposely choosing sample locations because of convenience or homogeneity of land cover, and implementing a complex selection protocol with numerous steps such that deriving inclusion probabilities is intractable.

The focus on probability sampling is different from the common admonition for representative sampling. By definition a probability sample is representative because such a sample can be used to

produce unbiased estimators of the population of interest. Although inclusion probabilities must be known, they are not required to be equal, so a sample can be representative in the sense of yielding unbiased estimators even if the sample elements are not selected with equal probability. Stratified sampling with equal allocation is a commonly used probability sampling design for which the inclusion probabilities are not equal for all elements of the population. In fact, a primary motivation for stratified sampling is to create such unequal inclusion probabilities for the purpose of increasing the sample size from rare map classes. It is thus a myth to believe that "the sample must represent the population" in the sense of the proportion of each class in the sample matching the proportion of each class in the population. As long as the analysis takes into account the known unequal inclusion probabilities, the estimates produced will reflect the actual parameters of the population (i.e., the estimators are unbiased) even though the sample itself may bear little resemblance to the population in terms of proportion of area of each class. Implementing an equal probability sampling design does offer the advantage of simplicity of analysis because the estimates are unweighted (i.e., weights are not needed to accommodate different inclusion probabilities). Sampling designs that have equal inclusion probabilities include simple random sampling, systematic sampling and stratified sampling with proportional allocation (i.e., the sample size in each stratum is proportional to the area of the stratum).

2.2. Choosing the sampling design

Deciding which sampling design to implement requires considering the objectives of the accuracy assessment and prioritizing desirable design criteria. The typical objectives for accuracy assessment focus on description of overall accuracy, class-specific accuracy, and proportion of area occupied by a class, and often these estimates are needed for subregions (e.g., states or provinces would be subregions within a national assessment). Desirable sampling design criteria include that the design: 1) satisfies the conditions defining a probability sampling design; 2) is easy to implement for both selecting the sample and producing estimates of accuracy and area; 3) is cost effective; 4) readily allows for increasing or decreasing the sample size; 5) is precise in the sense that estimates have small standard errors; 6) has an unbiased estimator of variance; and 7) is spatially well distributed across the study area. Different sampling designs achieve these criteria to varying degrees (Stehman 2009a). Criterion 1 is critical to achieve a statistically rigorous assessment, whereas criteria 2, 3, and 4 address practical concerns. In particular, criterion 4 addresses the common occurrence that the sample size may need to be adjusted, as, for example, when the actual data collection cost is greater or less than the cost initially budgeted for in planning. The ability to decrease the sample size (or, in rare circumstances

increase the sample size) while still maintaining the probability sampling feature of the design is very desirable in some applications. Criteria 5 through 7 pertain to the reliability of the assessment as small standard errors (i.e., precise estimates) translate to narrow confidence intervals thus reflecting greater reliability of the accuracy assessment. Criterion 7 is relevant to precision because a spatially well-distributed sample, such as from systematic sampling, often yields smaller standard errors (see Section 2.4).

The decision of what sampling design to use is not a matter of choosing between what is practical versus what is statistically rigorous, but rather choosing what is practical and rigorous (Stehman and Czaplewski 1998; Stehman 2001). The process for deciding which sampling design to implement focuses on three major questions: 1) Should clusters be used? 2) Should strata be used? and 3) What selection protocol, simple random or systematic should be used? Regarding the first two questions, the recommendation is to stratify for objectives, cluster for costs. Strata are generally used in the sampling design when the objectives include precise estimates by class (e.g., user's accuracy) or by region. Stratifying by map class achieves the former objective, and stratifying by region achieves the latter objective. Cluster sampling is typically implemented when substantial savings in cost of imagery or cost to travel to field sample locations are achieved by grouping sampling units together (e.g., Mayaux et al. 2006; Stehman and Selkowitz 2010; Zimmerman et al. 2013). The cost savings of cluster sampling must be tempered by the fact that because of intracluster correlation (i.e., correlation among secondary sampling units within clusters), each cluster may be equivalent to only a modest number of unclustered sample points or units (Gallego 2012; Gallego and Stibig 2013). Yang et al. (2018) provide a contrasting option to reduce cost by defining strata constructed to represent cost of access and then increasing the sample size in the low cost strata.

The need for reproducibility (section 1.2) dictates that the sampling design must be clearly documented. Unfortunately the sample design is often inadequately described. Ye et al. (2018) revealed the severity of the problem as they found that only one-third of the 209 articles they examined in their review of object-based accuracy assessment described how the sample was selected. Key features of the sampling design that must be documented to satisfy the reproducibility criterion are: 1) describe the randomization implemented in the sample selection protocol; 2) specify the inclusion probabilities or the information needed to derive the inclusion probabilities; 3) if stratified sampling was implemented, describe how the strata were constructed, provide the proportion of area in each stratum, specify the sampling design implemented within each stratum, and state the sample size allocated to each stratum along with the rationale for this allocation; 4) if cluster sampling was implemented, define the primary

sampling unit (PSU) and the secondary sampling unit (SSU), state whether one-stage or two-stage sampling was implemented (in one-stage sampling all SSUs within each sampled PSU are observed, whereas in two-stage sampling a sample of SSUs is selected within each sampled PSU), and specify the sampling design implemented at each stage.

Registering or archiving the original sample locations prior to collecting the reference data would be a valuable contribution to reproducibility. Gallego (2011) and Waldner et al. (2019) expressed the concern that except for systematic sampling designs, it is impossible to trace whether the originally selected sample locations were actually the ones visited to obtain reference data or whether some of the sample locations had been moved or simply not used. This concern with "traceability" would be greatly diminished if the original sample locations are registered or archived in the public domain prior to beginning data collection. To address the transparency criterion, any problems related to missing data (i.e., non-response due to difficult access on the ground or due to clouds in the reference imagery) should be documented and quantified (e.g., percent of sample units that were not obtainable). If sample units are moved or replaced, these modifications of the original sample selection should be clearly highlighted as part of an honest, transparent portrayal of the actual implementation of the sampling design. Practical issues often may impact the sample selection protocol, but the remedies implemented to address these practical challenges must still adhere to principles of statistical rigor (Stehman 2001).

2.3. Sampling design for accuracy assessment and area estimation of land-cover change

Applications mapping land cover change have become increasingly common and there are many methods for change detection (Singh 1989; Mas 1999). Here we focus on land-cover conversions and do not address accuracy of land-cover modifications that are not manifested as a change in the land-cover class label. Choosing a sampling design in land-cover change studies requires additional considerations not present in single date assessments. Not all of these issues have been fully resolved as the investigation of sampling designs for long-term, high temporal frequency land-cover monitoring is still at a nascent stage. Because change is typically a rare characteristic of the landscape, stratified sampling is often implemented to ensure a specified minimum sample size from the change classes (Olofsson et al. 2013, 2014). Stratification is commonly employed in single date land cover assessments as well, but for land-cover change assessments the number of from-to change types may be so large that defining every from-to change possibility as a stratum would require an impractically large sample size. In such cases, more general from-to categories can be defined as strata, as for example cropland loss to include all

changes from cropland to any other class and urban gain to include changes from all other classes to urban (Wickham et al. 2013, 2017).

Multiple change periods may be of interest in land-change monitoring. For example, suppose the objective is to monitor annual wetland change over a 20-year period using Landsat imagery. In such applications, and assuming a stratified design will be employed, is it better to have a single, permanent set of sample units observed for each year throughout the 20-year period, or is it better to select a different sample for each annual change estimate? Permanent sample locations may offer some efficiency in terms of the response design when using archived satellite imagery via software such as TimeSync (Cohen et al. 2010) or CollectEarth (Bey et al. 2016). Further, the ability to view the entire 20-year time series for each sample pixel may offer the interpreter greater temporal context for deciding whether change had occurred in a particular year of the time series and thus a permanent sample could conceivably yield a more accurate reference classification. If the change process is cyclical (e.g., forest management that produces losses followed by gains in tree cover), then permanent sample locations may be preferred to capture this change dynamic. But if the dynamic is predominantly unidirectional change (e.g., conversion of forest to developed), then a different sample for each time period of interest may be preferred because it would be possible to stratify the sampling design specific to that time period and this would likely yield better precision than permanent sample units. Such an outcome was observed in a study estimating forest change on a biannual basis in which Arevalo et al. (2019) found that selecting a new sample for each biannual change period (with strata defined specifically for each biannual period) was more precise than selecting a single, permanent sample based on strata defined for the entire time period monitored.

Assessing the accuracy of a burned area map presents a special case of change accuracy assessment. The objectives include assessing both the spatial and temporal features of the burned area map (i.e., do the map and reference data agree on whether the sample unit burned and agree on when it burned). The rarity of burned area exists in both space and time, so stratified sampling is motivated to intensify the sample in the space and during the time for which the classifier maps burned area. Boschetti et al. (2016) and Padilla et al. (2017) address this problem by representing the population using voxels defined by a time interval (e.g., a 16-day Landsat acquisition interval) and a two-dimensional spatial unit (e.g., a Thiessen Scene Area shown in Fig. 1 of Boschetti et al. 2016). These voxels can then be assigned to strata and the sample intensified in strata for which burned area is more likely to be found. The difference between this approach and a design in which a new sample is selected

for each time period is that the objective for the burned area application does not include estimating accuracy for each time period.

2.4. Spatial correlation and sampling design

Spatial correlation is frequently raised as a concern regarding sampling design for accuracy assessment. The presence of spatial correlation has no impact on design-based inference in the sense that estimators of accuracy and their associated variance estimators remain unbiased regardless of the magnitude of spatial correlation (de Gruijter and ter Braak 1990; Stehman 2000). Thus there is no need to space sample units apart to legitimize design-based inference. The primary impact of spatial correlation is that it may affect the precision of the estimates. For example, positive spatial correlation will increase standard errors for cluster sampling because pixels within each cluster will be more similar to each other. For systematic sampling, positive spatial autocorrelation leads to smaller standard errors relative to simple random sampling and this precision advantage is a primary motivation for having a spatially well-distributed sample (Sec. 2.3).

The desire to avoid sampling spatially proximate units has led to development of spatially balanced probability sampling designs that spread sample units apart (Benedetti et al. 2017; Stevens and Olsen 2004). Systematic sampling is the most familiar spatially balanced sampling design. A simple but effective alternative to systematic sampling for achieving spatial balance is to stratify the region spatially and to sample one unit (e.g., one pixel or one cluster) per spatial stratum (Fattorini et al. 2015). Two example applications of this design are Fichet et al. (2014), in which the spatial strata were 20 km x 20 km cells, and Stehman and Selkowitz (2010), in which the spatial strata were biomes in Alaska, USA. Benedetti et al. (2017) review a variety of other spatially balanced probability sampling designs, some of which are quite complex. The decision of whether to use these more complex designs will require choosing between two desirable design criteria, spatial balance versus ease of implementation and analysis.

A more problematic consequence of the desire for a spatially balanced sample is when *ad hoc* methods are used to avoid sampling neighboring or nearby units. For example, if a stratified random sample is initially selected and then sample units discarded or moved if they are within a certain distance of another sample unit, the inclusion probabilities of this design will be very complicated. Unless these complicated inclusion probabilities are derived, the sampling design will not satisfy the criteria defining a probability sample. Elaborate algorithms based on minimizing the spatial correlation among sample locations can be created but if such techniques are used it is incumbent on the

developers of the algorithm -to derive the inclusion probabilities for the design. If a spatially balanced sample is implemented for accuracy assessment, we recommend using one of the probability sampling designs currently documented in the peer-reviewed literature (Benedetti et al. 2017).

2.5. Sample size planning and allocation to strata

The enduring question of "What sample size is needed?" must be addressed when planning the sampling design. Although formulas for sample size planning yield deterministic, objective outcomes, the subjective component of choosing the sample size requires judicious selection of values to input to these sample size formulas. The reality for many applications is that the sample size chosen will be constrained by availability of resources to collect the sample data. Pragmatically, *a priori* sample size calculations are effectively an exercise in forecasting whether the sample size afforded by the project's budget will meet the desired precision requirements. Yet another difficulty is that typically several objectives are of interest, and sample size planning becomes more complex when attempting to meet precision requirements for multiple estimates.

A commonly used sample size formula that can be applied to overall accuracy for simple random sampling and to user's accuracy for stratified random sampling is:

$$n = \frac{z^2 p(1-p)}{d^2} \quad (1)$$

where z=1.645 for a 90% confidence interval or z=1.96 for a 95% confidence interval, d is the desired half-width of the confidence interval (i.e., the confidence interval is estimated accuracy ±d), and p is the anticipated overall accuracy in the case of simple random sampling, or p is anticipated user's accuracy in the case of stratified sampling where we determine the sample size n_h for each stratum. For example, if the anticipated user's accuracy is p=0.80 and we would like a 95% confidence interval (z=1.96) to have a half-width of d=0.05, Eq. (1) yields n=246. If instead we anticipate a user's accuracy of p=0.5, the sample size formula yields n=384. The choice of d may vary depending on the importance of a class, so for example a high priority class might have d=0.02 because we want a more narrow interval for this important class, whereas a class regarded as less important to the objectives might have d=0.05 reflecting tolerance of a wider interval. Because the true overall accuracy or user's accuracy is unknown at the sample planning stage, p is subjectively chosen. Therefore, it is often useful to try several values of p to assess the range of n that might be needed for various potential outcomes of overall or user's accuracy. If in reality it turns out that p in Eq. (1) is different from the value used to plan the sample size, the width of the confidence interval d specified at the planning stage will not be the width of the confidence interval estimated from the sample data. The maximum sample size from Eq. (1) occurs

when p=0.5 so it is possible to establish an upper bound on the sample size for a specified z and d. Eq. (1) does not normally apply to estimating producer's accuracy because in practice the reference classes cannot be used for stratification (this would require complete coverage reference data).

In practice, the answer to the question, "Was the sample size large enough?" is observable from the standard errors and width of the confidence intervals obtained for the accuracy and area estimates. Consequently, when addressing the question of whether the sample size was adequate, the sample size planning details are not as important as the magnitude of the standard errors or confidence intervals associated with the estimates. For example, if the 95% confidence interval for user's accuracy of cropland is 78% to 85%, would a user interested in cropland have a different opinion of the quality of the map if true user's accuracy is 78% versus 85%? If not, the sample size was adequate to generate a confidence interval that is sufficiently narrow for this user's purpose. Conversely, if the confidence interval is 60% to 85%, we might expect a user to have a different opinion of map quality for cropland if the user's accuracy is 60% (the lower bound of the confidence interval) compared to if the user's accuracy is 85% (the upper bound). For this latter interval, the range of values is likely too broad to meet the user's needs. If the standard errors are unacceptably large, the option exists to augment the sample size to reduce the standard errors and thereby narrow the associated confidence interval (see Sec. 2.2).

The sample size allocation to strata is a critical feature of stratified sampling designs. The three primary allocation options are equal, optimal, and proportional. Equal allocation is typically used when the objectives specify precise estimation of user's accuracy for each class. Equal allocation is justified if the classes are all considered equally important and all classes have the same user's accuracy (i.e., p in Eq. (1) is the same for all classes). If classes differ in importance, priority classes should have more narrow confidence intervals which can be achieved by decreasing d in Eq. (1). The effect of decreasing d will be to increase the sample size n. Optimal allocation is justified if the primary objective is to minimize the standard error of a single estimate such as overall accuracy or area of one class. If many estimates are of interest, an algorithm to determine the optimal allocation for multiple estimates may not be available. For proportional allocation the sample size n_h is proportional to population size N_h in stratum h. Proportional allocation has the benefit of simplicity of analysis because it is an equal probability sampling design, so estimation formulas for accuracy and area are the same as for simple random sampling (the standard error formulas are different). However, the practical utility of proportional allocation is limited because nearly the same precision of estimates can be achieved via poststratified estimation applied to a simple random or systematic sample (see Sec. 4.2).

The sample allocation decision becomes complicated when multiple objectives are of interest, as for example when estimating area and accuracy for several classes and/or for multiple dates. The challenge is that different sample allocations benefit precision of different estimates, so choosing an allocation that is very beneficial to one objective may be detrimental to other objectives in terms of the resulting standard errors. Olofsson et al. (2014) provide some *ad hoc* guidelines for sample allocation when the objectives are to estimate accuracy and area, and Wagner and Stehman (2015) provide a spreadsheet program for objectively determining the optimal allocation that simultaneously minimizes the sum of the variances of the estimates of user's accuracy, producer's accuracy, and area (based on the reference classification) of a single target class. Additional research on sample allocation options is needed because multiple objectives are the norm for many present day land cover studies. The sample size question is revisited in the context of comparing accuracy in Section 4.6. A final important concept regarding sample size is that it is the absolute size of the sample not the percent of the population sampled that drives the precision of the estimates. That is, the standard errors for the accuracy and area estimators are determined by how large n and n_h are. Even if n and n_h are a very small percent of the pixels present in the study region, the standard errors of the estimators may be acceptably small (Stehman 2001, p. 730).

2.6. Examples of sampling designs used in practice

A wide variety of sampling designs have been used to assess accuracy of large-area land-cover maps (Table 2). The examples listed in Table 2 have been purposely chosen to illustrate this diversity of options. One of the most commonly used sampling designs is stratified random sampling where the strata are the map classes. Olofsson et al. (2014) recommend this design as a good practice option. Stratification is a common feature of these examples with regional (geographic) stratification employed for the objective of estimating accuracy by region and stratification by land-cover class implemented to ensure that rare classes are represented by sufficient sample sizes to address the objective of estimating user's accuracy. Cluster sampling is also relatively prominent in these applications. The different sampling designs implemented for the series of National Land Cover Datasets (NLCD) of the United States (Wickham et al. 2004, 2010, 2013, and 2017) track the evolution of advances in technology and mapping objectives (Table 2). For example, the 1992 NLCD design (Wickham 2004) used cluster sampling because of the cost savings associated with collecting the reference data which required obtaining hard copy aerial photographs. The reference class labels for more recent NLCD assessments were interpreted from freely available Google Earth and Landsat imagery. Consequently, cluster

sampling was not used because the cost issue associated with purchase of aerial photographs was no longer a consideration. Stratification was a prominent feature of all NLCD sampling designs. However, the more recent NLCD products included the objective of assessing accuracy of change so the number of strata was increased and the strata re-defined to accommodate the objective of estimating the accuracy of change. Sampling designs that use the reference classification to define strata are conspicuously absent from these examples (Table 2). This is because the usual implementation of stratified sampling requires that all pixels in the region be assigned to a stratum, and reference class labels are not available for all pixels in the entire population. If complete coverage reference data were available we would have no need for sampling to conduct the accuracy assessment.

The high cost of obtaining accuracy reference data has motivated development of methods to improve efficiency of reference data collection. Having a common repository of reference data applicable to assessing accuracy of several map products would be a valuable addition to land cover science (Loveland et al. 2000; Wulder et al. 2018, p. 4266; Zhao et al. 2014). Olofsson et al. (2012) and Stehman et al. (2012) described a global sampling design tailored to this purpose, and Tsendbazar et al. (2018) implemented a modified version of this design in Africa. Efforts to implement sampling designs for collecting reference data that can be used for multiple maps should continue to be pursued as the current practice of collecting reference data for every individual mapping project is clearly an inefficient use of resources. Crowdsourcing and Volunteered Geographic Information (VGI) offer appealing possibilities for decreasing the cost of obtaining the reference classifications. Stehman et al. (2018) elucidated the sampling design and inference issues encountered when combining sample VGI and authoritative data, and Waldner et al. (2019) implemented a global stratified systematic sample to evaluate methodology for augmenting expert interpretation with VGI. Extracting reference data from ongoing land-cover monitoring sampling programs is another possibility to enhance efficiency of reference data collection. For example, the Land Use/Cover Area-frame sample (LUCAS) in Europe has been used for accuracy assessments (Gallego 2011) and Kempeneers et al. (2013) used sample data from several national forest inventories for accuracy assessment. Obtaining reference data from two or more sampling programs that were not specifically designed to provide accuracy assessment information introduces several challenging issues for use of these data (Stehman et al. 2000), but if these issues are addressed data from these programs may enhance accuracy assessment and area estimation.

2.7. *Sampling design for training and validation data*

Using the same data for training and validation can lead to optimistically biased accuracy assessments (Hammond and Verbyla 1996). Consequently, the training sample and the validation sample need to be independent of each other which can be achieved by appropriately dividing a single sample of reference data or, perhaps more commonly, by acquiring separate samples for training and testing. Moreover, the goals and needs of training are different from those of accuracy assessment and hence the optimal sample data for training could be markedly different from that for accuracy assessment. For example, purposive sampling and pragmatic site selection can be perfectly acceptable in training a classifier, whereas a probability sampling design is a good practice requirement for accuracy assessment.

Independence of the accuracy assessment sample from the training can be assured simply by implementing a separate probability sample for accuracy assessment. That is, all pixels in the region of interest (ROI) have a known and non-zero probability of being included in the reference sample for accuracy assessment. Often the reference sample used for accuracy assessment is selected after the map to be assessed has been produced. This allows use of stratified sampling where the strata are defined by the map classes. Constructing the strata from the map does not violate the independence requirement as the strata affect only precision of the estimates and using the map for stratification does not introduce bias in the accuracy estimators.

In some applications training and validation data are collected simultaneously from a common sampling design followed by random assignment of each sampled pixel to the training or validation set. In such applications, a probability sampling design must still be implemented to use design-based inference in a rigorous accuracy assessment, even though probability sampling may not be a necessity for training data. If cluster sampling is the probability sampling design implemented, the random separation of sample units to training or accuracy assessment should be done at the cluster level rather than the pixel level (i.e., if a cluster is randomly assigned to training data, then all sampled pixels within that cluster would also be assigned to training data). If pixels within a sampled cluster are randomly assigned to both training and accuracy assessment, the independence of the training and accuracy assessment samples is violated. Several studies have found that optimistic estimates of accuracy are produced if individual pixels rather than clusters are the units randomly separated to training and validation (Friedl et al. 2000; Zhen et al. 2013; Lyons et al. 2018).

Table 2. Example sampling designs used for accuracy assessment (listed in chronological order of publication). The three questions determining a sampling design (Sec. 2.2) are: 1) Are there strata? 2) Are there clusters (and if so, what are the primary sampling units (PSU) and secondary sampling units (SSU)? 3) What selection protocol was implemented?

Source	Strata	Clusters	Selection Protocol
Fuller (1994)	32-class environmental stratification	No	Simple random within strata
Edwards et al. (1998)	1) 3 ecoregions 2) Road, off-road within each ecoregion	PSU: USGS quadrangle SSU: 1 ha block	1st: Simple random within strata 2nd: Simple random
Scepan (1999)	Map land cover	No	Simple random within strata
Nusser and Klaas (2003)	1) 5 geographic strata 2) Land-cover class	PSU: 7.5' quadrangle SSU: pixel	1st: Systematic within strata 2nd: Systematic
Wickham et al. (2004)	1) 10 mapping regions 2) 60 x 30 km cell 3) map land cover	PSU: aerial photograph SSU: pixel	1st: One random PSU per grid cell 2nd: Simple random within strata
Wickham et al. (2010)	1) 10 mapping regions of US 2) 120 x 120 km cell 3) map land cover	PSU: 12 km x 12 km SSU: pixel	1st: One random PSU per grid cell 2nd: Simple random within strata
Olofsson et al. (2012) and Stehman et al. (2012)	Regional related to biogeography	PSU: 5 km x 5 km SSU: variable	Simple random within strata
Wickham et al. (2013)	1) Mapping regions of US 2) Map land cover	No	Simple random within strata
Gong et al. (2013) and Zhao et al. (2014)	7000 hexagons globally, approximately equal area	No	Simple random within strata
Zimmerman et al. (2013)	1) Three regional strata (PSUs) 2) Disturbance classes (SSUs)	PSU: 1/8 x 1/8° quadrangles SSU: map pixel	1st: Simple random within strata 2nd: Stratified random
Fichet et al. (2014)	20 x 20 km cell	PSU: 2 km x 2 km SSU: point	1st: One random PSU per stratum 2nd: simple random
Linke et al. (2017)	1) 5 biomes 2) Change type within biome	No	Simple random within strata
Wickham et al. (2017)	1) East and west of US 2) Change (e.g., forest loss) and stable classes	No	Simple random within strata
Badjana et al. (2017)	Stable and change classes	No	Simple random within strata
Waldner et al. (2019)	Four strata representing mis-classification probabilities	No	Systematic (specialized)

3. Response design

The response design defines how a decision on the agreement between the predicted (map) class label and the reference class label is made. Although this may seem a simple task the response design entails fundamental issues that greatly influence an accuracy assessment, including specifying the spatial unit of the assessment, deciding what ground data or imagery to collect, determining the labelling protocol for the reference data, and deciding how thematic agreement will be defined. Key general principles of the response design are identified in this section, but because details of the response design are very specific to each land-cover application, we do not delve into individual case study examples. Congalton and Green (1999, Chapter 4) and Olofsson et al. (2014) are good sources for additional description of response design options.

A variety of reference data sources have been used in accuracy assessments. Popular sources include data acquired directly in the field and imagery with a spatial resolution considerably smaller than that used to produce the land cover product being evaluated. The former may include fieldwork undertaken specifically for the accuracy assessment (e.g., Nusser and Klaas 2003) and rigorous authoritative studies such as forest inventories (e.g., Wulder et al. 2006b; Kempeneers et al. 2013). The latter data source may be aerial photographs (e.g., Wickham et al. 2004), airborne video (e.g., Marsh et al. 1994) or fine spatial resolution satellite sensor data such as are available via Google Earth (Gorelick et al. 2017; Wickham et al. 2017; Badjana et al. 2017; Tsendbazar et al. 2018).

In some applications, the same source of imagery used to determine the accuracy assessment reference classification may have also been used to produce the map classification. This typically occurs when it is not possible to obtain better imagery for the accuracy assessment. For example, land-cover change studies may be dependent on Landsat imagery both for accuracy assessment and for classifier training. In such applications, it is necessary to obtain the reference classification in a manner that ensures better quality than can be obtained from the classifier used to produce the map. Often better quality is achieved by focusing attention on a much smaller area (i.e., the reference sample) and employing visual interpretation of the imagery to obtain the reference classification (Cohen et al. 2010). Determining the reference classification via a different protocol from that used to produce the map, even if both employ common imagery, is acceptable to maintain independence of the reference and map classifications. Moreover, interpreters need not be different from the operator(s) who developed the classification as often the team that produced a classification has the most incentive, the most resources, and the greatest expertise for obtaining the most accurate reference interpretation. A major benefit of disclosing the sample locations, a feature of reproducibility of the sampling design (Sec. 2.2),

is that if there is any doubt regarding impartiality or expertise of the interpreters, the opportunity would exist to independently verify the reference class interpretations.

Even though gold standard reference data are rarely feasible to obtain, it is still essential that the reference dataset be more accurate than the map to be evaluated. The selection of reference data source, which will vary with data availability and cost constraints, can have important impacts on reference data quality. For example, the quality of the data obtained from fine resolution imagery may vary in relation to their date of acquisition relative to that of the data used to generate the land cover map (the closer the two data sets are in time the better). The quality of the data obtained via field work or image interpretation may vary with the expertise of the staff involved. Interpreters trained for a specific application are often used to obtain the reference class labels, but reference data derived via crowdsourcing have become increasingly available (Fritz et al. 2009; Fonte et al. 2015). Regardless of the reference labelling protocol, the quality of the reference data set is important as it can be a source of substantial mis-estimation of accuracy and land cover change (Foody 2010, 2013).

The reference data labelling protocol addresses the conversion of the reference data into a class label for comparison against the land cover product to be evaluated. This includes issues such as the minimum mapping unit and rules to allocate a class label. The minimum mapping unit is a fundamental variable that should be selected with care; if the unit is too large it may not adequately represent highly fragmented land cover mosaics. The rules to label an observation also require careful selection. There are many definitions of land cover classes (Ahlqvist 2005; Comber et al. 2005). The map relevant criterion (Sec. 1.2) requires that the definition of classes used for the reference data must be the same as those used in mapping, and an implicit assumption in cross-tabulating map and reference data is that the classes are exactly the same (Congalton and Green 1999).

Accuracy assessment based on the analysis of an error matrix is site-specific in that the label in the map is compared to a reference label and the spatial unit defines the fundamental geographical region upon which the comparison is based. Commonly assessments are undertaken on a pixel, group of pixels (e.g. block), or an object basis (Radoux et al. 2011; Stehman and Czaplewski 1998, Table 1; Stehman and Wickham 2011). Pixels are convenient as the basic building block of digital images but are artificial units whose boundaries need not relate in any meaningful way to the land cover mosaic on the ground. Conversely, objects are (potentially) natural spatial units that relate to real-world land cover patches, but the quality of objects is often highly dependent on the method to produce them. The analyses employed in a conventional accuracy assessment (Section 4) are typically constructed based on the assumption that the spatial unit represents homogeneous cover of a single class. Unfortunately

mixed pixels and mixed objects often occur, representing a source of ambiguity and potential error in accuracy assessment, and alternative analyses must be considered in such circumstances (Section 6).

The quantification of accuracy requires rules to define agreement between the map and reference classifications. If each classification provides a single label for an assessment unit (e.g., a pixel) and the two labels agree the unit would be regarded as being correctly classified. Should the labels disagree the map label is deemed to be incorrect, an assertion predicated on the assumption of perfect reference data. Accordingly, all error in the assessment is placed upon the map which may be unfair because issues such as geo-reference error in addition to reference label error may be present (Foody 2008; Verbyla and Hammond 1995). Sometimes the assessment is more difficult, for example, if a spatial unit can belong to more than one class. Here, agreement might focus on the dominant class or estimated class composition (e.g., Latovic and Olthof 2004) in the map and reference data set.

Deviations from the common circumstances assumed in the discussion of the core good practice may occur such as when the map and reference data do not use the same set of classes or spatial units. It is possible to adapt the accuracy assessment to accommodate such circumstances while maintaining the statistical rigor and credibility. For example, when the set of classes differ an accuracy assessment can be undertaken by bringing them to a common class scheme. Alternatively methods for non-square matrices such as entropy-based average mutual information content (Finn 1993; Foody 2006) may be used. Similarly, the use of different spatial units can be recognized in the response design of the accuracy assessment allowing adaptation from typical per-pixel based assessments for other scenarios such as the use of blocks or objects (Stehman and Wickham 2011; Ye et al. 2018).

Given that many options are available for each of these aspects of protocol in the response design, it is critical that a detailed description of the response design is provided to satisfy the reproducibility criterion. Specifically, the following key features of the response design should be documented: 1) definition of spatial assessment unit; 2) definition of classes; 3) sources of reference data (e.g., imagery, field visit, date of imagery or field visit); 4) specific information collected from each source (e.g., direct assignment of a reference class label versus collection of an assortment of attributes that are later converted to a reference class label); 5) rules for assigning reference class label(s) based on the reference information collected; and 6) specification of how agreement between the map and reference classifications is defined.

Documentation of the response design also requires considerable information to satisfy the transparency criterion, such as: 1) if the reference label was determined by human interpreters, describe characteristics of the interpreters (i.e., their background, how they were trained); 2) state

whether interpreters were unaware of (i.e., "blind" to) the map class of the sample units they were interpreting; 3) state whether interpreters were oblivious to information regarding sample allocation to strata that might consciously or unconsciously lead to trying to produce reference interpretations that would correspond to this known distribution of the sample by strata (e.g., if interpreters know that the sample design was equally allocated to strata with 50 pixels per stratum, it is plausible that interpreters would produce reference labels that were close to equal per class); and 4) provide information regarding the date of reference imagery or reference ground data to allow evaluating whether discrepancies in time between data for accuracy assessment and map production could be responsible for some portion of classification error. Land-cover studies that provide thorough explanations contribute to transparency and acquire credibility attributable to full disclosure and a self-critical evaluation. Acknowledging potential limitations of the assessment provides map users with a more informed understanding of the accuracy results and this is preferred to less transparent reporting that would give the impression of a stronger assessment than what took place in reality.

4. Analysis

The analysis component focuses on: 1) how to organize and summarize information to quantify accuracy, and 2) how to estimate accuracy and area from the sample data. Historically, the analysis protocol for accuracy assessment has depended on an error matrix and accuracy measures derived from the error matrix, specifically overall, user's, and producer's accuracies. The error matrix also includes the information needed to estimate the proportion of area of each class based on the reference classification. Because of its ease of interpretation and valuable descriptive information, the error matrix should remain a cornerstone of the analysis protocol. In practice, the analysis should be map relevant and statistically rigorous, focusing on easily interpretable accuracy measures. To be map relevant, the error matrix and accuracy estimates must reflect the proportional areal representation of the study region (e.g., as shown in Table 3, the row totals of the error matrix represent the proportion of area mapped of each class). The primary requirements for the analysis to satisfy the criterion of statistical rigor are to employ consistent estimators and to quantify the variability of the accuracy and area estimates by reporting standard errors or confidence intervals (Sections 4.2 and 4.3). Deviating from these exacting requirements for statistical rigor is sometimes justified by an argument for pragmatism, but such deviations carry with them the risk that the resulting accuracy and area estimates will have poor credibility. The analysis should also include provisions to assess quality of reference data and to evaluate the potential impact that bias and variability of the reference classification may have on

the accuracy and area estimates. Finally, it is important to note that the error matrix supplies a global (entire map) level of assessment, but cannot provide information regarding the spatial distribution of classification errors. Consequently, a spatial display of classification error and/or uncertainty may often be beneficial (Sec. 4.7).

4.1. Error matrix

The error matrix has traditionally provided an effective basis for organizing and summarizing agreement between the map and reference classifications (Sec. 1.1). A population error matrix representing a census of reference condition, and the corresponding notation defining population parameters used to describe accuracy are displayed in Table 3. In this error matrix, the rows represent the map classification and the columns represent the reference classification. The cell entry p_{ij} is the population proportion of area for which the map class is *i* and the reference class is *j*. The main diagonal cells of the error matrix represent correct classifications, whereas the off-diagonal cells indicate which classes are confused, providing valuable information for improving the map and for users to evaluate how specific types of misclassification might affect their applications. The population error matrix simplifies greatly to a 2x2 matrix for the common case of a binary legend (e.g., change and no change, burned area and not burned, or water and not water).

Table 3. Population error matrix for a classification with four classes (1-4), where the rows (*i*) represent the map classification and the columns (*j*) represent the reference classification; p_{ij} is the population proportion of area with map class *i* and reference class *j*. The row (p_{i+}) and column (p_{+j}) marginal totals are the sum of the p_{ij} values in each row and column.

Map	1	2	3	*4*	Total	*User's accuracy*
1	p_{11}	p_{12}	p_{13}	p_{14}	p_{1+}	p_{11}/p_{1+}
2	p_{21}	p_{22}	p_{23}	p_{24}	p_{2+}	p_{22}/p_{2+}
3	p_{31}	p_{32}	p_{33}	p_{34}	p_{3+}	p_{33}/p_{3+}

4	p_{41}	p_{42}	p_{43}	p_{44}	p_{4+}	p_{44}/p_{4+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+4}	1	
Producer's accuracy	p_{11}/p_{+1}	p_{22}/p_{+2}	p_{33}/p_{+3}	p_{44}/p_{+4}		

The row and column totals of the population error matrix are important because they quantify the area distribution (e.g., fractional cover) of the classes in the ROI. Given the layout of Table 3, the row totals represent the proportion of area of each class according to the map classification and the column totals represent the proportion of area according to the reference classification. The estimated row and column totals resulting from the analysis protocol must represent the reality of the ROI to satisfy the map relevant criterion (Sec. 1.2). Normalizing an error matrix (Congalton et al. 1983) is an egregious violation of the map relevant criterion as it results in an error matrix in which all row and column totals are forced to equal to 1/*c* (*c*=number of classes). Thus a normalized matrix has no basis in the reality of the map or ground condition as rarely will all classes be present in equal proportions (Stehman 2004). Further, normalizing an error matrix results in user's and producer's accuracies being equal and this also will almost never be a real feature of a land-cover map.

The typical suite of accuracy measures computed from the population error matrix includes overall, user's, and producer's accuracies. Overall accuracy is the sum of the entries on the main diagonal of the error matrix,

$$O = \sum_{i=1}^{c} p_{ii} \qquad (2)$$

Overall accuracy has a direct interpretation in terms of area of the ROI as it represents the proportion of area correctly classified. Overall accuracy provides a very coarse assessment because it aggregates information over all map classes thereby obscuring important class-specific information. For example, if the map is a binary change/no change classification and change is rare (e.g., 1 or 2% of the total area), a naïve map that shows no change will have very high overall accuracy simply because most of the area is unchanged. In this example, class specific information such as user's and producer's accuracies or quantity disagreement and allocation agreement (described in subsequent equations) would be more informative than reporting just overall accuracy. Overall accuracy does not "overweight" more prevalent classes or "underweight" less common classes, but instead incorporates classes proportionally to their areal representation in the ROI. The limitation of overall accuracy is not how it weights or

represents class-specific information, but rather that it simply provides no class-specific information. Consequently, it is strongly recommended to provide estimates of user's and producer's accuracies or to provide the error matrix when reporting an accuracy assessment so that these class-specific measures can be estimated by others if desired. User's accuracy for class *i*, defined as

$$U_i = p_{ii}/p_{i+} \quad (3)$$

quantifies accuracy conditional on the area mapped of each class, the row total p_{i+}. The complement of user's accuracy (1-U_i) is the commission error rate of class *i*. Producer's accuracy for class *j*, defined as

$$P_j = p_{jj}/p_{+j} \quad (4)$$

provides a view of accuracy conditional on the reference area of each class, the column total p_{+j}. The complement of producer's accuracy ($1 - P_j$) is the omission error rate of class *j*.

Having two measures of class-specific accuracy has caused some consternation when the objective is to compare accuracy or decide which of two classifiers or maps is better, leading to omnibus measures that provide a single value to characterize accuracy of a given class. Liu et al. (2007) reviewed a variety of such measures, including the Dice coefficient, Jaccard's index of similarity, the average of user's and producer's accuracy and the harmonic mean of user's and producer's accuracy (also called Hellden's Index and F-score). These omnibus measures treat the off-diagonal cell proportions for a given class (p_{ij} and p_{ji}) as equally important. An example in which it would be reasonable to make no distinction between the off-diagonal entries would be when comparing two analysts who provide independent reference interpretations for a common set of pixels in an evaluation of analyst consistency. Because the identity of the two analysts would be interchangeable, it would be reasonable to create a single measure of agreement per class between the two analysts. However, in accuracy assessment the map and reference classifications are not interchangeable as errors of omission and commission are not interchangeable. Consequently, combining omission error and commission error into one measure obscures valuable information. Omission and commission errors may not be equally problematic for the objectives of a given application of the map so having the separate estimates of user's and producer's accuracies is often critical. If omnibus measures are reported, they should be accompanied by the class-specific measures.

Non-site-specific accuracy is defined as the difference between the row and column proportions of each class,

$$p_{k+} - p_{+k} \quad (5)$$

(Congalton and Green 1999). Because omission and commission errors may negate each other, non-site-specific accuracy may be close to 0 even if both omission and commission error rates are high. Non-

site-specific-accuracy has a direct connection to the objective of area estimation (Sec. 4.3). For area estimation, the target parameter is p_{+k} because the reference classification is considered the best assessment of ground condition. Consequently, non-site-specific accuracy may also be viewed as "map bias" in that the difference $p_{k+} - p_{+k}$ represents the degree to which the proportion of area mapped as class *k* differs from the proportion of area of class *k* based on the reference classification. Map bias is thus the bias attributable to using pixel counting to estimate the area of each class (i.e., count the number of pixels for each class and multiply by the area per pixel).

Pontius and Millones (2011) proposed partitioning the overall differences between the map and reference classifications into two mutually exclusive components, quantity and allocation differences. Quantity disagreement of class *k* is defined as the absolute value of Eq. (5),

$$q_k = |p_{k+} - p_{+k}| \tag{6}$$

and overall quantity disagreement (aggregating the class-specific disagreement) is defined as

$$Q = \sum_{k=1}^{c} q_k / 2 \,. \tag{7}$$

Allocation disagreement for class *k* is defined as

$$a_k = 2min[p_{+k} - p_{kk} \,,\; p_{k+} - p_{kk}] \tag{8}$$

and overall allocation disagreement as

$$A = \sum_{k=1}^{c} a_k / 2. \tag{9}$$

Lastly, if total disagreement is defined as

$$D = 1 - \sum_{k=1}^{c} p_{kk} \tag{10}$$

overall agreement can be partitioned as the sum of overall quantity disagreement and overall allocation disagreement,

$$D = Q + A. \tag{11}$$

Pontius and Santacruz (2014) and Pontius (2019) have further developed this approach to describing map quality by partitioning the allocation difference into two components, exchange in which a pair of pixels swap class labels (e.g., a pixel classified as A by the map has reference class B and the paired pixel is classified as B by the map but has reference class A), and shift in which the allocation difference is not exchange. These components of the difference between the map and reference classifications may provide insights into map quality to supplement the understanding gleaned from the error matrix and user's and producer's accuracies.

4.2. Estimating the error matrix and accuracy from a sample

In practice, the error matrix and associated accuracy measures are estimated from the sample of reference data. The formulas for estimating the cell proportions of the error matrix and overall, user's, and producer's accuracies depend on the sampling design used. Satisfying the reproducibility criterion requires that the formulas used for estimating accuracy and area and their associated standard errors must be clearly documented. Pragmatically, the estimation formulas must incorporate a weight for each sampled pixel *u*, where the weight is $1/\pi_u$ (Stehman et al. 2018). In those cases for which the inclusion probabilities are not the same for all sampled pixels, the estimator formulas applied to the sample data do not necessarily match the formulas defining the population parameters presented in Sec. 4.1. For example, stratified sampling formulas (Eqs. 12-18) accommodate the fact that pixels sampled from different strata may have different inclusion probabilities, π_u, and the sample data from different strata must be weighted. The dependence of the estimator formulas on the sampling design is the essence of the property of consistent estimation (Särndal et al. 1992, Sec. 5.3; Overton and Stehman 1995) that serves as a necessary ingredient of a statistically rigorous accuracy assessment. Simply put, the consistency property is met by using estimator formulas specific to the sampling design implemented.

The majority of accuracy assessment studies employ simple random, systematic, or stratified random sampling. Stratified estimation formulas (Card 1982) are particularly important because of their general utility to all three of these designs. Stratified estimation formulas are, of course, always applied to a stratified sampling design. However, even when the sampling design is simple random or systematic, stratified estimation formulas can be applied, a protocol called poststratification or poststratified estimation, where "post" indicates that the strata information is employed after the sample has been obtained at the estimation (analysis) stage, but strata are not used in the sampling design. Thus, poststratification uses stratified estimation formulas applied to sample data from an unstratified design. The same information required to implement a stratified sampling design is required to implement poststratified estimation. That is, we must know the stratum to which each pixel in the ROI belongs, and we must know the proportion of area of the ROI each stratum occupies.

We present the stratified estimators because of their common use in practice. For these sample estimators we assume that the strata are defined as the classes mapped (e.g., a pixel with map class *i* is assigned to stratum *i*), and that all pixels have the same area. The stratified estimator (estimators are denoted by placing a ^ over the parameter being estimated) for the p_{ij}=proportion of area in cell (*i*, *j*) of the error matrix is

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i+}} \tag{12}$$

where W_i is the proportion of area mapped as class *i*, n_{ij} is the number of sample pixels in cell (*i*, *j*) of the error matrix, and n_{i+} is the sample size from stratum *i*. Estimators for other parameters characterizing the error matrix can be produced by substituting $\hat{p}_{ij}$ for p_{ij} in the formulas for each parameter. The estimator of overall accuracy is then

$$\hat{O} = \sum_{i=1}^{c} \hat{p}_{ii} = \sum_{i=1}^{c} W_i \frac{n_{ii}}{n_{i+}} \tag{13}$$

the estimator of user's accuracy of class *i* is

$$\hat{U}_i = \hat{p}_{ii}/\hat{p}_{i+} \tag{14}$$

and the estimator of producer's accuracy of class *j* is

$$\hat{P}_j = \hat{p}_{jj}/\hat{p}_{+j} \tag{15}$$

where $\hat{p}_{+j} = \sum_{i=1}^{c} \hat{p}_{ij}$ is the estimated proportion of area in reference class *j*. The standard errors of the stratified estimators are obtained by taking the square root of the estimated variance, where the variance estimators are as follows:

$$\hat{V}(\hat{O}) = \sum_{i=1}^{c} W_i^2 \hat{U}_i(1-\hat{U}_i)/(n_{i+}-1) \tag{16}$$

$$\hat{V}(\hat{U}_i) = \hat{U}_i(1-\hat{U}_i)/(n_{i+}-1) \tag{17}$$

$$\hat{V}(\hat{P}_j) = \frac{1}{\hat{N}_{+j}^2}\left[\frac{N_{j+}^2(1-\hat{P}_j)^2\hat{U}_j(1-\hat{U}_j)}{n_{j+}-1} + \hat{P}_j^2 \sum_{i\neq j}^{c} N_{i+}^2 \frac{n_{ij}}{n_{i+}}\left(1-\frac{n_{ij}}{n_{i+}}\right)/(n_{i+}-1)\right] \tag{18}$$

where $\hat{N}_{+j} = \sum_{i=1}^{c} \frac{N_{i+}}{n_{i+}} n_{ij}$ is the estimated total number of pixels of reference class j, N_{j+} is the total number of pixels of map class *j* in the entire map, and n_{j+} is the sample size of stratum *j*. For simple random and systematic sampling these are poststratified variance estimators following the recommendation of Särndal et al. (1992, Sec. 7.10.2) to "condition" the estimators on the actual sample sizes n_{i+} for each stratum. Further, for systematic sampling the variance estimators presented above are approximations that usually result in overestimation of variance.

In practice, situations arise in which the map class of a pixel is not the same as the stratum to which it belongs, so in such cases the strata do not correspond exactly to the map classes. For example, suppose that a stratified sample is selected using two strata defined by change and no change as determined from Map Y. After the reference sample data have been obtained, the producers of Map Z want to use the same reference data for their change/no change classification. However, some of the pixels that were labelled as change by Map Y are not change in Map Z, so for Map Z the stratum to which a pixel was assigned when the sample was selected is not necessarily the same as the Map Z class

label of the pixel. These reference sample data may still be used to assess accuracy of Map Z, but the conventional stratified formulas (Eqs. 12-18) must be replaced by different formulas based on indicator functions (Stehman 2014). Another common situation for which the strata do not correspond to the map classes is when a hierarchical classification legend is employed. For example, suppose stratified sampling is implemented and three of the strata are deciduous forest, coniferous forest, and mixed forest. If a general forest class is defined as the combination of these three classes, this forest class is not equivalent to a stratum for this sampling design so we cannot simply combine sample counts from the three forest-related strata. Again an indicator function approach provides the needed general framework for estimation that can be applied to this situation (Stehman 2014).

Estimation formulas for cluster sampling depend on whether one-stage or two-stage cluster sampling is implemented and on the specific sampling design employed at each stage. It is critically important that the variance estimation formulas account for the fact that the sample units are clustered because otherwise variances may be substantially underestimated. Stehman (1997a) presents formulas relevant to one-stage cluster sampling in which the clusters are selected using simple random sampling. Estimation formulas for various two-stage cluster sampling designs may be found in Zimmerman et al. (2013), Edwards et al. (1998), Nusser and Klaas (2003), Stehman et al. (2003), and Potapov et al. (2014). The variance estimation formulas may become complex when the cluster sampling design includes stratification at one or both stages.

4.3. Area estimation

Accuracy assessment and area estimation have been closely linked dating back at least to Card (1982) as area estimation is a common objective of land cover studies. Although the area of each class can be readily computed from a map (i.e., "pixel counting"), classification errors likely result in bias of the map area obtained from pixel counting (Eq. 5). Consequently, Olofsson et al.'s (2014) good practice recommendation is that area estimation should be based on the reference classification because it is the best possible assessment of ground condition (although see Waldner and Defourny (2017) for another perspective). The reference classification is available only for the pixels in the sample so it is necessary to estimate area from the sample. In this review, we limit attention to area estimation conducted in conjunction with accuracy assessment. Gallego (2004) and Stehman (2009b) provide expanded overviews of area estimation that encompass applications in which accuracy assessment is not an objective.

The proportion of area of each class based on the reference classification can be estimated directly from the error matrix if the cells of the error matrix are reported in terms of proportion of area (Table 3). For the commonly used sampling designs of simple random, systematic, or stratified random, an unbiased estimator of the proportion of area of class *k* (p_{+k}) is the sum of the estimated error matrix cell entries for column *k*,

$$\hat{p}_{+k} = \sum_{i=1}^{c} \hat{p}_{ik} \quad (19)$$

where $\hat{p}_{ik}$ is given by Eq. (12) and *c*=number of classes. Note, as explained below, this calculation accounts for the effect of errors of omission and commission on the area estimates for a class. The variance estimator for $\hat{p}_{+k}$ is

$$\hat{V}(\hat{p}_{+k}) = \sum_{i=1}^{c} W_i^2 \hat{p}_i(1-\hat{p}_i)/n_{i+} \quad (20)$$

where $\hat{p}_i = \hat{p}_{ik}/p_{i+}$ is the estimated proportion of area of class *k* in stratum *i* (Eqs. 19 and 20 assume that the strata are equivalent to the map classes). For simple random and systematic sampling designs, Eqs. (19) and (20) represent poststratified estimation of area (Sec. 4.2).

4.3.1 Role of the map in area estimation

Although the fundamental data used for area estimation are the reference classifications for the sampled pixels, the map still plays the important role of reducing standard errors of these area estimates. In the case of stratified sampling design or poststratified estimation, the map provides the information to construct the strata (e.g., Gallego and Bamps 2008). As a general rule, the more accurate the map, the greater the reduction in the standard errors of the area estimates compared to a strategy that does not use strata determined from the map.

Understanding that the role of the map is to reduce standard errors is critical because the terminology used for area estimation is sometimes confusing in this regard. The terms “error-adjusted” (Olofsson et al. 2013) or “bias-adjusted” estimator (Stehman 2013) are a source of this confusion. These phrases are based on the intuitively appealing idea that the bias of the pixel counting estimate of area can be compensated for by adjusting the area using a sample-based estimate of this bias. From the population error matrix (Table 3), the proportion of area of class *k* according to the reference classification can be expressed as

$$p_{+k} = p_{k+} - \sum_{j \neq k} p_{kj} + \sum_{i \neq k} p_{ik} \quad (21)$$

which upon inspection can be viewed as the proportion of area mapped as class *k* from pixel counting (p_{k+}) and then adjusting this area by subtracting the proportion of area of commission error and adding

the proportion of area of omission error of class *k*. Rather than estimate the three separate components of the righthand side of (21), we simply estimate p_{+k} directly using Eq. (19).

An alternative estimator of proportion of area directly originating from an adjustment of the map proportion has been proposed (McRobers and Walters 2012). That is, the bias of the pixel counting map proportion p_{k+} is given by $(p_{+k} - p_{k+})$. This bias can be estimated from the sample data, and then used to adjust the map proportion p_{k+} via

$$\hat{p}_{+k} = p_{k+} + \left(\hat{p}_{+k} - \hat{p}_{k+}\right) \tag{22}$$

where the term in parentheses is an estimator of the map bias given by Eq. (5), also known as non site-specific accuracy. The estimator (22) thus starts with the map proportion of area from pixel counting and adjusts that area for map bias estimated from the sample. Referring to this estimator as bias-adjusted mis-directs the focus of the role of the map in the process because the phrase bias-adjusted implies that the map supplies the fundamental data for the area estimator. In reality the reference sample data are the basis for the estimator of area, and the role of the map is to provide the auxiliary information that contributes to reducing the standard error of the area estimator via either stratified sampling design or poststratified estimation. The key consideration regarding the map is not whether pixel counting leads to bias but rather how much benefit does the map provide in terms of reducing the standard error of the area estimators. An additional consideration is that the so-called bias adjusted estimator (Eq. 22) is not equivalent to the stratified estimator (Eq. 19) (unless the sampling design is stratified random). The stratified estimator (Eq. 19) is generally preferred to the estimator given by Eq. (22) because of smaller standard error (Stehman 2013).

The sample provides estimates of the total area or proportion of area of each class, whereas the map provides the spatially explicit information on how the area of each class is distributed across the landscape. Consequently both the sample-based area estimates and the map spatial depiction are critically important. For consistency, the area mapped for each class would ideally match the area estimated from the sample. It would thus be desirable to adjust the map to create this correspondence with class area sample estimates (Song et al. 2017; Healey et al. 2018) or, for example, inventory data such as from agriculture and forestry surveys (Ramankutty et al. 2008). How best to make such area adjustments merits further study. The sample estimates of area have associated uncertainty as quantified by the confidence intervals, so techniques to adjust a map to match sample-based estimates of area would also need to incorporate the inherent uncertainty of the area estimates.

4.3.2 Pixel counting versus sample-based estimation of area

The good practice recommendation to estimate area using the reference classification (Olofsson et al. 2013, 2014) is based on the premise that the reference classification is obtained with negligible error (Sec. 6.1) and variability (Sec. 6.3). The former may introduce bias of the area estimator and the latter will increase the variance of the area estimator. In reality, reference data error and variability thus affect whether estimating area based on a sample of reference data is superior to the pixel counting approach to area estimation. Mean square error (MSE) can be used to decide which approach to area estimation, sample-based or pixel counting, is better, where

$$\text{MSE} = \text{Variance} + \text{Bias}^2 \quad (23)$$

and smaller MSE is preferable because it represents smaller uncertainty (Stehman 2005). When calculating area from the map (pixel counting), the bias is attributable to map classification error (Eq. 5), whereas reference data error would be the only source of bias for the sample-based approach to estimating area. In terms of the variance contribution to MSE, the sample-based estimator would include contributions from sampling variability, as quantified by the standard error, and variability resulting from the reference classification (i.e., interpreter variability, Sec. 6.3). It is less clear what contributions to variance should be considered when calculating area from the map. The map could be considered a census of the ROI in which case there would be no variance contributed from sampling variability (Stehman 2005). However, it would seem reasonable to consider variability of the mapping process as relevant, for example variability over different sets of training data or variability over different technicians implementing the classification.

Area estimation plays an important role in land cover science and a variety of considerations impact the utility of area estimates produced from pixel counting versus sample-based estimation. Uncertainty of the area estimators is one aspect of utility, and quality of reference data, both accuracy and consistency, have received greater attention as contributors to the overall uncertainty of area estimates. The conceptual foundation of the comparison of uncertainty requires further development to help resolve the question of which approach, pixel counting or sample-based estimation, is better in a given application.

4.4. Reporting accuracy and area

Early presentations of the error matrix often focused on sample counts by displaying the number of sample pixels that fell in each cell of the matrix (Congalton and Green 1999, pp. 46-47; Congalton 1991, p.36). This practice is justifiable for equal probability sampling designs such as simple

random and systematic sampling. However, if the sampling design is not equal probability, an error matrix comprised of sample counts does not provide the correct representation of the ROI in terms of the area proportions for the cells of the error matrix, thereby violating the map relevant criterion. Further, accuracy estimators produced from an error matrix of sample counts will be biased. Even when the sampling design is equal probability, poststratified estimation should typically be implemented and this will result in different estimated proportions $\hat{p}_{ij}$ from those obtained using the sample counts (see the numerical example in the following paragraph). The recommended reporting format is consequently an error matrix expressed in terms of percent or proportion of area (as in Table 3), with an error matrix of sample counts provided as an appendix or supplemental information.

A numerical example illustrates the basic principles of a statistically rigorous, map relevant analysis. In this example, the error matrix presented as sample counts is provided in Table 4. Upon inspection of Table 4, our attention should immediately be drawn to the fact that the class percentages represented by the row totals are all equal, raising a concern regarding the map relevant criterion. Is this mapped region a highly unusual case in which all classes are equally common, or is there an alternate explanation? Given that there are four classes and all row totals are 25%, the most likely explanation is that the sampling design used to collect the reference data was stratified by the map classes with an equal sample size specified for each class. The accuracy assessment documentation, as per the reproducibility criterion, should be explicit regarding these details of the sampling design to avoid the potential for mis-interpretation of the matrix or mis-calculation of accuracy measures. Assuming the sampling design was stratified with equal allocation, a simple diagnostic evaluation of whether the analysis of the error matrix was conducted appropriately can be obtained by examining the overall accuracy reported from the error matrix (Table 4). If overall accuracy is reported as 75%, based on summing the diagonal and dividing by the sample size of n=100, the analysis is likely wrong because this calculation of overall accuracy ignores the necessary weighting of different strata (Eq. 13) when equal allocation of the sample size to strata is used. Similarly, if producer's accuracy is reported as the diagonal entry divided by the corresponding column total, this naïve calculation of producer's accuracy is also likely incorrect because again the proper stratified estimation formula (Eq. 15) has not been used.

The results of the analysis using the stratified estimation formulas are provided in Table 5. The map relevant criterion specifies that the row and column proportions should be consistent with expectations of class composition for the ROI. In the case of stratified sampling or poststratified estimation, the row totals should match closely the proportion of each class computed from the entire map (i.e., the row totals in Table 5 are consistent with W_k specified in Table 4). The Table 5 error matrix

results based on using the statistically rigorous, consistent estimators may differ substantially from the results of Table 4. Because user's accuracy is estimated only from sample data within each stratum, the unweighted estimates based on the sample counts from Table 4 are correct. However, overall and producer's accuracies are estimated from sample data combined over all strata so employing the stratified estimation formulas (i.e., the consistent estimators for the stratified design) is a necessity for correct estimation. In this example, the correct (i.e., consistent) estimator of producer's accuracy yields estimates (Table 5) that differ considerably from the naive estimates shown in Table 4 (e.g., for class A the correct estimate from Table 5 is 91% versus 70% from Table 4, and for class D the correct estimate is 43% from Table 5 compared to 73% from Table 4). This example demonstrates that adherence to correct methodology is important as naïve estimation based on the sample counts of Table 4 mis-represents the reality that is expressed by the results in Table 5.

Table 5 also illustrates a recommended format for reporting error matrices. Cell entries are reported as proportion (or percent) of the total area of the ROI which allows for computing estimates of overall, user's, and producer's accuracies directly from the error matrix. Reporting standard errors (SE) is important because an approximate 95% confidence interval can be constructed as the estimated accuracy or area ±1.96*SE to quantify the uncertainty (precision) of the estimates thus providing a key indicator of reliability of the estimates. The row and column totals provide direct representation of the composition of the ROI in terms of proportion of area of the map classification and reference classification allowing easy confirmation of the map relevant criterion. Further, map bias (Eq. 5) for each class is readily estimated as the difference between the corresponding map and column proportions (or percents). Row and column sample sizes, which strongly impact the standard errors, are provided to indicate the number of sample observations that were used for estimating user's and producer's accuracies as well as proportion of area. If area estimation is an objective, the standard error for the estimated proportion of area of each class based on the reference classification should also be included (as it is in Table 5). Because the proportion of area based on the map classification is known from the map and does not have to be estimated from the reference sample, it is unnecessary to include a standard error in association with the map proportion of area (row total).

Table 4. Error matrix presented in terms of sample counts (n_{ij}). Because the total sample size is 100, the cell numbers can also be viewed as percent of area. W_k is the percent of area mapped as class *k* in the ROI. Naïve producer's accuracy, calculated as (n_{kk}/n_{+k})*100%, and naïve overall accuracy, calculated as the sum of the diagonal entries (75%), are the estimates obtained if the stratified estimation formulas are not used. User's accuracy computed as n_{kk}/n_{k+} from the sample counts is an unbiased estimator.

	Reference						
Map	A	B	C	D	n_{k+}	User's(%)	W_k(%)
A	21	3	1	0	25	84	60
B	4	18	0	3	25	72	25
C	0	2	20	3	25	80	10
D	5	2	2	16	25	64	5
n_{+k}	30	25	23	22	100		
Naïve Prod(%)	70	72	87	73			

Table 5. Estimated error matrix based on sample data in Table 4 and assuming stratified random sampling with equal allocation. Cell entries represent percent of area. Standard errors are presented in parentheses for the user's and producer's accuracy estimates. Estimated overall accuracy is 79.6% with a standard error of 5.1% which would yield a 95% confidence interval of 69.5% to 89.7%.

	Reference				Total	User%	
Map	A	B	C	D	Area%	(SE)	*n*
A	50.4	7.2	2.4	0.0	60.0	84.0 (7.5)	25
B	4.0	18.0	0.0	3.0	25.0	72.0 (9.2)	25
C	0.0	0.8	8.0	1.2	10.0	80.0 (8.2)	25
D	1.0	0.4	0.4	3.2	5.0	64.0 (9.8)	25
TotalArea% (SE)	55.4(4.9)	26.4(4.6)	10.8(2.6)	7.4(1.9)			
Prod% (SE)	91.0(3.2)	68.2(10.8)	74.1(16.7)	43.2(11.1)			
n	30	25	23	22			

4.5. Land-cover change

The analysis for change accuracy assessment has much in common with single date map accuracy assessment as an error matrix and associated accuracy measures are still applicable (Congalton and Green 1999). A starting point for describing accuracy of change is an error matrix for the binary

change / no change classification. Of course, a full transition error matrix (van Oort 2007) that displays all class-specific transitions would provide more complete information. However, for *c* land-cover classes the error matrix of all possible transition and steady state types is c^2 x c^2, so a full transition error matrix may be too unwieldy to be broadly useful in practice. To reduce the volume of results reported, user's and producer's accuracies could be limited to more general transitions such as "any class to urban" (i.e., urban gain) and "cropland to any class" (i.e., cropland loss) (e.g., Wickham et al. 2017).

Additional accuracy metrics may be desirable to assess the accuracy of land-cover change trajectories over multiple dates. The approach in the previous paragraph could be used for change between any two dates within the time series, but such analyses would be less effective for evaluating a long time series trajectory. Cohen et al. (2010) developed several accuracy measures applicable to a time series trajectory in which segments partitioned the time series into periods of forest stability, disturbance, and recovery. These measures were based on the number of segments in the trajectory, the labels of the segments, and match scores representing the proportion of the trajectory in which the reference and map condition agreed. Pontius et al. (2017) proposed several ways to summarize change information that could be used in accuracy assessment: 1) number of incidents, defined as the number of times a pixel experiences a change across all time intervals; 2) number of states, defined as the number of different categories that the pixel represents over all time points; and 3) flow matrices which express transitions of one category to a different category between two time points. These change accuracy metrics could be obtained from the reference data for each pixel in the reference sample and compared to the corresponding map change information to provide the data for estimating accuracy metrics. Simplicity and ease of interpretation are still guiding principles for choosing how to characterize accuracy of change products. Quantifying accuracy of multi-date land-cover products remains a challenge and a subject of ongoing research (Wulder et al. 2018, Sec. 4.2).

4.6. Use of accuracy data to compare maps

Land cover products obtained from remote sensing are often compared to evaluate different mapping methods (Khatami et al. 2016). For example, a series of classifiers may be applied to a single remotely sensed dataset to determine which yields the most accurate classification, or different approaches to classifier training may be used with a single classifier to yield recommendations of protocols for the design of the training stage of a particular classifier. In such studies, the relative accuracy of the maps obtained is used to indicate the quality of the mapping method. Such analyses are typically undertaken on the assumption that the maps being compared are perfectly co-registered

spatially and that the same set of classes are used. Although seemingly straightforward the comparison must adhere to protocols that ensure a rigorous and credible comparison can be achieved. Further, map accuracy comparisons may be motivated by different objectives, for example choosing a classifier to use for a particular application or evaluating classification procedures under more general conditions of application. These objectives may require different study designs and different inference frameworks (Stehman 2006). In particular, if the results of the comparison are intended to be broadly applicable to a range of conditions, more than a single test site needs to be evaluated.

Because map accuracy is typically estimated from a sample, it is inappropriate to simply compare the estimated accuracy (e.g., overall accuracy) directly. Instead, the comparison should take into account the variances of the estimates which may be achieved by a test based on use of the standard z score (Foody 2009b). In essence this is essentially ensuring that the confidence interval fitted to each estimate is considered. This is important because the confidence intervals fitted to two dissimilar accuracy estimates may overlap indicating that, at the relevant level of confidence, the accuracy parameters (i.e., population values) do not differ statistically from each other. The width of the confidence intervals can therefore be important and the analyst has some influence over this as the width is inversely related to sample size. The latter should be selected with care as it is quite possible to obtain an exceptionally large sample but find that trivial non-meaningful differences appear statistically significant (Fleiss et al. 2003; Foody 2009b). Conversely, the use of a sample that is too small may result in confidence intervals that are so wide that finding a statistically significant difference is unlikely due to lack of statistical power. Fleiss et al. (2003) and Foody (2009c) provide guidance on choosing the sample size in relation to the aim of the study.

A further consideration in the comparison of accuracy values relates to the data used in their derivation. Considering a basic pairwise comparison, a key issue is whether the samples used to generate the accuracy assessments are independent or related. Independent samples are often assumed and standard equations suitable for such situations are presented (Congalton and Green 1999). However, it is common in remote sensing studies to use a single reference dataset in the assessment of each classification. For such related or "paired" samples the equations need to be modified to account for covariance (Stehman 1997) or an alternative method such as a McNemar test should be used (Foody 2004).

In relation to the aim of the comparison, the test to be undertaken often has a directional component. Thus, rather than simply looking for a difference in accuracy the *a priori* objective might be to determine if the accuracy of one map is greater than another. In such circumstances a test of non-

inferiority or superiority should be undertaken (Fleiss et al. 2003; Foody 2009b). In addition sometimes the desire is to show that two maps are of equivalent accuracy. This should not be assessed by showing no difference but by showing equivalence. These various tests, whether for difference, non-inferiority, superiority or equivalence, can all be undertaken on the basis of the confidence intervals obtained for the classifications to be compared (Fleiss et al. 2003; Foody 2009b).

4.7. Spatial description of accuracy

The error matrix analyses described in Section 4.1 provide only a global evaluation of thematic map accuracy but do not indicate how accuracy may vary spatially across the mapped region (McGwire and Fisher 2001). A guide to the spatial variation in map quality can be generated from some classifiers by mapping the uncertainty of class allocation. With a standard maximum likelihood classification, for example, it is possible to map the probabilities of class membership, likelihood and typicality, on a per-case basis to indicate the spatial pattern of class allocation uncertainty in addition to the most likely class label for each pixel (Foody et al. 1992; Steele et al. 1998). Similar approaches to representing the uncertainty of allocation can be made from fuzzy classifiers and machine leaning approaches. While representations of the uncertainty of class allocation do help convey useful information on classification quality they do not necessarily reflect accuracy; a classifier may, for example, confidently mis-classify cases.

Several approaches to reveal spatial patterns of accuracy have been developed. Accuracy could, for example, be assessed for sub-regions of the map (Foody 2005). For this approach to be viable, however, a large reference data set would typically be required. Spatial modelling techniques have been applied to depict accuracy spatially. For example, Kyriakidis and Dungan (2001) used a geostatistical, stochastic simulation approach and Park et al. (2016) employed indicator kriging. Comber et al. (2012) and Comber (2013) applied geographically weighted logistic regression to estimate local accuracy, and Tsutsumida and Comber (2015) extended this approach to include the temporal dimension of accuracy. Khatami et al. (2017) developed a method for producing maps of accuracy based on the spectral feature space. For large-area land-cover products, these methods that exploit spatial correlation to predict accuracy at unsampled locations may be challenged by the sparse density of the reference sample data that is common in practice. Further, the computing time for some spatial prediction methods may be excessive when working with a large area such as a national or global map.

4.8. Kappa

The kappa coefficient expressed as a population parameter is

$$\kappa = \frac{O - \sum_{i=1}^{c} p_{i+} p_{+i}}{1 - \sum_{i=1}^{c} p_{i+} p_{+i}} \quad (24)$$

Despite Pontius and Millones (2011) provocative "Death to Kappa" proclamation, a cursory review of recent published accuracy assessments reveals that reports of kappa's demise are premature. Kappa is a good illustration of the principle of primacy in which an idea or method introduced in the early stages of a developing field will persist despite evidence contraindicating its use. Although the notion of correcting for chance agreement has some intuitive appeal, the criterion of map relevance raises the question of how chance agreement is related to the reality of the map. For example, it is not possible to identify actual pixels classified correctly by random chance because chance agreement is a model construct and different models yield different chance agreement. Kappa also fails the map relevant criterion because as noted by Ye et al. (2018), a random classification is usually not a realistic alternative method to create a map.

Liu et al. (2007) showed that kappa was highly correlated with overall accuracy, which is evident from Eq. (24), so reporting both measures is redundant. In most cases, chance agreement is inconsistently applied as kappa is typically reported as an overall value (analogous to overall accuracy) but chance-corrected agreement is rarely reported at the individual class level via conditional kappa (i.e., chance-corrected analogs to user's and producer's accuracies). Pontius and Millones (2011) stated that use of kappa rarely if ever has changed the interpretation or conclusion of an accuracy assessment. In practice, kappa seems to be reported more out of a sense of obligation than to provide enlightenment of map accuracy. Thus kappa is to accuracy assessment what the appendix is to the human body – it may cause no serious harm if you have it and pay little attention to it, but it does not fulfil a necessary function.

5. Statistical inference and accuracy assessment

Statistical inference is the process of generalizing from sample data to produce estimates of population parameters. Key elements of inference include how variability and bias of estimators are defined. Design-based inference has traditionally been the inference framework invoked in accuracy assessment (Stehman 2000, 2001). In design-based inference, uncertainty is attributed to the randomization present in the sample selection, whereas the observations obtained on each sample unit are regarded as fixed constants, not random variables. Because probability sampling is a necessity when using design-based inference, it is consequently a key element contributing to the statistical rigor of

accuracy assessment (Stehman 1995; Stehman and Czaplewski 1998). The emphasis on probability sampling as a good practice recommendation (Strahler et al. 2006; Olofsson et al. 2014) is an important development in the evolution of accuracy assessment methods.

As a matter of terminology, "design-based" refers to a method of inference, not a method of sampling or estimation. That is, there is no such thing as a "design-based sample" or a "design-based estimator. Although a probability sample is required to implement design-based inference, data from a probability sample could be used in any mode of inference, so "design-based sampling" should not be used as a synonym for probability sampling. Similarly estimators such as the sample proportion and the stratified estimators of Eqs. (12-15) are not inherently design-based but the manner in which bias and variance of these estimators is defined determines whether it is being used in a design-based or other inference setting.

An assumption of design-based inference is that measurement error is negligible, which in the accuracy assessment setting translates to these reference data being correct. Therefore, design-based inference does not directly accommodate the likely reality that some sample units are assigned an incorrect reference label or that variability in the assignment of reference labels likely exists (e.g., inconsistency among multiple interpreters). Measurement error models (Särndal et al. 1992, Chapter 16) are needed to accommodate these features of reference data uncertainty in the estimation methods and quantification of variability.

Two other approaches to inference are model-based (Valliant et al. 2000) and Bayesian (Green and Strawderman 1994; Denham et al. 2009; Magnussen 2009). As indicated by its name, model-based inference requires positing a model that treats the observation recorded for a sampling unit as a random variable. For example, the observation that a pixel is correctly classified would take the form of a model predicting whether the observation was y=1 if correctly classified and y=0 otherwise. The model must specify a form for the variance of this random variable and may also specify a functional relationship between the target variable of interest for estimation and other auxiliary variables observable on the sampling unit. One of the challenges of model-based inference for accuracy assessment is that a model would need to be specified for each accuracy estimate produced. Further, the model assumptions would need to be verified as plausible. An advantage of model-based inference is that it does not require a probability sampling design, although it is suggested that probability sampling should still be employed to convey the advantage of objectivity (Valliant et al. 2000, p. 19). Model-based inference offers a viable option for using reference data that were not obtained via a probability sampling design. Steele et al. (2003) provide an example of model-based inference applied

to accuracy assessment and McRoberts (2006) applied model-based inference to area estimation. Magnussen (2015) provides an informative quantitative assessment comparing the variance of estimators derived from model-based and design-based inference.

Bayesian inference is based on a probability distribution (called the posterior distribution) for the parameter of interest where this distribution is conditional on the sample data observed (Denham et al. 2009). The Bayesian perspective that a parameter is characterized by a probability distribution is in contrast to the design-based approach in which a parameter is viewed as a fixed, albeit unknown constant characterizing a population. The Bayesian approach allows incorporating prior knowledge into the assessment, as for example a completed accuracy assessment of a neighboring area (Denham et al. 2009). This prior information may enhance the precision of the accuracy estimates. Denham et al. (2009) state, "This approach [Bayesian] may appear to involve more work than a standard analysis, but any analysis, Bayesian or frequentist, should involve a similar amount of effort in testing assumptions and interpreting the results." In reality, one of the main advantages of design-based inference is the absence of assumptions and so it requires virtually no effort to verify assumptions. The primary assumption of design-based inference is that the reference data are correct, but Bayesian inference, as well as model-based inference, would similarly need to make some accommodation to account for error and variability in the reference class labels.

All three inference options, design-based, model-based, and Bayesian offer a statistically rigorous basis for inference. Design-based inference requires the fewest assumptions, but is dependent on probability sampling. Model-based inference can be applied when a probability sample has not been implemented or when the sample size is so small that design-based inference lacks adequate precision (e.g., accuracy estimates for small sub-areas). Bayesian inference offers an alternative perspective of how parameters are viewed (i.e., the posterior distribution) and the potential advantage to use prior information to enhance precision.

6. Imperfect reference data

In the context of this article, accuracy assessment is the process of determining the quality of the mapped representation of land cover obtained via remote sensing. In essence, an accuracy assessment is based on the assessment of error, the misclassifications where the class label in the mapped representation differs from that observed in reality. Strictly, gold-standard, error-free, reference data depicting the ground condition perfectly are required to undertake an accuracy assessment. In reality, the reference data set is just another classification and may contain error. In

recognition of this situation many have called for the reference data set to be of a higher quality than the map it is being used to assess (Olofsson et al. 2014). This may be a pragmatic approach to adopt in an accuracy assessment but it does not exonerate us from having to consider that the reference classification may be imperfect. In the site-specific approach to accuracy assessment that is used widely, disagreements between the map and reference class labelling are taken to indicate an error in the map obtained via remote sensing. However, the mapped label could actually be correct and the error lies in the reference label, just one way in which accuracy assessments in remote sensing can be viewed as being harsh (Foody 2008).

Determining the reference condition is a challenging aspect of accuracy assessment. Imperfections in reference data may include: 1) reference data error (i.e., an incorrect reference class label); 2) reference class ambiguity in which a single class label does not adequately characterize the reference condition of the pixel; and 3) inconsistent reference class labels (i.e., different interpreters assign a different reference class to the same pixel). These aspects of imperfect reference data are discussed in the following three subsections. Congalton and Green (1993; 1999, Chapter 4), Powell et al. (2004), and Defourny et al. (2012) provide additional discussion of specific sources of uncertainty associated with reference data.

6.1. Reference data error

Several studies have explored the effect of reference data error on classification and estimation (e.g., Verbyla and Boles 2000; Carlotto 2009; Foody 2010; McRoberts et al. 2018). In the conventional approach to accuracy assessment that is stressed in this article, a major concern is that, even if small, reference data error may lead to substantial biases in accuracy and area estimators, for example, deflating the apparent accuracy for a class and leading to over-estimation of its areal extent (Foody 2009a). Taking one simple scenario presented in Foody (2013) for a binary classification in which both the map being evaluated and the reference data are very accurate, with producer's accuracy values of 90% and 95% respectively, the effect of the reference error was to overestimate class abundance (area) by nearly a factor of 6. The magnitude of the bias varies with the accuracies of the data sets and the abundance of the classes. For example, Foody (2013) provided an example in which the areal extent of a rare class was overestimated by a factor of nearly 40 because of reference data error even when accuracy was well within the range of results typically reported in the literature. Reference data error also has the undesirable effect of introducing a prevalence dependency into accuracy assessment based on measures that are (when using a gold standard reference) prevalent-independent (Foody 2010).

Given that a binary error matrix is often used in studies of land cover change and that the class of interest, change, is often rare, the mis-estimation may be so large that the effect of reference data error should not be ignored (Foody 2013), especially as there are methods to address the issue. For example, if the accuracy of the reference dataset is known it is possible to compute the real accuracy of the map rather than the apparent accuracy suggested by naïve interpretation of the error matrix (Staquet et al. 1981; Foody 2010). If the reference class accuracy is not well known but there are a set of labels for each case, perhaps arising from a series of classifications of the same region or multiple interpreters, a latent class analysis may be used to estimate the actual accuracy of the maps (Foody 2012). Latent class analysis is an example of a model-based approach to accuracy assessment that may have considerable potential as an alternative to the standard design-based approaches when reference data error is problematic.

6.2. Reference class ambiguity

Accommodating ambiguity in the reference class of a sample unit has justifiably received a great deal of attention. As a minimum, the percent of ambiguous sample pixels should be documented and it should be stated whether such pixels have been included in the estimation of the error matrix and accuracy measures. The recommended approach is that ambiguous pixels should be included in the analysis with accommodations to address their potential impact. In some studies ambiguity of reference class labels is addressed by having the interpreters rate their confidence in each interpretation. Accuracy results can then be produced for sample subsets defined by these confidence ratings to assess if accuracy differs by confidence rating. For example, if low confidence is associated with greater reference data ambiguity, accuracy likely will decrease as confidence declines. Another option for accommodating reference class ambiguity is to assign a primary and an alternate reference label, where the alternate label is included when a single label does not suffice to characterize the sample unit. Agreement can then be defined as a match between the map class and either the primary or alternate reference label (e.g., Wickham et al. 2004; Olofsson et al. 2014).

The use of primary and alternate labels is closely related to the more elegantly formulated linguistic scale fuzzy reference labelling protocol developed by Gopal and Woodcock (1994) in which the information recorded for each land cover class of a sample unit would be: 1 = absolutely wrong, 2 = understandable but wrong, 3 = reasonable or acceptable answer, 4 = good answer, and 5 = absolutely right. Based on this linguistic scale reference data, two fuzzy accuracy measures are the MAX operator in which the map is considered correct if the map label matches the reference class with the greatest

value on the linguistic scale, and the RIGHT operator in which the map is considered correct if the map class matches a reference class that would be deemed an acceptable answer on the linguistic scale (e.g., 3, 4, or 5). Accuracy results based on defining agreement using the primary and alternate reference labels will closely approximate the results of the MAX and RIGHT operators (Stehman et al. 2003). That is, agreement based on only the primary reference label will mimic the MAX operator and agreement with either the primary or alternate reference label will approximate fuzzy accuracy based on the RIGHT operator.

Several other extensions have been developed to address the problem of reference class ambiguity. For example, Sarmento et al. (2013) used fuzzy intervals to account for uncertainty in linguistic scale reference data. Hagen (2003) and Hagen-Zanker (2006) developed methods that take into consideration spatial uncertainty of the reference data in addition to reference class ambiguity. When the reference data for each sampled pixel is recorded as the proportion of area of each class, several options exist for quantifying agreement, including methods that could be applied to a soft-classified map (Finn 1993; Foody 1996; Lewis and Brown 2001; Latifovic and Olthof 2004; Pontius and Cheuk 2006; Pontius and Connors 2009). A potentially challenging aspect of some of these approaches is that sample-based estimation formulas have only been provided for equal probability sampling, and even for such designs formulas for estimating standard errors may not have been published.

6.3. Inconsistent reference class labelling

When the reference classification is obtained by human interpreters, it is highly likely that the interpreters will disagree on some proportion of the sample units (Powell et al. 2004). Interpreter variability may impact the analysis in several ways. If several interpreters view the same sample unit, it is common practice to resolve disagreements among interpreters by having the interpreters reach a consensus decision or to have an expert interpreter make a final decision on the reference class. In addition to the well recognized issue of how to resolve these interpreter inconsistencies (Sec. 3), there is also the impact of variability among interpreters on the standard errors of the accuracy and area estimates. McRoberts et al. (2018) demonstrated that standard errors of area estimates are inflated by interpreter variability, and incorporating this added uncertainty into standard error estimates may be necessary for a valid assessment of reliability. Greater attention to and quantification of interpreter consistency may motivate new analysis methods to address this longstanding challenge of quality of reference data.

7. Future needs and directions

This article has focused on basic methods for conducting a rigorous and credible evaluation of map quality. There is considerable scope for additional analyses, to reveal a richer and more informative assessment of accuracy. There are many possible ways in which the basic approach to accuracy assessment can be usefully extended. We discuss several such possibilities in the following subsections.

7.1. Technology and reference data

Recent technological developments may provide a means to help acquire reference data to facilitate accuracy assessments. Three technological developments in particular have revolutionised the generation of reference data: citizen sensing, unmanned aerial vehicles, and resources such as Google Earth that distribute free high resolution imagery. Citizen science has a long history and is described by a range of terms such as crowdsourcing and VGI (See et al. 2016). But the proliferation of inexpensive location aware devices and web 2.0 technology that fosters sharing of information have made it simple for potentially anyone to provide geolocated information on land cover. In addition, this acquisition of data can be an active process in which citizens are steered to sites selected (e.g., following a probability sampling design) or a passive one in which data acquired by citizens are mined for reference data. The former approach has been used to acquire reference data in a range of studies, especially using collaborative internet projects (Fritz et al. 2017; Laso Bayas et al. 2017; Foody et al. 2018). If the active process does steer the volunteers to a locations selected by a probability sampling design, these data can be used in design-based inference. For example, the Degree Confluence Project (Iwao et al. 2006) provides photographs over a systematic grid which can be treated as a probability sample. The passive approach has also been used for acquisition of reference data, often involving visual interpretation of photography uploaded to internet sites (Iwao et al. 2006; Foody and Boyd 2013; Antoniou et al. 2016). Typically these photographs are acquired opportunistically and not from a probability sampling design so are not ideally suited to a design-based inference approach to accuracy assessment. Such data can, however, be usefully integrated with reference data acquired by a probability sample to enhance the accuracy assessment (Stehman et al. 2018).

A common concern raised with the use of crowdsourced data such as that made available by citizen scientists is its quality, clearly an important issue given the impacts of using imperfect reference data noted elsewhere (Sec. 6). However, the accuracy of such data can be evaluated (Foody 2012) and in some instances once characterized can be used to aid accurate estimation (Foody et al. 2015). The accuracy of citizen derived data can also be comparable to if not superior to authoritative data (e.g.,

Dorn et al. 2015). A variety of strategies to enhance the accuracy of labelling by citizens have also been proposed (Foody et al. 2018; Prelec et al. 2017; Navajas et al. 2018).

The development of drones or unmanned aerial vehicles (UAVs), and especially inexpensive systems capable of carrying a basic sensor such as a camera, has revolutionised the acquisition of data that could be used as a reference in accuracy assessment (Pla et al. 2017). Critically, a UAV based sensor can be deployed to acquire images that may be interpreted to yield the reference classification. This has many advantages over ground based data collection. For example, UAVs can be used to acquire data for sites that may be dangerous to visit on foot (e.g., swamp), and the ability of UAV-based sensors to rapidly collect data allows acquisition of large data volumes. Because the location of deployment is in the control of the pilot the data acquisition can be steered to sites selected by a probability sampling design if desired.

In many mapping studies the reference data do not arise from field based work but from analyses of imagery with a much finer spatial resolution than that used to produce the map being evaluated. Commonly, for example, Landsat sensor data may be used to provide reference data for a map derived from coarse resolution MODIS data. As with other data sources concerns with data quality arise, the classification of the fine resolution image will contain error and this will impact negatively on the accuracy assessment (Foody 2008). However, in many studies a very fine spatial resolution image can be an excellent source of high quality reference data. Such activity was boosted by the launch of a series of fine spatial resolution systems such as IKONOS in 1999 that provided multispectral images with a resolution of <5 m and further extended by the launch of systems such as Digital Globe's WorldView systems that provide multispectral images with sub-metre resolution (Toutin 2009). Additionally, access to such data was substantially increased by their use in resources such as Google Earth. The latter provides global coverage of imagery at a variety of resolutions. Most critically in the context of this review article it provides fine resolution images, typically from systems such as WorldView, across the world in a manner that is easy to use (Gorelick et al. 2017), and such resources are widely used to generate reference data to support accuracy assessments (Cha and Park 2007).

Although the temporal and spatial availability of very high resolution (VHR) imagery is uneven (Lesiv et al. 2018), we still advocate that a probability sampling design be implemented for the full region (i.e., the sample should not be limited to only those areas for which VHR imagery is available) so that design-based inference can be invoked. However, the transparency criterion of the response design will require clear documentation of which sample pixels had VHR available. Further, the analysis may include estimates for the subsets of the sample for which VHR was and was not available to gain

some insight into the impact of using different reference data sources for different sample pixels. In the ideal situation the reference classification would be obtained by the same protocol for all sample pixels, but under the guiding principle that the best reference data should be used, we would recommend using VHR where available even if it means forfeiting uniformity in how the reference class labels are obtained. Clearly employing different sources of reference data for different pixels in the sample may create inconsistencies both spatially and temporally in the reference classification, and methods to resolve these inconsistencies are needed.

7.2. Importance of classification errors

In the conventional approach to accuracy assessments all errors are treated equally. Yet some errors may be more serious than others (DeFries and Los 1999; Smits et al. 1999), and thus in these studies an accuracy assessment could be enhanced by accounting for differences in error severity. Some classes may be more (dis)similar to others for a range of reasons. For example, different classes of forest may be more similar in terms of their role in the hydrological cycle than artificial surfaces. In such a situation, misclassifications between forest classes would be less serious than those involving a forest class and an artificial class. Mayaux et al. (2006, Table IV) formalize this concept using a matrix of thematic distance that can then be incorporated to adjust accuracy estimates based on dissimilarity of the classes. Also, sometimes classes are defined by dividing up a continuum (Ahlqvist 2005) and confusion between neighboring classes on the continuum is less severe than confusion of classes located at the end points. If the feature being mapped is continuous the analysis protocol must be modified from the standard methods used to assess categorical, discrete features. Fortunately, methods to assess accuracy for classifications that range from those with variable degrees of error severity (Woodcock and Gopal 2000) through to continuous classes exist (Foody 1996; Riemann et al. 2010) but may benefit from additional research to facilitate issues such as rigorous comparison. This would be valuable in some areas of topical interest such as the study of land cover changes that are less severe than transformations of cover, for example modifications such as forest thinning that change the character of the land cover but not its class type.

It is common for the results of an accuracy assessment to be referred to when a map is used. However, an accuracy assessment can be more useful than just a descriptive statement of map quality, it can be used to enhance studies that use the map. The impact of different classification errors will impact the value of a map for use in a particular application (Stehman 1999b), and identifying the cost or loss function of different classification errors on estimation of a target value can be used to compare

the expected value of different maps for a given application (de Bruin et al. 2001). Several recent examples illustrate how the results of an accuracy assessment of a map can be used to enhance the analysis and interpretation of land cover studies using that map. Estes et al. (2017) provided a comprehensive assessment of how errors in cropland maps could impact downstream analyses of estimation of carbon stocks, simulation of evapotranspiration, disaggregation of crop yield and production, and simulation of household dynamics. In addition to assessing how quantitative estimates were influenced by land cover map errors, Estes et al. (2017) also noted the importance of evaluating how map errors impact the ability to locate specific features of interest, such as areas of high cropland cover. Tsendbazar et al. (2016) compared the fitness for use of several global land cover maps for applications such as general circulation models, agriculture assessments, and biodiversity assessments. Their approach was to produce weighted accuracy estimates based on incorporating similarity matrices where similarity between classes was determined specific for each application of the global maps. Foody (2015) illustrates another important use of accuracy assessment results by demonstrating how valuations of ecosystem services derived from land cover class areal extent can differ substantially when adjustments are made for classification error. Olofsson et al. (2013) used confidence intervals for the area estimates obtained from an accuracy assessment to conduct a sensitivity analysis of a carbon flux model, illustrating another way in which an accuracy assessment may inform applications of the map.

7.3. Other dimensions of map quality

As well as an increasing array of data sources to inform an accuracy assessment there is scope to look at other map quality dimensions. The core focus of this article has been on thematic accuracy but other indicators of map quality such as consistency, reliability, trust, and precision may be considered (Fonte et al. 2015). Also, it may be helpful to explore issues of locational accuracy or completeness. The various aspects of data quality may also gainfully be used together. For example, a high locational accuracy is implicitly assumed with standard site-specific accuracy assessments, so a small mis-location error, such as one pixel error, could substantially degrade a per-pixel accuracy assessment of a region of heterogeneous land cover. Some means to accommodate locational error within the accuracy assessment may provide a degree of tolerance that makes the approach less harsh (Foody 2008). For example, accuracy results can be reported for the subset of homogeneous locations, where homogeneity is defined as all pixels in a 3x3 window centered on the sample pixel having the same map class. The difference in accuracy estimates for the homogeneous subset compared to the full sample

provides an indication of the potential impact of spatial misregistration between the map and reference data as sample pixels within the homogeneous subset should not be greatly affected by geolocation error. The non-homogeneous subset will generally have lower accuracy because of the possibility of geolocation error but also because this subset includes all edge pixels which are generally more difficult to classify correctly.

7.4. Object-based accuracy assessment

Object-based image analysis (OBIA) has become an increasingly popular approach in land-cover mapping (e.g., Blaschke et al. 2014). Ye et al. (2018) provide a comprehensive critique of OBIA map accuracy assessment based on their review of 209 articles published between 2003 and 2017. Assessing the accuracy of OBIA maps introduces new dimensions to all three component protocols of accuracy assessment because of the variable sizes of the objects mapped (Stehman and Czaplewski 1998). For example, depending on how the sample is selected, the inclusion probabilities may be a function of the area of the objects. That is, if objects are sampled based on whether they are intersected by points of a systematic grid, the inclusion probability will be proportional to the area of the object, so objects of greater area will have greater inclusion probabilities. Such a design still yields a probability sample but accounting for these unequal inclusion probabilities adds complexity to the analysis. To avoid this unequal inclusion probability feature, a list of all objects could be constructed and the sampled objects selected with equal inclusion probabilities from this list. An object-based assessment also introduces response design issues such as choice of assessment unit (map polygon, reference polygon, or pixel), heterogeneity of land cover within a polygon, and definition of agreement. Ye et al. (2018) state that based on their review of the literature there was no straightforward way to match mapped polygons with reference polygons. Further, when the objects differ in size accuracy measures will need to take object area into account if an area-based accuracy assessment is desired. Radoux et al. (2011), Radoux and Bogaert (2017), Stehman and Wickham (2011) and Whiteside et al. (2014) addressed various aspects of accuracy assessment of OBIA maps. Ye et al. (2018) concluded that "… the literature shows no obvious method to assess the accuracy of OBIA maps in a manner that appreciates that OBIA maps consist of objects of various shapes and sizes." The good practice recommendations of Radoux and Bogaert (2017) provide progress in this direction as protocols continue to be developed and tested in practice.

7.5. Eliminating bad practice

It is hoped that in the future not only will good practices (Olofsson et al. 2014) be followed but aspects of bad practice eliminated. Following earlier discussion, three examples of bad practice that are widespread are the often unjustified use of 85% target accuracy, normalisation of the error matrix, and chance correction of agreement. While a target accuracy is a desirable feature in an accuracy assessment, the target should be selected for the application at-hand. The 85% target value had a clear and well-justified place in the literature, linked to Anderson (1971) for large area mapping of broad land cover classes, but it is not and should not be considered universally applicable (Foody 2008). The key accuracy metrics and their targets should be identified in advance of the analysis and these targets should be application specific to indicate fitness for purpose. Thus the 85% threshold has no universal status despite sometimes being used as such. Normalisation of the error matrix is touted as an aid to the interpretation of an error matrix but its use is undesirable not least because it has the effect of equalizing the producer's and user's accuracy which may be meaningfully different (Stehman 2004). Last but not least, the community should cease to correct for chance agreement. The latter may have some value in assessing the performance of a classifier but it has no place in the assessment of map accuracy. The source of misclassifications or correct allocations is not a necessary component of an accuracy assessment. Thus widely used measures such as the kappa coefficient of agreement should not be used. The latter measures provide only an indication of accuracy that is highly correlated with overall accuracy, from which it is calculated, and conveys no useful new information on map accuracy. Numerous calls for kappa to be removed from the community's toolbox (Foody 1992; Stehman 1997b; Pontius and Millones 2011) need finally to be heeded.

8. Conclusions

Accuracy assessment has a long history and its origins and development have paralleled the development of remote sensing methods to map land cover. The error matrix and associated user's, producer's, and overall accuracies have remained core elements of accuracy assessment. Having served well in this capacity for nearly half a century, the error matrix and estimates of accuracy and area derived from the error matrix should continue to be relied upon as standard practice. Accuracy assessment has expanded beyond the error matrix to account for imperfect reference data and to provide better spatial representation of error. The critical role of sampling to obtain the reference data was recognized early on (Hord and Brooner 1976; Hay 1979; Fitzpatrick-Lins 1981), and the recognition of design-based inference as the traditional approach for inference in this sampling context formalized the meaning of "statistically rigorous" when applied to accuracy assessment (Stehman 2000).

Specifically, statistical rigor can be achieved by implementing a probability sampling design and applying statistically consistent estimators (i.e., formulas specific to the sampling design used to collect the reference data).

Advances in the availability and quality of reference data (e.g., very high resolution imagery, internet access to Google Earth) have transformed response designs from dependence on hard copy aerial photographs and expensive ground visits to intensive human interpretation of multiple data sources at a desktop computer. The development of tools such as TimeSync (Cohen et al. 2010), Collect Earth (Bey et al. 2016) and LACO-wiki (See et al. 2017) for convenient and efficient acquisition of imagery and other sources of reference data for use in the response design offers a major advance over the early days of accuracy assessment. Citizen science has opened up potentially new options for obtaining reference data that will surely continue to be explored. Response design methodology has advanced beyond simply recognizing that reference data are not "ground truth" to specifying the impacts of reference data error and variability and having options for accounting for reference error and for quantifying the contribution of reference data variability to the total variance.

Advances in remote sensing technology have enhanced the spatial and temporal scale of issues that can be studied and reinforced the importance of rigorous and map relevant accuracy assessments (Wulder et al. 2018). National, continental, and global land-cover mapping efforts have led to development and implementation of accuracy assessments for these very expansive map coverages. The greater availability and ease of access to imagery has created the opportunity for monitoring land cover at a much denser temporal frequency. Accordingly, accuracy assessment methodology must progress to develop methods capable of assessing these long time series, land-cover change products. The increasing popularity of object-based mapping methods has brought about yet other needs for methodological advances in accuracy assessment.

Much has been accomplished in the nearly 50 years of theory, methods, and applications of accuracy assessment and area estimation in the study of land cover. We can no doubt expect advances to continue as new technology as well as new opportunities and challenges arise in the study of Earth's land cover. In the meantime, it is critical that practitioners implement current methodology in a manner that ensures a statistically rigorous and map relevant accuracy assessment. Documentation of accuracy assessments must certainly improve to enhance transparency and facilitate reproducibility of methodology and results. Greater attention to quality assurance of reference data will also contribute to the overall reliability of accuracy assessments. Practitioners should continue to aspire to resolve the

present day challenges of accuracy assessment as we move into the next half century of studying land cover.

Acknowledgments

Funding was provided by United States Geological Survey grant G12AC20221 and NASA Carbon Monitoring System program grant NNX13AP48G (S.Stehman). We thank the editor and referees for numerous helpful suggestions to improve the manuscript.

References

Ahlqvist, O. (2005). Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science*, 19, 831-857.

Anderson, J. R. (1971). Land use classification schemes. *Photogrammetric Engineering*, 37, 379–387.

Antoniou, V., Fonte, C. C., See, L., Estima, J., Arsanjani, J. J., Lupia, F., Minghini, M., Foody, G., and Fritz, S. (2016). Investigating the feasibility of geo-tagged photographs as sources of land cover input data. *ISPRS International Journal of Geo-Information* 5, 64; doi:10.3390/ijgi5050064.

Anuta, P. E., and MacDonald, R. B. (1971). Crop surveys from multiband satellite photography using digital techniques. *Remote Sensing of Environment*, 2, 53-67.

Arevalo, P., Woodcock, C. E., and Olofsson, P. (2019). Continuous monitoring of land change activities and post-disturbance dynamics from Landsat time series: a test methodology for REDD+ reporting. *Remote Sensing of Environment*, doi: 10.1016/j.rse.2019.01.013.

Aronoff, S. (1982a). Classification accuracy: a user approach. *Photogrammetric Engineering & Remote Sensing*, 48, 1299-1307.

Aronoff, S. (1982b). The map accuracy report: a user's view. *Photogrammetric Engineering & Remote Sensing*, 48, 1309-1312.

Badjana, H. M., Olofsson, P., Woodcock, C. E., Helmschrot, J., Wala, K., and Akpagana, K. (2017). Mapping and estimating land change between 2001 and 2013 in a heterogeneous landscape in West Africa: Loss of forestlands and capacity building opportunities. *International Journal of Applied Earth Observation Geoinformation*, 63, 15-23.

Benedetti, R., Piersimoni, F., and Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85, 439-454.

Bey, A., Diaz, A. S.-P., Maniatis, D., Marchi, G., Mollicone, D., et al. (2016). Collect Earth: Land use and land cover assessment through augmented visual interpretation. *Remote Sensing* 8(10), 807.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., et al. (2014). Geographic object-based image analysis – towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 180-191.

Boschetti, L., Stehman, S. V., and Roy, D. P. (2016). A stratified random sampling design in space and time for regional to global scale burned area product validation. *Remote Sensing of Environment*, 186, 465-478.

Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62, 394-407.

Bryan, M. L. (1975). Interpretation of an urban scene using multi-channel radar imagery. *Remote Sensing of Environment*, 4, 49-66.

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering & Remote Sensing*, 48, 431-439.

Carlotto, M. J. (2009). Effect of errors in ground truth on classification accuracy. *International Journal of Remote Sensing*, 30, 4831-4849.

Castilla, G. (2016). We must all pay more attention to rigor in accuracy assessment. *Remote Sensing*, 8, 288; doi:10.3390/rs8040288.

Cha, S.Y. and Park, C. H. (2007). The utilization of Google Earth images as reference data for the multitemporal land cover classification with MODIS data of North Korea. *Korean Journal of Remote Sensing*, 23, 483-491.

Chapin, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., et al. (2000). Consequences of changing biodiversity. *Nature*, 405, 234-242.

Comber, A., Fisher, P. and Wadsworth, R. (2005). What is land cover? *Environment and Planning B: Planning and Design*, 32, 199-209.

Cohen, W. B., Yang, Z., and Kennedy, R. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - Tools for calibration and validation. *Remote Sensing of Environment,* 114, 2911-2924.

Comber, A., Fisher, P., Brunsdon, C., and Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, 127, 237–246.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C. and Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 37-48.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37, 35-46.

Congalton, R. G., and Green, K. (1993). A practical look at the sources of confusion in error matrix generation. *Photogrammetric Engineering & Remote Sensing*, 59, 641-644.

Congalton, R. G., and Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, CRC Press, Boca Raton, FL, 137 pp.

Congalton, R.G., Oderwald, R.G., and Mead, R.A. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering & Remote Sensing*, 49, 1671-1678.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., et al. (1997). The value of the world's ecosystem services and natural capital. *Nature*, 387, 253-260.

de Bruin, Bregt, A., and van de Ven, M. (2001). Assessing fitness for use: the expected value of spatial data sets. *International Journal of Geographical Information Science*, 15, 457-471.

Defourny, P., Mayaux, P., Herold, M., and Bontemps, S. (2012). Global land-cover map validation experiences: Toward the characterization of uncertainty. In: *Remote Sensing of Land Use and Land Cover: Principles and Applications*, Giri, C.P. (Editor), Boca Raton: Taylor & Francis, pp. 207 - 224.

DeFries, R.S. and Los, S.O. (1999). Implications of land-cover misclassification for parameter estimates in global land-surface models: an example from the simple biosphere model (SiB2). *Photogrammetric Engineering & Remote Sensing*, 65, 1083-1088.

De Gruijter, J. J., and Ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, 22, 407-415.

Denham, R, Mengersen, K., and Witte, C. (2009). Bayesian analysis of thematic map accuracy data. *Remote Sensing of Environment*, 113, 371-379.

Dorn, H., Törnros, T., and Zipf, A. (2015). Quality evaluation of VGI using authoritative data - A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information*, 4, 1657-1671.

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

Estes, L., Chen, P., Debats, S., Evans, T., Ferreira, S., Kuemmerle, T., Ragazzo, G., Sheffield, J., Wolf, A., Wood, E., and Caylor, K. (2017). A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. *Global Change Biology*, 24, 322-337.

Fattorini, L., Corona, P., Chirici, G., and Pagliarella, M. C. (2015). Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use, and land cover estimation. *Environmetrics*, 26, 216-228.

Fichet, L-V., Sannier, C., Makaga, E. M. K., and Seyler, F. (2014). Assessing the accuracy of forest cover map for 1990, 2000 and 2010 at national scale in Gabon. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7, 1346-1356.

Finn, J.T. (1993). Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Science*, *7*, 349-366.

Fitzpatrick-Lins, K. (1981). Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering & Remote Sensing*, 47, 343-351.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, 3rd edition, New Jersey: Wiley.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., et al. (2005). Global consequences of land use. *Science*, 309, 570-574.

Fonte, C. C., Bastin, L., See, L., Foody, G., and Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, 29, 1269-1291.

Foody, G. M. (1996). Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, 17, 1317-1340.

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

Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, 70, 627-633.

Foody, G. M. (2005). Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *International Journal of Remote Sensing*, 26, 1217-1228.

Foody, G. M. (2006). What is the difference between two maps? A remote senser's view. *Journal of Geographical Systems*, 8, 119-130.

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, 29, 3137–3158.

Foody, G. M. (2009a). The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, 30, 3275–3281.

Foody, G. M. (2009b). Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113, 1658-1663.

Foody, G. M. (2009c). Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, 30, 5273-5291.

Foody, G.M. (2010). Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, 114, 2271–2285.

Foody, G.M. (2012). Latent class modeling for site-and non-site-specific classification accuracy assessment without ground data. *IEEE Transactions on Geoscience and Remote Sensing*, 50, 2827-2838.

Foody, G.M. (2013). Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sensing Letters*, 4, 783-792.

Foody, G. M. (2015). Valuing map validation: The need for rigorous land cover map accuracy assessment in economic valuations of ecosystem services. *Ecological Economics*, 111, 23-28.

Foody, G. M., and Boyd, D. S. (2013). Using volunteered data in land cover map validation: Mapping West African forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6, 1305-1312.

Foody, G.M., and Mathur, A. (2006). The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment*, 103, 179-189.

Foody, G.M., Pal, M., Rocchini, D., Garzon-Lopez, C.X., and Bastin, L. (2016). The sensitivity of mapping methods to reference data quality: Training supervised image classifications with imperfect reference data. *ISPRS International Journal of Geo-Information*, *5*(11), p.199.

Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D.S. and Comber, A. (2015). Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. *The Cartographic Journal*, 52, 336-344.

Foody, G., See, L., Fritz, S., Moorthy, I., Perger, C., Schill, C., and Boyd, D. (2018). Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. *ISPRS International Journal of Geo-Information*, *7*(3), p.80.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., and Obersteiner, M. (2009). Geo-Wiki.Org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1, 345-544.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., et al. (2017). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data*, 4, p.170075.

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

Gallego, F. J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, 25, 3019-3047.

Gallego, F. J. (2011). Validation of GIS layers in the EU: getting adapted to available reference data. *International Journal of Digital Earth*, 4 (Supplement 1), 42-57.

Gallego, F. J. (2012). The efficiency of sampling very high resolution images for area estimation in the European Union. *International Journal of Remote Sensing,* 33, 1868-1880.

Gallego, J., and Bamps, C. (2008). Using CORINE land cover and the point survey LUCAS for area estimation. *International Journal of Applied Earth Observation and Geoinformation*, 10, 467-475.

Gallego, F. J. and Stibig, H. J. (2013). Area estimation from a sample of satellite images: the impact of stratification on the clustering efficiency. *International Journal of Applied Earth Observation and Geoinformation*, 22, 139-146.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., et al. (2013). Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing* 34, 2607-2654.

Gopal, S., and Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering & Remote Sensing*, 60, 181-188.

Gordon, S. I. (1980). Utilizing LANDSAT imagery to monitor land-use change: A case study in Ohio. *Remote Sensing of Environment*, 9, 189-196.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18-27.

Green, E. J., and Strawderman, W. E. (1994). Determining accuracy of thematic maps. *The Statistician*, 43, 77-85.

Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science* 17, 235-249.

Hagen-Zanker, A. (2006). Map comparison methods that simultaneously address overlap and structure. *Journal of Geographical Systems*, 8, 165-185.

Hammond, T. O., and Verbyla, D. L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, 17, 1261-1266.

Haralick, R.M., Caspall, F., and Simonett, D. S. (1969-1970). Using radar imagery for crop discrimination: A statistical and conditional probability study. *Remote Sensing of Environment*, 1, 131-142.

Hay, A. M. (1979). Sampling designs to test land-use map accuracy. *Photogrammetric Engineering & Remote Sensing*, 45, 529-533.

Healey, S. P., Cohen, W. B., Yang, Z., Brewer, C. K., Brooks, E. B., Gorelick, et al. (2018). Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment*, 204, 717-728.

Herold, M., Mayaux, P., Woodcock, C.E., Baccini, A. and Schmullius, C. (2008). Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment*, 112, 2538-2556.

Hollmann, R., Merchant, C. J., Saunders, R., Downy, C., Buchwitz, M., et al. (2013). The ESA climate change initiative: Satellite data records for essential climate variables. *Bulletin of the American Meteorological Society*, *94*, 1541-1552.

Hord, R. M., and Brooner, W. (1976), Land-use map accuracy criteria. *Photogrammetric Engineering & Remote Sensing*, 42, 671-677.

Iwao, K., Nishida, K., Kinoshita, T., and Yamagata, Y. (2006). Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters*, 33, L23404.

Janssen, L.L.F., and van der Wel, F.J.M. (1994). Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering & Remote Sensing*, 60, 419-426.

Jensen, J. R., Christensen, E. J., and Sharitz, R. (1984). Nontidal wetland mapping in South Carolina using airborne multispectral scanner data. *Remote Sensing of Environment*, 16, 1-12.

Kempeneers, P., McInerney, D., Sedano, F., Gallego, J., Strobl, P., et al. (2013). Accuracy assessment of a remote sensing-based, pan-European forest cover map using multi-country national forest inventory data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6, 54-65.

Khatami, R., Mountrakis, G., and 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.

Khatami, R., Mountrakis, G., and Stehman, S. V. (2017). Mapping per-pixel predicted accuracy of classified remote sensing images. *Remote Sensing of Environment*, 191, 156-167.

Kyriakidis, P. C., and Dungan, J. L. (2001). A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, 8, 311-330.

Laso Bayas, J.-C., See, L., Fritz, S., Sturn, T., Perger, C., et al. (2017). Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. *Remote Sensing* 8(11), 905.

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

Lesiv, M., See, L., Laso Bayas, J. C., Sturn, T., Schepaschenko, D., Karner, M., Moorthy, I., McCallum, I., and Fritz, S. (2018). Characterizing the spatial and temporal availability of very high resolution satellite imagery in Google Earth and Microsoft Bing Maps as a source of reference data. *Land*, 7, 118.

Lewis, H. G., and Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22, 3223-3235.

Linke, J., Fortin, M-J., Courtenay, S., and Cormier, R. (2017). High-resolution global maps of 21st-century annual forest loss: Independent accuracy assessment and application in a temperate forest region of Atlantic Canada. *Remote Sensing of Environment*, 188, 164-176.

Liu, C., Frazier, P., and Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606-616.

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., and Merchant, J. W. (2007). Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, 21, 1303-1330.

Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28, 823-870.

Magnussen, S. (2009). A Bayesian approach to classification accuracy inference. *Forestry*, 82, 211-226.

Magnussen, S. (2015). Arguments for a model-dependent inference? *Forestry*, 88, 317-325.

Marsh, S. E., Walsh, J. L., and Sobrevila, C. (1994). Evaluation of airborne video data for land-cover classification accuracy assessment in an isolated Brazilian forest. *Remote Sensing of Environment*, 48, 61-69.

Mas, J. F. (1999). Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing*, 20, 139-152.

Mather, P. and Tso, B., (2016). *Classification Methods for Remotely Sensed Data*. CRC press.

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

McGwire, K. C., and Fisher, P. (2001). Spatially variable thematic accuracy: Beyond the confusion matrix, In *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications* (Hunsaker, C. T., Goodchild, M. F., Friedl, M. A., and Case, T. J.), Springer, New York, pp. 308-329.

McRoberts, R. E. (2006). A model-based approach to estimating forest area. *Remote Sensing of Environment*, 103, 56-66.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, 115, 715–724.

McRoberts, R. E., Stehman, S. V., Liknes, G. C., Næsset, E., Sannier, C., and 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.

McRoberts, R. E., and Walters, B. F. (2012). Statistical inference for remote sensing-based estimates of net deforestation. *Remote Sensing of Environment*, 124, 394-401.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B. and Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), p.126.

Nusser, S. M., and Klaas, E. E. (2003). Survey methods for assessing land cover map accuracy. *Environmental and Ecological Statistics*, 10, 309-331.

Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., et al. (2012). A global land-cover validation data set, part I: fundamental design principles. *International Journal of Remote Sensing*, 33, 5768-5788.

Olofsson, P., Foody, G. M., Stehman, S. V., and 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.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42-57.

Overton, W. S., and Stehman, S. V. (1995). The Horvitz-Thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. *American Statistician*, 49, 261-268.

Padilla, M., Olofsson, P., Stehman, S. V., Tansey, K., and Chuvieco, E. (2017). Stratification and sample allocation for reference burned area data. *Remote Sensing of Environment*, 203, 240-255.

Pal, M. and Foody, G. M. (2012). Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5, 1344-1355.

Park, N-W., Kyriakidis, P. C., and Hong, S-Y. (2016). Spatial estimation of classification accuracy using indicator kriging with an image-derived ambiguity index. *Remote Sensing* 8, 320.

Pla, M., Duane, A. and Brotons, L. (2017). Potential of UAV images as ground-truth data for burn severity classification of Landsat imagery: approaches to a useful product for post-fire management. *Revista de Teledetección*, 49, 91-102.

Pontius, R. G., Jr. (2019). Component intensities to relate difference by category with difference overall. *International Journal of Applied Earth Observation and Geoinformation*, 77, 94-99.

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

Pontius, R. G., Jr., and Connors, J. (2009). Range of categorical associations for comparison of maps with mixed pixels. *Photogrammetric Engineering & Remote Sensing*, 75, 963-696.

Pontius, R. G., Jr., Krithivasan, R., Sauls, L., Yan, Y., and Zhang, Y. (2017). Methods to summarize change among land categories across time intervals. *Journal of Land Use Science*, 12:4, 218-230.

Pontius, R. G., Jr., and Millones, M. (2011). Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32, 4407-4429.

Pontius, R. G., Jr., and Santacruz, A. (2014). Quantity, exchange, and shift components of difference in a square contingency table. *International Journal of Remote Sensing*, 35, 7543-7554.

Potapov, P. V., Dempewolf, J., Talero, Y., Hansen, M. C., Stehman, S. V., et al. (2014). National satellite-based humid tropical forest change assessment in Peru in support of REDD+ implementation. *Environmental Research Letters* 9(2014) 124012 (13pp).

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

Prelec, D., Seung, H.S. and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532.

Radoux, J., and Bogaert, P. (2017). Good practices for object-based accuracy assessment. *Remote Sensing*, 9, 646.

Radoux, J., Bogaert, P., Fasbender, D., and Defourny, P. (2011). Thematic accuracy assessment of geographic object-based image classification. *International Journal of Geographical Information Science*, 25, 895-911.

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A. (2008). Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22, GB1003.

Riemann, R., Wilson, B. T., Lister, A., and Parks, S. (2010). An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sensing of Environment*,114, 2337-2352.

Sarmento, P., Fonte, C. C., Caetano, M., and Stehman, S. V. (2013). Incorporating the uncertainty of linguistic-scale reference data to assess accuracy of land-cover maps using fuzzy intervals. *International Journal of Remote Sensing*, 34, 4008-4024.

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

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

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-information*, *5*(5), p.55.

See, L., Laso Bayas, J. C., Schepaschenko, D., Perger, C., Dresel, C., Maus, V., Salk, C., Weichselbaum, J., Lesiv, M., McCallum, I., and Moorthy, I. (2017). LACO-Wiki: A new online land cover validation tool demonstrated using GlobeLand30 for Kenya. *Remote Sensing*, *9*(7), p.754.

Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10, 989-1003.

Smits, P. C., Dellepiane, S. G., and Schowengerdt, R. A. (1999), Quality assessment of image classification algorithms for land-cover mapping: a review and a proposal for a cost-based approach, *International Journal of Remote Sensing* 20, 1461-1486.

Song, X.-P., Potapov, P. V., Krylov, A., King, L., Di Bella, C. M., et al. (2017). National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. *Remote Sensing of Environment*, 190, 383-395.

Staquet, M., Rozencweig, M., Lee, Y. J., and Muggia, F. M. (1981). Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases*, 34, 599-610.

Steele, B. M., Winne, J. C., and Redmond, R. L. (1998). Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment*, 66, 192-202.

Steele, B. M., Patterson, D. A., and Redmond, R. L. (2003). Toward estimation of map accuracy without a probability test sample. *Environmental and Ecological Statistics*, 10, 333-356.

Stehman, S. V. (1995). Thematic map accuracy assessment from the perspective of finite population sampling. *International Journal of Remote Sensing*, 16, 589-593.

Stehman, S. V. (1997a). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62, 77-89.

Stehman, S. V. (1997b). Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, 60, 258-269.

Stehman, S. V. (1999a). Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, 20, 2423-2441.

Stehman, S. V. (1999b). Comparing thematic map accuracy based on map value. *International Journal of Remote Sensing,* 20, 2347-2366.

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

Stehman, S. V. (2001), Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering & Remote Sensing*, 67, 727-734.

Stehman, S. V. (2004). A critical evaluation of the normalized error matrix in map accuracy assessment. *Photogrammetric Engineering & Remote Sensing*, 70, 743-751.

Stehman, S. V. (2005). Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, 96, 466-474.

Stehman, S. V. (2006). Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: a special case of map comparison. *Journal of Geographical Systems,* 8, 209-226.

Stehman, S. V. (2009a). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, 30, 5243-5272.

Stehman, S. V. (2009b). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sensing of Environment*,113, 2455-2462.

Stehman, S. V. (2013). Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, 132, 202-211.

Stehman, S. V. (2014). Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing*, 35, 4923-4939.

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

Stehman, S. V., Czaplewski, R. L., Nusser, S. M., Yang, L., and Zhu, Z. (2000). Combining accuracy assessment of land-cover maps with environmental monitoring programs. *Environmental Monitoring and Assessment*, 64, 115-126.

Stehman, S. V., Fonte, C. C., Foody, G. M., and See, L. (2018). Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover. *Remote Sensing of Environment*, 212, 47-59.

Stehman, S. V., Olofsson, P., Woodcock, C. E., Herold, M., and Friedl, M. A. (2012). A global land cover validation dataset, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. *International Journal of Remote Sensing* 33, 6975-6993.

Stehman, S. V., and Selkowitz, D. J. (2010). A spatially stratified, multi-stage cluster sampling design for assessing accuracy of the Alaska (USA) National Land-Cover Data (NLCD). *International Journal of Remote Sensing*, 31, 1877-1896.

Stehman, S. V., and 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.

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

Stevens, D. L., and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99, 262-278.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., et al. (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, EUR 22156 EN – DG, Office for Official Publications of the European Communities, Luxembourg, 48 pp.

Tchuenté, A. T. K., Roujean, J. L. and De Jong, S. M. (2011). Comparison and relative quality assessment of the GLC2000, GLOBCOVER, MODIS and ECOCLIMAP land cover data sets at the African continental scale. *International Journal of Applied Earth Observation and Geoinformation*, 13, 207-219.

Toll, D. L. (1985). Effect of Landsat Thematic Mapper sensor parameters on land cover classification. *Remote Sensing of Environment*, 17, 129-140.

Toutin, T. (2009). Fine spatial resolution optical sensors. In: *The SAGE Handbook of Remote Sensing*, Warner, T. A., Nellis, M. D., and Foody, G. M. (Eds),Sage, London, pp. 139-150.

Tsendbazar, N-E., de Bruin, S., Mora, B., Schouten, L., and Herold, M. (2016). Comparitive assessment of thematic accuracy of GLC maps for specific applications using existing reference data. *International Journal of Applied Earth Observation and Geoinfomation*, 44, 124-135.

Tsendbazar, N-E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., et al. (2018). Developing and applying a muti-purpose land cover validation dataset for Africa. *Remote Sensing of Environment*, 219, 298-309.

Tsutsumida, N., and Comber, A. J. (2015). Measures of spatio-temporal accuracy for time series land cover data. *International Journal of Applied Earth Observation and Geoinformation* 41, 46-55.

Turner B.L., Lambin, E. F., and Reenberg, A. (2007). The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences* 104, 20666-20671.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, Inc., New York.

Van Oort, P. A. J. (2007). Interpreting the change detection error matrix. *Remote Sensing of Environment*, 108, 1-8.

Verbyla, D. L., and Boles, S. H. (2000), Bias in land cover change estimates due to misregistration, *International Journal of Remote Sensing*, 21, 3553-3560.

Verbyla, D. L., and Hammond, T. O. (1995). Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *International Journal of Remote Sensing*, 16, 581-587.

Vitousek, P. M., Mooney, H. A., Lubchenco, J., and Melillo, J. M. (1997). Human domination of earth's ecosystems. *Science,* 277*,* 494-499.

Wagner, J. E. and Stehman, S. V. (2015). Optimizing sample size allocation to strata for estimating area and map accuracy. *Remote Sensing of Environment*, 168, 126-133.

Waldner, F., and Defourny, P. 2017. Where can pixel counting area estimates meet user-defined accuracy requirements? *International Journal of Applied Earth Observation and Geoinformation*, 60, 1-10.

Waldner, F., Schucknecht, A., Lesiv, M., Gallego, J., See, L., et al. (2019). Conflation of expert and crowd reference data to validate global binary thematic maps. *Remote Sensing of Environment*, 221, 235-246.

Walsh, S. J. (1980). Coniferous tree species mapping using LANDSAT data. *Remote Sensing of Environment*, 9, 11-26.

Whiteside, T. G., Maier, S. W., and Boggs, G. S. (2014). Area-based and location-based validation of classified image objects. *International Journal of Applied Earth Observation and Geoinformation*, 28, 117-130.

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

Wickham, J. D., Stehman, S. V., Fry, J. A., Smith, J. H., and Homer, C. G. (2010). Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. *Remote Sensing of Environment*, 114, 1286-1296.

Wickham, J. D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., and Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment*, 130, 294-304.

Wickham, J., Stehman, S. V., Gass, L., Dewitz, J. A., Sorenson, D. G., et al. (2017). Thematic accuracy assessment of the 2011 National Land Cover Database (NLCD). *Remote Sensing of Environment*, 191, 328-341.

Woodcock, C.E. and Gopal, S. (2000). Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, 14, 153-172.

Wulder, M. A., Coops, N. C., Roy, D. P., White, J. C., Hermosilla, T. (2018). Land cover 2.0. *International Journal of Remote Sensing*, 39, 4254-4284.

Wulder, M. A., Franklin, S. E., White, J. C., Linke, J., and Magnussen, S. (2006a). 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.

Wulder, M. A., White, J. C., Luther, J. E., Strickland, G., Remmel, T. K., and Mitchell, S. W. (2006b). Use of vector polygons for the accuracy assessment of pixel-based land cover maps. *Canadian Journal of Remote Sensing*, 32, 268-279.

Yang, L., Brus, D. J., Zhu, A-X., Li, X., and Shi, J. (2018). Accounting for access costs in validation of soil maps: A comparison of design-based sampling strategies. Geoderma 315, 160-169.

Ye, S., Pontius, R.G., Jr., and Rakshit, R. (2018). A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches. ISPRS Journal of Photogrammetry and Remote Sensing, 141, 137-147.

Zhao, Y., Gong, P., Yu, L., Hu, L., Li, X., et al. (2014). Towards a common validation sample set for global land-cover mapping. *International Journal of Remote Sensing*, 35, 4795-4814.

Zimmerman, P. L., Housman, I. W., Perry, C. H., Chastain, R. A., Webb, J. B., & Finco, M. V. (2013). An accuracy assessment of forest disturbance mapping in the western Great Lakes. *Remote Sensing of Environment*, 128, 176-185.