Comprehensive Remote Sensing. Article title: 092 Accuracy and Area Estimation

Pontus Olofsson, April 8, 2025
NASA Marshall Space Flight Center
320 Sparkman Dr NW, Huntsville, AL 35805
pontus.olofsson@nasa.gov

Abstract

The quality, quantity, and availability of remote sensing data and products are greater now than in any other point in time. Free data policies and increasing computing power have made it easy and inexpensive to acquire satellite imagery and make maps. Still, while the need to better understand a changing Earth is greater than ever, if remote sensing is to reach its full potential, greater attention to map quality, bias, and uncertainty using protocols rooted in statistical inference theory is required. In this context, maps are not simply subject to an accuracy assessment where the end-product is a measure of map accuracy but integral components of an inference framework that yields estimates of study area parameters such as areas of thematic categories and area bias and uncertainty. Hence, maps are essential in the protocols that generate estimates of parameters rather than providing such parameters directly. This chapter outlines the development of statistical inference within the remote sensing science and application communities. It reviews various approaches to statistical inference, accuracy assessment, and area estimation.

Contents

1. Introduction

A key strength of remote sensing, and one of the main reasons for its usage, is the provision of spatially exhaustive (wall-to-wall) coverage of a region of interest. Classification and interpretation of remote sensing data allow for mapping of thematic features present in the region. However, mapping complex and often spatially continuous surface conditions into a set of discrete map categories is bound to result in some map units being erroneous. The magnitude of errors will determine the reliability, usage, and interpretation of the map; consequently, map users and producers have a direct interest in communicating and understanding the quality of maps. This is the primary reason for the tradition within the remote sensing community of conducting map accuracy assessments. The basis of an accuracy assessment is the comparison of the map to observations of reference conditions at locations selected by probability sampling. The sample is selected by probability sampling if the locations of the sampling units are selected such that the likelihood of a unit (a pixel for example) being included in the sample is known and greater than zero (Stehman, 2001). A probability sample allows for design-based inference of parameters of the entire study area. For example, consider the following common scenario: a land-cover map has been constructed over specific region and a sample of map units (such as pixels) has been selected by simple random sampling. By identifying land-cover reference conditions at the locations of the units in the sample, the overall accuracy of the map is easily computed as the ratio of correctly classified sample units to the total number of units in the sample. The overall accuracy is a measure of the probability that a random map unit is correctly classified – not just a random map unit in the sample, but of all units of the map. This holds true because the sample was selected by probability sampling. As explained in Section 2 below, accuracy measures specific to the individual map categories could also be easily computed in this scenario.

Accuracy assessments of maps grew in importance during the 1980s when digital imagery and greater processing power became readily available to the larger community, and maps based on manual photointerpretation – often accepted as correct – became rarer (Congalton, 1991). The assessments were largely focused on overall map accuracy and less attention was paid to class-specific accuracy (Congalton, 2004). Overall accuracy provides important information on map quality but does not communicate the accuracy of individual map categories, especially smaller categories. For example, consider mapping forest cover loss in a region with 98% stable forest and 2% forest loss; a “map” where each pixel is classified as forest would have a 98% overall accuracy despite being useless for any realistic application. For these reasons, literature from the 1980s and early 1990s highlighted the need for class-specific accuracy such as user’s and producer’s accuracy (Card, 1982; Congalton, Oderwald, & Mead, 1983; Foody, 1992) – an era that has been referred to as the fourth “age of accuracy assessment” (Congalton, 2004). However, an important message of this book chapter is that while an analysis of class-specific accuracy provides a more comprehensive investigation of the map quality, accuracy measures are not the end of the analysis. If the analysis ends with measures of accuracy – being it overall or class-specific – it merely shows that the map is more or less incorrect (excluding the very unlikely event of achieving 100% accuracy). Any attempt to obtain the area of a specific map category directly from the map will result in an erroneous area, as the effects of classification errors are unaccommodated. Methods for producing area estimates that sum map units assigned to certain map categories are referred to as “pixel-counting” (GFOI, 2016). Measures of class-specific accuracy provide information on the magnitude and nature of map classification errors, but they do not directly adjust for such errors, nor do they produce the information required to estimate areas and uncertainty such as confidence intervals. A situation where errors cancel each other out such that the bias is removed in areas obtained by pixel-counting is possible but highly unlikely and cannot be assumed.

Instead of pixel-counting in maps to estimate areas of land-surface features or activities, it is recommended to use methods (1) that neither over- nor under-estimates and (2) that allows for uncertainty quantification. While this recommendation applies to remote sensing analyses in general, these criteria have been explicitly called out in the guidance for implementing large-scale efforts aimed at reducing deforestation and forest degradation (GFOI, 2016). An analysis based on pixel-counting, even if accompanied by an assessment of class-specific accuracy, does not satisfy these criteria. The first criterion relates to the concept of bias, which is the property of an estimator. Bias is the difference between the expected value of the estimator and the population parameter of interest; the estimator is called unbiased if the difference (i.e. bias) is zero (Casella & Berger, 2002). Therefore, an estimate of the area of an activity (deforestation for example) produced by applying an unbiased estimator to sample data satisfies criterion (1). Criterion (2) relates to uncertainty, which can be quantified in the form of a confidence interval. A confidence interval is a random interval that contains the parameter when calculated from sample data with some specified probability (normally 0.95) (Rice, 1995).

If relying on pixel-counting, the mapped area of the feature of interest might be considerably different from the actual area (Olofsson et al., 2013). Still, at least up until 2010, bias and uncertainty were largely ignored in remote sensing applications: an assessment of all articles related to mapping of land change published in *Remote Sensing of Environment* and *International Journal of Remote Sensing* 2005-2010, showed that all but a few articles failed to include this information (Olofsson, Foody, Stehman, & Woodcock, 2013). While no formal assessment has been performed on articles published after 2010, it is likely that the situation has changed. Several articles, published in remote sensing journals since 2010, has described the need, use and guidance of estimation protocols (e.g. McRoberts, 2011; Olofsson et al., 2014; Stehman, 2013). A quick glance at recently published articles related to land change gives the impression that estimation protocols are now a rather established practice (e.g. Potapov et al., 2015; Richards & Friess, 2016; Curtis et al., 2018; Souza Jr, et al., 2020; Murray et al., 2022). Following the terminology in Congalton (2004), the era after 2010 could be referred to as the fifth age of accuracy assessment.

Even if the efforts related to highlighting the importance of estimation protocols in remote sensing applications have been successful in increasing awareness, previous work recognizing the importance of area estimation needs acknowledgement. As early as 1982, Card

(1982) introduced the stratified estimator – originally published by Cochran (1977) in a social science context – in a remote sensing application for area estimation. Card's article was complete with equations and a numerical example to allow readers to implement the stratified estimator. The review of approaches to area estimation in Gallego (2004) provides an excellent summary of many of the estimation options. It includes a critique of the above-mentioned pixel-counting approach and an overview of estimators combining ground and remote sensing information. Additional important contributions that deserve recognition are Foody (2002), McRoberts, (2010), Stehman, (2000) among many others.

2. Design-based inference

There are two main approaches to inference: design-based and model-based. The difference between the frameworks has been discussed in detail by many authors in a variety of academic disciplines (as cited in Stehman, 2009) and only short descriptions are provided here. Model-based inference assumes that a model is used to make predictions of all units in the population, and that the observation for a population unit is a random variable (instead of a constant value as in a design-based framework). While model-based estimators are advantageous in some aspects – they produce maps as by-products, estimates are consistent with the maps, and importantly, they do not require a probability sample – they are not necessarily unbiased and often computationally intense (McRoberts, 2010). These disadvantageous are likely to explain why design-based inference is more frequent in remote sensing applications. Design-based inference assumes that the observation on a sampling unit is a fixed constant, and the uncertainty in the inferred information about the population is attributed to the randomization of the sampling design. The bias and variance of an estimator – statistical properties often central to remote sensing applications related to land cover as discussed above – are determined by the set of all possible sample sets from the sampling design used to select the sample used for inference (Stehman, 2009). Design-based inference is the classical approach to inference and this chapter is concerned only with a design-based estimation protocol. A design-based protocol consists of three major components: sampling design, response design, and analysis (Stehman & Czaplewski, 1998).

2.1 Sampling design

Because the expression of the estimator, and its properties, bias and variance, depend on how the sample data was selected, the choice of sampling design is a critical issue; or as expressed in Stehman (2009, page 5263): "the sampling design is everything."

Sampling design is the protocol for selecting the sample from the population. In the context of remote sensing, this typically means selecting a subset of map units (e.g. pixels or segments) from the study area. The selection can be achieved in several ways but in a design-based framework, probability sampling is a requirement. Probability sampling is defined by two conditions: (1) the inclusion probability (the likelihood of a population unit being included in the sample) must be known for each unit selected in the sample; (2) and the inclusion probability must be greater than zero for all population units (Stehman, 2001). Still, there are several designs that fulfill these criteria. To assist in the selection of a sampling design, discussing and answering the following three questions are helpful: Whether to use clusters? Whether to use strata? Whether to use systematic or simple random selection?

To answer the first question, we need to define cluster sampling. Cluster sampling is based on selection of clusters of one or more basic assessment units instead of selecting units as individual entities. Even if clusters of units are selected, they are still analyzed as individual units (if analyzed as a single entity the cluster would be referred to as a block assessment unit). Cluster sampling has one main advantage: cost saving. If, for example, the aim of a study is to estimate population parameters for a large country and the selected sample units are interpreted using costly high-resolution satellite imagery, limiting the extent of the sampling to a few larger clusters would save cost compared to covering the whole country with high-resolution imagery. An example of how cluster sampling was used to save cost when purchasing high resolution imagery for estimating deforestation rates is provided by Potapov et al. (2014). There are a few drawbacks though. Of importance is the correlation among units within a cluster that often reduces precision relative to a simple random sample of equal size. Also, it may complicate the use of strata if they are based on map categories and the assessment unit is a pixel (which is a rather common scenario); and the analysis is typically more complicated (Olofsson et al., 2014). These factors must be weighed against the cost savings before choosing a cluster-based design. It is worth noting that several large-scale remote sensing applications, published in the last couple of years, have successfully employed two-stage cluster sampling together with a ratio estimator (e.g., Pickens et al, 2020; Song et al., 2021; Li et al., 2023).

The question of using strata is often easier to answer. Strata are partitions of the population such that each population unit belongs to exactly one stratum. The two main reasons for using strata are (1) a need to estimate parameters for subregions of the population and (2) to increase precision of estimates. The latter reason is often a strong argument for using strata. This is especially true in change studies where the land-surface activities of interest occupy a small proportion of the study area. Without a stratification of the landscape, a very large sample might be required to produce an estimate. However, if a change map of the study area exists, using the change categories as strata will ensure that a sufficient sample in areas of activity is selected for precise estimation. The same precision could be achieved with a simple random sample (i.e. without using strata) but at a larger cost as a considerably larger sample is likely to be required. Olofsson et al. (2014) recommended stratified random sampling to estimate area and accuracy of land change for these reasons.

Finally, the question of whether to use simple random or systematic sampling. The latter is considered random as the starting location is randomly selected and consecutive units are selected with a fixed distance between sample locations. Both can be used as selection protocols in cluster-based and stratified designs, and unbiased estimators are available for both designs.

While some differences related to precision and variance estimation exist, the choice between systematic and simple random usually depends on whether reference observations are collected by field visits (systematic preferable) or by satellite or aerial imagery (simple random preferable).

2.2 Response design

In many social science applications, the term response design is intelligible as the analyst of the survey designs the response options. In a map context, it is less intuitive, but the response design has a similar meaning in that we often design the response options of sample units – "did this unit experience deforestation or not?" for example. In other words, the response design is the protocol that leads to a decision regarding agreement of the reference and map labels. The sampling design selects the locations of the units that are interpreted using reference data for collection of reference observations. The sample of reference observations (also referred to as the reference classification) is the data to which an estimator is applied for inference of population parameters such as accuracy and area. This makes the response design an important component of the estimation protocol. The first step of the response design is the selection of a spatial assessment unit. This is often a pixel but can also be a polygon (typically a segment) or a block of pixels. Regardless of assessment unit, the analyst should aim to have reference data with units of equal or finer level than the map units. The choice of unit will also impact the analysis; Stehman & Wickham (2011) provide a comprehensive analysis of the considerations of various spatial assessment units.

Another critical aspect of the response design is the source of reference data. There is a variety of viable sources of reference data ranging from field inventories to satellite imagery. A requirement is that the reference classification must be of higher quality than the map classification. This requirement can be ensured by selecting data of higher quality than the data used to create the map. If using the same reference data for both map and reference classifications, the process of collecting reference observations must be more accurate than the map making process (Olofsson et al., 2014). Since the opening of the Landsat archive in 2008 (Woodcock et al., 2009), and the launch of Sentinel-2 in 2015, time series of satellite observations of surface reflectance provide a powerful source of reference data. Various research groups have developed software tools that facilitate access and use of time series data for this purpose; examples include Collect Earth Online (https://www.collect.earth/) and the tools in Arévalo et al. (2020).

The final step of the response design is the labelling protocol and definition of agreement. If land categories and mapping and assessment units are well-defined, and the map and response legends are unambiguous, this is a rather straightforward task. But this is seldom the case; heterogeneous landscapes, mixed pixels, vague land category definitions, geolocation errors, varying experience, confidence and effort among interpreters, and clerical errors are common problems that need attention. A discussion of these issues is provided in Olofsson et al. (2014), who also provides a set of recommendations for dealing with the afore-mentioned problems.

2.3 Analysis

2.3.1 The error matrix

The analysis of sample data is the final step of the estimation protocol. Central to the analysis is the application of an estimator to sample data to obtain estimates of accuracy and area. Note that there is an important distinction between the *estimator* and the *estimate* – the estimate (a number) is a realized value of an estimator (a mathematical formula) (Casella & Berger, 2002). Before applying the estimator, it is common practice in remote sensing studies to construct an error matrix, which is also called a contingency or confusion matrix (Card, 1982). The error matrix is a cross-tabulation of map and reference labels for each location in the sample. The rows of the matrix normally represent the map labels and the columns the reference labels. Table 1 shows a simple example of an error matrix with two categories.

Table 1. Example error matrix with two categories.

		Reference (*j*)		
		Category 1	Category 2	Total
Map (*i*)	Category 1	p_{11}	p_{12}	p_{1+}
	Category 2	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

The elements in the error matrix in Table 1 (p_{ij}) are the area proportions such that p_{12} is the proportion of the map area that was mapped as category 1 but identified as belonging to category 2 in the reference data. In an error matrix, the diagonal values represent the correctly classified area whereas the off-diagonal elements represent the incorrectly mapped area. The matrix element p_{12} is the commission error of Category 1 and p_{21} is the omission error, both expressed as area proportions of the total map area. The elements of the matrix are estimated from the sample data, which requires construction of a sample-based estimator ($\hat{p}_{ij}$). The formula of the estimator depends on the sampling design; for stratified random sampling (STR) in which the strata correspond to the map classes, the area proportion is estimated as (Olofsson et al., 2013, Eq. 1)

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i+}} \tag{1}$$

where n_{ij} is the number of units in the sample classified as category i and identified as category j in the reference data; n_{i+} is the total number of sample units in map category *i*; and W_i is the weight of stratum *i* (often indexed using the letter *h* in the literature). For simple random and systematic sampling, Eq. (1) is a post-stratified estimator of p_{ij} (Card, 1982). For designs that are more complex such as cluster-based and multi-stage designs or STR designs when strata are different from map classes, the construction of an error matrix is more complicated (the formulas are available in Stehman, 2013, 2014; Strahler et al., 2006).

2.3.2 Accuracy measures

Once the area proportions have been estimated and the error matrix populated, overall, user's and producer's accuracy are easily calculated. The overall accuracy is the diagonal total which is equivalent to the estimated proportion of the area correctly mapped. In an assessment of a map of q classes, the overall accuracy (O) is estimated as (Olofsson et al., 2013, Eq. 6)

$$\hat{O} = \sum_{j=1}^{q} \hat{p}_{jj} \tag{2}$$

The user's accuracy of category i (U_i) is estimated as (Olofsson et al., 2013, Eq. 7)

$$\hat{U}_i = \frac{p_{ij}}{p_{i+}} \tag{3}$$

while the producer's accuracy of category j (P_j) is estimated as (Olofsson et al., 2013, Eq. 8)

$$\hat{P}_j = \frac{p_{ij}}{p_{+j}} \tag{4}$$

Variance estimators are available for each of the measures. They are not provided here because of the complexity of the computations but can be found in Olofsson et al. (2014).

2.3.2 Area estimates

Often more important than map accuracy is estimation of area. As discussed above, this requires construction of an estimator. In a design-based inference framework, the estimator should be unbiased and must correspond to the sampling design used to select the sample. For example, if the sample was selected by simple random (SRS) or systematic sampling, the sample mean is an unbiased estimator of the population mean (Cochran, 1977, Theorem 2.1):

$$\hat{\mu}_{SRS} = \frac{\sum_i y_i}{n} \tag{5}$$

with an unbiased variance estimator (Cochran, 1977, Theorem 2.4):

$$\hat{\mathrm{V}}(\hat{\mu}_{SRS}) = \frac{\sum_i (y_i - \hat{\mu}_{SRS})^2}{n-1} \tag{6}$$

In Eqs. 4 and 5, y_i is the ith reference observation of a sample of a total of n observations. For example, if a sample of 500 reference observations of land-surface conditions has been collected and 72 instances of deforestation were observed, the estimated area proportion of deforestation is (Eq. 4): $\hat{\mu}_{SRS} = \sum_i y_i \div n = 72 \div 500 = 0.14$.

With stratified random sampling (STR), Eqs. 4 and 5 need to be modified to account for the different sampling intensities within strata. For q number of strata, an unbiased estimator, often called a stratified estimator (Cochran, 1977, Eq. 5.1) is

$$\hat{\mu}_{STR} = \sum_{i=1}^{q} W_i \hat{\mu}_i \tag{7}$$

where $\hat{\mu}_i$ is the sample mean for stratum *i*:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n} \tag{8}$$

In this case, y_{ij} is the *j*th reference observations in stratum *i*, n is the total sample size and n_i is the sample size of stratum *i*. Again, note that Cochran (1977) is not using *i* to index strata but *h*. A stratified variance estimator is given by Cochran (1977, Eq. 5.7 and 5.11) as

$$\hat{V}(\hat{\mu}_{STR}) = \sum_{i=1}^{q} W_i^2 \frac{\sigma_i^2}{n_i}; \tag{9}$$

$$\sigma_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{q} \left(y_{ij} - \hat{\mu}_i\right)^2 \tag{10}$$

The stratified estimator is typically used when stratified random sampling is implemented but the estimator can be applied to simple random and systematic sampling as well. The stratified estimator is then referred to as a post-stratified estimator as the strata are incorporated via the estimator instead of via the sample selection as is the case for a stratified design (Cochran, 1977; Särndal, Svensson, & Wretman, 1992). Post-stratification may increase precision in estimates when the sample is based on permanent plots (a national forest inventory for example) and stratified sampling is not possible (McRoberts, Næsset, & Gobakken, 2013).

A common alternative to the stratified estimator is the model-assisted regression estimator (Särndal et al., 1992). Just as with the stratified estimator, it has been used in remote sensing studies for estimation of area (e.g. McRoberts & Walters, 2012; Sannier, McRoberts, Fichet, & Makaga, 2013). The model-assisted regression estimator is constructed from auxiliary data that is not the specific estimation target; (for example, a map or satellite imagery could be auxiliary data when using the estimator to estimate the area of deforestation) and includes an explicit estimate of the bias (McRoberts, 2010, Eq. 10 and 11):

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i - \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \tag{11}$$

and

$$\hat{V}(\hat{\mu}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 \tag{12}$$

where $\varepsilon_i = \hat{y}_i - y_i$ and $\bar{\varepsilon} = \frac{1}{n} \sum_i \varepsilon_i$. The second term in Eq. 11 is the estimated bias that is subtracted from the mapped value of the parameter of interest. Regarding which estimator to use, the stratified estimator is typically more efficient when working with categorical categories (e.g. deforestation or not) while the model-assisted regression estimator has been found to produce higher precision when used with continuous reference observations (e.g. percent deforestation) (GFOI, 2016).

For cluster and multi-stage sampling designs, the estimators get more complicated. Estimators for various situations that include multiple stages and clusters are available in Chapters 9 and 10 in Cochran (1977).

While the estimators explained in this subsections are presented similar to the format in the textbooks of Cochran (1977), Casella & Berger (2002) and Särndal et al. (1992), it is also possible to extract the area estimates directly from the error matrix. For example, the area estimate produced by a stratified estimator of Category 1 in Table 1, is simply the sum the first column. This is also called a direct estimator and is written as (Stehman, 2013, Eq. 1)

$$\hat{p}_{+j} = \sum_i \hat{p}_{ij} \tag{13}$$

Similar to the model-assisted regression estimator is the a bias-adjusted estimator which would produce the same estimate as the direct estimator but would include an explicit expression of the bias (Stehman, 2013, Eq. 6):

$$\hat{p}_{+j} = p_{i+} + (\hat{\varepsilon}_{omission} - \hat{\varepsilon}_{commission}) \tag{14}$$

If using this expression to estimate the area of Category 1 in Table 1 we would get $\hat{p}_{+1} = p_{1+} + (\hat{p}_{12} - \hat{p}_{21})$, i.e. we would add the estimated omission error ($\hat{p}_{12}$) and subtract the estimated commission error ($\hat{p}_{21}$) from the mapped area of Category 1 (p_{1+}). A comprehensive analysis of how to estimate area from the information in an error matrix is provided by Stehman (2013).

3. Additional considerations

A couple of recent developments in this field are worth mentioning.

First, errors of omission (defined in the list of terms below) have in certain situations been found to introduce considerable uncertainty in parameter estimates (e.g., Arévalo et al., 2019; Bullock et al., 2018; Potapov et al., 2017). The situation arises when following recommended practices of using stratified random sampling to estimate an activity that is a small proportion of the study area relative a large single stratum such as forest. This is a common situation in studies of activities such as deforestation and forest degradation in tropical countries where the relative area of intact forest tends to be large. If in such situations, with strata constructed that correspond to mapped deforestation and intact forest, any observations of deforestation at sample locations in the forest stratum (i.e. instances of deforestation mapped as intact forest) will carry a large area weight due to the area proportion/weight of the forest stratum. The weight of the errors can thus be reduced if the size of the stratum in which the error was observed is reduced. Under stratified sampling, any strata can be constructed provided they are exhaustive and non-overlapping; this implies that any available information on the likely location of omission errors can be used to define strata. Because omissions of change typically occur near areas of mapped change, while areas mapped as stable forest at larger distances from mapped change are unlikely to contain omissions, creating a stratum corresponding to a spatial buffer around mapped change has been reported in the literature; the topic is discussed in detail by Olofsson et al. (2020).

Second, there are attempts in the literature to bridge wall-to-wall maps and point estimates which deserve attention (e.g. Song et al., 2017; Li et al., 2023). The spatial information of a map is often of high value to subsequent analyses and decision-making, but such map usage is problematic if a sample analysis shows that the map categories of interest have considerable area bias. To mitigate this discrepancy between spatial information and sampling-based estimates, methods have been published that aim to conform maps to estimates. For example, Li et al. (2023) mapped various crops in optical satellite data and estimated associated areas and map accuracy using a regression estimator applied to sample data selected under two-stage sampling. To match the map to the sampling-based area estimates, intermediate map output was used as an auxiliary variable when constructing the regression estimator, and the mapping exercise was constrained by the area estimates such that map-based areas (i.e. pixel-counting) matched the sampling-based area estimates. Greater adoption by the community of such methods could help elevate the use of remote sensing in decision-making.

4. Terminology

Accuracy

A measure of “correctness”; in the context of this chapter, an estimate of accuracy is the degree to which the map agrees with a sample of reference observations.

Accuracy Assessment

The process and steps involved in estimating measures of map accuracy from a sample of reference observations.

Bias

A property of an estimator; the bias of an estimator $\hat{\mu}$ of a population parameter μ is the difference between μ and the expected value of $\hat{\mu}$ over all possible samples; that is, $\mathrm{Bias}(\hat{\mu}) = \mathrm{E}(\hat{\mu}) - \mu$ (Casella & Berger, 2002, p. 330). Note that because variance and bias are properties of estimators, and “because an estimate is a number, it has no variation and no bias. The term biased estimate is [not recommended but is] nevertheless used occasionally.” (Särndal et al., 1992, p. 41).

Cluster

A “sampling unit [that] consists of group or *cluster* of smaller units” (Cochran, 1977, p. 233).

Commission error

Commission error is the proportion or percentage of the area mapped as the category of interest that is erroneously predicted based on comparison to the reference classification. Commission error is the complement of user's accuracy.

Confidence interval

A 95% confidence interval for a population parameter, μ, expresses uncertainty in the parameter estimate, $\hat{\mu}$, and is calculated using the sample data. Confidence intervals are often, but not necessarily, in the form $\hat{\mu} \pm a \times \mathrm{SE}(\hat{\mu})$ where $\mathrm{SE}(\hat{\mu})$ is the standard error of the estimate and a is a statistic related to the desired confidence level (see z-score). Among the aggregate set of confidence intervals constructed using all samples that could be realized using the sampling design, 95% of such intervals are expected to include the true value of the population parameter μ, although which intervals do and which do not include μ is generally unknown. Typically, 95% confidence intervals are reported in the literature.

Design-based inference

The process of drawing an inference for a population parameter by analyzing a probability sample selected from the population. With design-based inference, the observation for a population unit is a constant value, apart from negligible measurement error. Design-based properties of estimators such as bias and variance are determined from the sampling distribution constructed from the aggregate set of population parameter estimates obtained from all possible samples that could be realized from the sampling design. See also Inference.

Estimate

The value obtained from the estimator when applied to a specific sample. An estimate of μ is denoted $\hat{\mu}$. Note that in the literature, the estimator is usually denoted in the same way as the estimate (e.g. Cochran, 1977, p. 11; Särndal et al., 1992, p. 40) because there will not likely be confusion regarding which use we intend.

Estimator

"The rule by which an estimate of some population characteristic μ is calculated from the sample results" (Cochran, 1977, p. 11). Note that that an estimator is not the same thing as an estimate: "An *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator (that is, a number) that is obtained when a sample is actually taken" (Casella & Berger, 2002, p. 312).

Inference

In a sampling framework, an inference expresses the relationship between the population parameter, μ, and its estimate, $\hat{\mu}$, in probabilistic terms, typically in the form of a confidence interval (Dawid, 1983).

Kappa

The kappa coefficient of agreement is often used as an overall measure of map accuracy. It allegedly incorporates an adjustment for random agreement, which has been questioned by the literature. For example, it has been reported "that these Kappa indices are useless, misleading and/or flawed for … practical applications in remote sensing" (Pontius & Millones, 2011).

Margin of error

A relative measure of the uncertainty in an estimate. Note that the definitions of margin of error are not all the same. Typically, it is calculated as the ratio of the half width of a 95% confidence interval to an estimate.

Model-assisted estimator

An estimator used in design-based inference framework that incorporates auxiliary information to increase precision by comparing a model's predictions, often in the form of map unit values, to a probability sample of reference observations (Särndal et al., 1992, p. 219). Model-assisted estimators can be particularly effective when the response variable is continuous (e.g., proportion forest or biomass) rather than categorical (e.g., forest/non-forest or forest change class).

Model-based inference

Inference based on the perspective that an observation for a population unit is a random variable and that the model used to make predictions for all units in the population has been adequately specified in the sense that there is no systematic lack of the model fit to data that represent the entire population (Stehman, 2009).

Omission error

An omission error is the proportion of area with the reference classification of the category of interest that is erroneously predicted (mapped) to be in other categories. Omission error is the complement of producer's accuracy.

Overall accuracy

The overall accuracy is the "overall proportion of area correctly classified" (Stehman, 1997, p. 79).

Population

"The aggregate from which the sample is chosen." In the context of this chapter, the population is typically a study area (Cochran, 1977, p. 5).

Population parameter

"Numerical characteristics, or parameters, of the population that we will estimate from a sample" (Rice, 1995, p. 186). (Example: the area deforested or a fixed value in a model.)

Population unit

An individual member of the set of elements that make up the population. Population units are referred to as elements in Särndal et al. (1992, p. 9).

Post-stratified estimation

Post-stratification refers to a stratification of the study area that is independent of the selection of the sample and applied subsequently to the sampling. For example, if a systematic sample of plots exists and we want to estimate the area of forest, post-stratifying the area using a forest/non-forest map is likely to increase precision in the area estimate, even if stratified random sampling is not used. In this case, a stratified estimator applied to the sample data is referred to as a post-stratified estimator.

Precision

In the context of estimation, Cochran (1977, p. 16) states that because “of the difficulty of ensuring that no unsuspected bias enters into estimates [sic], we will usually speak of the precision of an estimate instead of its accuracy. Accuracy refers to the size of deviations from the true mean μ, whereas precision refers to the size of deviations from the mean m obtained by repeated application of the sampling procedure.” In the context of this document, we often characterize the precision of an estimate with a 95% confidence interval – the larger the interval the less the precision (and greater the uncertainty).

Probability sample

A sample drawn from a population using a randomization mechanism such that “the inclusion probability for each element of the sample is known, and the inclusion probabilities are non-zero for all elements of the population.” (Stehman, 1999).

Producer’s accuracy

From Stehman (1997, p. 79): “Producer’s accuracy for [category] j [is] the conditional probability that an area classified as category j by the reference data is classified as category j by the map.” When expressed in terms of area, producer’s accuracy is the proportion of area that has the reference classification of the category of interest that is correctly predicted (mapped) as that category. Producer’s accuracy is the complement of omission error.

Reference data

Data characterizing the most accurate available assessment of the true condition at the sample location (example: fine-resolution satellite imagery).

Reference classification

The most accurate available assessment of the true condition of a population unit (example: deforestation).

Reference observations

The reference classification applied to the collection of all sample units.

Sample

A subset of units selected from the population.

Sampling unit

Entities that make up the sampling frame (Särndal et al., 1992, p.5). In the literature, sometimes there is no distinction made between population units and sampling units (e.g. Cochran, 1977, p. 6). However, population and samplings units are different entities because the sampling frame and the population are sometimes different entities (in many situations though, the sampling frame is equivalent to the population).

Simple random sampling

"A method for selecting n units out of the N such that every one of [the sets of n specified units] has an equal chance of being drawn." (Cochran 1977, p. 18).

Spatial assessment unit

"Spatial unit for comparing the map and reference classifications" (Stehman & Wickham, 2011, p. 3044); typically a pixel, block of pixels or polygon.

Standard deviation

The standard deviation of a random variable X is the square root of the variance: $\mathrm{SD}(X) = \sqrt{\mathrm{V}(X)}$ (Rice, 1995) and is often denoted by $\mathrm{S}(X)$, $\mathrm{SD}(X)$, $\mathrm{D}(X)$, or σ_X. Because of the square root, the standard deviation has the same unit as the random variable as opposed to the variance. A standard deviation calculated from sample data is sometimes referred to as the sample standard deviation to distinguish it from the population standard deviation. The standard deviation of an estimate is referred to as its standard error.

Standard Error

The standard error is the standard deviation (i.e. square root of the variance) of an estimator (Rice, 1995, p. 192). For example, consider the situation in the explanation of variance below: we want to estimate the mean $\bar{Y}$ of a population of size N with variance σ^2. To do this, we select a simple random sample of n units $(y_1, \dots, y_n)$; $\bar{y}$ is an estimate of $\bar{Y}$ and the estimated variance of the sample mean is $\hat{\mathrm{V}}(\bar{y}) = s^2 \div n$ (the sample variance s^2 is defined under *Variance* below). The standard deviation of $\bar{y}$ is referred to as its standard error. Because the sample variance is an unbiased estimator of the population variance, we can estimate the standard error using the sample variance: $\mathrm{SE}(\bar{y}) = \sqrt{\hat{\mathrm{V}}(\bar{y})} = s \div \sqrt{n}$. Note that because a standard error is a standard deviation of an estimator, all standard errors are also standard deviations, but not all standard deviations are standard errors. This sometimes causes confusion even though the definitions of standard error and standard deviation are consistent. The confusion is exacerbated by the common use of the letter S to denote both standard errors and standard deviations.

Strata

Strata are "subpopulations that are nonoverlapping, and together they compromise the whole population" (Cochran, 1977, p. 89).

Unbiased estimator

"An estimator $\hat{\mu}$ of μ given by a sampling plan is unbiased if the mean value of $\hat{\mu}$, taken over all possible samples provided by the plan, is equal to μ" (Cochran, 1977, p. 11); or in other words, the estimator is characterized as unbiased if it produces an "estimate [that] is correct 'on the average'" (Rice, 1995, p. 192).

User's accuracy

From Stehman (1997, p. 79): "User's accuracy for [category] i [is] the conditional probability that an area classified as category i by the map is classified as category i by the reference data". When expressed in terms of area, user's accuracy is the proportion of the area that has the predicted class of the category of interest that is correctly classified as determined by comparison to the reference classification. User's accuracy is the complement of commission error.

Variance

The formal definition of the variance of a random variable X with expected value $\mathrm{E}(X)$, is $\mathrm{V}(X) = \mathrm{E}(X - \mathrm{E}(X))^2$ to provide "a measure of the degree of spread of a distribution around its mean" (Casella & Berger, 2002, p. 59). But this definition is not very relevant in geography/remote sensing context. Instead, we are concerned with situations where a sample has been selected from a population (e.g. all pixels of a country) with the objective of estimating a certain population parameter, μ (e.g. the area of deforestation). For example, let's say we have a population of N units $(y_1, \dots, y_N)$ with mean $\bar{Y}$ and variance σ^2 from which a sample of n units $(y_1, \dots, y_n)$ has been selected under simple random sampling. We have collected reference observations for the n sample units. The sample variance is $s^2 = \sum(y_i - \bar{y})^2 \div (n-1)$ (Cochran, 1977, p. 26). Again, this is also not very helpful as we are usually not interested in the sample variance but in the variance of an estimate (e.g. the area of deforestation). For simple random sampling designs, the sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$, and the sample variance s^2 is an unbiased estimator of the population variance σ^2, such that $\mathrm{E}(\bar{y}) = \bar{Y}$ and $\mathrm{E}(s^2) = \sigma^2$ (Cochran, 1977, p. 22, 26). The variance of $\bar{y}$ is $\mathrm{V}(\bar{y}) = \sigma^2 \div n$ (assuming a small n relative N); the estimated variance of $\bar{y}$ substitutes the sample variance s^2 for the population variance σ^2 to create the variance estimate of $\bar{y}$ as $\mathrm{V}(\bar{y}) = s^2 \div n$. This is the variance that is of primary interest to us, and that allows us to calculate a standard error and a confidence interval of the estimated population parameter of interest.

Z-score

A z-score (also referred to as standard score), $z_{1-\alpha/2}$ where $1 - \alpha$ is the confidence level between 0 and 100%, is a constant such that the area under the standard normal density function in between $\pm z_{1-\alpha/2}$ is $1 - \alpha$ (Rice, 1995, p. 202). For example, a confidence interval at the 95% confidence level for an estimate $\hat{\mu}$ would be computed as $\hat{\mu} \pm z_{1-0.025} \times \mathrm{SE}(\hat{\mu}) = \hat{\mu} \pm 1.96 \times \mathrm{SE}(\hat{\mu})$. A condition for using z to compute a confidence interval is that that the sampling distribution for $\hat{\mu}$ is approximately a normal distribution, which we can assume for large sample sizes (Särndal et al.,1992).

References

Arévalo, P., Olofsson, P., & Woodcock, C. E. (2020). Continuous monitoring of land change activities and post-disturbance dynamics from Landsat time series: A test methodology for REDD+ reporting. *Remote Sensing of Environment*, *238*, 111051.

Arévalo, P., Bullock, E. L., Woodcock, C. E., & Olofsson, P. (2020). A suite of tools for continuous land change monitoring in google earth engine. *Frontiers in Climate*, *2*, 576740.

Bullock, E. L., Woodcock, C. E., & Olofsson, P. (2020). Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis. *Remote Sensing of*

Environment, *238*, 110968.

Card, D. H. (1982). Using map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, *49(12)*(3), 431–439.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

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

Congalton, R. G. (1991). A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment*, *37*, 35–46.

Congalton, R. G. (2004). Putting the Map Back in Map Accuracy Assessment. In R. S. Lunetta & J. G. Lyon (Eds.), *Remote Sensing and GIS Accuracy Assessment* (pp. 1–12). Boca Raton: CRC Press.

Congalton, R. G., Oderwald, R. G., & Mead, R. A. (1983). Assessing Landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering and Remote Sensing*, *49*(12), 1671–1678.

Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A., & Hansen, M. C. (2018). Classifying drivers of global forest loss. *Science*, *361*(6407), 1108-1111.

Foody, G. M. (1992). On the compensation or chance agreement in image classifcation accuracy assessment. *Photogrammetric Engineering and Remote Sensing.*, *58*(10), 1459–1460.

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

Gallego, F. J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, *25*(15), 3019–3047.

GFOI. (2016). *Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests: Methods and Guidance from the Global Forest Observations Initiative* (2nd ed.). Rome: Food and Agriculture Organization.

Li, H., Song, X. P., Hansen, M. C., Becker-Reshef, I., Adusei, B., Pickering, J., ... & Justice, C. (2023). Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation. *Remote Sensing of Environment*, *294*, 113623.

McRoberts, R. E. (2010). Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment*, *114*(5), 1017–1025.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, *115*(2), 715–724.

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment*, *128*, 268–275.

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

Murray, N. J., Worthington, T. A., Bunting, P., Duce, S., Hagger, V., Lovelock, C. E., ... & Lyons, M. B. (2022). High-resolution mapping of losses and gains of Earth's tidal wetlands. *Science*, *376*(6594), 744-749.

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

Olofsson, P., Foody, G. M., Stehman, S. V, & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, *129*, 122–131.

Olofsson, P., Arévalo, P., Espejo, A. B., Green, C., Lindquist, E., McRoberts, R. E., & Sanz, M. J. (2020). Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, *236*, 111492.

Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., ... & Sherani, Z. (2020). Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. *Remote Sensing of Environment*, *243*, 111792.

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International journal of remote sensing*, *32*(15), 4407-4429.

Potapov, P. V., Dempewolf, J., Talero, Y., Hansen, M. C., Stehman, S. V, Vargas, C., … Zutta, B. R. (2014). National satellite-based humid tropical forest change assessment in Peru in support of REDD+ implementation. *Environmental Research Letters*, *9*(12), 124012.

Potapov, P. V., Turubanova, S. A., Tyukavina, A., Krylov, A. M., McCarty, J. L., Radeloff, V. C., & Hansen, M. C. (2015). Eastern Europe’s forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. *Remote Sensing of Environment*, *159*, 28–43.

Potapov, P., Siddiqui, B. N., Iqbal, Z., Aziz, T., Zzaman, B., Islam, A., ... & Hansen, M. C. (2017). Comprehensive monitoring of Bangladesh tree cover inside and outside of forests, 2000–2014. *Environmental Research Letters*, *12*(10), 104015.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Richards, D. R., & Friess, D. A. (2016). Rates and drivers of mangrove deforestation in Southeast Asia, 2000–2012. *Proceedings of the National Academy of Sciences*, *113*(2), 344–349.

Sannier, C., McRoberts, R. E., Fichet, L. V., & Makaga, E. M. K. (2013). Using the regression estimator with Landsat data to estimate proportion forest cover and net proportion deforestation in Gabon. *Remote Sensing of Environment*, *151*, 138–148.

Särndal, C. E., Svensson, B. H., & Wretman, J. H. (1992). *Model assisted survey sampling*. Springer.

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

Song, X. P., Hansen, M. C., Potapov, P., Adusei, B., Pickering, J., Adami, M., ... & Tyukavina, A. (2021). Massive soybean expansion in South America since 2000 and implications for conservation. *Nature sustainability*, *4*(9), 784-792.

Souza Jr, C. M., Z. Shimbo, J., Rosa, M. R., Parente, L. L., A. Alencar, A., Rudorff, B. F., ... & Azevedo, T. (2020). Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, *12*(17), 2735.

Stehman, S. V. (2000). Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment. *Remote Sensing of Environment*, *72*(1), 35–45.

Stehman, S. V. (2001). Statistical Rigor and Practical Utility in Thematic Map Accuracy Assessment. *Photogrammetric Engineering and Remote Sensing.*, *67*, 727=734.

Stehman, S. V. (2009). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sensing of Environment*, *113*(11), 2455–2462.

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

Stehman, S. V. (2014). Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing*, *35*(13), 4923–4939.

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

Stehman, S. V, & Wickham, J. D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, *115*(12), 3044–3055.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., … Woodcock, C. E. (2006). Global Land Cover Validation: Recommendations For Evaluation And Accuracy Assessment Of Global Land Cover Maps. European Commission.

Verbesselt, J., Hyndman, R., Newnham, G., & Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, *114*(1), 106–115.

Woodcock, C. E., Allen, R., Anderson, M., Belward, A. S., Bindschadler, R., Cohen, W. B., … Wynne, R. H. (2009). Free Access to Landsat Imagery. *Science*, *320*, 1011.

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment*, *144*, 152–171.