The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Olofsson, Pontus and Foody, Giles M. and Herold, Martin and Stehman, Stephen V. and Woodcock, Curtis E. and Wulder, Michael A. (2014) Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148 . pp. 42-57. ISSN 1879-0704

Preprint submitted to Remote Sensing of Environment *May 30, 2013*

Good Practices for Estimating Area and Assessing Accuracy of Land Change

Pontus Olofsson[a,*], Giles M. Foody[b], Martin Herold[c], Stephen V. Stehman[d], Curtis E. Woodcock[a] and Michael A. Wulder[e]

[a] *Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, USA*

[b] *School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD, UK*

[c] *Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, Droevendaalsesteeg 3, 6708 Wageningen, The Netherlands*

[d] *Department of Forest and Natural Resources Management, State University of New York, 1 Forestry Drive, Syracuse, NY 13210, USA*

[e] *Canadian Forest Service (Pacific Forestry Centre), Natural Resources Canada, Victoria, BC, 12 V8Z 1M5, Canada*

Key words: accuracy assessment, sampling design, response design, area estimation, land change, remote sensing

*Corresponding author

Email olofsson@bu.edu
URL people.bu.edu/olofsson
Phone +1-617-353-9734
Fax +1-617-353-8399

Abstract

The remote sensing science and applications communities have developed increasingly reliable, consistent, and robust approaches for capturing land dynamics to meet a range of information needs. Statistically robust and transparent approaches for assessing accuracy and estimating area of change are critical to ensure the integrity of land change information. We provide practitioners with a set of "good practice" recommendations for designing and implementing an accuracy assessment of a change map and estimating area based on the reference sample data. The good practice recommendations address the three major components: ~~of the process including the~~ sampling design, response design and analysis. The primary good practice recommendations for assessing accuracy and estimating area are: (i) implement a probability sampling design that is chosen to achieve the priority objectives of accuracy and area estimation while also satisfying practical constraints such as cost and available sources of reference data; (ii) implement a response design protocol that is based on reference data sources that provide sufficient spatial and temporal representation to accurately label each unit in the sample (i.e., the "reference classification" will be considerably more accurate than the map classification being evaluated); (iii) implement an analysis that is consistent with the sampling design and response design protocols; (iv) summarize the accuracy assessment by reporting the estimated error matrix in terms of proportion of area and estimates of overall accuracy, user's accuracy (or commission error), and producer's accuracy (or omission error); (v) estimate area of classes (e.g., types of change such as wetland loss or types of ~~no change~~persistence such as stable forest) based on the reference classification of the sample units; (vi) quantify uncertainty by reporting confidence intervals for accuracy and area parameters; (vii) evaluate variability and potential error in the

reference classification; and (viii) document deviations from good practice that may substantially affect the results. An example application is provided to illustrate the recommended process.

1. Introduction

Land change maps quantify a wide range of processes including wildfire (Schroeder et al., 2011), forest harvest (Olofsson et al., 2011), forest disturbance (Huang et al., 2010), land use pressure (Drummond and Loveland, 2010) and urban expansion (Jeon et al., 2013). Map users and producers are acutely interested in communicating and understanding the quality of these maps. Accordingly, guidance on how to assess accuracy of these maps in a consistent and transparent manner is a necessity. The use of remote sensing products depicting change for scientific, management, or policy support activities~~,~~ all require quantitative accuracy statements to buttress the confidence in the information generated and in any subsequent reporting or inferences made. Area estimation, whether of change in land cover/use or of status of land cover/use at a single date, is a natural value-added use of land change maps in many local, national and global land accounting applications. For example, the amount of land area allocated for a specific use is a key country reporting requirement to the United Nations (UN) Food and Agriculture Organization (FAO) statistics and the global forest resources assessment (FAO, 2010) ~~and~~ as well as for countries reporting under the Kyoto protocol and the evolving activities for the UN Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation – UN-REDD (UN-REDD, 2008; Grassi et al., 2008). Estimates of forest extent or deforestation are often derived via remote sensing (cf. Achard et al., 2002; DeFries et al., 2002; Hansen et al., 2010) , and area estimation also plays a prominent role in ongoing efforts to establish scientifically valid protocols for forest change monitoring in the context of specific accounting applications to policy approaches for reducing greenhouse gas emissions from forests (DeFries et al., 2007; GOFC-GOLD, 2011).

Area estimation also plays a prominent role in ongoing efforts to establish scientifically valid protocols for forest change monitoring in the context of specific accounting applications to policy approaches for reducing greenhouse gas emissions from forests (DeFries et al., 2007; GOFC-GOLD, 2011). ~~One approach to quantifying greenhouse gas emissions from forests, an important component of carbon accounting, is based on estimating the area of forest change and then applying emissions factors associated with these changes to translate the area changes into emissions (Herold and Skutsch, 2011). Thus, understanding the uncertainty in area change estimates is one key factor determining the accuracy of the overall emission and for assessing the performance and impact of climate change mitigation activities to reduce these emissions (GOFC-GOLD, 2011; Herold et al., 2011). Furthermore, the efforts of the UN-REDD clearly call for area estimates of deforestation and degradation with known uncertainty (UN-REDD, 2008). The reporting obligations of national governments also benefit from a capacity to quantitatively report on accuracy of products and to build confidence in the reported outcomes (Wulder et al., 2007). Forest certification programs, aimed at ensuring sustainable forest management practices, also require scientifically accepted means for monitoring land-based changes in a transparent and quantifiable manner.~~

A key strength of remote sensing is that it enables spatially exhaustive, wall-to-wall coverage~~,~~ of the area of interest. ~~But~~However, as might be expected with any mapping process, the results are rarely perfect. Placing spatially and categorically continuous conditions into discrete classes ~~will~~ may result in confusion at the categorical transitions. Error can also result from the change mapping process, the data used, and analyst biases (Foody, 2010). Change detection and mapping approaches using remotely sensed data are increasingly robust, with improvements aimed at the mitigation of these sources of error. However, any map made from remotely sensed data can be assumed to contain some error, with the areas calculated from the map (e.g., pixel counting) also potentially subject to bias. An accuracy assessment identifies the errors of the classification, and the sample data can be used for estimating both accuracy and area along with the uncertainty of these estimates. While the notion of accuracy assessment is well-established within the remote sensing community (Foody, 2002; Strahler et al., 2006), studies of land change routinely fail to assess the accuracy of the final change maps and few published studies of land change make full use of the information obtained from accuracy assessments (Olofsson et al., 2013).

1.1 Good Practice Recommendations

In this article, we synthesise the current status of key steps and methods that are needed to complete an accuracy assessment of a land change map and to estimate area of land change. ~~The~~ This article addresses the fundamental protocols required to produce scientifically rigorous and transparent estimates of accuracy and area. The set of good practice recommendations provides guidelines to assist both scientists and practitioners in the design and implementation of accuracy assessment and area estimation methods applied to land change assessments using remote sensing. The accuracy and area estimation objectives are linked via a map of change. A change map provides a spatially explicit depiction of change and this spatial information can be readily aggregated to calculate the total mapped area or the proportion of mapped area of change for the region of interest (ROI). Accuracy assessment addresses questions related to how well locations of mapped change correspond to actual areas of change. A fundamental premise of the recommended good practices methodology is that the change map will be subject to an accuracy assessment based on a sample of higher quality change information (i.e., the reference classification). The higher quality reference classification is compared to the map classification on a location-specific basis to quantify accuracy of the change map and to estimate area. Although it is possible to estimate area of change without producing a change map (Achard et al., 2002; FAO, 2010; Hansen et al., 2010), we will assume that a map of change exists (although there will not necessarily be a map for each date). The focus for this document is change between two dates.

~~At the outset b~~Before any detailed planning of the response and sampling designs is undertaken, a basic visual assessment should be conducted to identify obvious errors and concerns in the remotely sensed product. This assessment provides an evaluation of the map's suitability for the intended application and should detect if a map is so unsuitable for use that there is no value in proceeding to a more detailed assessment. The visual assessment should also highlight errors that are easy to remove enabling the map to be refined prior to initiating a detailed assessment or confirm that no obvious concerns exist and the map is ready for further rigorous evaluation.

We separate the accuracy assessment methodology into three major components, the response design, sampling design, and analysis (Stehman and Czaplewski, 1998). The response design encompasses all aspects of the protocol that lead to determining whether the map and reference classifications are in agreement. Because it is often impractical to apply the response design to the entire ROI, a subset of the area is sampled. The sampling design is the protocol for selecting that subset of the ROI. The analysis includes protocols for defining how to quantify accuracy along with the formulas and inference framework for estimating accuracy and area and quantifying uncertainty of these estimates. A separate section of this guidance document is devoted to each of these three major components of accuracy assessment methodology. These sections are followed by an example of the recommended workflow.

1.2 Context of Good Practice Recommendations

The good practice recommendations are intended to represent a synthesis of the current science of accuracy assessment and area estimation. We fully anticipate that improved methods will be developed over time. As the designation of "best practice" implies a singular approach, we prefer the use of "good practice" to indicate that "best" is relative and will vary, with one hard-coded approach not always appropriate. In communicating good practices, desirable features and selection criteria can be followed to ensure that the protocol applied satisfies – as thoroughly as possible – the accuracy and area estimation recommendations. The good practices recommendations do not preclude the existence of other acceptable practices, but instead represent protocols that, if implemented correctly, would ensure scientific credibility of the results. Furthermore, the recommendations presented herein allow flexibility to choose specific details of the different components of the methodology. For example, while the general recommendation for the sampling design is to implement a probability sampling protocol, there are numerous sampling designs that meet this criterion (Stehman, 2009). Similarly, the response design protocol allows flexibility to use a variety of different sources for determining the reference classification and multiple options exist for defining agreement between the map and reference classifications. The good practices recommendations represent an ideal to strive for, but it is likely that most projects will not satisfy every recommendation. Documenting and justifying deviations from good practices are expected features of many accuracy assessment and area estimation studies. For the most part, the good practice recommendations consist of methods for which there is considerable experience of practical use in the remote sensing community.

These good practice recommendations for area estimation and accuracy assessment of land change build on earlier guidelines for single-date land-cover maps described by Strahler et al.

(2006). Strahler et al. (2006) presented general guiding principles of good practices with less emphasis on details of methodology. In the intervening years since Strahler et al. (2006), additional theory and practical application related to accuracy assessment and area estimation have been accumulated, and this current document avails upon these developments to delve more deeply into methodological details. We do not attempt to provide an exhaustive description of methods given the range of issues and the highly application-specific nature of the topic. Instead, our purpose is to focus upon the main issues needed to establish a common basis of good practice methodology that will be generally applicable and result in transparent methods and rigorous estimates of accuracy and area. A list of recommendations for all components of the process (sampling design, response design, and analysis) is presented in the Summary (Section 6).

Estimating area and accuracy of change maps introduces additional methodological challenges that were not within the scope addressed by Strahler et al. (2006). In particular, the area estimation objective was not addressed at all by Strahler et al. (2006). Accuracy assessment of change highlights many unique challenges, including the dynamic nature of the reference data, and aspects of the change features including type, severity, persistence, and area~~, as examples~~. Another challenge is that change is usually a rare feature over a given landscape. The accuracy of a map and the area estimates derived with its aid are a function of the land-cover mosaic under study, the underlying imagery and the methods applied. Accuracy and area estimates for the same region will, for example, vary if using a per-pixel or object-based classification or if the spatial resolution of the imagery is altered~~ and different methods vary in value for a given application~~ (cf. Duro et al., 2012; Baker et al., 2013; Johnson, 2013).

~~The~~ Our recommendations also focus on methods for providing robust estimates of land (area) change and its uncertainties. A primary use of such estimates is in analysis and accounting frameworks such as national inventories. In evolving frameworks compensating for successful climate change mitigation actions in the forest sector (such as REDD+, DeFries et al., 2007), the consideration of uncertainties are likely linked with financial incentives and are subject to critical international political negotiations on reporting and verification (Sanz-Sanchez et al., 2013). Understanding and management of uncertainties in area change is essential, ~~in~~ particularly ~~since~~ because data and capacity gaps in forest monitoring are large in many developing countries (Romijn et al., 2012). Accuracy assessments should also focus on identifying and addressing error sources, and prioritize on capacity development needs to provide continuous improvements and reduce uncertainties in the estimates over time. This also includes assessing the value of data streams from evolving monitoring technologies (de Sy et al., 2012; Pratihast et al., 2013) where the ultimate impact on lower uncertainties need to be proven in operational contexts. Thus, the methods of good practice presented here are generic for providing robust estimates, and having agreed-upon tools to do so will provide the saliency and legitimacy for using them in quantifying improvements in monitoring systems, and for dealing with uncertainties in financial compensation schemes (e.g., for climate change mitigation actions).

This article synthesizes key steps and methods needed to complete an accuracy assessment of a change map and to estimate area and accuracy of the map classes. It addresses the protocols required to produce scientifically rigorous and transparent estimates of accuracy and area.

2. Sampling Design

The sampling design is the protocol for selecting the subset of spatial units (e.g., pixels or polygons) that will form the basis of the accuracy assessment. Choosing a sampling design requires ~~taking into~~ a consideration of the specific objectives of the accuracy assessment and a prioritized list of desirable design criteria. The most critical recommendation is that the sampling design should be a probability sampling design. An essential element of probability sampling is that randomization is incorporated in the sample selection protocol. Probability sampling is defined in terms of inclusion probabilities, where an inclusion probability relates the likelihood of a given unit being included in the sample (Stehman, 2000). The two conditions defining a probability sample are that the inclusion probability must be known for each unit selected in the sample and the inclusion probability must be greater than zero for all units in the ROI (Stehman, 2001).

A variety of probability sampling designs are applicable to accuracy assessment and area estimation, with the most commonly used designs~~,~~ being simple random, stratified random, and systematic (Stehman, 2009). Non-probability sampling protocols include purposely selecting sample units (e.g., choosing units that are convenient to access ~~units~~), restricting the sample to homogeneous areas, and implementing a complex or *ad hoc* selection protocol for which it is not possible to derive the inclusion probabilities. The condition that the inclusion probabilities must be known for the units selected in the sample must be adhered to. These inclusion probabilities are the basis of the estimates of accuracy and area, so if they are not known, the probabilistic basis for design-based inference (see Section 4.2) is forfeited. It is difficult to envision a circumstance in which a deviation from this condition of probability sampling (i.e., known inclusion probabilities) would be acceptable in rigorous scientific research.

In practice, it is not always possible to adhere perfectly to a probability sampling protocol (Stehman, 2001). For example, if the response design specifies field visits to sample locations, it may be too dangerous or too expensive to access some of the sample units. Conversely, persistent cloud coverage or lack of useable imagery for portions of the ROI may prevent obtaining the reference classification for some sample units. The reference data are often derived from another set of imagery and the spatial and temporal coverage of reference data might be different from the coverage of the imagery used to create the map. If the reference classification for a sample unit cannot be obtained, the inclusion probability is zero for that unit. All deviations from the probability sampling protocol should be documented and quantified to the greatest extent possible. For example, the proportion of the selected sample units for which cloud cover prevented assessment of the unit should be reported, or the proportion of area of the ROI for which the reference imagery is not available should be documented. Whereas probability sampling ensures representation of the population via the rigorous probabilistic basis of inference established, when a large proportion of the ROI is not available to be sampled, the question of how well the sample represents the population must be addressed by subjective judgment.

2.1. Choosing the Sampling Design

The major decisions in choosing a sampling design relate to trade-offs among different designs in terms of advantages to meet specified accuracy objectives and priority desirable design criteria. The objectives commonly specified are to estimate overall accuracy, user's accuracy (or commission error), producer's accuracy (or omission error), and area of each class (e.g., area of each type of land change). Estimates for subregions of the ROI are also often of interest (cf. Scepan, 1999). Desirable sampling design criteria include: probability sampling design~~;~~, ease~~y~~ and practicality of ~~to~~ implementation~~;~~, cost effectiveness~~;~~, representative spatial~~ly~~ ~~well~~

distribution~~ed~~ across~~over~~ the ROI~~;~~, small standard errors in the~~yields~~ accuracy and area estimates, ~~that have small standard errors;~~ ease~~y~~ ~~to~~ of accommodating~~e~~ a change in ~~sample size at~~ any step in the implementation of the design~~;~~, and availability of an approximately unbiased estimator of variance. Determining whether ~~certain~~ any or all of these desirable design criteria have been satisfied by the chosen sampling design may be subjective. For example, determining what constitutes a small standard error will depend on the application and may vary for different estimates within the same project. There are also precedents for defining an accuracy target and desired error bounds as a means for determination of sample size using standard statistical theory (Wulder et al., 2006a) (see also Section 5.1.1).

Stehman and Foody (2009) provide an overview and comparison of the basic sampling designs typically applied to accuracy assessment. Stehman (2009) provides a more expansive review of sampling design options and discusses how these designs fulfill different objectives and desirable design criteria. A variety of sampling designs will satisfy good practice guidelines so the key is to choose a design well suited for a given application. Three key decisions that strongly influence the choice of sampling design are whether to use strata, whether to use clusters, and whether to implement a systematic or simple random selection protocol (Stehman, 2009). Each of these decisions will be discussed in the following subsections.

2.1.1. Strata

There is ~~Often~~ often ~~there is~~ a desire to partition the ROI into discrete, mutually exclusive subsets or strata (e.g., a global map could be stratified geographically by continents). Stratification is a partitioning of the ROI in which each assessment unit is assigned to a single stratum. The two most common attributes used to construct strata are the classes determined from the map and geographic subregions within the ROI. Stratification is implemented for two primary purposes. The first purpose is when the strata are of interest for reporting results (e.g., accuracy and area are reported by land-cover class or by geographic subregion). The second use of stratification is to improve the precision of the accuracy and area estimates. For example, when strata are created for the objective of reporting accuracy by strata, the stratified design allows specifying a sample size for each stratum to ensure that a precise estimate is obtained for each stratum. Land change often occupies a small proportion of the landscape, so a change stratum can be identified and the sample size allocated to this stratum can be large enough to produce a small standard error for the change user's accuracy estimate.

The practical reality is that limited resources will likely be available for the reference sample and this constraint will strongly impact sample allocation decisions because different allocations favour different estimation objectives. For example, allocating equal sample sizes to all strata favours estimation of user's accuracy over estimation of overall and producer's accuracies (Stehman, 2012). Conversely, the standard errors for estimating producer's and overall accuracies are typically smaller for proportional allocation (i.e., the sample size allocated to each stratum is proportional to the area of the stratum) relative to equal allocation. As a compromise between favouring user's versus producer's and overall accuracies, the allocation recommended is to shift the allocation slightly away from proportional allocation by increasing the sample size in the rarer classes, but the sample size for the rare classes should not be increased to the point where the final allocation is equal allocation (see Section 5 for an example). The sample size allocation decision can be informed by calculating the anticipated standard errors (see Sections 4.3 and 4.4) for different sample sizes and different allocations. An ineffective allocation of sample size to strata will not result in biased estimators of accuracy or area, but it may result in larger standard errors (see Section 5 for an example).

When stratified sampling is applied to a single date land-cover map, it is usually feasible to define a stratum for each land-cover class (Wulder et al., 2007). Identifying an effective stratification for change can be more challenging. A common approach is to use a map of change to identify the strata, and such strata are effective for estimating user's accuracy of change precisely. However, the number of different types of change may be so large that defining every change type as a stratum is not advisable. For example, in a post-classification comparison of two land-cover maps~~,~~ that each include ~~with a map legend that includes~~ 8 land-cover classes, there are 56 possible types of change in the final change map. If each stratum must receive a relatively large sample to achieve a precise user's accuracy estimate, the overall sample size may be unaffordable.

The trade-offs between precision of user's accuracy, producer's accuracy, and area estimates from different sample size allocations become exacerbated as the number of strata increases. Some types of change may be very unlikely to occur and consequently could be eliminated as strata. To further reduce the number of strata, strata could be defined on the basis of generalized change categories (Wickham et al., 2013). For example, a stratum could be change from any class to urban (i.e., urban gain), and another stratum could be change to any class from forest (i.e., forest loss). These generalized or aggregated change strata are obviously less focused on all possible individual change types. For example, the forest loss stratum could include forest to developed, forest to water, or forest to cropland. These generalized change strata would allow for specifying the sample size allocated to different general change types, but within one of the generalized strata, the sample size allocated to the individual change types would be proportional to the area of that change type. For example, if the most common type of forest loss is to cropland and the least common change is forest loss to water, many more of the sample units within the forest loss stratum will be forest-to-cropland-conversion. Strahler et al. (2006, Fig. 5.2, p. 32) provides additional examples of aggregated change classes that could be used as strata.

The desire to limit the number of strata motivates discussion of subpopulation estimation as it relates to sampling design. A subpopulation is any subset of the ROI, for example a particular type of change or a particular subregion. Subpopulations can be defined as strata, but it is not necessary for a subpopulation to be defined as a stratum to produce an estimate for that subpopulation. For example, when aggregating multiple types of change into a generalized change stratum, it would still be possible to estimate accuracy of each of the subpopulations representing the individual types of change making up the aggregated change stratum. However,~~But~~ if these subpopulations are not defined as strata, the sample size representing the subpopulation may not be large enough to obtain a precise estimate. Resources available for accuracy assessment may require limiting the number of strata used in the design, so prioritizing subpopulations may be necessary to establish which subpopulations are defined as strata.

It is sometimes the case that several maps will be assessed based on a common accuracy assessment sample. This forces a decision on whether the strata should be based on a single map (and if so, which map) or if the strata should be defined by a combination of the multiple maps. Once strata are defined and the sample is selected using these strata, the strata become a fixed feature of the design because the analysis is dependent on the estimation weight associated with each sample unit and this weight is determined by the sampling design. Fortunately, whatever the decision is to define strata when multiple maps are to be assessed, the sample reference data are still valid to assess any of the maps, even if the strata are defined on the basis of a single map. The principles of estimation outlined in the Analysis Section (Section 4) must be adhered to, and this simply requires using the estimation weights for the sample units determined by the original stratified selection protocol. The impact of the choice of strata will be reflected in the standard errors of the estimates. Olofsson et al. (2012) and Stehman et al. (2012) discuss sampling design issues associated with constructing a reference validation database that would allow assessment of multiple maps.

To summarize the recommendations related to the important question of whether to incorporate stratification in the sampling design, stratifying by mapped change and by subregions is justified to achieve the objective of precise class-specific accuracy and to report accuracy by subregion. If the overall sample size is not adequate to support both class-specific and subregion accuracy estimates, the subregional stratification may be omitted and accuracy by subregion relegated to the status of subpopulation estimation. The recommended allocation of sample size to the strata defined by the map classes is to increase the sample size for the rarer classes making the sample size per stratum more equitable than what would result from proportional allocation, but not pushing to the point of equal allocation. The rationale for this recommendation is that user's accuracy is often a priority objective and we can control the precision of the user's accuracy estimates by the choice of sample allocation. However, the trade-off is that a design allocation chosen solely for the objective of user's accuracy precision (i.e., equal allocation) may be detrimental to precision of estimates of overall accuracy, producer's accuracy, and area, so a compromise allocation is in order. Lastly, defining aggregations of change types as strata may be necessary if the number of strata needs to be limited, and accuracy and area estimates for the individual change types would be obtained as subpopulation estimates.

2.1.2. Cluster Sampling

A cluster is a sampling unit that consists of one or more of the basic assessment units specified by the response design. For example, a cluster could be a 3 x 3 block of 9 pixels or a 1 km x 1 km cluster containing 100 1 ha assessment units. In cluster sampling, a sample of clusters is selected and the spatial units within each cluster are therefore selected as a group rather than selected as individual entities. Each of the spatial units within a cluster is still interpreted as a separate unit even though it is selected into the sample as part of a cluster. For example, a 3 x 3 pixel cluster would require obtaining the reference classification for individual pixels within the cluster.

The primary motivation for cluster sampling is to reduce the cost of data collection. For example, if field visits are required to obtain the reference classification, transit time and costs may be reduced if the sample units are grouped spatially into clusters. Zimmerman et al. (2013) used cluster sampling to reduce the number of raster images (i.e., clusters) required because the primary cost of the sampling protocol was associated with processing the very high resolution images used for reference data. As another example, Stehman and Selkowitz (2010) used a 27 km x 27 km cluster sampling unit to constrain sample locations to a single day of flight time per cluster when the reference data were collected by aircraft. Cluster sampling may also be motivated by the objectives of an accuracy assessment. For example, a cluster sampling unit becomes necessary to assess accuracy at multiple spatial supports (e.g., single pixel, 1 ha unit, and 1 km^2 unit).

The cost savings gained by cluster sampling should be substantial before choosing this design because the correlation among units within a cluster (i.e., intracluster correlation) often reduces precision relative to a simple random sample of equal size. Focusing on the specific example of estimating land-cover area in Europe, Gallego (2012) showed that a 10 km x 10 km sampling unit produced equivalent information to that of a simple random sample of only 25 points or fewer. The low yield of information per cluster diminishes the cost advantage of cluster sampling if the intracluster correlation is high. Another potential disadvantage of cluster sampling is that it complicates stratification when the strata are the map classes and the assessment unit is a pixel. In the simplest setting, each cluster would be assigned to a stratum, but rules have to be established for assigning a cluster to a stratum when the cluster includes area of several different classes. Cluster sampling can be combined with stratification of pixels by the map class of each pixel in a two-stage stratified cluster sampling approach (Stehman et al., 2003, 2008), but such designs require more complex analysis and implementation protocols than what are required of a stratified design without clusters. Because of the added complexity ~~of~~ cluster sampling introduces for sampling design (e.g., accommodating stratification within a cluster sampling design) and estimation (e.g., estimating standard errors), we recommend this design only in cases for which the objectives require a cluster sampling unit or in which the cost savings or practical advantages of cluster sampling are substantial.

2.1.3. Systematic vs. Random Selection

The two most common selection protocols implemented in accuracy assessment are simple random and systematic sampling (we define “systematic” as selecting a starting point at random with equal probability and then sampling with a fixed distance between sample locations). Both protocols can be implemented to select units from within strata or to select clusters, and both can be applied to a ROI that is not partitioned into strata or clusters. Unbiased estimators of the various accuracy parameters are available from either systematic or simple random selection, so the bias criterion is not a basis for choosing between these options. Instead, the choice of simple random versus systematic depends on how each selection protocol satisfies the priority desirable design criteria (Stehman, 2009). For example, systematic sampling is often simpler to implement when the response design is based on field visits, but the greater convenience of systematic versus simple random is diminished when working with imagery or aerial photographs as a source of the reference data. Typically, systematic selection will yield more precise estimates than simple random selection, but systematic sampling requires use of a variance approximation so if unbiased variance estimation is a priority criterion, simple random is preferred. Simple random selection also is advantageous if it is likely that the sample size will need to be modified during the course of the accuracy assessment (Stehman et al., 2012). A scenario in which systematic selection opportunistically arises is when accuracy assessment reference data can be simultaneously obtained in conjunction with another field sampling activity. For example, many national forest inventories employ a systematic sample of field plots (Tomppo et al., 2010) and these field plot data may be an inexpensive, high quality source of reference data. In general, the simple random selection protocol will better satisfy the desirable design criteria and is the recommended option. However, systematic selection is also nearly always acceptable.

2.2. A Recommended Good Practice Sampling Design

Stratified random sampling is a practical design that satisfies the basic accuracy assessment objectives and most of the desirable design criteria. Stratified random sampling affords the option to increase the sample size in classes that occupy a small proportion of area to reduce the standard errors of the class-specific accuracy estimates for these rare classes. Thus this design addresses the key objective of estimating class-specific accuracy. In regard to the desirable design criteria, stratified random sampling is a probability sampling design and it is one of the easier designs to implement. Stratified sampling is commonly used in accuracy assessment so it has an advantage of being familiar to the remote sensing community (cf. Mayaux et al., 2006; Cakir et al., 2006; Huang et al., 2010; Olofsson et al., 2011). Increasing or decreasing the sample size after the data collection has begun is readily accommodated by stratified random sampling, and unbiased variance estimators are available thus avoiding the need to use variance approximations. An assumption implicit in this recommendation is that change between two dates is of interest. Little work has been done to investigate~~ing~~ the effective use of strata for multiple change periods. ~~Stratifying by a change map also assumes that it is possible to obtain the reference data for the initial date of the change period given that the change map will not be available until the end date of the change period. If this is not possible, stratification is still an option but the strata would need to be constructed on the basis of predicted change.~~In the case of stratification based on a change map, it is assumed that reference data for the sampled locations exists for the initial date of the change period (e.g., archived imagery or aerial photography is available). If the reference data must be obtained in real time (e.g., via ground visit), it would not be possible to stratify by a change map that does not yet exist at the initial date. An alternative would be to stratify by anticipated change or predicted change, with the effectiveness of such strata dependent on how well the predicted change matched with the ensuing reality of change.

3. Response Design

For the accuracy assessment objective, the response design encompasses all steps of the protocol that lead to a decision regarding agreement of the reference and map classifications. For area estimation, the response design provides the best available classification of change for each spatial unit sampled. The ~~F~~four major features of the response design are the spatial unit, the source or sources of information used to determine the reference classification, the labelling protocol for the reference classification, and a definition of agreement. Each of these major features is discussed in the following subsections.

3.1. Spatial Assessment Unit

The spatial unit that serves as the basis for the location-specific comparison of the reference classification and map classification can be a pixel, polygon (or segment), or block (Stehman and Wickham, 2011). The ROI is partitioned based on the chosen spatial unit (i.e., the region is completely tiled by these non-overlapping spatial units). Commonly, the pixel is selected as the spatial unit. The pixel is an arbitrary unit defined mainly by the properties of the sensing system used to acquire the remotely sensed data or a function of the grid used to sub-divide space in a raster based data set. A polygon is defined as a unit of area, perhaps irregular in shape, representing a meaningful feature of land cover. For example, a polygon may be delineated from a map such that the area within the polygon has the same map classification (e.g., the entire polygon is stable forest or the entire polygon represents an area of change from forest to urban). Polygons defined on the basis of a map will be called “map polygons.” Alternatively, a polygon could be delineated on the basis of the reference classification as an area within which the reference class is the same. A polygon delineated on the basis of the reference classification will be called a “reference polygon”. A “block” spatial assessment unit is defined as a rectangular array of pixels (e.g., a 3 x 3 block of pixels). Irrespective of the spatial unit selected, it is important to note that some spatial units may be impure, ~~that is~~i.e., they represent an area of more than one class. Mixed pixels are~~, for example~~ common, especially in coarse spatial resolution data. Similarly, it is~~, for example,~~ possible that a map polygon is not internally homogeneous in terms of the reference classification, and a reference polygon may not be internally homogeneous in terms of the map classification. A polygon defined by a segmentation algorithm would not necessarily be homogeneous in terms of either the map or the reference classifications.

Pixels, polygons, or blocks can be used as the spatial unit in accuracy assessment. Regardless of the unit chosen, a critical feature of the response design protocol is that the spatially explicit character of the accuracy assessment must be retained. Practitioners should aim to have reference data with an equal or finer level of detail than the data used to create the map, but we make no recommendation ~~is made~~ regarding the choice of spatial assessment unit. However, once the spatial assessment unit has been chosen, there will be good practice recommendations associated with that specific unit and the choice of spatial unit also has implications on the sampling design (Stehman and Wickham, 2011) and analysis. Estimates of accuracy and area derived from the same map but through the use of different spatial units may be unequal.

3.2. Sources of Reference Data

The reference classification can be determined from a variety of sources ranging from actual ground visits to the sample locations or the use of aerial photography or satellite imagery. There are two ways to~~To~~ ensure that the reference classification is of higher quality than the map classification:, ~~either~~ the reference source has to be of higher quality than what was used to create the map classification, and 2)~~or~~ if using the same source material for both the map and reference classifications, the process to create the reference classification has to be more accurate than the process used to create the classification being evaluated. ~~(e.g.~~For example, if Landsat imagery is used to create the map and Landsat is the only available imagery for the accuracy assessment, then the process for obtaining the reference classification has to be more accurate than the process for obtaining the map classification~~)~~. ~~Further~~Additionally, other spatial data may be used to improve the quality of the reference classification, such as forest inventory data or some form of vector data (e.g., roads, pipelines, or crop records). In this subsection, different potential sources of reference data for assessing accuracy of change are identified and strengths and weaknesses of these sources are described.

Possible reference data sources include field plots, aerial photography, forest inventory data, airborne video, lidar, and satellite imagery (Table 1). Additional sources of freely accessible reference data may also be opportunistically available from data mining and crowdsourcing (Iwao et al., 2006; Foody and Boyd, 2013).~~, and silvicultural records (Hyyppä et al., 2000; Wulder et al., 2006a).~~

<< TABLE 1 HERE >>

Practical considerations regarding costs often influence the selection of reference data, or the use of existing data. While existing or lower cost data may be desirable from a purchase perspective, the use of disparate data sources will result in additional effort by project analysts to deal with exceptions and inconsistencies. A key to using disparate data sources is to have the reference data that are actually used in the accuracy assessment be, as much as possible, invariant to source. For example, the creation of attributed change polygons makes the polygon the common denominator, rather than the source data. Creating polygonal change units in a portable format and populating a minimum set of fields to support a consistent labelling protocol is desirable. The information to be recorded for each change unit is itemized in Table 2.

<< TABLE 2 HERE >>

Ideally a data source is available for the entire ~~with uniform likelihood over the~~ ROI, representing the change types and dates of interest, at a low cost. The realities versus the ideal result in a series of considerations are detailed in Table 3. For instance, if the ROI is small, the costs may be less of an issue and access may not be relevant. For large area projects over poorly monitored areas, existing data sources are not often available so data purchase and interpretation costs become the dominant criteria. The ease of interpretation and consistency of source reference data permits economies in the project flow for the analysts and also promotes automation of repeated activities. Further, the development of a well documented and consistent change validation data set will have utility for multiple projects and purposes.

<< TABLE 3 HERE >>

Both high- and very high spatial resolution satellite data are viable candidates for reference data. Imagery is typically considered as very high spatial resolution (VHSR) with a spatial resolution of ~~when pixels are sided~~ < 1 m and high spatial resolution (HSR) with a spatial resolution of < 10 m. Both data sources provide information that is finer than the data used in most large area monitoring projects, which would typically have ~~use imagery with~~ a spatial resolution of greater than 10 m. At the fine spatial resolution of satellite-borne VHSR imagery, panchromatic is often the only spectral information collected. The typical 400 to 900 nm panchromatic data with small pixels (0.50 m in the case of WorldView-1) closely resemble large scale aerial photography and can be interpreted using established aerial photograph interpretation techniques (Wulder et al., 2008a) or subject to digital analyses (cf. Falkowski et al., 2009). Both the SPOT Image® and

DigitalGlobe® archives can be accessed through Google Earth™, with the image extents by year portrayed. The presence of freely accessible high spatial resolution imagery online~~, freely accessible,~~ through Google Earth™ also presents low cost interpretation options. Limitations of this approach include a lack of data prior to the initiation of the high spatial resolution satellite commercial era (circa 2000), spatial distribution of available imagery, and the actual temporal revisit of the images available. The reported temporal revisit can be on the order of days based upon an ability to point the sensor head. For instance, IKONOS has off-nadir revisit of 3 to 5 days, with 144 days required for nadir revisit (Wulder et al. 2008b). The implication is that when the sun-surface-sensor viewing geometry changes the structure captured changes, such that trees evident on one image may be occluded in another. For a given on-line accessible source of satellite imagery, it should not be expected that historical, archival, global coverage from launch to present exist ~~should not be expected~~. Regardless, the ability to view images from multiple years can help determine that date when a change (e.g., a disturbance) occurred. The additional context provided around particular change events aids with interpretation of change type (e.g., determination of harvesting versus forest removal in support of agricultural expansion). ~~Development and sharing of a change data base, once interpreted and attributed following defined procedures, leveraging Google Earth™ is a consideration for global or large area accuracy assessment activities.~~

There are few, if any, reference data sources that are available with a uniform likelihood globally. There are some archival datasets with wide global coverage (e.g., Kompsat); although, the utility of these data sets may be limited. The utility of any given ~~data~~ reference data source when used to capture and relate change is the date or represented by ~~vintage of~~ the data. While less of an issue with satellite data, air photos and maps may not be of a known vintage.

Acquisition dates of historic photos are often lost, plus maps are often representative of a period, not a singular date. Knowing the conditions that previously existed may not be helpful if the date of change occurrence is not known.

Over some regions, land use change and silvicultural records may also be available to inform on the land-~~-~~cover change. Note that forest harvesting is a land-~~-~~cover change relating a successional stage, rather than a land use change (which implies a permanent change in how a particular parcel of land is used – e.g., forestry to agriculture). ~~The~~ This distinction is important for both monitoring and reporting purposes as the permanent removal of forests has differing carbon consequences than ~~a~~ forest harvesting (Kurz, 2010).

While the good practice guidelines advocate for use of reference data of finer spatial resolution than the map product, this is especially so for single date interpretations of the reference data. Following the opening of the Landsat archive by the USGS (Woodcock et al., 2008), time series of imagery ~~creates~~ created new opportunities for using imagery of the same spatial resolution (e.g., Landsat) when archival data are available. Simple visual approaches may be applied, such as in Figure 1, where a change event (fire) that is evident in 2010 can be timed quite precisely by the evidence captured (smoke plume) showing when the fire ~~is~~ occurred~~ing~~. This type of change dating is rather opportunistic and not to be commonly expected.

<<FIGURE 1 HERE>>

Figure 1. Landsat data can be used for the visual dating of change, with the fire event in progress in Inset A, August 3, 2010, with the burned forest outcome evident in Inset B, September 20, 2010, Yukon, Canada (Landsat Path 55, Row 18).

A more reliable means for determining the timing of change events can be from developing and interrogating time series of images (Kennedy et al., 2010). To ensure the quality of time series transitions developed, Cohen et al. (2010) created a logic and tool for determining the timing and nature of changes captured (TimeSync, http://timesync.forestry.oregonstate.edu/). Based upon the image collection and archiving protocols present through the history of Landsat, the spatial and temporal coverage of imagery is not uniform. The temporal precision possible for dating changes based upon time series analysis is likely weaker for locations that already have a paucity of data. This situation is due to the historic practices followed at given Landsat receiving stations through to the commercial era (during the 1980s) when fewer images were collected and archived (Wulder et al., 2012). It should not be assumed that the temporal density possible for the conterminous United States is possible for all other regions (Schroeder et al., 2011).

Another critical aspect of the response design is that the change period represented by the reference classification must be synchronous with the change period of the classification. Consider a map representing change between 2000 and 2010. To capture ~~near anniversary dates (within year) and a~~the northern hemisphere peak photosynthetic period, the imagery used for this hypothetical project was collected July 15, 2000, and 10 years later, July 15 2010. The reference data should be collected in 2010, but ideally not after July 15 (assuming similar satellite overpass times) to avoid confusion. Data collected after July 15, 2010 will have to be vetted to ensure the change present in the reference data did not occur after the product date of the change map. Imagery from the same year is desired but may not always be possible. As such, it is required that the change reference data ~~includes~~ approximates the date the change occurred as precisely as ~~possible~~available. Multiple images help refine the timing of the change event. Mismatched change periods between the map and reference classifications would be a major source of reference data error.

3.3. Reference Labelling Protocol

The labelling protocol refers to the steps in the response design that take the information provided by the reference data and convert that information to the label or labels constituting the reference classification. Labelling is far from trivial with numerous definitions for land-~~-~~cover classes in use (cf. Comber et al., 2008) although recent developments such as the FAO's Land Cover Classification system (LCCS) may act to enhance interoperability (Ahlqvist, 2008). The labelling protocol should also include specification of a minimum mapping unit (MMU) for the reference classification. The MMU can have important implications for accuracy assessment and area estimation. For example, increasing the size of the MMU will lead to a reduction in the representation of classes that occupy small, often fragmented, patches (Saura, 2002). Changing the MMU can also impact ~~on~~ accuracy estimates, although the effect is most apparent when a large change is made (Knight and Lunetta, 2003). ~~Clearly, s~~Small patches present a challenge to mapping (cf. He et al., 2011) and the accuracy of their mapping will degrade as the MMU is increased. However, ~~but~~ it is possible that overall map accuracy may increase with a larger MMU, making it is important to ensure that attention is focused on an appropriate measure of accuracy for the application in-hand. The precise effects of the MMU will vary as a function of the land-~~-~~cover mosaic under study and the imagery used. The MMU specified for the response design does not necessarily have to match the MMU specified for the map. In fact, if the reference classification is intended to apply to a variety of maps, it would be likely that the MMU of the reference classification does not match the map classification for all maps that might be assessed. Often the reference imagery or information will permit distinguishing smaller patches or features than can be distinguished from the map so a smaller MMU will be possible for the reference classification.

The easiest case for the labelling protocol occurs when the assessment unit is homogeneous and a single reference class label can be assigned (the reference class could be a type of change). ~~But o~~Often, however, the situation will be more complex making class labelling less certain. For example, the assessment unit may contain a mixture of classes, and even if the unit is homogeneous, it may be difficult to assign a single label (e.g., change type) because the unit is not unambiguously one of the classes in the legend but instead falls between two of the discrete class options in the legend (i.e., land-~~-~~cover classes are a continuum represented on a discrete scale). A variety of options exist for labelling a unit when a single reference label does not adequately represent the uncertainty of a unit. One or more alternate reference class labels can be assigned to account for ambiguity in the reference classification. Another option when defining agreement is to construct a weighted agreement based on how closely the different classes are related. For example, in the GlobCover assessment, a “matrix” of class relationships was established (Mayaux et al., 2006, GLC2000). A fuzzy reference labelling protocol may also be employed, ~~for example~~such as the linguistic scale devised by Gopal and Woodcock (1994) or a fuzzy membership vector in which the reference label for a unit specifies a membership value for each class (Foody, 1996; Binaghi et al., 1999). Another option for mixed units is to specify the proportion of area of each class present in the unit (Foody et al., 1992; Lewis and Brown, 2001). A different characterization of uncertainty in the reference classification is obtained by assigning a confidence rating that represents the interpreter’s perception of uncertainty in the reference classification for that unit. For example, low, moderate and high confidence ratings would indicate increasing confidence on the part of the interpreter that the reference classification is correct. Typically this information can then be used in the analysis to subset results by confidence rating (Powell et al., 2004; Wickham et al., 2001, Table 4).

The response design should include protocols to enhance consistency of the reference class labelling. For example, interpretation keys should be created if visual assessment is used to obtain the reference classification (Kelly et al., 1999) and specific instructions to translate quantitative field data into reference labels should be provided and documented. If multiple interpreters are used, training interpreters to ensure consistency is critical. Interpreters should be in communication throughout the process to discuss and review difficult cases and to agree upon a common approach to labelling such cases. Difficult cases should be noted for future reference and consensus development (e.g., the imagery is retained and accessible, and the decision process leading to the reference label of the case is documented). Rather than solely visual approaches, entire high spatial resolution images can be classified, with the underlying imagery also maintained and accessible as support information to the accuracy assessment (that is, to gain/ensure confidence in the categories selected for a given location).

3.4. Defining Agreement

Once the map and reference classifications have been obtained for a given spatial unit, rules for defining agreement must be specified before proceeding to the analyses that quantify accuracy. In the simplest case, a single class label is present for the map and a single label is provided by the reference classification. If these labels agree, the map class is correct for that unit~~,~~ ;~~and~~ if the labels disagree, the type of misclassification is readily identified. Defining agreement becomes more complex if the assessment unit is not homogeneous or if more than a ~~single~~ one class label is assigned by the map or reference classification. For example, if the reference classification provides a primary and secondary reference label, agreement can be defined as a match between the map label and either the primary or secondary reference label. If the reference classification consists of a vector of proportions of area of the classes present in the assessment unit (e.g., the area proportions of the classes are 0.2, 0.5, and 0.3), agreement can be defined as the proportion of area for which the map and reference labels are the same. The critical feature of the protocol for defining agreement is that it allows construction of an error matrix in which the elements of the matrix represent proportion of area of agreement and disagreement between the map and reference classifications. These proportions (in terms of area) achieve the necessary spatially explicit assessment of map accuracy and the requirements for area estimation.

3.5. Reference Classification Uncertainty: Geolocation and Interpreter Variability

In an ideal case, the reference classification is based on a reference data set of such quality that the sample labels represent the ground truth (i.e. a "gold standard" reference data set). However, the reference classification is subject to uncertainty, and an assessment of this uncertainty should be conducted. Small errors in the reference data set can lead to large biases of the estimators of both classification accuracy and class area (Foody, 2010; 2013). Two potential sources of uncertainty in the reference classification are the uncertainty associated with spatial co-registration of the map and reference location (Pontius, 2000) and uncertainty associated with the interpretation of the reference data (Pontius and Lippitt, 2006).

Geolocation error is defined as a mismatch between the location of the spatial assessment unit identified from the map and the location identified from the reference data. The response design should be constructed to minimize geolocation error. For instance, it is common for plots to have a GPS position. The quality of the GPS position can be related ~~by~~ to the type of instrument used, which can provide an indication of spatial precision. The length of time, number of position measures to resolve the location, and the number of satellites are also aspects that can be recorded. The magnitude of geolocation error should be characterized by documenting the spatial location quality of the map and reference data sources (e.g., GPS units, aerial photography, or satellite imagery). If airborne imagery is to be used, aircraft positioning and pointing information should be collected. The GPS location of the aircraft does not necessarily indicate the position of the point on the ground that is captured in photographic or video data. A slight roll of the aircraft can create a mismatch between the recorded and actual positions. Error in the classification may be incorrectly indicated due to these spatial mismatches, especially for smaller change events or rare classes.

Interpreter uncertainty can be separated into two parts: 1) interpreter bias is defined as an error in the assignment of the reference class to the spatial unit; 2) interpreter variability is a difference between the reference class assigned to the same spatial unit by different interpreters (i.e., interpreter variability is the complement of among interpreter agreement). ~~Although i~~Ideally an assessment of both interpreter bias and interpreter variability would be conducted~~,~~; in practice, assessing only interpreter variability may be feasible. The difficulty hindering assessment of interpreter bias is whether a “gold standard” of truth exists against which the interpreted reference classification can be compared. For example, on-the-ground reference data may serve to establish the gold standard of truth for land cover at a single date, but a gold standard for change based on field visits would be much more difficult and costly to establish. Comparison of interpreters to an “expert” interpreter is a practical but less satisfying option for quantifying interpreter bias and the success of this approach depends on how closely the expert classification mimics the gold standard. A distinction between the accuracy assessment of land cover and change does exist, whereby the continuous nature of land cover benefits more from field visits. Depending on the change categories of interest, field visits may not be as informative. For example, slower continuous changes may benefit from field visits, but rapid stand replacing disturbances may not. The date of change, if not captured in silvicultural records or fire maps, may actually be better captured from imagery of known vintage than through field visits (Cohen et al., 2010).

~~If multiple interpreters or interpreter teams are providing the reference classification, interpreter variability can be assessed by having interpreters classify a common sample of locations. Ideally, the sample would include locations covering a variety of classes to allow evaluating how interpreter variability differs by class (e.g., do interpreters consistently agree for some classes, but not others). The quality of the interpreters in terms of the accuracy of their labelling may also be assessed directly from the data generated (Foody et al., 2013). If this evaluation sample is selected using a probability sampling design (see Section 2), estimates of interpreter variability will have a strong inferential basis and results from the sample can be rigorously inferred to the population of all interpretations. If multiple interpreters operating independently are employed to determine the reference classification for each sample location, a number of considerations may affect the decision of how many interpreters are used. Wulder et al. (2007) who used seven interpreters in a land cover labelling protocol, detail the issues that arise when using multiple interpreters, noting common disagreement between interpreters, especially for more refined and rare classes. Ensuring that consensus is reached, rather than an aggregation of independent interpretations, is also possible. Also using airborne video data, Powell et al. (2004) required five interpreters to agree upon a specific class, with the outcome then treated as a "gold standard". While some disagreement could be linked to difficulty in identifying the vegetation in the video, other sources of disagreement included data entry error and misreading of sample labels. These are sources of error that can be mitigated by using~~

~~intelligent data management and entry tools. Wulder et al. (2007), recommend the use of an independent evaluation protocol, followed by cross-calibration, and the revisit of problematic classes. This would allow for the use of fewer resources and interpreters yet still gain the benefit of multiple interpreters.~~

A number of issues arise when using multiple interpreters to obtain the reference classification (Wulder et al. 2007). Disagreements among interpreters evaluating the same sampling unit are likely. These disagreements may be resolved by a consensus agreement on the reference class; for example, Powell et al. (2004) required five interpreters to agree upon a specific class, with the outcome then treated as a “gold standard”. Constant communication among the multiple interpreters to discuss and document difficult cases is important to foster enhanced consistency and accuracy of the reference labeling process (Wickham et al. 2013).

The response design protocols described in this section have ~~has~~ focused on land-~~-~~cover changes that can be characterised by a complete change in class type: conversions of cover. In some studies attention is focused on more subtle changes or modifications of land cover, as changes in land cover can be considered as processes (Gomez et al., 2011) with ~~depletions~~ gains and ~~accruals~~ losses in vegetation captured and possible to assign a label (Kennedy et al., 2010). Cohen et al. (2010) show how investigation of time series of satellite imagery supported by period photography can illuminate ~~on~~ subtle changes in forest conditions ~~(~~such as decline due to insects or water stress and conversely recovery of forests following disturbance~~)~~. ~~The importance of the ability to capture and label subtle changes is dependent upon the goals of the change classification. The interest in quantifying emissions of CO_2 to the atmosphere, a full accounting of subtle changes is increasingly desired, with capture of degradation (FAO, 2011) – while difficult – of interest for averting and related documentation of deforestation (UN-REDD, 2008).~~

The response design protocols presented also do not address the situation in which the map provides information as a continuous variable. Although many of the basic concepts underlying the good practice recommendations would apply to a continuous variable, the details of ~~methodology of~~ the accuracy assessment methodology (cf. Riemann et al., 2010) and area estimation would likely be considerably different from the methods presented herein.

4. Analysis

The analysis protocol specifies the measures to be used to express accuracy and class area as well as the procedures to estimate the selected measures from the sample data ~~acquired~~. In the context of studies of land change, there are two key objectives of the analysis: 1) accuracy~~the~~ assessment ~~of the accuracy~~ of the change classification, and 2) estimation ~~the provision~~ of ~~information on the~~ area of change. The confusion or error matrix (hereafter noted as the error matrix) plays a central role in meeting both the accuracy assessment and area estimation objectives (Foody, 2013; Stehman, 2013).

4.1 The Error Matrix

The error matrix is a simple cross-tabulation of the class labels allocated by the classification of the remotely sensed data against the reference data for the sample sites. The error matrix organizes the acquired sample data in a way that summarizes key results and aids the quantification of accuracy and area. The main diagonal of the error matrix highlights correct classifications while the off-diagonal elements show omission and commission errors. The cell entries and marginal values of the error matrix are fundamental to both accuracy assessment and area estimation. Table 4 illustrates a four-class example error matrix of the type often used in studies of land change.

<< TABLE 4 HERE >>

The rows of the error matrix represent the labels shown in a map derived from the classification of the remote sensing data and the columns represent the labels depicted in the reference data. This layout is not a universal requirement and some may wish to reverse the contents of the rows and columns. In the matrix, p_{ij} represents the proportion of area for the population that has map class *i* and reference class *j*, where "population" is defined as the full region of interest, and p_{ij} is therefore the value that would result if a census of the population were obtained (i.e., complete coverage reference classification).

Accuracy parameters derived from a population error matrix of *q* classes include overall accuracy

$$O = \sum_{j=1}^{q} p_{jj} \quad (1)$$

user's accuracy of class *i* (the proportion of the area mapped as class *i* that has reference class *i*)

$$U_i = p_{ii}/p_{i\cdot} \quad (2)$$

or its complementary measure, commission error of class *i*, $1 - p_{ii}/p_{i\cdot}$, and producer's accuracy of class *j* (the proportion of the area of reference class *j* that is mapped as class *j*),

$$P_j = p_{jj}/p_{\cdot j} \quad (3)$$

or its complementary measure, omission error of class j, $1 - p_{jj}/p_{\cdot j}$. A variety of other measures of accuracy has been used in remote sensing (Liu et al., 2007). A commonly used measure is the kappa coefficient of agreement (Congalton and Green, 2009). The problems associated with kappa include but are not limited to: 1) the correction for hypothetical chance agreement produces a measure that is not descriptive of the accuracy a user of the map would encounter (kappa would underestimate the probability that a random selected pixel is correctly classified); 2) the correction for chance agreement used in the common formulation of kappa is based on an assumption of random chance that is not reasonable because it uses the map marginal proportions of area in the definition of chance agreement and these proportions are clearly not simply random; and 3) kappa is highly correlated with overall accuracy so reporting kappa is redundant with overall accuracy."~~However, kappa has numerous problems not least an incorrect and unnecessary "correction" for chance agreement~~ (Foody, 1992; Stehman, 1997; Liu et al., 2007; Pontius and Millones, 2011). Consistent with the recommendation in Strahler et al. (2006), the use of kappa is strongly discouraged as, despite its widespread use, it actually does not serve a useful role in accuracy assessment or area estimation.

4.2 General Principles of Estimation for Good Practice

The ~~core nature of the~~ analysis protocol is designed to achieve the objectives of estimating ~~produce estimates of~~ accuracy and area from the sample data. Analysis thus requires statistical inference as the underlying scientific support for generalizing from the sample data to the population parameters and for quantifying uncertainty of the sample-based estimators. We recommend design-based inference (Särndal et al., 1992) as the framework within which estimation is conducted. A fundamental tenet of design-based inference is that the specific estimators for accuracy, area, and the variances of these estimators depend on the sampling design implemented; different estimators are appropriate for different sampling designs. ~~It is,~~ T~~t~~herefore, it is essential that only unbiased or consistent estimators should be used. In practical terms, this means that only formulas for estimating parameters and variances that account for the inclusion probabilities associated with the sampling design implemented should be used. All recommended good practice estimators meet this condition, but the versions of the estimators presented are usually forms where the individual inclusion probabilities do not appear explicitly.

4.3 Estimating Accuracy

The cell entries of the error matrix and the population parameters derived from it must be estimated from a sample. Suppose the sample-based estimator of p_{ij} is denoted as $\hat{p}_{ij}$. Once $\hat{p}_{ij}$ is available for each element of the error matrix, parameters can be estimated by substituting $\hat{p}_{ij}$ for p_{ij} in the formulas for the parameters. Accordingly, the error matrix should be reported in terms of these estimated area proportions, $\hat{p}_{ij}$, and not in terms of sample counts, n_{ij}. The specific formula for estimating p_{ij} depends on the sampling design used. For equal probability sampling designs (e.g., simple random and systematic sampling) and stratified random sampling in which the strata correspond to the map classes,

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i\cdot}} \qquad (4)$$

where W_i is the proportion of area mapped as class i. For simple random and systematic sampling, Eq. (4) is a poststratified estimator of p_{ij} (Card, 1982) and for these sampling designs the poststratified estimator is recommended because it will have better precision than the estimators commonly used (cf. Stehman and Foody, 2009). Substituting $\hat{p}_{ij}$ of Eq. (4) into Eqns. 1-3 yields estimators of overall, user's, and producer's accuracies. These formulas are simpler special cases of a more general estimation approach described in Strahler et al. (2006, Eqn. 3.1).

The sampling variability associated with the accuracy estimates should be quantified by reporting standard errors. The variance estimators are provided below, and taking the square root of the estimated variance results in the standard error of the estimator. For overall accuracy, the estimated variance is

$$\hat{V}(\hat{O}) = \sum_{i=1}^{q} W_i^2 \hat{U}_i (1 - \hat{U}_i)/(n_{i\cdot} - 1) \quad (5)$$

For user's accuracy of map class i, the estimated variance is

$$\hat{V}(\hat{U}_i) = \hat{U}_i (1 - \hat{U}_i)/(n_{i\cdot} - 1) \quad (6)$$

For producer's accuracy of reference class $j = k$, the estimated variance is

$$\hat{V}(\hat{P}_j) = \frac{1}{\hat{N}_{\cdot j}^2} \left[\frac{N_{j\cdot}^2 (1-\hat{P}_j)^2 \hat{U}_j (1-\hat{U}_j)}{n_{j\cdot}-1} + \hat{P}_j^2 \sum_{i \neq j}^{q} N_{i\cdot}^2 \frac{n_{ij}}{n_{i\cdot}} \left(1 - \frac{n_{ij}}{n_{i\cdot}}\right) / (n_{i\cdot} - 1) \right] \quad (7)$$

where $\hat{N}_{\cdot j} = \sum_{i=1}^{q} \frac{N_{i\cdot}}{n_{i\cdot}} n_{ij}$ is the estimated marginal total number of pixels of reference class j, $N_{j\cdot}$ is the marginal total of map class j and $n_{j\cdot}$ is the total number of sample units in map class j. These are the usual variance estimators applied to the stratified sampling, and the estimators would be viewed as poststratified variance estimators for simple random and systematic sampling. For systematic sampling, the variance estimators are approximations that usually result in overestimation of variance. These variance estimators are also based on assumptions that the assessment unit for the response design is a pixel and each pixel has a hard classification for the map and a hard classification for the reference data. The variance estimators would not apply to a polygon assessment unit or to a mixed pixel situation.

4.4 Estimating Area

The error matrix also provides the basis for estimating the area of classes such as those representing change and no-change. ~~Indeed, t~~The population error matrix (Table 4) provides two different approaches for estimating the proportion of area. Suppose we are interested in estimating the proportion of area of class *k*. The row and column totals are the sums of the p_{ij} values in the respective rows and columns. Thus, the row total $p_{k\cdot}$ represents the proportion of area mapped as class *k* (e.g., if *k* is a change class such as forest loss then $p_{k\cdot}$ is the proportion of area mapped as forest loss) and the column total $p_{\cdot k}$ represents the proportion of area of class *k* as determined from the reference classification (e.g., $p_{\cdot k}$ would be the proportion of area of forest loss as determined from the reference classification).

The two area proportion parameters for class *k* (i.e., $p_{k\cdot}$ and $p_{\cdot k}$) are unlikely to have the same value, so a decision arises as to which parameter should be the focus. Once a change map is complete, $p_{k\cdot}$ is known, but because the reference classification is available only for a sample, $p_{\cdot k}$ must be estimated from the sample. Consequently, the need to estimate $p_{\cdot k}$ introduces uncertainty in the form of sampling variability, whereas $p_{k\cdot}$ is not subject to sampling variability (Stehman, 2005).The map-based parameter $p_{k\cdot}$ is known with certainty but likely biased because of classification error. Conversely, $p_{\cdot k}$ is determined from the reference classification.~~, and,~~

~~t~~Therefore, $p_{\cdot k}$ should have smaller bias than $p_{k\cdot}$ (i.e., the bias attributable to reference data error is smaller than the bias attributable to map classification error). The "good practice" guidelines are founded on the premise that the reference classification is of superior quality to the map classification and that the sampling design implemented yields estimates with small standard errors. Consequently, we recommend that area estimation should be based on $p_{\cdot k}$, the proportion of area derived from the reference classification.

A variety of estimators has been proposed for estimating $p_{\cdot k}$ from the error matrix. For any sampling design and response design leading to an estimated error matrix with p_{ij} in terms of proportion of area, a direct estimator of the proportion of area of class *k* is

$$\hat{p}_{\cdot k} = \sum_{i=1}^{q} \hat{p}_{ik} \qquad (8)$$

This estimator is simply the sum of the estimated area proportions of class *k* as determined from the reference classification (i.e., the sum of column *k* of the estimated error matrix). If the sampling design is simple random, systematic, or stratified random with the map classes defined as the strata, Eq. (8) would be computed using $\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i\cdot}}$ leading to the often used special case estimator

$$\hat{p}_{\cdot k} = \sum_{i=1}^{q} W_i \frac{n_{ik}}{n_{i\cdot}} \qquad (9)$$

This estimator is a poststratified estimator for simple random and systematic sampling, and it is the direct stratified estimator of $p_{\cdot k}$ for stratified random sampling when the map classes are the strata. For these sampling designs, the stratified estimator (Eq. 9) generally has better precision than a variety of alternative estimators of area (Stehman, 2013) and consequently the stratified estimator is recommended.

For the stratified estimator of proportion of area (Eq. 9), the standard error is estimated by

$$S(\hat{p}_{\cdot k}) = \sqrt{\sum_i W_i^2 \frac{\frac{n_{ik}}{n_{i\cdot}}\left(1-\frac{n_{ik}}{n_{i\cdot}}\right)}{n_{i\cdot}-1}} = \sqrt{\sum_i \frac{W_i \hat{p}_{ik} - \hat{p}_{ik}^2}{n_{i\cdot}-1}} \qquad (10)$$

where n_{ik} is the sample count at cell (i,k) in the error matrix, W_i is the area proportion of map class i, and the summation is over the q classes. For systematic sampling, Eq. (10) is an approximation that is typically an overestimate for the actual standard error of systematic sampling. The estimated area of class k is $\hat{A}_k = A \times \hat{p}_{\cdot k}$, where A is the total map area. The standard error of the estimated area is given by

$$S(\hat{A}_k) = A \times S(\hat{p}_{\cdot k}) \qquad (11)$$

An approximate 95% confidence interval is obtained as $\hat{A}_k \pm 1.96 \times S(\hat{A}_k)$.

5. Example of Good Practices: Estimating Area and Assessing Accuracy of Forest Change

The following hypothetical example illustrates the workflow of assessing accuracy of a forest change map and estimating area. Consider a change map for 2000 to 2010 consisting of two change classes and two stable classes: deforestation, forest gain, stable forest and stable non-forest. The map was produced by supervised classification of data from Landsat ETM+ with the objective of estimating the gross rates of forest loss and gain. The first step in the assessment was to visually inspect the change map and identify obvious errors by comparing the classified results to the Landsat data of 2000 and 2010. Misclassified regions were relabelled before proceeding to the rigorous evaluation of the map. After obvious errors were removed, the areas of the map classes were 200,000 Landsat pixels (18,000 ha) of deforestation, 150,000 pixels (13,500 ha) of forest gain, 3,200,000 pixels (288,000 ha) of stable forest, and 6,450,000 pixels (580,500 ha) of stable non-forest. The two change classes thus occupy 3.5% of the total map area. The accuracy assessment was designed for the objectives of estimating overall and class-specific accuracies, areas of the individual classes (as determined by the reference classification), and confidence intervals for each accuracy and area parameter. The spatial assessment unit in this example is a Landsat pixel (30 m × 30 m).

5.1 Sampling Design

A stratified random sampling design with the four map classes as strata adheres to the recommended practices outlined in Section 2.3 and satisfies the accuracy assessment and area estimation objectives. In the next two subsections, we present sample size and sample allocation planning calculations for the stratified design. Sample size planning is an inexact science because it is dependent on ~~information on~~ accuracy and area information that must be speculative prior to conducting the actual accuracy assessment. Nevertheless, these planning calculations can provide informative insight into the choices of sample size and sample allocation to strata.

5.1.1 Determining the Sample Size

For simple random sampling and targeting overall accuracy as the estimation objective, Cochran (1977, Eq. 4.2) suggests using a sample size of

$$n = \frac{z^2 O(1-O)}{d^2} \quad (12)$$

where O is the overall accuracy expressed as a proportion, z is a percentile from the standard normal distribution ($z = 1.96$ for a 95% confidence interval, $z = 1.645$ for a 90% confidence interval), and d is the desired half-width of the confidence interval of O. Eq. (12) provides a starting point for assessing sample size for the limited scope of estimating overall accuracy.

For stratified random sampling, Cochran (1977, Eq. 5.25) provides the following sample size formula (the cost of sampling each stratum is assumed the same):

$$n = \frac{(\sum W_i S_i)^2}{[S(\hat{O})]^2 + (1/N)\sum W_i S_i^2} \approx \left(\frac{\sum W_i S_i}{S(\hat{O})}\right)^2 \quad (13)$$

where N = number of units in the ROI, $S(\hat{O})$ is the standard error of the estimated overall accuracy that we would like to achieve, W_i is the mapped proportion of area of class i, and S_i is the standard deviation of stratum i, $S_i = \sqrt{U_i(1-U_i)}$ (Cochran, 1977, Eq. 5.55). Because N is typically large (e.g., over 10 million pixels in this example), the second term in the denominator of Eq. (13) can be ignored. We specify a target standard error for overall accuracy of 0.01. Suppose from past experience with similar change mapping efforts we know that errors of commission are relatively common for the change classes while the stable classes are more accurate (e.g., Olofsson et al., 2010; 2011). Consequently, we conjecture that user's accuracies of the two change classes will be 0.70 for deforestation and 0.60 for forest gain, and user's accuracies of the stable classes will be 0.90 for stable forest and 0.95 for stable non-forest. The resulting sample size from Eq. (13) is n = 641. These sample size calculations should be repeated for a variety of choices of $S(\hat{O})$ and U_i before reaching a final decision.

5.1.2. Determine Sample Allocation to Strata

Once ~~we have chosen~~ the overall sample size is chosen, we determine the allocation of the sample to strata ~~needs to be determined~~. It is important that the sample size allocation results in precise estimates of accuracy and area. Stehman (2012) identifies four different approaches to sample allocation: proportional, equal, optimal and power allocation. In proportional allocation, the sample size per map class is proportional to the relative area of the map class. In this example, and which is usually the case when mapping land change, the mapped areas of change are small relative to other classes so proportional allocation will lead to small sample sizes in the rare classes (unless n is very large) and imprecise estimates of user's accuracy for these rare classes. Allocating an equal sample size to all strata targets estimation of user's accuracy of each map class but equal allocation is not optimized for estimating area and overall accuracy. Neyman optimal allocation (Cochran, 1977) can be used to minimize the variance of the estimator of overall accuracy or the estimator of area, but optimal allocation becomes difficult to implement when multiple estimation objectives are of interest as will be the case when estimating accuracy and area of several land-cover classes or land-cover change types.

We suggest the following simplified approach to sample size allocation. Allocate a sample size of 50-100 for each change strata using the variance estimator for user's accuracy (Eq. 6) to decide the sample size needed to achieve certain standard errors for the assumed estimated user's accuracy for that class. ~~The sample size allocated to these rare class strata will also be affected by the total sample size, *n*, available to allocate.~~ A small overall sample size~~*n*~~ might allow for only 50 sample units per rare class stratum. Suppose that n-r sample units remain after a sample size of *r* units has been allocated to the rare class strata. The sample size of *n*-*r* is then allocated proportionally to the area of each remaining stratum. The anticipated estimated variances can then be computed (based on the sample size allocation) for user's and overall accuracy and area using Eqs. (5), (6) and (10). The sample size allocation process can be iterated until an allocation is found that yields satisfactory anticipated standard errors for the key accuracy and area estimates. The effect of the choice of sample allocation will be observed in the standard errors of the estimates, however, a poor allocation of sample size to strata will not result in biased estimators.

In this example, we know the mapped areas of the four map classes (W_i), we have conjectured values of user's accuracies and standard errors of the strata, and we have estimated a total sample size of 641 (Table 5). The resulting sample sizes for proportional and equal allocation are shown in Table 5. As described above, neither of these is optimal and we want to find a compromise between the two. We start by allocating 100 sample units each to the change classes and then allocate the remainder of the sample size proportionally to the stable classes. This gives the allocation in column "Alloc1". Since the recommendation is to allocate between 50 and 100 sample units in the change strata, we introduce two additional allocations with 75 and 50 sample units in the change strata, respectively ("Alloc2" and "Alloc3"). To determine which of these allocations to use, we need to examine the standard errors of the estimated user's accuracy, estimated overall accuracy, and estimated areas using Eq. (5), (6) and (10).

<< TABLE 5 HERE >>

It is necessary to speculate the outcome of the accuracy assessment to compute the anticipated standard errors for each sample allocation considered. The hypothesized error matrix in Table 6 reflects the anticipated outcome that the change classes will be rare and have lower class-specific accuracies than the two stable classes. The population error matrix was also constructed to yield the hypothesized accuracies input into the sample size planning calculations of the previous section. When creating the hypothesized error matrix used for sample size and sample allocation planning, we should draw upon any past experience for insight into the accuracy of the map to be produced.

<< TABLE 6 HERE >>

Table 7 shows the standard errors of the user's and overall accuracies and estimated areas of both deforestation and stable forest for each of the five sample allocations in Table 5 and the hypothetical population error matrix of Table 6. No single allocation is best for all estimation objectives, so a choice among competing objectives is necessary. The emphasis on prioritizing objectives during the planning stage (Section 2) becomes particularly relevant to the decision of sample allocation because different allocations favour different estimation objectives. For example, equal allocation gives the smallest standard error of the user's accuracy of deforestation but a high standard error of the estimated area of deforestation. Proportional allocation will result in smaller standard errors of overall accuracy and area of stable forest but the standard error for estimated user's accuracy of deforestation is two to four times larger than the corresponding standard errors for other sample allocations. In this case, "Alloc1-3" provide allocations that generate relatively small standard errors for the different estimates. We will choose “Alloc2” with 75 sample units in the two change classes.

<< TABLE 7 HERE >>

5.2 Estimating Accuracy, Area and Confidence Intervals

To create the reference classification for labelling each sample unit, a combination of Landsat data from the USGS open archive together with GoogleEarth™ provides a source of cost free reference data. Our hypothetical map was produced using Landsat, and the good practice recommendations stipulate that if using the same data for creation of both the map and reference classifications, the process of creating the latter should be of higher quality than the map-making process. The process of labelling the sample units thus has to be more accurate than supervised classification. A manual inspection by three analysts of each of the sample units using a set of Landsat images together with GoogleEarth™ imagery acquired around the same time as the images used to make the map is assumed to be a more accurate process than supervised classification. ~~Suppose t~~The error matrix resulting from this response design and sample is presented in terms of the sample counts displayed in Table 8, and the computations for the accuracy and area estimates are detailed in the following two subsections.

<< TABLE 8 HERE >>

5.2.1. Estimating Accuracy

Because the sampling design is stratified random using the map classes as strata, the cell entries of the error matrix are estimated using Eq. (4).

<< TABLE 9 HERE >>

We can now estimate user's accuracy $\widehat{U}_i = \frac{\hat{p}_{ii}}{\hat{p}_{i\cdot}}$; producer's accuracy $\widehat{P}_j = \frac{\hat{p}_{jj}}{\hat{p}_{\cdot j}}$; and overall accuracy $\widehat{O} = \sum_{j=1}^{q} \hat{p}_{jj}$ using the estimated area proportions. Variances for these accuracy measures are estimated using Eq. (5)-(7). 95% confidence intervals are estimated as $\pm 1.96\sqrt{\widehat{V}(\widehat{U}_i)}$ (replace $\widehat{U}_i$ with $\widehat{P}_j$ and $\widehat{O}$ for the producer's and overall accuracies). In this case, the estimated user's accuracy (± 95% confidence interval) is 0.88 ± 0.07 for deforestation, 0.73 ± 0.10 for forest gain, 0.93 ± 0.04 for stable forest, and 0.96 ± 0.02 for stable non-forest. The estimated producer's accuracy is 0.75 ± 0.21 for deforestation, 0.85 ± 0.23 for forest gain, 0.93 ± 0.03 for stable forest, and 0.96 ± 0.01 for stable non-forest. The estimated overall accuracy is 0.95 ± 0.02.

5.2.2. Estimating Area and Uncertainty

The next step is to use the estimated area proportions in Table 9 to estimate the area of each class. The row totals of the error matrix in Table 9 give the mapped area proportions (which are also given by W_i) while the column totals give the estimated area proportions according to the reference data. Multiplying the latter by the total map area gives the stratified area estimate of each class according to the reference data. For example, the estimated area of deforestation according to the reference data is $\hat{A}_1 = \hat{p}_{\cdot 1} \times A_{tot} = 0.024 \times 10{,}000{,}000$ pixels = 235,086 pixels = 21,158 ha. The mapped area of deforestation ($A_{m,1}$) of 200,000 pixels was thus underestimated by 35,086 pixels or 3,158 ha.

The second step is to estimate a confidence interval for the area of each class. From Eq. (10), $S(\hat{p}_{\cdot 1}) = 0.0035$ and the standard error for the estimated area of forest loss is $S(\hat{A}_1) = S(\hat{p}_{\cdot 1}) \times A_{tot} = 0.0035 \times 10{,}000{,}000 = 34{,}097$ pixels. The margin of error of the confidence interval is $1.96 \times 34{,}097 = 68{,}418$ pixels = 6,158 ha. We have thus estimated the area of deforestation with a 95% confidence interval: $21{,}158 \pm 6{,}158$ ha. The area estimate with a 95% confidence interval of the forest gain class is 11,686 ± 3,756 ha; stable forest is 285,770 ± 15,510 ha and stable non-forest 581,386 ± 16,282 ha.

This example has illustrated the workflow of assessing accuracy, and estimating area and confidence intervals of area of the classes of a change map. While this is fairly straightforward once the error matrix has been constructed, the example highlights the need to consider different objectives when designing the sample.

A tool for estimating unbiased accuracy measures and areas with 95% confidence intervals can be downloaded from www.people.bu.edu/olofsson/ (click ‘Research’ > ‘Accuracy/Uncertainty’). The tool is implemented in Matlab™.

6. Summary

Conducting an accuracy assessment of a land change map serves multiple purposes. In addition to the obvious purpose of quantifying the accuracy of the map, the reference sample serves as the basis of estimates of area of each class where area is defined by the reference classification.~~, and~~ ~~t~~ The accuracy assessment sample data also contribute to estimates of uncertainty of the area estimates. Without an accuracy assessment, there is no way to communicate map quality in a quantitative and meaningful fashion. We acknowledge that there is no singular “best” approach and the recommendations provided do not preclude the existence of other acceptable practices.

However, by following the “good practice” recommendations presented by this paper, scientific credibility of the accuracy and area estimates is ensured. The “good practice” recommendations are summarized as follows, organized by the three major components of the accuracy assessment methodology, the sampling design, response design, and analysis:

6.1 General

- Visually inspect the map and correct obvious errors before conducting the accuracy assessment
- Accuracy and area estimates will be determined from a classification (i.e., the reference classification) that is of higher quality than the land change map being evaluated
- A sampling approach is needed because the cost of obtaining the reference classification for the entire region of interest will be prohibitive
- The sample used for accuracy assessment and area estimation is separate from (independent of) the data used to train or develop the classification

6.2 Sampling design

- Implement a probability sampling design to provide a rigorous foundation via design-based sampling inference
- Document and quantify any deviations from the probability sampling protocol
- Choose a sampling design on the basis of specified accuracy objectives and prioritized desirable design criteria
- Sampling design guidelines
    - Stratify by map class to reduce standard errors of class-specific accuracy estimates

    - If resources are adequate, stratify by subregions to reduce standard errors of subregion-specific estimates
    - Use cluster sampling if it provides a substantial cost savings or if the objectives require a cluster unit for the assessment
    - Both simple random and systemic selection protocols are acceptable options
- The recommended allocation of sample size to strata (assuming the map classes are the strata) is to increase the sample size for rare change classes to achieve an acceptable standard error for estimated user's accuracies and to allocate the remaining sample size roughly proportional to the area occupied by the common classes
- Use sample size and optimal allocation planning calculations as a guide to decisions on total sample size and sample allocation
- Evaluate the potential outcome of sample size and sample allocation decisions on the standard errors of accuracy and area estimates for hypothetical error matrices based on the anticipated accuracy of the map
- Stratified random sampling using the map classification to define strata is a simple, but generally applicable design that will typically satisfy most accuracy and area estimation objectives and desirable design criteria

6.3 Response design

- Reference data should be of higher quality than the data used for creating the map, or if using the same source, the process of creating the reference classification should be more accurate than the process of creating the map
- High overhead cost may eliminate field visits as a source of reference data

- The reference data should provide sufficient temporal representation consistent with the change period of the map
- Data from the Landsat open archive in combination with high spatial resolution imagery provide a low-cost and often useful source of reference data (national photograph archives, satellite photo archives (e.g., Kompsat), and the collections available through Google Earth™ are possible high resolution imagery sources)
- Specify protocols for accounting for uncertainty in assigning the reference classifications
- Assign each sample unit a primary and secondary label (secondary not required if there is highly confidence~~t~~ in the primary label)
- Include an interpreter specified confidence for each reference label (e.g., high, medium, or low confidence)
- Implement protocols to ensure consistency among individual interpreters or teams of interpreters
- Specify a protocol for defining agreement between the map and reference classifications that will lead to an error matrix expressed in terms of proportion of area

6.4 Analysis

- Report the error matrix in terms of estimated area proportions
- Report the area (or proportion of area) of each class as determined from the map
- Report user’s accuracy (or commission error), producer’s accuracy (or omission error), and overall accuracy (Equations 1-3)
- Avoid use of the kappa coefficient of agreement for reporting accuracy of land change maps

- Estimate the area of each class according to the classification determined from the reference data
- Use estimators of accuracy and area that are unbiased or consistent
- For simple random, systematic, and stratified random sampling when the map classes are defined as strata, use stratified estimators of accuracy (Eqs. 5-7) and a stratified estimator of area (Eq. 9)
- Quantify sampling variability of the accuracy and area estimates by reporting standard errors or confidence intervals
- Use design-based inference to define estimator properties and to quantify uncertainty
- Assess the impact of reference data uncertainty on the accuracy and area estimates

The recommendations provided are intended to serve as guidelines for choosing from among options of sampling design, response design, and analysis that will yield rigorous and defensible accuracy and area estimates. But good practice is not static. As improvements in technology become available and new methods are developed, good practice recommendations will evolve over time. Also, as practical experience accumulates with using new technology and methodologies~~y~~, good practice recommendations will be further amended to provide even more efficient yet still rigorous methods to estimate accuracy and area of land change.

References

Achard, F., Eva, H., Stibig, H.-J., Mayaux, P., Gallego, J., Richards, T., and Malingreau, J.-P. (2002). Determination of deforestation rates of the world's humid tropical forests. *Science*, 297, 999-1002.

Ahlqvist, O. (2008). In search of classification that supports the dynamics of science: The FAO Land Cover Classification System and proposed modifications. *Environment and Planning B: Planning and Design*, 35, 169-186.

Baker, B. A., Warner, T. A., Conley, J. F., and McNeil, B. E. (2013). Does spatial resolution matter? A multi-scale comparison of object-based and pixel-based methods for detecting change associated with gas well drilling operations. *International Journal of Remote Sensing*, 34, 1633-1651.

Binaghi, E., Brivio, P. A., Ghezzi, P., and Rampini, A. (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters*, 20, 935-948.

Cakir, H. I., Khorram, S., and Nelson, S. A. C. (2006). Correspondence analysis for detecting land cover change. *Remote Sensing of Environment*, 102, 306–317.

Card, D. H. (1982). Using map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 49, 431–439

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons

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. J., Wadsworth, R. A., and Fisher, P. F. (2008). Using semantics to clarify the conceptual confusion between land cover and land use: the example of ‘forest’. *Journal of Land Use Science*, 3, 185-198.

Congalton, R., and Green, K. (2009). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (2nd ed.). Boca Raton: CRC/Taylor & Francis

DeFries, R., Houghton, R. A., Hansen, M., Field, C., Skole, D. L., and Townshend, J. (2002). Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 90s. *Proceedings of the National Academy of Sciences*, 99, 14256-14261.

DeFries, R., Achard, F., Brown, S., Herold, M., Murdiyarso, D., Schlamadinger, B., and Souza, C. M. (2007). Earth observations for estimating greenhouse gas emissions from deforestation in developing countries. *Environmental Science and Policy*, 10, 385-394.

de Sy, V., Herold, M., Achard, F., Asner, G. P., Held, A., Kellndorfer, J., and Verbesselt, J. (2012). Synergies of multiple remote sensing data sources for REDD+ monitoring. *Current Opinion in Environmental Sustainability*, 4, 696–706.

Drummond, M. A., and Loveland T. R. (2010). Land-use pressure and a transition to forest-cover loss in the eastern United States. *BioScience*, 60, 286-298.

Duro, D. C., Franklin, S. E., and Duba, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118:259-272.

Falkowski, M. J., Wulder, M. A., White, J. C., and Gillis, M. D. (2009). Supporting large-area, sample-based forest inventories with very high spatial resolution satellite imagery. *Progress in Physical Geography*, 33, 403-423.

FAO (2010). Global Forest Resources Assessment 2010. Food and Agriculture Organization of the United Nations.

FAO (2011). Food and Agriculture Organization of the United Nations. Assessing forest degradation. Towards the development of globally applicable guidelines. Forest Resources Assessment Working Paper 177.

Foody, G. M. (1992). On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, 58, 1459-1460

Foody, G. M., Campbell, N. A., Trodd, N. M., and Wood, T. F. (1992). Derivation and applications of probabilistic measures of class membership from the maximum likelihood classification. *Photogrammetric Engineering and Remote Sensing*, 58, 1335-1341.

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

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 Observation and Remote Sensing*, in press. DOI: 10.1109/JSTARS.2013.2250257

~~Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., and Boyd, D. S. (2013). Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS*, in press.~~

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*, in press.

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.

GOFC-GOLD (2011). A sourcebook of methods and procedures for monitoring and reporting anthropogenic greenhouse gas emissions and removals caused by deforestation, gains and losses of carbon stocks in forests remaining forests, and forestation. GOFC-GOLD Report version COP17-1, (GOFC-GOLD Project Office, Natural Resources Canada, Alberta, Canada).

Gómez, C., White, J. C., and Wulder, M. A. (2011). Characterizing the state and processes of change in a dynamic forest environment using hierarchical spatio-temporal segmentation. *Remote Sensing of Environment*, 115, 1665-1679.

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

Grassi, G., Monni, S., Federici, S., Achard, F., and Mollicone, D. (2008) Applying the conservativeness principle to REDD to deal with the uncertainties of the estimates. *Environmental Research Letters*, 3, 3.

Hansen, M. C., Stehman, S. V., and Potapov, P. V. (2010). Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences*, 107, 8650-8655.

He, Y. H., Franklin, S. E., Guo, X. L., and Stenhouse, G. B. (2011). Object-orientated classification of multi-resolution images for the extraction of narrow linear forest disturbance. *Remote Sensing Letters*, 2, 147-155.

Herold, M., and Skutsch, M. (2011). Monitoring, reporting and verification for national REDD + programmes: two proposals. *Environmental Research Letters* 6 014002.

Herold, M., Román-Cuesta, R.M., Mollicone, D., Hirata, Y., Van Laake, P., Asner, G.P., Souza, C., Skutsch, M., Avitabile, V., and Macdicken, K. (2011). Options for monitoring and estimating historical carbon emissions from forest degradation in the context of REDD+. *Carbon balance and management*, 6, 13

Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., and Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, 114, 183–198.

~~Hyyppä, J., Hyyppa, H., Inkinen, M., Engdahl, M., Linko, S., and Zhu, Y.-H. (2000). Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, 128, 109-120.~~

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

Jeon, S. B., Olofsson, P., and Woodcock, C. E. (2013). Land use change in New England: a reversal of the forest transition. *Journal of Land Use Science* DOI: 10.1080/1747423X.2012.754962

Johnson, B. A. (2013). High-resolution urban land-cover classification using a competitive multi-scale object-based approach. *Remote Sensing Letters*, 4, 131-140.

Kelly, M., Estes, J. E., and Knight, K. A. (1999). Image interpretation keys for validation of global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, 65, 1041-1050.

Kennedy, R., Yang, Z., and Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr – Temporal segmentation algorithms. *Remote Sensing of Environment*, 114, 2897-2910.

Knight, J. F., and Lunetta R. S. (2003). An experimental assessment of minimum mapping unit size. *IEEE Transactions on Geoscience and Remote Sensing*, 40, 2132-2134.

Kurz, W. A. (2010). An ecosystem context for global gross forest cover loss estimates. *Proceedings of the National Academy of Science*, 107, 9025-9026.

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.

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

Lindberg, E., Olofsson, K., Holmgren, J., and Olsson, H. (2012). Estimation of 3D vegetation structure from waveform and discrete return airborne laser scanning data. *Remote Sensing of Environment*, 118, 151-161.

Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., Naumov, S., De Miranda, E. E., Di Bella, C. M., Ordoyne, C., Kopin, Y., and Roy, P. S. (2006). Validation of the Global Land Cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, 44, 1728-1739.

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

Olofsson, P., Torchinava, P., Woodcock, C. E., Baccini, A., Houghton, R. A., Ozdogan, M., Zhao, F., and Yang, X. (2010). Implications of Land Use Change on the National Terrestrial Carbon Budget of Georgia. *Carbon Balance and Management*, 5, 4.

Olofsson, P., Kuemmerle, T., Griffiths, P., Knorn, J., Baccini, A., Gancz, V., Blujdea, V., Houghton, R.A., Abrudan, I.V., and Woodcock C.E. (2011). Carbon implications of forest restitution in post-socialist Romania, *Environmental Research Letters,* 6, 045202.

Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., Newell, J. D., Friedl, M. A., and Herold, M. (2012). A global land cover validation dataset, 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

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

Pontius, R. G., and Lippitt, C. D. (2006). Can Error Explain Map Differences Over Time? *Cartography and Geographic Information Science*, 33, 159-171.

Pontius, R. G., 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.

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

Pratihast, A. K., Herold, M., de Sy, V., Murdiyarso, D., and Skutsch, M. (2013). Linking community-based and national REDD+ monitoring: a review of the potential. *Carbon Management*, 4, 91–104

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.

Romijn, J. E., Herold, M., Kooistra, L., Murdiyarso, D., and Verchot, L. (2012). Assessing capacities of non-Annex I countries for national forest monitoring in the context of REDD+. *Environmental Science and Policy*, 20, 33-48.

Sanz-Sanchez, M., Herold, M., and Penman, J. (2013). REDD+ related forest monitoring remains key issue: a report following the recent UN climate convention in Doha. *Carbon Management*, 4, 125-127.

Särndal, C., Swensson, B., and Wretman, J. (1992). Model assisted survey sampling. New York: Springer.

Saura, S. (2002). Effects of minimum mapping unit on land cover data spatial configuration and composition. *International Journal of Remote Sensing*, 23, 4853-4880.

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

Schroeder, T. A., Wulder, M. A., Healey, S. P., and Moisen, G. G. (2011). Mapping wildfire and clearcut harvest disturbances in boreal forests with Landsat time series data. *Remote Sensing of Environment*, 115, 1421-1433.

Skirvin, S. M., Kepner, W. G., Marsh, S. E., Drake, S. E., and Maingi, J. K., Edmonds, C. M., Watts, C.J., and Williams D. R. (2004). Assessing the accuracy of satellite-derived land-cover classification using historical aerial photography, digital orthophoto quadrangles, and airborne video data. In R. S. Lunetta and J. G. Lyon (Eds.), *Remote Sensing and GIS Accuracy Assessment*. Boca Raton: CRC Press.

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

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. *Photogrammetric Engineering and Remote Sensing*, 67, 727–734.

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

Stehman, S. V. (2012). Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sensing Letters*, 3, 111-120.

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

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., and Foody, G.M. (2009). Accuracy Assessment. In T. A. Warner, M. D. Nellis, and G. M. Foody (Eds.) *The SAGE Handbook of Remote Sensing*. London: Sage Publications.

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., Sohl, T. L., and Loveland, T. R. (2003). Statistical sampling to characterize recent United States land-cover change. *Remote Sensing of Environment*, 86, 517-529.

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

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V., and Woodcock, C. E. (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.

Tomppo, E. O., Gschwantner, T., Lawrence, M., and McRoberts, R. E. (2010). *National Forest Inventories: Pathways for Common Reporting*, Springer, New York

UN-REDD (2008). UN Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD). FAO, UNDP, UNEP Framework Document.

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

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., and Wynne, R. (2008). Free access to Landsat imagery. *Science*, 320, 1011

Wulder, M. A., Franklin, S., 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, L. 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.

Wulder, M. A., White, J. C., Magnussen, S., and McDonald, S. (2007). Validation of a large area land cover product using purpose-acquired airborne video. *Remote Sensing of Environment*, 106, 480-491.

Wulder, M. A., White, J. C., Hay, G. J., and Castilla, G. (2008a). Towards automated segmentation of forest inventory polygons on high spatial resolution satellite imagery. *The Forestry Chronicle*, 84, 221-230.

Wulder, M. A.; White, J. C.; Coops, N. C., and Butson, C. R. (2008b). Multi-temporal analysis of high spatial resolution imagery for disturbance monitoring. *Remote Sensing of Environment*. 112, 2729-2740.

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., and Woodcock, C.E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, 122, 2-10.

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

Table 1. Possible reference data sources

Reference data source	*Exemplar citation*
Field plots	Hyyppä et al. 2000
Air photography	Skirvin et al. (2004)
Forest inventory data	McRoberts (2011); Wulder et al. (2006b)
Airborne video	Wulder et al. (2007)
Lidar	Lindberg et al. (2012)
Satellite imagery	Scepan (1999); Cohen et al. (2010)
Crowdsourcing	Iwao et al. (2006); Foody and Boyd (2013)

Table 2. Example characteristics to record for each change polygon. Some attributes can be generated in the GIS; others will need to be entered by the analyst. Notion is that information is captured and carried to provide insights and a record regarding the changes captured. The aim is that the change polygons can be used in a manner that is invariant to source, but that metadata is captured to explain or better understand any data related anomalies that may emerge.

Attribute	*Definition / comments.*
Change Area	Area changed, e.g., polygon size in hectares
Change Perimeter	Perimeter of polygon, in meters
Change Type	Notation of change type, harvest, fire, insect, urban expansion, agricultural development
Change Date	As possible, note the change date. May be available from other records, e.g., when a fire occurred, or the acquisition date of the image or photography used.
Data Source	Note the data source from which the change polygon is made
Analyst	Name or code to denote the interpreter
Date Interpreted	Note the date when the interpretation occurred

Table 3. Elements for consideration when selecting reference data

Element	*Considerations*
Cost	What is the budget? What amount per unit of reference data can be purchased? Is the interpretation / labelling protocol efficient?
Ease of access	Varies by data type. Can field visits be made? Is archival image data available?
Ease of use	Is the data produced in a consistent fashion? Is it in formats that are commonly used?
Opportunity for consistency	Can protocols be developed and applied in a systematic and repetitive fashion? Can some tasks be automated?
Vintage – temporal representation	Is the data representative of a time or time period that is relevant to the change product under consideration?
Spatial coverage	Are there opportunities for multiple reference sites from a given reference data source?
Interpretability of change types	Does the data source capture and portray the change types of interest? E.g., is the spatial resolution sufficiently fine to enable interpretation?
Geolocation	Can the candidate reference data source be assumed to be accurately positioned? Will additional geolocation activities be required?

Table 4. Population error matrix of four classes with cell entries (p_{ij}) expressed in terms of proportion of area as suggested by good practice recommendations.

		Reference				
		Class 1	Class 2	Class 3	Class 4	Total
Map	Class 1	p_{11}	p_{12}	p_{13}	p_{14}	$p_{1\cdot}$
	Class 2	p_{21}	p_{22}	p_{23}	p_{24}	$p_{2\cdot}$
	Class 3	p_{31}	p_{32}	p_{32}	p_{34}	$p_{3\cdot}$
	Class 4	p_{41}	p_{42}	p_{43}	p_{44}	$p_{4\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	$p_{\cdot 4}$	1

Table 5. Information needed to decide allocation of sample size to strata. The information includes the mapped area proportions (W_i), conjectured values of user's accuracies (U_i) and standard deviations (S_i) of the strata. Columns 5-9 contain five different allocations.

Strata (i)	W_i	U_i	S_i	Equal	Alloc1	Alloc2	Alloc3	Prop
1 Deforestation	0.020	0.700	0.458	160	100	75	50	13
2 Forest gain	0.015	0.600	0.490	160	100	75	50	10
3 Stable forest	0.320	0.900	0.300	160	149	165	182	205
4 Stable non-forest	0.645	0.950	0.218	160	292	325	358	413

Table 6. Hypothetical population error matrix expressed in terms of proportion of area (see Section 4) used for sample size and sample allocation planning calculations.

		Reference					
	-	*Defore-Station*	*Forest gain*	*Stable forest*	*Stable non-forest*	Total (W_i)	U_i
Map	*Deforestation*	0.014	0	0.003	0.003	0.020	0.70
	Forest gain	0	0.009	0.003	0.003	0.015	0.60
	Stable forest	0.002	0	0.288	0.030	0.320	0.90
	Stable non-forest	0.004	0.002	0.025	0.614	0.645	0.95
	Total	0.020	0.011	0.319	0.650	1	

Table 7. Standard errors of selected accuracy and area estimates for different sample size allocations to strata (Table 5) and the hypothetical population error matrix (Table 6). Standard errors are shown for estimated overall accuracy, estimated user's accuracy for the rare class deforestation ($i = 1$) and the common class stable forest ($i = 3$), and estimated area (in units of hectares) of deforestation and area of stable forest.

Allocation	$S(\hat{O})$	$S(\hat{U}_1)$	$S(\hat{U}_3)$	$S(\hat{A}_1)$	$S(\hat{A}_3)$
Equal	0.013	0.036	0.024	4035	11,306
Alloc1	0.011	0.046	0.025	3307	9,744
Alloc2	0.011	0.053	0.023	3138	9,270
Alloc3	0.010	0.065	0.022	3125	8,860
Proportional	0.010	0.132	0.021	3600	8,614

Table 8. Description of sample data as an error matrix of sample counts, n_{ij} (see Table 9 for recommended estimated error matrix used to report accuracy results).

		Reference						
	-	*Defore-station*	*Forest gain*	*Stable forest*	*Stable non-forest*	Total	$A_{m,i}$ [pixels]	W_i
Map	*Deforestation*	66	0	5	4	75	200,000	0.020
	Forest gain	0	55	8	12	75	150,000	0.015
	Stable forest	1	0	153	11	165	3,200,000	0.320
	Stable non-forest	2	1	9	313	325	6,450,000	0.645
	Total	69	56	175	340	640	10,000,000	1

Table 9. The error matrix in Table 8 populated by estimated proportions of area.

		Reference					
	-	*Defore-station*	*Forest gain*	*Stable forest*	*Stable non-forest*	Total (W_i)	$A_{m,i}$ [pixels]
Map	*Deforestation*	0.0176	0	0.0013	0.0011	0.020	200,000
	Forest gain	0	0.0110	0.0016	0.0024	0.015	150,000
	Stable forest	0.0019	0	0.2967	0.0213	0.320	3,200,000
	Stable non-forest	0.0040	0.0020	0.0179	0.6212	0.645	6,450,000
	Total	0.0235	0.0130	0.3175	0.6460	1	10,000,000

Figure 1