Summary of Country experiences and critical issues related to estimation of activity data

4th April 2018

1 Purpose

To clarify MGD guidance related to the estimation of uncertainty of activity data generated from remote sensing data sources through the elaboration of real world experiences.

2 Generating activity data from remote sensing data

The IPCC definition of good practice requires that greenhouse gas inventories should satisfy two criteria: (1) neither over- nor under-estimates so far as can be judged, and (2) uncertainties reduced as far as is practicable (IPCC, 2003; preface).

Good practice requires the use of a 95 percent confidence interval which expresses estimates in a probability context that includes the estimate, the variance and a degree of probability (or confidence). Confidence intervals may also be expressed as a percentage of the mean estimate (IPCC, 2006). The width of a confidence interval is closely related to precision, a measure of the uncertainty addressed by the second IPCC criterion. Good practice requires that biases be identified and corrected to the extent possible, so that the uncertainty analysis can focus on quantification of the random errors with respect to the mean estimate. As such confidence intervals constructed using unbiased estimators satisfy both IPCC good practice criteria specified above (see section 5.1.5; GFOI, 2016).

Methods that produce estimates of activity data as sums of areas of map units assigned to map classes are characterized as pixel counting. This census approach to generating activity data does not apply a sample-based estimator and therefore there is no estimator bias. Map classification errors are the source of the bias in pixel counting approaches.

Confusion or error matrices and map accuracy indices from pixel counting methods do not produce the information necessary to construct confidence intervals (Box 1).

BOX 1 – Challenges in addressing IPCC good practice guidance when using map estimates (or pixel counts) as activity data.

The confusion matrix of a specific case in Costa Rica is represented below. This confusion matrix does not show the sample counts but rather represents the proportion of area (Olofsson et al. 2014, Table 4), see also section 5.1.5.3, GFOI, 2016. In this case, the confusion matrix indicates that a particular mapped area has an overall accuracy of 0.8500 and in the context of deforestation, a user’s accuracy of 0.6190 and a producer’s accuracy of 0.4907 is achieved. An overall accuracy of 0.8500 would be considered an acceptable and the user’s accuracy of 0.62 would also be considered as acceptable for change detection.

The user’s accuracy value for the map class of deforestation indicates that 62% (i.e. 0.6910) of the area classified as deforested was actually deforested, while the producer’s accuracy indicates that 49% (i.e. 0.4907) of the actual deforestation that occurred in the region of interest was captured by the map.
In this specific case, using the map as the basis for estimating activity data would underestimate deforestation by 0.012; i.e. difference between the proportion given by the map (0.0445, within grey box) and the proportion given by the reference data (0.0561 within grey box), or by 26% in relative terms. Additionally, the use of the map would not enable the generation of associated confidence intervals required to report uncertainty according to IPCC guidelines and good practice guidance.

Table 1: Confusion matrix or proportion of area matrix of a case in Costa Rica

where DF = Deforestation AF = Afforestation F=Forest NF=Non Forest

Therefore, as seen in this example, **pixel-counting methods provide no assurance that estimates are “neither over- nor under-estimates” or that “uncertainties are reduced as far as practicable”**.

The role of reference data, also characterized as accuracy assessment data, is to serve as the basis of the area estimates. That is, if the reference data are the best assessment of the ground condition, then these data provide the best available information to estimate area. The primary role of the map in this context is to reduce the standard errors (uncertainty) of area estimates based on the reference sample data. Further, the area estimates produced from the reference data are accompanied by estimates of uncertainty, thereby providing the information necessary for construction of confidence intervals for compliance with IPCC good practice guidance.

Countries are progressively incorporating this guidance and using sampling to estimate areas. Below is a compilation of the approach followed by countries in their Forest Reference Levels submitted to the UNFCCC or to the Forest Carbon Partnership Facility (FCPF). Currently, 17 out of 35 submissions use sampling to estimate activity data, and the number of programs using the sampling approach has been increasing.
3 Sampling design

3.1 Simple random, stratified random: What sampling design should I use?

Samples for assessing the accuracy of map products are usually drawn from area frames, as opposed to list frames, to provide a better representation of the population (Gallego, 2004). In an area frame, sample units can be points, lines (often referred to as transects) or areas (often referred to as segments). The first step is to define the geographical area for which the activity data is to be estimated and the type of sample units. Point or pixel samples tend to be used in the majority of cases, but areas or segments have also been used in some cases (Sannier et al (2014)).

Probability sampling designs are summarised in Table 2. When selecting a sampling design the sample size for each activity must be large enough to produce sufficiently precise estimates of the area of the activity, in accordance to the policy requirements while being cost effective (see section 3.7 below).
Table 2: Considerations of various probability sampling designs

<table>
<thead>
<tr>
<th>Sampling design</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRS</td>
<td>SRS and SYS designs produce sample sizes for individual activities that are approximately proportional to their occurrence (these designs are examples of “equal probability” sampling designs). These designs are <strong>usually adopted in cases where there is more than one variable of interest</strong> (i.e. data on different land cover classes) <strong>or when stratification might not be suitable</strong> (e.g. stratification could constrain future monitoring). <strong>However, it is important to keep in mind that a large sample size is needed</strong> to achieve an acceptable precision for all variables of interest. This is relevant in the case of variables of interest that have little occurrence, i.e. area of deforestation. Examples: Examples of SYS are, Papua New Guinea’s FREL(^1), Mozambique’s REDD Program(^2), Bastin et al. (2017).</td>
</tr>
<tr>
<td>SYS</td>
<td>The objectives play a prominent role in deciding whether to use strata (e.g., objectives specify subnational estimates as well as a national estimate or estimates by each activity). <strong>A primary motivation for STR is when estimates are needed for different variables of interest or activities</strong> (e.g. data on different land cover change classes). If a very large overall sample is obtained, then SRS or SYS may produce large enough sample sizes for individual activities to produce estimates of sufficient precision. However, unless the overall sample size is large, sample sizes for activities representing small proportions of the total area may be too small to satisfy the precision criterion. Thus, given the likely rarity of some events (e.g. afforestation or even deforestation in some cases) and the potentially large costs associated with large sample sizes, <strong>serious consideration should be given to stratified sampling for which the strata correspond to map activity classes.</strong> Example: Sri Lanka’s FREL(^3), Potapov et al. (2017),</td>
</tr>
</tbody>
</table>
| STR             | The choice of whether to use clusters is motivated almost exclusively by the potential cost reduction sometimes gained by cluster or multi-stage sampling. Cluster sampling aggregates sampling units (e.g., pixels) within clusters (also called primary sampling units) so that the sample is constrained to a relatively limited number of clusters. **The cost savings achieved by sampling pixels in closer spatial proximity is the primary motivation for cluster sampling.** For example, if the primary sampling unit is a 5 km x 5 km block, the sample of secondary sampling units (pixels) is constrained spatially to the sample of blocks. If sampling is conducted to obtain sample units which are subject to image interpretation, high costs for acquiring high-resolution imagery for interpretation may justify cluster sampling. If free-of-charge imagery is available, then sample unit location is unlikely to affect the costs. With two-stage sampling, initial primary sampling units are chosen, several secondary sample units are selected within the first-stage primary sampling units. Although the motivation is often to reduce sampling costs, several factors must be considered when planning a two-stage sampling design. If distances between pairs of second-stage sampling units are less than the geographic range of spatial correlation, then observations will tend to be similar and the sampling will be less efficient. Consequently, the cost savings of cluster sampling must be such that the increase in the sample size (number of pixels) relative to an non-clustered design must be sufficient to overcome the loss of precision attributable to the within cluster

---

\(^1\) http://redd.unfccc.int/files/png_frl__submission-15.01.2017.pdf#page=18

\(^2\) https://www.forestcarbonpartnership.org/sites/fcp/files/2017/Aug/ZILMP_ERPD_ADVANCED%20DRAFT_to%20be%20posted%20online\_20Aug31_bettergraphs.pdf#PAGE=159

\(^3\) http://redd.unfccc.int/files/sri_lanka___s_forest_reference_level_submission_to_the_unfccc-06jan2017.pdf#page=13

\(^4\) from acquiring very high or high resolution imagery
Sampling design | Considerations
--- | ---
**correlation** (i.e., similarity of pixels within each cluster). Further, the analysis of the sample is often more complex than if analysing a sample selected by SRS, SYS or STR designs. Estimating standard errors can be particularly complex with two-stage cluster sampling designs.

Example: Guyana Forestry Commission Guyana REDD+ Monitoring Reporting & Verification System (MRVS) Year 5 Interim Measures Report 1 January 2014 – 31 December 2014[^1]; Sannier et al. (2014) in which single stage random/systematic and two stage cluster sample are compared; and Næsset et al. (2016) in which systematic single stage cluster sample was used for forest area estimation with different types of remotely sensed data (lidar, RapidEye, TanDEM-X and global map products from Landsat and ALOS PALSA).

STR designs are the most common in submissions made to the UNFCCC and the FCPF (Figure 2).

![Figure 2. Number of FRELS submitted to UNFCCC and FCPF using STR and SYS by 2016](https://www.forestry.gov.gy/wp-content/uploads/2015/10/MRVS-Interim-Measures-Report-Year-5-Ver-11.pdf)

### 3.2 Relative efficiency of using maps for estimates

Relative efficiency is a measure of the improvement in precision obtainable by using map data and reference data in combination to produce area estimates as opposed to using reference data only (see section 4.1.7; GFOI, 2016). Consideration of relative efficiency can help with cost effective design decisions such as, the cost of collecting more reference observations versus establishing a national mapping capability, and for example, costs of

establishing the relationship between global maps and national forest definitions. Assessing relative efficiency can assist to evaluate the most efficient sampling design, or within a design, which is the best strategy (e.g. which stratification map or which stratification criteria will further reduce standard errors). If sampling reference data and estimates are already available, relative efficiency could be a good basis to understand the best strategies to reduce standard errors in a cost-effective manner.

Sannier et al. (2014)⁶ presents examples of the application of relative efficiency, suggesting that collecting sample reference data and using an existing dataset for stratification (in this case a global map), might serve as good pilot sampling. This approach can provide data to estimate relative efficiencies and decide the best strategies for stratification purposes. It is important to keep in mind that maps may be generated and are useful for many purposes other than emissions estimation.

### 3.3 Sampling design should be matched to reporting objectives

Before specifying the sampling design it is important to keep in mind your objectives in terms of variables of interest to be reported (area of deforestation, area of degradation, deforestation of secondary forest, deforestation of primary forest and forest gain) and the reporting scope (geographical and temporal scope) the target/required precision (i.e. as specified by stakeholders/program requirements) and sometimes the combination of scope and variable of interest (e.g. area of deforestation per Province, or area of deforestation for two sub-periods) (see Appendix C; GFOI, 2016).

For instance, the design would differ if a variable of interest is the overall accuracy of a map for a region of interest, opposed to the area of deforestation in that region of interest. Also the design would differ if the interest is in reporting the variable for sub-areas of the region of interest or within different sub-periods with a certain required precision.

**Reporting objectives strongly influence decisions regarding strata (Box 2).** If the geographic region or variable of interest represents a small proportion of the area of the entire region (i.e. deforestation), it may be necessary to define the region or class type as a stratum. This is done to ensure that a sufficient sample size is obtained to yield the desired precision of the estimate for that region or class type. For instance, the sampling design would differ if the variable of interest is the overall accuracy of a map for a region of interest (simple random or systematic sampling), compared to the variable of interest being the area of deforestation (stratified random sampling) in that same region of interest.

The same applies to specific periods of analysis. For example, if the intent is to obtain estimates for different sub-periods within a historical period, then a way to ensure enough precision per sub-period is by considering this as a stratification criterion. In the planning process, a general rule is that every objective will require some allocation of the sample; therefore as the number of objectives increases so does the required sample size.

---

For a fixed total sample size, the **trade-off between the number of objectives and the precision of the estimators may need to be addressed**. If the resources available permit a total sample size that is sufficient to satisfy precision requirements for only a limited set of objectives, then the objectives should be prioritised so that sampling resources are not directed to lower priority objectives (e.g. required precision is achieved for the estimate of area of deforestation but not for the area of afforestation/reforestation). It is important to recognize that estimates for geographic subregions or class types can be produced even if they are not defined as strata. The challenge is that the precision of these estimates may not be satisfactory because the sample size is too small.

**Box 2 - Example Democratic Republic of Congo (DRC): Attending to different objectives at national and subnational level**

DRC is currently developing its Forest Reference Level to be submitted to the UNFCCC. DRC will report on GHG emissions from deforestation in different types of forests (i.e. swamp forests, **Terra Firma** Forests, Miombo forests and secondary forests). DRC has also developed a REDD-plus program in the Province of Mai Ndombe, in similar forest types, seeking payments from the FCPF Carbon Fund. The FCPF Carbon Fund also requires reporting of GHG emissions and removals from additional variables of interest including forest degradation and enhancement of carbon stocks in new and existing forests.

Since the objectives are different at the national and sub-national level (i.e. different REDD-plus activities are reported at each level), the DRC will apply at a national level a STR design, with simple random sampling within strata, where the stratification criteria will be the forest type, the change class (deforestation or stable) and the Province. This will allow reporting at the scale of the Province or adopting different sampling methods per Province. At the sub-national level, since the variables of interest are multiple, and the intention is to give the same priority to all of them, a SYS design will be applied instead. In this case, data on deforestation, degradation and enhancement of carbon stocks will be collected. Moreover, the precision requirements are more stringent under the Carbon Fund; this combined with the lack of stratification at the Provincial level requires a higher intensity grid to be applied in this case.

3.4 What sampling unit considerations should I be aware of?

The objective of sampling is to make inference for a parameter of interest within a population which is defined as the totality of all elements. The sample should be drawn from the **sampling frame** which may be a list of all elements of the population that can be selected as part of statistical sampling. The elements that we select and observe are the **sample units**.

One approach is the **finite population approach** for which the population is defined as the aggregate of units that cover the region of interest completely. When maps are used for stratification, the map would represent the population that is composed of a discrete number of population elements (pixels, polygons). In this case, we would select each element of the map according to a probabilistic design and its reference condition would be observed. The usefulness of this approach is that if the reference data and the map have the same spatial support, it is possible to estimate accuracy indicators of maps (used for stratification) while estimating areas or activity data.

A different approach is the **infinite population approach** for which the population can be imagined as an infinite number of points within an area sampling frame. In this case the point possesses a reference condition that is being observed, but this observation is based on the point itself or on the area around it, i.e. spatial support.
Considering the different design options of the previous section and these two approaches, we might have a menu of different options combining design and sampling unit: SRS/SYS, STR, CLU following the finite or infinite population approaches. A summary of some of these options, excluding CLU is presented in Figure 3:

Figure 3: Sampling Design Options

Although it is possible to estimate activity data for any sampling unit (e.g., point, pixel, or Polygon), the majority of experiences to date have been with pixel-based assessments, so this approach has an advantage of greater familiarity to practitioners. From the standpoint of sampling design, numerous options have been described and used in practice for area estimation based on a sample of pixels (Stehman 2013; Olofsson et al. 2014; GFOI 2016).
However, one of the disadvantages of using points or pixels as sampling unit is that the reference and map data can be misaligned leading to disagreements between the two datasets that are purely related to geometry even though the thematic content would agree otherwise. A second disadvantage is that the presence of mixed pixels in fragmented landscapes leads to erroneous thematic classification. Both issues can be overcome by using landscape feature polygons or objects as sample units. In planned landscapes this tends to be quite efficient such as the agricultural survey conducted in the USA by the USDA NASS\(^7\). However, in more natural landscapes with irregular boundaries, certain complications arise with object-based assessments. For polygon based assessments, the primary challenge is how to deal with the fact that objects differ in size and shape between time 1 and time 2. For example, if the sampling protocol first generates a set of random points (either completely at random or the result of placing a systematic grid at a randomized starting location) and then the objects intercepted by these points selected for the sample, the objects will be sampled with probability proportional to their area.

While such a protocol is a probability sampling design, the unequal inclusion probabilities will need to be accounted for in the area estimators as well as the standard error estimators. Such estimators are typically complex. An equal probability sampling design can be implemented for both the entire population or within strata by creating a list frame of all objects\(^8\). Further, if a list frame of all objects is created, it would also be simple to implement STR with the strata based on attributes of the objects.

However, this can be easily overcome by using square areas or segments that are uniform over the study area. This has been extensively used for collecting agricultural statistics in Europe (Gallego, 1995) and other parts of the world and was also used for forest area estimates by Sannier et al. (2014).

An example is shown below in which polygons of Forest are digitised (in yellow in the figure) considering the MMU within the square area (i.e. groups of trees smaller than minimum forest area are not included). The proportion of Forest cover can then be extracted from each square area or segment and compared with the same area proportion in the map to develop a regression estimate.


\(^{8}\) Stratified random samples may be equal probability within strata but probably not across the entire population.
3.5 Type of inference

Design-based inference is the classical approach to inference (Cochran, 1977; Särndal et al., 1992; Stehman, 2000), and the approach exemplified in version 1 and 2 of the MGD (GFOI 2014; 2016). It is an attractive approach in the context of remote sensing-based mapping which often involves a desire to infer collective properties, such as overall map accuracy or the area of a certain land cover class, from a population of all units of a particular map.

By randomly selecting a subset of units from the map (i.e. a sample, typically pixels but not necessarily as explained in Section 4.1) such that reference conditions are observed at each sample location, estimates of parameters such as accuracy and area can be calculated by application of an unbiased estimator to the sample data.

A requirement of design-based inference is that the sample is selected by probability sampling which requires that the inclusion probability is known for each sampling unit, and that the inclusion probabilities are non-zero for all population units (Stehman, 1999). Because the sample is often a small subset of the population, the sample can include many different combinations of selected pixels; if several samples are selected and an estimate is calculated for each one, a range of estimates would be produced. This variability or uncertainty, often expressed as a variance or standard error of the estimate, is easily computed for most sampling designs. Such properties make design-based approaches attractive in a mapping context. Further, implementing a design-based estimation protocol is rather straightforward as illustrated in GFOI (2016). As a result, examples in the remote sensing literature most often use design-based inference.

Other viable alternatives to design-based inference exist such as model-based inference. With this approach, the population is considered a realization of a model instead of a tangible entity. In a remote sensing context, this is not a farfetched idea as one can easily consider a
map to be a single realization of a classifier (i.e. a model). If we are interested in the performance of various classifiers, the assumption that maps are realizations of classifiers is logical and attractive; we would not be interested in the collective properties of one single realization as in the design-based example above. However, this is rarely a desirable objective in the context of activity data and carbon emissions. The typical situation instead involves a single map of activity data that will be used in the process of estimating area and sample-based variability.

There is one situation when design-based inference is not possible: if sample data were collected by non-probability sampling (opportunistic sampling, sampling along roads or rivers, etc.), model-based inference is required. While model-based inference is a viable alternative, there are fewer examples in the remote sensing literature (e.g. McRoberts, 2006). Implementation requires expert knowledge and is recommended only if such knowledge is available. **In the absence of such expert knowledge, selecting a probability sample and implementing design-based inference is recommended.**

### 3.6 Use of auxiliary information

#### 3.6.1 What is the right number and type of strata (use of buffers, pre and post stratification)?

Since the publication of documents like, McRoberts (2011), Olofsson et al. (2013; 2014), Stehman (2013), GFOI (2016), and others, estimation protocols have become common components of most remote sensing-based change studies and national forest monitoring systems. With better guidance and familiarity, the focus has shifted from implementation of protocols to approaches for how to increase estimator precision. **A problem often encountered in efforts aimed at estimating activity data, is the impact of omitted activities present in strata that occupy a large proportion of the study area.** A few sample units in a large stratum, for example stable forest, identified as deforestation in the reference data could have a profound impact on the precision of the area estimator of deforestation.

Omission errors will have this impact when sample units identified as belonging to a stratum \( x \) corresponding to mapped activity data, appear in a stratum \( y \) that is much larger (i.e. \( w_{h=y} \gg w_{h=x} \)) but that was sampled with lesser intensity (i.e. \( w_{h=x}n_{h=x} \gg w_{h=y}n_{h=y} \) ) (Box 3).

**Box 3: Example of the impact of omission errors in relatively small variables of interest.**

The confusion matrix shown below is from a specific example in the Republic of Congo. Contrary to the example shown in Box 1, this confusion matrix shows the pixel counts not the proportions of area. The columns show the reference condition of the sample, while the rows show the classification from the stratification map or forest cover change map. As shown below, a single sample unit of deforestation (column DF) is in the stable forest stratum (row FF).

This combination results in a low proportion of deforestation (i.e. \( 1/512 = 0.00195 \)) within a large weight stratum, leading to very high standard error. The resulting relative margin of error at 95% confidence level of the proportion of deforestation estimated with equation 10 of Olofsson et al. (2014) is 54%. **Had there not been this one sample pixel of DF omission in the map FF stratum, the margin of error would be reduced to 25% at 95% confidence level.** This shows the impact of such a single sample pixel of omission in a large stratum.
This single error of omission carries a large area weight because of the very large size of the FF stratum compared to DF stratum; from Eq. 14 in MGD-2 (GFOI, 2016), the error represents an estimated area, expressed as a proportion the of the total map area, of \(0.93 \times \left( \frac{1}{512} \right) = 0.0018\). Such large errors will decrease map accuracy and precision of area estimates. If a buffer stratum, as explained below, would capture this error, its weight would be greatly reduced. For example, a buffer stratum with a weight of 0.03 would reduce the estimated area of the error from 0.18% to 0.0056% of the total map area.

Table 3: Confusion matrix or proproportion of area matrix of a case in Republic of Congo where DF = Deforestation DG = Degradation FF=Stable Forest NF=Non Forest

<table>
<thead>
<tr>
<th>Reference classification</th>
<th>FF</th>
<th>NF</th>
<th>DF</th>
<th>DG</th>
<th>total</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF</td>
<td>505</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>512</td>
<td>0.93</td>
</tr>
<tr>
<td>NF</td>
<td>13</td>
<td>103</td>
<td>0</td>
<td>3</td>
<td>119</td>
<td>0.03</td>
</tr>
<tr>
<td>DF</td>
<td>39</td>
<td>11</td>
<td>36</td>
<td>14</td>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>DG</td>
<td>56</td>
<td>0</td>
<td>2</td>
<td>42</td>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>Buffer</td>
<td>82</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>Total</td>
<td>695</td>
<td>122</td>
<td>42</td>
<td>72</td>
<td>931</td>
<td></td>
</tr>
</tbody>
</table>

Because omissions of activities in the map tend to appear around areas mapped correctly as activity data, the aim of the creation of a new stratum, \(b\), corresponding to a buffer of one or a few pixels around stratum \(x\) is that it will contain the omission errors. Stratum \(b\) will be small in size relative to the study area and preferably sampled at greater intensity; any omissions of activities that would have appeared in stratum \(y\), but that is now in stratum \(b\) will have a substantially reduced impact on estimates of activity data. This issue also highlights the importance of stratification, sample size and allocation. The more accurate the stratification, the more precise the stratified estimator; a greater sampling intensity in large stratum (i.e. proportional allocation) will benefit estimation of area compared to an equal allocation of the sample; but a proportional allocation will run the risk of not getting a representative sample in smaller strata. If a target standard error and anticipated accuracy of the stratification are known, sample size can be determined using the variance estimator of the selected design solved for \(n\) (e.g. Eq. 5.25 in Cochran, 1997). Drawing upon any past experience for insight into the accuracy of the map used for stratification will allow an investigation prior to sample selection of how sample size and sample allocation will impact estimates.

Examples in the literature of constraining omission errors by buffering are still scarce but a few recent studies are worth noting. Potapov et al. (2017) constructed stratified estimators in a study of the forest loss and gain dynamics in Bangladesh 2000-2014 in support of a national forest monitoring system. Because forest cover loss and gain were just a few percent of the total study area, sample data were collected by stratified random sampling using strata based on a nation-wide map of forest cover change. The stratification included a one-pixel buffer around areas mapped as forest loss and forest gain to reduce the impact of omitted forest disturbance in the map. Arevalo et al. (2018) in a study of activity data (more specifically, conversions between IPCC land categories) across the Colombian Amazon was faced with a
forest stratum occupying 80% of the study area and annual rates of activities of a fraction of a percent. To reduce the impact of omissions of activity data in the forest stratum, a three-pixel buffer around areas mapped as forest to pasture conversion was created and introduced in at the pre-stratification stage.

3.6.2 Country experiences have shown that MGD allocation of sample size works well and that maps are useful to reduce standard errors. A review of completed assessments allows a retrospective assessment of the effectiveness of stratification and sample allocation decisions. Specifically, standard errors of area estimates can be compared for four approaches: (1) the “Dir” approach entails use of simple random sampling and the direct estimator which uses no auxiliary map information; (2) the “Post” approach entails use of simple random sampling and stratified estimators with map-based strata applied subsequent to the sampling, a practice characterized as post-stratification; (3) the “Actual” approach entails stratified sampling using map-based strata with the country’s allocation of sample units to strata followed by use of stratified estimators; and (4) the “Opt” approach is the same as the “Actual” approach except an optimized allocation of sample units to strata is used. With the “Opt” approach, within-stratum sample sizes are estimated using the Neyman optimal allocation formula (Cochran, 1977, p. 99). Of importance, the “Opt” standard error is based on the true but unknown \( p_h \) for each stratum and will be smaller than the standard error realized in practice using estimates of \( p_h \).

The standard error and variance estimators that underpin this retrospective assessment are,

\[
SE(\hat{p}) = \sqrt{V(\hat{p})},
\]

where

\[
V(\hat{p}) = \sum_h W_h^2 p_h (1 - p_h)/n_h,
\]

and \( n_h \) is the within-stratum sample size, \( W_h \) = proportion of area of the population in stratum \( h \), and \( p_h \) = proportion of the sample from stratum \( h \) comprised of sample units in target class (for example, if the strata are map classes, \( p_i \) would be the proportion of the sample from stratum 1 that has the target class as the reference label). Eight country examples of ER programs submitted to the FCPF Carbon Fund are examined, with the focus only on estimating area of deforestation. The eight countries, the number of strata used in the sampling design, the total sample size (\( n \)), the estimated percent area of deforestation, and the lower and upper bounds of a 95% confidence interval for percent area of deforestation are reported in Table 4. The number of strata, sample sizes, and estimates of percent area of deforestation all vary considerably by country.
Table 4: Countries evaluated in retrospective assessment of sampling designs.

<table>
<thead>
<tr>
<th>Country</th>
<th>Change period (years)</th>
<th>Number of strata</th>
<th>n</th>
<th>Area estimate (%)</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower bound</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>13</td>
<td>4</td>
<td>639</td>
<td>1.48</td>
<td>0.26</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>10</td>
<td>4</td>
<td>639</td>
<td>5.60</td>
<td>4.12</td>
</tr>
<tr>
<td>Madagascar</td>
<td>11</td>
<td>6</td>
<td>594</td>
<td>5.21</td>
<td>3.89</td>
</tr>
<tr>
<td>Congo</td>
<td>9</td>
<td>5</td>
<td>931</td>
<td>0.70</td>
<td>0.32</td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>15</td>
<td>8</td>
<td>2600</td>
<td>10.07</td>
<td>9.33</td>
</tr>
<tr>
<td>Vietnam</td>
<td>5</td>
<td>6</td>
<td>536</td>
<td>2.98</td>
<td>2.38</td>
</tr>
</tbody>
</table>

Relative efficiencies (RE, Section 3.2), calculated as ratios of variances, are reported in Table 5 where REs greater than 1 favour the approach expressed by the denominator. For each country, the variances for the approaches are calculated using the same sample size (n) with REs for any two approaches only negligibly affected by the actual sample size.

The general conclusion from the assessment of these eight countries is that with the exception of Ethiopia, the sampling and estimation strategy actually used by countries produced substantially smaller variances and standard errors than the “Dir” strategy (see RE for “Direct/Actual”); thus, using map information via a stratified sampling design was effective in reducing variances and corresponding standard errors. The benefit of the map was usually substantial. For example, RE=1.51 for Costa Rica means that the overall sample size for use with the “Dir” approach which does not use map information would have to be increased by 51% to achieve the same variance or standard error that was actually achieved using the stratified approach that used the map information to facilitate stratified sampling.

The “Post/Actual” RE assesses the impact of using the map-based strata in the sampling design as per the “Actual” approach to control within-stratum sample allocation versus the “Post” approach which uses the map strata only for estimation. With the exception again of Ethiopia, the “Actual” approach was superior to the “Post” approach indicating that use of a stratified sampling design with specific within-stratum allocation of sample sizes produced smaller variances and standard errors than using the strata only for estimation. Predictably, the “Opt” approach yielded variances and standard errors smaller than achieved by the “Actual” approach as indicated by the “Optimal/Actual” RE being less than 1. In practice the reduction in variances and standard errors would not be as great as indicated because the optimal allocation was estimated based on the true $p_h$ values which will not be known in practice. Ethiopia and Vietnam were the countries for which the “Opt” approach would have yielded the most benefit relative to the “Actual” approach. Lastly, the “Post/Optimal” RE indicates the maximum benefit that could be achieved via the “Opt” approach which makes optimal use of the map-based strata in the sampling design relative to the “Post” approach which used the
strata only for estimation. With the exception of Ethiopia and Costa Rica, the maximum potential benefit of optimal allocation is substantial, but whether this maximum can be realized operationally depends on the degree to which the true \( p_h \) values can be predicted prior to sampling. In practice, the predicted \( p_h \) values would likely be the map class proportions, meaning that the degree to which the potential can be realized is directly related to map accuracy.

Table 5: Comparison of approaches using map-based strata for sampling and estimation of area of deforestation.

<table>
<thead>
<tr>
<th>Country</th>
<th>Relative efficiency (RE)</th>
<th>Direct/Actual</th>
<th>Post/Actual</th>
<th>Optimal/Actual</th>
<th>Post/Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethiopia</td>
<td>0.62</td>
<td>0.59</td>
<td>0.55</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Costa Rica</td>
<td>1.51</td>
<td>1.08</td>
<td>0.92</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>Madagascar</td>
<td>1.90</td>
<td>1.19</td>
<td>0.77</td>
<td>1.59</td>
<td></td>
</tr>
<tr>
<td>Congo</td>
<td>2.02</td>
<td>1.46</td>
<td>0.81</td>
<td>1.82</td>
<td></td>
</tr>
<tr>
<td>Ghana</td>
<td>2.46</td>
<td>1.00</td>
<td>0.69</td>
<td>1.44</td>
<td></td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>6.71</td>
<td>1.35</td>
<td>0.88</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td>4.00</td>
<td>1.51</td>
<td>0.44</td>
<td>3.42</td>
<td></td>
</tr>
</tbody>
</table>

Of importance, this retrospective assessment focused on a single parameter, proportion of area of deforestation; different results should be expected for different target parameters.

3.7 Sample size

3.7.1 How should I estimate the sample size to estimate areas?

Practitioners, especially REDD-plus countries, usually estimate the sample size of their STR designs using Equation (13) in the Olofsson et al. (2014). This equation provides a sample size formula for estimating overall accuracy from a stratified sampling design:

\[
 n = \left( \frac{\sum W_i S_i^2}{SE(\hat{\theta})} \right)^2
\]

(4; Equation 13 in Olofsson, 2014)

where \( S_i = \text{standard deviation per stratum} = \sqrt{U_i(1 - U_i)} \) and \( U_i \) is the user’s accuracy of class \( i \), \( W_i = \text{proportion of area in stratum } i \), and \( SE(\hat{\theta}) \) is the standard error of the estimated overall accuracy.

This formula has been applied to some activity data applications. Practitioners should note that Olofsson et al. (2014) presented this formula in the context of applications in which both accuracy assessment and area estimation were equally important objectives. If the dominant objective is estimating area, the use of Equation 4 above (Equation 13, in Olofsson, 2014) should include two modifications.
First, the stratum standard deviations $S_i$ should be based on the proportion of area of the target class in each stratum instead of user’s accuracy. Specifically, if $p_i$=proportion of stratum $i$ that is the target class, then:

$$S_i = \sqrt{p_i(1 - p_i)}.$$  \hspace{1cm} (5)

Second, the standard error of interest is $SE(\hat{P})$ where $\hat{P}$ is the estimated proportion of area of the target class, so $SE(\hat{O})$ should be replaced by $SE(\hat{P})$ in equation (4).

The change in the standard error inserted in the denominator of equation (4) may have a substantial impact on the resulting sample size. For example, when equation (4) is applied to estimating overall accuracy, it would be reasonable to specify a target standard error for overall accuracy of $SE(\hat{O})=0.02$ because we expect overall accuracy to be in the neighborhood of 0.85 to 0.90 (if not higher). However, a standard error of 0.02 for estimating proportion of area of deforestation or degradation would generally not be satisfactory given that the target estimate (proportion of area) may only be 0.01 or 0.02 (even smaller in some cases). The target $SE(\hat{P})$ must align with the magnitude of the expected proportion estimate, $\hat{P}$, so it is likely that $SE(\hat{P})$ could be very small, for example 0.002, if the proportion of area is expected to be just 0.01 or 0.02.

**Box 3: Sample size estimation for the REDD-plus program of Madagascar presented to the FCPF Carbon Fund.**

The following table shows the weights per stratum provided by the forest cover change map, and the proportions for three possible variables of interest: deforestation in modified natural forest (DMNF), deforestation in primary forest (DP) and afforestation/reforestation (AR). These proportions are estimates given by a pilot sampling conducted so as to inform the sampling design.

<table>
<thead>
<tr>
<th>Strata</th>
<th>Weight per stratum (Wi)</th>
<th>Proportion ($p_i$)</th>
<th>DMNF</th>
<th>DP</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deforestation Modified Natural Forest (DMNF)</td>
<td>0.03</td>
<td>0.642</td>
<td>0.053</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Deforestation Primary Forest (DP)</td>
<td>0.06</td>
<td>0.396</td>
<td>0.238</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Afforestation/reforestation (AR)</td>
<td>0.02</td>
<td>0.044</td>
<td>0.000</td>
<td>0.286</td>
<td></td>
</tr>
<tr>
<td>Modified Natural Forests</td>
<td>0.11</td>
<td>0.048</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Non-Forest</td>
<td>0.54</td>
<td>0.022</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Primary Forest</td>
<td>0.24</td>
<td>0.020</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Other</td>
<td>0.00</td>
<td>0.020</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The standard errors of the proportions of each variable of interest are estimated with the formula $S_i = \sqrt{p_i(1 - p_i)}$ described in the previous section. The standard errors for the proportion $\hat{P}$ for each stratum are provided in the following table.
Assuming that the target standard error is 10% of the stratified estimate, i.e. $10\% \cdot \sum W_i \cdot p_i$ and entering the above values in the following formula $n = \left( \frac{\sum W_i \cdot s_i}{SR(p)} \right)$ described in the previous section, the sample size for each variable of interest would be:

<table>
<thead>
<tr>
<th>Strata</th>
<th>$W_i$ (DMNF)</th>
<th>$S_i$</th>
<th>DMNF</th>
<th>DP</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deforestation Modified Natural Forest (DMNF)</td>
<td>0.03</td>
<td>0.4794</td>
<td>0.2240</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Deforestation Primary Forest (DP)</td>
<td>0.06</td>
<td>0.4891</td>
<td>0.4259</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Afforestation/reforestation (AR)</td>
<td>0.02</td>
<td>0.2051</td>
<td>0.0000</td>
<td>0.4519</td>
<td></td>
</tr>
<tr>
<td>Modified Natural Forests</td>
<td>0.11</td>
<td>0.2138</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Non-Forest</td>
<td>0.54</td>
<td>0.1467</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Primary Forest</td>
<td>0.24</td>
<td>0.1400</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.00</td>
<td>0.1400</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

As indicated previously, the sample size needed to achieve the same target standard error, would differ depending on the objective of the sampling, i.e. the variable of interest chosen. Half the sample size needed to estimate deforestation of modified forest would be required to estimate the area of primary deforestation, while one third would be required to estimate the area of afforestation/reforestation. The largest number will be assumed in this case as deforestation in modified natural forest is one of the variables with highest impacts on GHG emissions and removals. The determination of the number of sample size per stratum ($n_i$) will depend on the type of allocation used, proportional or optimal.

<table>
<thead>
<tr>
<th>Strata</th>
<th>$n_i$ (DMNF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportional</td>
<td>Optimal</td>
</tr>
<tr>
<td>Deforestation Modified Natural Forest (DMNF)</td>
<td>64</td>
</tr>
<tr>
<td>Deforestation Primary Forest (DP)</td>
<td>125</td>
</tr>
<tr>
<td>Afforestation/reforestation (AR)</td>
<td>32</td>
</tr>
<tr>
<td>Modified Natural Forests</td>
<td>229</td>
</tr>
<tr>
<td>Non-Forest</td>
<td>1108</td>
</tr>
<tr>
<td>Primary Forest</td>
<td>486</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>2048</td>
</tr>
</tbody>
</table>

In this case, there would be substantial differences depending on the type of allocation. Following the guidance from Olofsson et al. (2014), the optimal allocation is used in this case, and a minimum sample size of 100 per stratum is retained which results in a sample of 2210 points.

It is important to note that the allocation would differ significantly if the variable of interest was the area of afforestation/reforestation. Most of the sample would be concentrated in the AR stratum as there are no observations beyond this stratum. Therefore, the design for DMNF will not be optimal for the estimation of area of AR.
3.8 Repeated/long term monitoring considerations

Combining wall-to-wall monitoring of changes from satellite time series with designed-based sampling to estimate area changes should consider requirements for repeated measurements. If the aim is for annual or bi-annual estimation periods, the areas changing will vary. The sampling design needs to be capable of providing estimates with varying change areas. Fixed sample designs with permanent plots and constant sampling intensities are often quite inefficient when the objective is to estimate change parameters such as area of deforestation, particularly when the area in which change occurs is itself changing. For such applications, a few approaches are available but they are relatively unexplored in the literature.

A straightforward solution is to acquire a sample of reference observations for each time interval for which activity data are required. Such an approach allows for full control of the allocation of the sample to strata to guarantee sufficient statistical representation of activities. The drawback is the cost and time involved in collecting the sample data. An alternative is to select a single sample by stratified random sampling using a stratification of the study area that represents conditions during the whole study period (i.e. between the time represented by the reference level and present). If “continuous” reference observations are collected at sample locations for the study period, such that a sample representing conditions for any time interval can be recreated, estimates of activity data can obtained for any interval using indicator functions (Stehman, 2012). Arevalo et al. (2018) compared the two approaches for estimating biannual activity data in Colombia 2000-2016 and found that selecting a sample every two years gave consistently higher precision in estimates of activity data. A single sample together with indicator functions did not yield estimates that were significantly different from zero for numerous two-year intervals. A third alternative is based on a sampling design featuring a fixed design component with permanent plots augmented with a dynamic design component with temporary plots. In this context, a dynamic design is one for which the locations of the plots change over time to accommodate a specific sampling objective.
Because annual or bi-annual reporting is often required, approaches to estimating activity data at higher temporal frequency are becoming increasingly important and need further investigation. It is an inherently complex task though as the area of activity data tend to be very small relative the study area at annual/bi-annual intervals. Arevalo et al. (2018) for example estimated the annual area proportion of primary forest converted to pastures in the Colombian Amazon to less than a tenth of a percent. Further research is needed to provide guidance on which approach to use in which situation, or if reporting at annual/bi-annual is even feasible. Still, based on experience and sampling theory, we hypothesize that if the additional cost of acquiring a new sample for each reporting period of bi-temporal reference observations of activity data is small relative to continuous observations of already existing sample units, it may be more cost-effective to establish an entirely new sample. A new sample designed according to the current land use situation may offer greater estimation efficiency than repeated observations of existing sample locations, which gradually will become less efficient as the land use changes.

4 Response design

4.1 Sampling unit: provide a list of issues to consider and point at examples – point / polygon / pixel / contextual

4.1.1 Forest definition MMU vs. reference sample interpretation unit
Once the sampling unit and the sampling design are selected, it is important to describe how the forest definition will be applied in practice. A review of existing FRELs, shows that practitioners and researchers have used different sampling units and different operationalization of forest definition. An example of the most common approaches and possibilities is provided below. It simulates various types of sample units with different spatial support in a case where the minimum area of the forest definition is 1 ha.

The most common case is where the sample unit corresponds to the pixel of the stratification map (e.g. a 30m x 30 m pixel or a 100 m x 100 m pixel).

If the sampling unit is a 30m x 30m square (e.g. a map pixel represented in the left as the smaller box), and the classification protocol requires the interpreter to only look the condition within the unit, i.e. spatial support is the 30 m x 30m square, the example to the left (small box) would be classified as forest. Therefore, this approach would overestimate forest cover and deforestation in this case as the patch shown does not have a minimum area of 1 ha.

However, if this same interpretation is based also on
the minimum area of forest, the interpreter would have to confirm that the unit has a minimum area of 1 ha, i.e. spatial support out of the square, and this sample unit would be classified as non-forest.

If a sampling unit of 100 m x 100 m (i.e. 1 ha) is used instead (e.g. a resampled map pixel represented in the left as the larger box) and the spatial support is the 100 m x 100 m square, different rules would have different consequences too, but if a majority rule is applied to the left example, this could be considered as non-forest.

This is the most common example used, but interpretation rules may differ.

In this case the sampling unit would be the point. In order to assess if the condition of the sample unit is forest, the interpreter would use as spatial support the area around the point and assess whether the point has fallen on an object (drawn in dashed lines) that complies with the forest definition.

In the example shown in the left, this would not be classified as forest as the object that intersects with the unit does not have this minimum area.


Another example is where square sampling units are selected using a selected appropriate sample method and forest is visually interpreted and visually delineated within the boundaries of the box using the parameters of the forest definition. In this case a 0.5 by 0.5 km sample unit is shown. This assessment allows to convert categorical data into continuous data for the purpose of estimating forest cover percentage which can then be used together with a GREG estimator explained in section 5 below.

Example: This method was used, but with a 2 km by 2 km unit, as part of the Sannier et al (2016) study in Gabon. Each sampling unit was visually interpreted using all available imagery and auxiliary data by a team of photo-interpreters working independently from the production team but using the same mapping rules and thus forest definition which in this case corresponded to the
following criteria for defining forest land: minimum area of 1 ha, tree crown cover of at least 30%, and minimum potential height at maturity of 5 m. In addition, to these quantitative criteria, the forest in Gabon refers to natural forest formations excluding commercial industrial plantations such as rubber or oil palm but potentially including agroforestry systems and urban tree vegetation if they meet the quantitative thresholds described above.

This assessment allows to convert categorical data into continuous data for the purpose of estimating forest cover percentage which can then be used together with a GREG estimator explained in section 5 below.

In this other case the sampling unit is the point, but a circular plot equivalent to the minimum area of the forest definition (e.g. 1 ha) is used as spatial support. Different rules may be applied to classify the sample such as a majority rule, in which the sample is classified as forest if more than 50% of its area is covered by forest. In this case, the sample would not be classified as forest according to such rule. In some cases this might lead to an overestimation of forest cover, but this should be compensated by other points where there is an underestimation.


In this other case the sampling unit is a point and the spatial support is a 1 ha square centered in it. Usually a grid is placed within a plot, in this case a box, in order to assess the tree canopy cover and the forest cover. This case could also correspond to the first case, if the point is located in the centre of the pixel and the size of the sampling unit is equal to the spatial support of the map.

Different rules are set by countries, for instance in Papua New Guinea’s FREL the example to the left would be classified as forest as at least 30% of the points are located on tree canopy. According to the defined protocols in its submission, if 20% of the points fall on Settlement or Cropland, the sampling unit would be classified respectively as Settlement
or Cropland as a way of moving beyond land cover to land use. Other countries are setting different rules following the guidance from Martinez and Mollicone (2012).

In other cases, such as in the assessment of extent of forests in drylands done by Bastin et al. (2017), the example on the left would be classified as forest as at least 20% of the points would fall on forestland which is characterized by at least a 10% of tree canopy cover.

In other cases, such as in Donoghue et al. (2015), the assessment is based on a majority rule.

In this case the sampling unit is a segment of the map that is selected randomly using sampling methods as described earlier¹⁰.

In the selected segments, a grid of points can be used to assess crown cover. The SEPAL tool applies this methodology. Thus, every segment selected into the sample can be assessed for crown cover in percent. Since a restriction on size (e.g. 0.5 ha) has already been imposed, crown cover% (and height, if that could be assessed) can be used to classify the sample unit as forest/non-forest (F/NF) according to the crown cover requirement of the national forest definition.

In the left, a segment has been selected as sampling unit, and the interpretation shows that 100% of the segment is forest. A regular grid of points can be distributed within the segments to assist interpretation e.g. if the crown cover support classification of forest e.g. > 20%.

The sample units can be used to estimate forest area using a direct sample-based estimator. However, the sample of reference segments can also be used to model F/NF (e.g. using logistic regression) with spectral and contextual metrics from the images as independent variables and then predicted F/NF for all segments in the area. This segment-based prediction can be used in a model-

¹⁰ The images for the area of interest are homogenously segmented with respect to forest cover as much as possible (object based classification). This can be achieved with various supervised or unsupervised methods, for example k-means clustering. A minimum size may be adopted in the segmentation to force all segments to be greater than the defined minimum, for example a size that conforms to the national forest definition (e.g. 0.5 ha). This can be an effective way to provide an areal support that conforms to the national forest definition, which can be difficult to obtain e.g. with pixel-based methods.
assisted estimation of forest area, which is expected to reduce uncertainty.

There can be land configurations where other methodologies may be biased and for which the segmentation-based approach can offer unbiased estimation. For example, in landscapes with long and narrow forest patches, 0.5 ha sample squares (the typical sample size adopted by Sepal) will tend to assign NF to the sample units if the patches are narrower than a width necessary to reach a cover% needed to qualify for F according to the national definition (e.g. 20% cover). However, such patches may very well be >>0.5 ha in size, and a segmentation approach will capture this regardless of the shape of the patches.

As it can be seen in the above table, different sampling units, different support units and different rules may be applied in the labelling protocol, leading to different results for similar cases. Some recommendations could be extracted from the above and previous section:

- In the case of STR, if the same spatial support as the map (e.g. pixels) should be used. Points may be used but won’t provide reliable statistics of map accuracy. Spatial supports that span across different strata should also be avoided, i.e. map of 30m x 30m and plot of 1 ha. Also a finite population approach would impede sampling units overlapping with the limits of the region of interest; such overlap would mean that areas in the boundary would have a different probability of being selected which has implications for the sampling design. For more information on strategies to address this refer to Gregoire and Valentine (2011).

- In the case of SYS, SRS, any sampling unit and spatial support could be used.

- If the intention is to use a GREG estimator (c.f. Section 5), continuous data, not categorical data, would be required. In this case the 3rd and 5th strategies shown above would provide continuous data per sampling unit (i.e. % of forest).

- In any case, classification rules including hierarchical rules should be clearly described and reported. The impact of these rules in the estimates (e.g. overestimation or underestimation of forest or forest change) should also be assessed and reported.

4.1.2 Source of reference data should be of greater quality
In the context of accuracy assessment of land cover and land cover change maps, reference data refer to data that represent the ground reference conditions. This may be achieved through direct observations of ground conditions by field crews or interpretations of aerial photography and satellite data are also used. However, data collection for area estimation through field inventory is often logistical difficult and/or cost-prohibitive and remote areas
may even be inaccessible. For these situations, reference data in the form of interpretations of satellite data and in some few cases airborne orthophotos are often used. The reference data must be of at least the same and preferably of greater quality with respect to both spatial resolution and accuracy than the map data. However, if finer resolution imagery are not available a careful and manually classified sampling unit should be more accurate than an automated classification, even if both are using the same source of data. Human interpretation may bring in information about spatial context and structure which is often difficult to incorporate in automated machine-based methods.

In any case, the quality of the reference data sets should be carefully controlled to ensure the greatest level of quality. As mentioned this can be achieved through the use of finer resolution and more detailed analysis of available source data. There should also be a proper quality assurance process in place to minimise both systematic and random interpreter error through a double interpretation process and calibration at the start of process which can be gradually reduced as the differences between individual interpreters reduce to the point when they can be considered similar.

However, in order to achieve this, it is important that interpreters use all resources available that would give them enough information and context to be able to classify the sample with confidence. For instance, looking at the time series of imagery available in archive (e.g. Landsat archive), even looking at imagery from before or after the period of analysis, should provide in most cases contextual information to be able to classify the sampling units with better accuracy than a map. Figure 5 shows an example of a sampling unit being assessed in the period 2007-2016. According to this example a change detection analysis with Landsat in the period 2007-2016 would not show changes when comparing the first and the last epoch, but when looking to imagery within the period, including Landsat 7 imagery, it can be confirmed that degradation occurred within the sampling unit during the period.

Regardless of the quality of the resources and the experience of the interpreters, some additional uncertainty will result from using interpretations as reference data.

![Figure 5: Comparison of different images in change detection for the period 2007 - 2016](image)

In any case, for area estimation to be valid for an area of interest using the familiar design- or probability-based framework (McRoberts, 2014), the reference data must be collected using a probability sampling design, regardless of how the training data used to classify for example a satellite image are collected.
4.1.3 Labelling protocol (transparent/consistent; land use vs land cover, cover degradation challenges; examples)

It is good practice to establish labelling protocols and include them in a Standard Operating Procedure (SOP). This will allow interpreters to have standardized set of rules for their interpretation, reducing the number of inconsistencies, and it will allow to apply the same rules in future monitoring. Experience, has shown that image interpretation can be a very significant source of error, in many cases larger than the sampling error. Powell et al. (2004)\textsuperscript{11} shows that five interpreters disagreed on the classification of almost 30% of the sampling units, mainly due to the subjectivity of labelling of continuous land-cover types. Powell et al. (2004) recommends that using multiple interpreters to produce the reference data classification increases reference data accuracy.

One important component of these labelling protocols is the definition and the operationalization of the land cover classification system. This is usually done with a set of decision trees that specify objective criteria used by the interpreter to take a decision on the class to assign and that it includes guidance for image interpretation. One important aspect of this is the operationalization of the forest definition and the definitions of deforestation and forest degradation. For instance, some experiences in the Congo basin have shown a large variation amongst interpreters in the classification of forest degradation as a result of lack of consistent definitions. For instance, some interpreters would classify a temporary logging road as forest degradation as they would not consider it as a land use change, while other interpreters would classify it as deforestation as it would be a total loss of forest cover, even if it is temporary.

4.1.4 Importance of QA/QC

It is good practice to implement Quality Assurance / Quality Control (QA/QC) procedures in the phases of design, implementation and analysis. QA/QC procedures contribute to improve transparency, consistency, comparability, and accuracy (IPCC, 2006). Quality Control (QC) is a system of routine technical activities to assess and maintain the quality of the areas that are estimated and it is performed by personnel compiling the inventory, while Quality Assurance (QA) is a planned system of review procedures conducted by personnel not directly involved in the process of estimating areas.

As indicated above, one of the main sources of errors or an issue that could affect the quality of the area estimates is linked to the sampling unit interpretation. Implementation of SOPs is a key QC procedure, but regular training on these SOPs, internal consistency checks or multiple interpretations per sampling unit are also important QC procedures. For instance, in some countries training is done on a weekly basis based on the results of QA procedures. In other countries, three or four interpreters classify each sampling unit and inconsistent classifications are checked by the team of interpreters so as to refine and consolidate the labelling protocols and calibrate interpreters. In other cases, as recommended by Olofsson et al. (2014), interpreters indicate the confidence of their interpretation as a way to flag these confidence

\textsuperscript{11} https://www.oef.berkeley.edu/~matzke(matzke_ev)/pubs/Powell_etal_2004_RSE_error_land-cover_maps_Amazonia.pdf
interpretations and revisit them later. It is important to note, that these low confidence sample observations should not be removed from the sample as this would introduce bias in the estimates; these sampling units should be revisited instead. Also important is to specify what “low confidence” means so as to ensure that different interpreters use it in a consistent way.

Classical QA procedures would be checking 5-10% of the sampling units by an expert interpreter in order to confirm that the SOPs and QC procedures are implemented adequately. This should be done in parallel to the interpretation work and not at the end of the data collection process, so as to identify corrective actions soon in the process and allow the continuous improvement of the system. These are just a few examples of QA/QC procedures. An example of control sampling, a QA procedure, may be found in the Swiss National Forest Inventory Manual12.

5 Analysis

The analysis protocol specifies the measures to be used to express the accuracy and area of the land surface categories of interest. The protocol also specifies the procedures to estimate the selected measures from the sample data. In the context of activity data, the most relevant objective of the analysis is the estimation of the area of each activity. Of importance is the specification of an estimator that corresponds to the sampling design used to collect the sample data.

When a map has been used in the sampling design or when accuracy is an objective of the analysis, a common outcome of the analysis is a cross-tabulation of map and reference labels at sample locations, commonly referred to as a confusion or error matrix. While the construction of an error matrix is not a requirement for achieving common estimation objectives, it summarizes key results and aids the quantification of accuracy and area estimate. The main diagonal of the error matrix highlights correct classifications while the off-diagonal elements show omission and commission errors of the map. An example is provided in Table 5. A map of four categories and a sample of reference observations of the same four categories are available for the same area. At the sample locations, the map labels (rows) are compared to the reference observations (columns) to produce the upper matrix. If the sample data were collected by a stratified design, the area weight of the cross-tabulated sample counts will depend on the strata in which they occur. Thus, a presentation of the error matrix expressed in terms of estimated area proportions is often more informative (lower table). The area estimator will depend on the sampling design; for simple random/systematic designs, the estimated area proportions are simply the sample count divided by the total sample size; for stratified designs, the area proportions need to incorporate the sampling intensity and area weight of strata.

---

12 https://www.lfi.ch/publikationen/publ/LFI2_Methoden.pdf#page127
Table 6. Errors matrices of four categories with map data represented by rows and reference data by columns.

<table>
<thead>
<tr>
<th></th>
<th>Non-forest</th>
<th>Forest</th>
<th>Forest loss</th>
<th>Forest gain</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{13}$</td>
<td>$n_{14}$</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>Non-forest</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_{23}$</td>
<td>$n_{24}$</td>
<td>$n_{2+}$</td>
</tr>
<tr>
<td>Forest loss</td>
<td>$n_{31}$</td>
<td>$n_{32}$</td>
<td>$n_{33}$</td>
<td>$n_{34}$</td>
<td>$n_{3+}$</td>
</tr>
<tr>
<td>Forest gain</td>
<td>$n_{41}$</td>
<td>$n_{42}$</td>
<td>$n_{43}$</td>
<td>$n_{44}$</td>
<td>$n_{4+}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>$n_{+3}$</td>
<td>$n_{+4}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\hat{p}_{11}$</th>
<th>$\hat{p}_{12}$</th>
<th>$\hat{p}_{13}$</th>
<th>$\hat{p}_{14}$</th>
<th>$\hat{p}_{1+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest</td>
<td>$\hat{p}_{21}$</td>
<td>$\hat{p}_{22}$</td>
<td>$\hat{p}_{23}$</td>
<td>$\hat{p}_{24}$</td>
<td>$\hat{p}_{2+}$</td>
</tr>
<tr>
<td>Non-forest</td>
<td>$\hat{p}_{31}$</td>
<td>$\hat{p}_{32}$</td>
<td>$\hat{p}_{33}$</td>
<td>$\hat{p}_{34}$</td>
<td>$\hat{p}_{3+}$</td>
</tr>
<tr>
<td>Forest loss</td>
<td>$\hat{p}_{41}$</td>
<td>$\hat{p}_{42}$</td>
<td>$\hat{p}_{43}$</td>
<td>$\hat{p}_{44}$</td>
<td>$\hat{p}_{4+}$</td>
</tr>
<tr>
<td>Forest gain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$\hat{p}_{+1}$</td>
<td>$\hat{p}_{+2}$</td>
<td>$\hat{p}_{+3}$</td>
<td>$\hat{p}_{+4}$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

Table 6. Errors matrices of four categories with map data represented by rows and reference data by columns.

Once error matrices such as those in Table 6 have been constructed, a variety of area estimators can be constructed using the information in the matrix. Stehman (2013) illustrate the construction of model-assisted and bias-adjusted estimators, stratified and post-stratified estimators, and ratio and simple regression estimators from the information in an error matrix. Producer’s and user’s accuracy of each map category, and overall map accuracy, are also easily estimated using the information in the matrix.

Regardless of the objective of the analysis, a confidence interval should be estimated for any estimate computed in the analysis of the sample data. A confidence interval is often used to express the uncertainty of estimates and is computed using the variance estimator corresponding to the sampling design. Accordingly, a crucial component of an analysis of sample data for activity data estimation is identifying correct and suitable area estimators and variance estimators. Numerous alternatives are available and documented in the literature; Table 6 is provided to guide readers to literature that illustrate the analysis and implementation of estimators in various situations (different objectives, maps, sampling designs, etc.).
Table 7. Literature that exemplify various estimation objectives, sampling designs and estimators.

<table>
<thead>
<tr>
<th>Analysis Objective</th>
<th>Map</th>
<th>Sampling design</th>
<th>Area estimator</th>
<th>Variance estimator?</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation of area of forest loss</td>
<td>One change map with discrete map categories</td>
<td>Stratified random</td>
<td>Stratified estimator</td>
<td>Yes</td>
<td>Box 24, p. 134-138 in GFOI (2016)</td>
</tr>
<tr>
<td>Estimation of area of forest loss</td>
<td>Three F/NF maps; continuous forest change</td>
<td>Two-stage [?]</td>
<td>Model-assisted regression estimator</td>
<td>Yes</td>
<td>Box 25, p. 139-140 in GFOI (2016); Sannier et al. (2014)</td>
</tr>
<tr>
<td>Estimation of user’s, prod.’s and overall accuracy</td>
<td>One change map with discrete map categories</td>
<td>Stratified random</td>
<td>Estimators of accuracy measures</td>
<td>Yes</td>
<td>Section 4.3 in Olofsson et al. (2014)</td>
</tr>
<tr>
<td>Estimation of the area of forest loss</td>
<td>Two F/NF maps; continuous forest change in segments</td>
<td>The US National Forest Inventory (FIA)</td>
<td>Model-assisted regression estimator; and model-based estimator</td>
<td>Yes</td>
<td>McRoberts (2014)</td>
</tr>
<tr>
<td>Estimation of area</td>
<td>One map with discrete map categories</td>
<td>Simple random; Stratified random</td>
<td>Post-stratified; Stratified; Model-assisted post-stratified, difference, ratio and simple regression estimators</td>
<td>Yes</td>
<td>Stehman (2013)</td>
</tr>
<tr>
<td>Estimation of accuracy and area of forest loss</td>
<td>One change map with discrete map categories</td>
<td>Two-stage cluster</td>
<td>Ratio estimator</td>
<td>Yes</td>
<td>Appendix in Potapov et al. (2014)</td>
</tr>
<tr>
<td>Estimation of area of forest degradation and loss</td>
<td>One map with discrete categories based on wall-to-wall lidar data</td>
<td>Systematic stratified; Post-stratification</td>
<td>Stratified estimator; Model-assisted regression estimator</td>
<td>Yes</td>
<td>Naesset et al. (2013)</td>
</tr>
<tr>
<td>Estimation of bi-annual accuracy and area of forest loss</td>
<td>Time series of change maps with discrete map categories</td>
<td>Stratified random</td>
<td>Stratified estimator</td>
<td>Yes</td>
<td>Arevalo et al. (2018)</td>
</tr>
<tr>
<td>Estimation of user’s, prod.’s and overall accuracy; and area</td>
<td>One map with discrete categories</td>
<td>Stratified random with strata being different from the map categories</td>
<td>Estimator using indicator functions; Ratio estimator</td>
<td>Yes</td>
<td>Stehman (2012)</td>
</tr>
</tbody>
</table>
6 Transparency / completeness

One important feature that practitioners should consider is the completeness and the transparency of the reported results. Both principles refer to the provision of supplying enough information to allow the reconstruction of the results by individuals or groups that were not involved in preparing the reported results. Completeness is required for submissions of FRELs to the UNFCCC (CP.19 / Decision12\textsuperscript{13}) whereas, transparency is one of the principles of the 2006 IPCC Guidelines\textsuperscript{14} that compilers of National GHG inventories should follow. Apart of being required under the UNFCCC, transparency is important to allow the replication of the same results by readers or the replication for the same methods in the future, i.e. monitoring.

After careful review of submissions of FRELs made to the UNFCCC and to the FCPF Carbon Fund, submissions in general do not provide enough information on the sampling plans. In most of the cases information was incomplete and did not allow the reader to reconstruct the methods used to estimate areas. This was particularly important in the sampling unit and labelling protocol, as it is not possible to understand how the reference data were collected in practice.

Countries and practitioners should report according to the outline classically presented in sampling designs. Some guidance on elements to report is provided below:

---

\textsuperscript{13} http://unfccc.int/resource/docs/2013/cop19/eng/10a01.pdf?page=34

\textsuperscript{14} http://www.ipcc-nggip.iges.or.jp/public/2006gl/pdf/1_Volume1/V1_1_Ch1_Introduction.pdf?page=7
Sampling design

- **Sampling Plan**: Should explain clearly the approach followed.
- **Parameter of interest**: Should specify the parameters of interest.
- **Sampling unit and sampling frame**: Should indicate clearly what is the sampling frame and unit considered, including the temporal boundaries.
- **Sampling design and estimator**: Should identify the design used, including a figure and table showing sample locations, and the type of estimator used, including the equation to be applied.
- **Use of auxiliary information**: Specify the stratification criteria used.
- **Sample size and allocation to strata**: Present clearly the calculation of the sample size and the allocation to strata.

Response Design

- **Reference data source**: Explain clearly the sources of data used and how different contextual information should be used in order to define the reference condition.
- **Labelling protocol**: Should provide definitions of the land cover classification system and its operationalization. This should include the forest definition and the operationalization of this definition and the definitions of deforestation and forest degradation.
- **QA/QC**: Should provide information on the QA/QC procedures applied, such as cross-validations, training sessions, consistency analysis, etc.

Analysis

- Should include in tabular form each step of the calculations so as to allow its reconstruction.
- In the case of STR, at least,
  - the number of observations of the parameter of interest per stratum,
  - the weight of the stratum,
  - the proportion per stratum,
  - the standard error per stratum,
  - the estimate and its standard error.
7 References


Intergovernmental Panel on Climate Change (IPCC) (2003). Penman J., Gytarsky M., Hiraishi T., Krug,


This module was prepared by:

Andres Espejo – Forest Carbon Partnership Facility World Bank

Carly Green – Global Forest Observations Initiative

Martin Herold – Wageningen University

Maria-José Sanz-Sanchez - BC3-Basque Centre for Climate Change

Inge Jonckheere - Food and Agriculture Organization of the United Nations

Erik Lindquist – Food and Agriculture Organization of the United Nations

Ronald McRoberts – U.S. Forest Service

Erik Næsset – Norwegian University of Life Sciences

Pontus Olofsson – Boston University

Christophe Sannier - Systèmes d'Information à Référence Spatiale, France