Food and Agriculture Organization of the United Nations

White paper

Good practices in sample-based area estimation

AIM4Forests

Cover illustration:
@FAO/Lorenzo Catena

White paper

Good practices in sample-based area estimation

Editors:

Inge Jonckheere

Randy Hamilton

José Maria Michel

Emily Donegan

Food and Agriculture Organization of the United Nations

Rome, 2024

REQUIRED CITATION:
Jonckheere, I., Hamilton, R., Michel, J.M. & Donegan, E., eds. 2024. *Good practices in sample-based area estimation. White paper*. Rome, FAO. https://doi.org/10.4060/cc9276en

ISBN 978-92-5-138531-9

Abstract

Reducing Emissions from Deforestation and Forest Degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries (REDD+), as well as greenhouse gas reporting for the agriculture, forestry and other land use sector, requires land use changes to be characterized to estimate the associated greenhouse gas emissions or absorptions. It is becoming increasingly common to generate these estimates using sample-based area estimation (SBAE). This technique has been widely used in recent years in the generation of activity data – particularly for estimating areas of deforestation – for REDD+ measuring, reporting and verification. However, implementing countries and agencies have repeatedly highlighted the lack of guidance on how to address certain frequently encountered issues with this approach. This paper responds to this need by addressing the most urgent technical issues faced by countries relating to SBAE, such as how to best monitor forest dynamics other than deforestation, how to account for variability between interpreters looking at the same sample unit, how to define the sample unit to use, and how many assessments are needed per sample unit. These issues were identified and prioritized based on a review of country experience and online expert consultations in March 2020. For each issue, a description and recommendations are provided. Existing good practices are consolidated, and new good practices are proposed as solutions where appropriate. The paper also indicates areas for future research which should be pursued to answer the remaining questions surrounding area estimation.

This paper seeks to enable donors, academia, and countries that currently use or want to use SBAE for generating activity data for REDD+ or for other national or international reporting purposes, to delve into current good practice and existing literature, as well as gain a better understanding of the most pressing research needs in the area. The paper moreover will give non-experts an overview of area estimation, as well as its applications and limitations.

Contents

Tables

Figures

Acknowledgements

This paper was written by the following National Forest Management (NFM) team members at the Food and Agriculture Organization of the United Nations (FAO) headquarters: Inge Jonckheere, José Maria Michel, Emily Donegan, Erik Lindquist, Rémi d'Annunzio and Andreas Vollrath. It was also written by Randy Hamilton from SilvaCarbon.

Furthermore, specific sections were written by the following external experts: Frédéric Achard, Charles Scott, Andrew Lister, Ron McRoberts, Steve Stehman, Pontus Olofsson, German Obando-Vargas, Christophe Sannier, Andres Espejo and Paul Patterson.

Editing of this white paper was done by Inge Jonckheere, Randy Hamilton, José Maria Michel, and Emily Donegan.

Copyediting was completed by Alex Gregor and Rachel Golder, with graphic design by Roberto Cenciarelli and Lorenzo Catena. Sara Maulo and Vanessa Vertiz supported the publication process. This paper was produced by FAO thanks to finance from the Global Forest Observations Initiative (GFOI), the World Bank and the Department for Energy Security and Net Zero of the United Kingdom of Great Britain and Northern Ireland.

Abbreviations

eSBAE	ensemble sample-based area estimation
FAO	Food and Agriculture Organization of the United Nations
FREL	Forest Reference Emission Level
GFOI	Global Forest Observations Initiative
IPCC	Intergovernmental Panel on Climate Change
LULC	land use and land cover
MGD	Methods and Guidance Document
NFI	national forest inventory
PSU	primary sampling unit
QA/QC	quality assurance/quality control
REDD+	Reducing Emissions from Deforestation and Forest Degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries
SBAE	sample-based area estimation
SRS	simple random sampling
STR	stratified random sampling
SYS	systematic sampling
UNFCCC	United Nations Framework Convention on Climate Change

Units and formulae

CV	coefficient of variation
h	stratum
ha	hectare
k	kilometre
k²	square kilometre
m	metre
m²	square metre
N	number of sample units
M	number of population units (points) within a sample unit
$\hat{p}$	estimated proportion of area
p_h	proportion of area of stratum (h) that is deforestation based on the reference classification
$\hat{P}_i$	proportion of points in the class of interest on plot
$\hat{P}_R$	estimator of the proportion of the attribute of interest
$\hat{P}_{RS}$	stratified or post-stratified sampling estimator
R	region
SU	sample unit
$V(\hat{p})$	potential impact of omission errors on the variance
$v(\hat{P}_R)$	variance estimator under simple random sampling
$v_s(\hat{P}_{RS})$	variance estimator for the stratified estimator
$v_{PS}(\hat{P}_{RS})$	variance estimator for the post-stratified estimator
W_h	proportion of area in stratum h for the study region

Introduction

Sample-based area estimation (SBAE) has been used increasingly in recent years to produce activity data estimates in the measuring, reporting and verification framework for Reducing Emissions from Deforestation and Forest Degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries (REDD+). However, implementing countries and agencies have frequently encountered issues during the implementation of activity data estimates and have required additional technical guidance to be made available.

In an effort to address the concerns, interviews were conducted with more than 15 country representatives and experts to obtain an initial list of issues. The issues were then grouped into four topics: **general design of monitoring systems, sampling design, response design, and quality assurance/quality control (QA/QC)**. FAO staff provided feedback and an online consultation was conducted by way of a series of surveys among experts in order to prioritize issues within each topic. Finally, three webinars were held to discuss the prioritized issues in further detail. Four issues were discussed in each. Ongoing dialogue with stakeholders and interested experts has led to the development of this white paper.

The objectives of this paper are to:

- highlight key issues that countries have experienced while implementing SBAE;
- consolidate existing know-how and techniques available to respond to those issues;
- propose new good practices of existing approaches (as appropriate); and
- identify opportunities for research and development to address the remaining issues and develop new approaches.

This paper is divided into four main sections, with each section corresponding to one of the four topics identified. Each section is then subdivided into the prioritized issues. In some cases, a co-author contributed an additional section or material not specifically discussed in the webinars. The treatment of each issue is structured in the following way:

- description of the issue with examples (when possible);
- overview of existing good practices that have been implemented or that can be implemented in order to address the issue, including examples and citations of relevant publications;
- new good practices based on expert knowledge and experience (as appropriate); and
- identification of knowledge gaps that need to be addressed by research (as appropriate).

The fifth, final section proposes the way forward. References for each issue can be found at the end of each section.

Section 1

General design of monitoring system

1

1.1 Role of sample-based area estimation in monitoring degradation, reforestation, and afforestation

by

José Maria Michel, Emily Donegan and Frédéric Achard

Description of the problem

The development of SBAE has typically focused on deforestation, and less so on degradation, afforestation and reforestation, and forest carbon stock enhancement (sustainable management of forests and conservation). However, countries often seek to include these other activities in their REDD+ reporting. Sample-based area estimation may be a good approach to quantify the extent of these activities, but some considerations need to be made to adequately measure, report, verify and monitor them.

Forest degradation, as noted by the Intergovernmental Panel on Climate Change (IPCC),[1] is defined in different ways by countries. One definition of forest degradation is that it is a human-induced loss in carbon stock in forestland remaining forestland. The IPCC notes however that current definitions often lack the quantitative and temporal aspects that would be necessary to meet the criteria for definitions under the Kyoto Protocol. Degradation is not an IPCC subcategory but rather a carbon stock change issue, and it is difficult to assess directly (biomass derived from radar and LIDAR sensors) or indirectly (through land cover maps) from remote sensing technology. Existing definitions do not facilitate the localization of degradation across an area of land. In current practice, many countries use proxies to localize degradation, such as the time series analysis of satellite images to estimate variations in tree crown cover, often in combination with field inventory data for validation and verification.

Simply using reduction in tree crown cover as a proxy can be inadequate, as it seems to underestimate degradation (Bullock *et al*., 2020). Degradation can apply to a broad range of different processes including forest fragmentation, selective logging and removal of undergrowth and litter. If the problem of defining and localizing degradation is adequately addressed, the next problem is how to monitor degradation consistently over time. Signs of degradation can be quickly obscured by vegetation regrowth, necessitating frequent (annual) monitoring. In addition, human induced degradation can be confused with natural disturbance processes such as windthrow. The question arises of whether all types of degradation can feasibly be monitored. In obtaining frequent time series data for monitoring, SBAE can increase efficiency, if suitable proxies are available, or if remote sensing data are used as auxiliary data to be combined with forest inventory data. Stratified sampling can also be applied, with stratification by degradation risk, as determined by past trends and proxies such as proximity to settlements and roads or changes in forest vegetation structure as a result of deforestation.

Defining and implementing adequate methodologies to monitor greenhouse gas emissions from forest degradation in a way that is acceptably accurate, easily validated and that does not run into temporal reporting issues can prove challenging. Soon-to-be-available imagery from new LIDAR and radar sensors should help, such as GEDI and BIOMASS (Dupuis *et al*., 2020), but depending on the availability of existing inventory data, may require setting up expensive permanent plots to define backscatter–biomass relationships. Further research is required to determine which remote sensing-based indicators or products are preferable for degradation. Current approaches and methodologies are dependent on definitions and the complexity of the landscape, such as mosaic of agriculture and forest.

Afforestation and reforestation are, by IPCC definition, practices implemented intentionally by people. Unlike forest degradation, the people practicing them would most frequently have no qualms in reporting on these activities. Therefore, extra effort to collect data on the location, timing and area of their occurrence could enable localization and monitoring.

[1] See https://www.ipcc.ch/site/assets/uploads/2018/03/Degradation.pdf

Overview of existing good practices to address the issue

The REDD Sourcebook (GOFC-GOLD, 2016), complemented by the Global Forest Observations Initiative (GFOI) Methods and Guidance Document (MGD) (2020), provides a good overview of methods for monitoring REDD+ activities including degradation. New approaches now exist to more accurately map forests that experience canopy cover disturbance over a reference period, to be used for stratification in the sampling design phase (Shimabukuro *et al*., 2014; Lima *et al*., 2019). Maniatis and Mollicone (2010) provide a framework for forestland classification that includes degradation, conservation, afforestation and reforestation, based on which a national forest inventory (NFI) for REDD+ could be developed. Stratified sampling approaches have been used in the quantification of degradation. A recent example is Bullock *et al*. (2020), quantifying degradation in a region of Brazil, as well as the Forest Reference Emission Levels (FRELs) of Equatorial Guinea and Liberia (Equatorial Guinea, 2020; Liberia, 2019). The use of a systematic grid may be a good approach for multisectoral monitoring because it provides representative coverage of all lands, and a common sampling frame from which to assess many types of variables. It may also help in a first rough assessment of the quality of a forest area change map to be used for stratification. Some temporal changes such as *[forest > shifting cultivation > forest]* may be closer to forest degradation than deforestation, and a systematic approach could help reduce this confusion when deforestation and degradation are hard to distinguish.

New good practices based on expert knowledge and experience

A sampling design based on a combination of a systematic approach, such as using a hexagonal equal area grid, and stratification can provide some flexibility to assess a range of parameters beyond deforestation. This type of multipurpose sampling design has been used in the Global Drylands Assessment (Bastin *et al*., 2017; FAO, 2019) and by FAO for the implementation of the Remote Sensing Survey of the Forest Resources Assessment 2020.[2]

The Democratic Republic of the Congo used a stratified sampling design for the determination of their FREL in the Mai-Ndombe province in the framework of the Emission Reductions Program of the World Bank (FCPF, 2016). About 37 000 sample plots were visually interpreted in Landsat imagery over six dates (over five periods for change assessment) using the following land cover classes: primary forest, secondary forest and non-forest. The visual interpretation of the secondary forest class and changes from and to this class proved to be challenging from Landsat imagery. More robust automated methods combined with reference data from field inventories or derived from very fine spatial resolution imagery are needed in such cases of low-impact selective logging.

What we have learned from national practices is that we can use SBAE to assess beyond deforestation, but it is necessary to have a robust interpretation (response) protocol to assess what the possible transition between a non-degraded forest and a degraded one is. For example, in the Democratic Republic of the Congo,[3] they separate the classes of primary forest and secondary forest and identify the transitions between them. Another approach can be by counting elements (points within the sampling plot), which can be related to canopy cover and its dynamics, as is the case in the Dominican Republic (FCPF, 2019b)[4] and Guatemala (FCPF, 2019a).[5] Another good practice is the example of Nepal,[6] where processes such as fragmentation and edge effects, known to be linked to degradation, are used to assist stratification, informed by landscape ecology and field observations when defining the buffers (FCPF, 2018).

[2] For the methodology of the FRA 2020 Remote Sensing Survey, see http://www.fao.org/forest-resources-assessment/remote-sensing/fra-2020-remote-sensing-survey/methodology/en

[3] See https://www.forestcarbonpartnership.org/system/files/documents/20161108%20Revised%20ERPD_DRC.pdf

[4] See https://www.forestcarbonpartnership.org/system/files/documents/Version%20ERPD%2014-08-2019%20Uncertainty%20correction-Trend%20in%20Ref%20level_rev.pdf

[5] See https://www.forestcarbonpartnership.org/system/files/documents/Guatemala_ERPD_11_05_2019.pdf

[6] See https://www.forestcarbonpartnership.org/system/files/documents/Nepal%20ERPD%2024May2018final_CLEAN_0.pdf

Identify knowledge gaps that need to be addressed by research and development

Guidance needs to be provided on response design for: (i) definition of forest degradation (reduction of tree cover within a forested minimum map unit – what are the criteria?), and (ii) assessment of forest degradation through visual interpretation of change in tree canopy percentage within a sample plot. New remote sensing-based indicators for low impact canopy disturbances in tropical forests are only emerging (Langner *et al.*, 2018) and need more research, in particular for the dryland domain. Upcoming radar sensors or other kinds of very high-resolution imagery should help but will require further research such as setting up permanent forest inventory field plots to define backscatter–biomass relationships.

Further research is needed on how we can ensure that the response design adequately captures degradation.

References

Bastin, J.-F., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C. *et al.* 2017. The extent of forest in dryland biomes. *Science*, 356(6338): 635–638. https://doi.org/10.1126/science.aam6527

Bullock, E.L., Woodcock, C.E. and Olofsson, P. 2020. Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis. *Remote Sensing of Environment*, 238: 110968. https://doi.org/10.1016/j.rse.2018.11.011

Dupuis, C., Lejeune, P., Michez, A. and Fayolle, A. 2020. How can remote sensing help monitor tropical moist forest degradation? – a systematic review. *Remote Sensing*, 12(7): 1087. https://doi.org/10.3390/rs12071087

Equatorial Guinea. 2020. *Forest Reference Emission Level*. REDD+ Web Platform. https://redd.unfccc.int/files/eg_frlsubmissions_2020_01_13.pdf

FAO (Food and Agriculture Organization of the United Nations). 2019. *Trees, forests and land use in drylands: the first global assessment – Full report*. FAO Forestry Paper No. 184. Rome. https://www.fao.org/3/ca7148en/CA7148EN.pdf

FCPF (Forest Carbon Partnership Facility). 2016. *Mai-Ndombe Emission Reductions Program, Democratic Republic of Congo*. Emission Reductions Program Document (ERPD). FCPF Carbon Fund. https://www.forestcarbonpartnership.org/system/files/documents/20161108%20Revised%20ERPD_DRC.pdf

FCPF. 2018. *People and Forests – a sustainable forest management-based emission reduction program in the Terai Arc Landscape, Nepal*. Emission Reductions Program Document (ERPD). FCPF Carbon Fund. https://www.forestcarbonpartnership.org/system/files/documents/Nepal%20ERPD%2024May2018final_CLEAN_0.pdf

FCPF. 2019a. *Guatemala National Program for the Reduction and Removal of Emissions*. Emission Reductions Programme Document (ERPD). FCPF Carbon Fund. https://www.forestcarbonpartnership.org/system/files/documents/Guatemala_ERPD_11_05_2019.pdf

FCPF. 2019b. Dominican Republic. Emission Reductions Programme Document (ERPD). FCPF Carbon Fund. https://www.forestcarbonpartnership.org/system/files/documents/Version%20ERPD%2014-08-2019%20Uncertainty%20correction-Trend%20in%20Ref%20level_rev.pdf

GOFC-GOLD (Global Observation of Forest and Land Cover Dynamics). 2016. *A sourcebook of methods and procedures for monitoring and reporting anthropogenic greenhouse gas emissions and removals associated with deforestation, gains and losses of carbon stocks in forests remaining forests, and forestation*. GOFC-GOLD Report version COP 22-1. Wageningen, Kingdom of the Netherlands, Wageningen University. http://www.gofcgold.wur.nl/redd/sourcebook/GOFC-GOLD_Sourcebook.pdf

Langner, A., Miettinen, J., Kukkonen, M., Vancutsem, C., Simonetti, D., Vieilledent, G., Verhegghen, A., Gallego, J. and Stibig, H.J. 2018. Towards Operational Monitoring of Forest Canopy Disturbance in Evergreen Rain Forests: A Test Case in Continental Southeast Asia. *Remote Sensing*, 10(4): 544. https://doi.org/10.3390/rs10040544

Liberia. 2019. *Liberia's Forest Reference Emission Level submission to the UNFCCC*. REDD+ Web Platform. https://redd.unfccc.int/files/liberia_frel_submission_december_2019.pdf

Lima, T.A., Beuchle, R., Langner, A., Grecchi, R.C., Griess, V.C. and Achard, F. 2019. Comparing Sentinel-2MSI and Landsat 8 OLI Imagery for Monitoring Selective Logging in the Brazilian Amazon. *Remote Sensing*, 11(8): 961. https://doi.org/10.3390/rs11080961

Maniatis, D. and Mollicone, D. 2010. Options for sampling and stratification for national forest inventories to implement REDD+ under the UNFCCC. Carbon Balance and Management, 5(1): 9. https://doi.org/10.1186/1750-0680-5-9

Shimabukuro, Y.E., Beuchle, R., Grecchi, R.C. and Achard, F. 2014. Assessment of forest degradation in Brazilian Amazon due to selective logging and fires using time series of fraction images derived from Landsat ETM+ images. *Remote Sensing Letters*, 5(9): 773–782. https://doi.org/10.1080/2150704X.2014.967880

1.2 Issues related to varying dates and qualities of imagery in sample-based area estimation

by

Erik Lindquist and Frédéric Achard

Description of the problem

Poor availability and quality of temporally and spatially consistent imagery are common issues encountered by countries estimating and reporting on change in area of land use and land cover (LULC) over time. Radiometric anomalies such as clouds and cloud shadow, spatial or spectral inconsistencies such as differing pixel sizes and poor image co-registration, and temporal differences in pixel acquisition dates[7] can all impact the ability to accurately detect, interpret, characterize, and precisely estimate area change. The effects of these characteristics must be accounted for and reduced as much as is practically possible in operational area estimates of land cover. In any area estimation exercise, data quality and availability can affect the stratification, such as classification and map-making, and interpretation of sample units, both of which must be as error-free as possible to obtain accurate and precise estimates of LULC.

When creating national land cover or land change maps, image pixels are generally being selected from varying acquisition dates and composited together to form a representative depiction of the land surface for a given time period. Typically, these composites give the best representation of the land in a unit of time, such as annual composites, and are used in further classification and change detection. The resulting maps are then used to stratify the area of interest for sample unit distribution in stratified area estimation. While the composite is often made up of images from a whole calendar year, for example, reference data collected at each sample unit come generally from a specific date when based on finer spatial resolution imagery (note that also other ancillary data sets might be used as reference data, namely time series indices). Though these dates can differ from sample unit to sample unit, the current best practice is to collect the reference data as close as possible from the year(s) referred to by the change analysis. For example, if analysing change from 2010 to 2015, reference data must be as close as possible to a target calendar date in 2010 and 2015 regardless of the vintage of stratification information.

In all cases, poor stratification, such as a bad map, poor reference data interpretation, and/or poor temporal consistency between stratification and reference data will lead to high variance (in case of poor stratification) or will have deleterious effects on area estimates. Typically, these effects show themselves in large differences between users' and producers' map accuracy, large differences in "mapped" area compared to "estimated" area, and increased confidence intervals about the "estimated" area.

One of the most critical issues to avoid when temporal inconsistencies are present are errors of omission of change within a large stratum of stable land cover. Since sample units typically have a relatively large sample weight in large strata, even a few omission errors of this kind can severely negatively impact the resulting area estimates and confidence intervals (see Section 2.1). This can occur when the multitemporal data used to make the stratification, such as a map, is assembled before or after a land cover change has taken place. The map classification of such an area may appear to be stable. If the change is detected in the reference data used to assess this same area, the reference sample unit will be classified as a change. This will result in an error of omission of change and must be carefully considered when selecting data for both the stratification and reference sample interpretation.

[7] This can be particularly problematic due to seasonality, especially in locations with pronounced dry or wet seasons in which cloud-free images may be more available in the dry season when deciduous or semi-deciduous trees have lost some or all of their leaves. Pixels from different seasons for the same land cover can be vastly different, causing confusion in classification, change detection, and visual interpretation.

In practical terms, these data issues, if not accounted for, can reduce or preclude a country's ability to properly report on verified reductions of emissions due to reduced deforestation and forest degradation, ultimately leading to a loss of credibility and confidence from the international community for these efforts.

An approach to minimize errors in these phases of change detection is illustrated in the following example. Country A is using satellite imagery to assess land cover change between two time periods to report their findings to the United Nations Framework Convention on Climate Change (UNFCCC). The country has decided to take the approach of using a stratified area estimation by creating a single map (for the stratification) representing stable forest, stable non-forest, and forest loss over the time period. The available imagery for Time 1 consists of a mix of fine and moderate spatial resolution imagery with a considerable amount of cloud cover. In Time 2, wall-to-wall fine-resolution imagery with little cloud cover is available, complemented by very fine-resolution imagery in some parts. Country A may decide to expand the resolution imagery in some parts and expand the Time 1 window of acceptable dates from one year to two or more years to achieve full coverage of the area of interest with the same spatial resolution data. If this is done, expert knowledge must be applied and documented to justify the expansion of the Time 1 window (for example, only include pixels from other time periods where change is unlikely to have occurred). In Time 2, for consistency purposes, it is important to use the same spatial resolution as the data used in Time 1 to construct the change map and to enable pixel-to-pixel comparisons with Time 1. The available very fine spatial resolution data should be reserved for reference data collection at the sample units used for area estimation.

Widening the date range for the change map reduces the quality of the stratification with an impact on the results of the final assessment (potentially leading to a larger variance). If we know the areas for which we may guarantee consistency in advance (before the sampling design), we may further stratify the change map by introducing two zones: the same years from those for which we need to widen the date range. The more uncertain strata (within a wide date range zone) would need to be attributed higher sample intensity to guarantee higher accuracy in terms of area estimates.

Overview of existing good practices to address issues caused by varying dates and quality of imagery

One of the simplest ways of reducing errors caused by temporal inconsistencies is to track the date of every pixel used in the satellite image-based stratification and, more importantly, in the reference data. This is especially important where reference data indicate a change in a large stable stratum. It is recommended to use reference data with dates that are as close as possible to the target reference dates. It is important to make sure that reference data come from the specific year or the specific period you are assessing; if it comes from a different year, you are actually assessing the status or change for those other years. If reference data cannot be found close to the target reference dates, then analysing a time series of reference data would be advisable to help determine the land use at the target dates.

Use of time series of satellite imagery

Current good practices for estimating area recommends using all available data to make the best possible interpretation of the sample units. This includes the recommendation to use all available sensors and time series of imagery to provide reference data. Consider that since 2016 with the combination of Landsat imagery (16 days repeat cycle) and Sentinel-2 imagery (five days repeat cycle), it is easier to obtain valid observations (cloud-free) close to the target dates for the parameter to be estimated (annual changes between 1 January and 31 December). Smaller-scale changes, such as logging activities, might be easier to detect from 2016 (see Section 1.3 for related issues). Making use of every available satellite acquisition when determining whether or not a given area has changed is advantageous and should be implemented whenever possible. Incorporating more information helps interpreters.

Using time series visualizations created from Landsat, Planet or Sentinel satellite imagery can aid sample unit interpretation by providing greater contextual information at each site, helping to better understand dynamics of land change. For instance, this can be used to determine if detected changes are simply due to seasonality (Breaks For Additive Season and Trend [BFAST], Landtrendr, or vegetation indices as Normalized Difference Vegetation Index [NDVI]).

In places with wet and dry seasons, a combination of imagery from both seasons is preferred, when possible, for accurate interpretation of tree canopy cover.

A good practice may be to evaluate all the images available in the reference period and not only the images of the start and end dates. This reduces the interpretation errors of change. It allows tackling the issue of seasonality (that can be characterized from a full year of imagery) or other phenomena like El Niño, flooded events, and of degradation due to droughts or fires. Additionally, change events with a short time duration and/or small area impact, such as selective logging in humid forests, may be more detectable using time series information; these events may be only visible over a few months (see Shimabukuro *et al*., 2014). In such cases, it will decrease omission errors in the assessment of such disturbances.

Temporal interpolation for date of imagery

When large sample units are used, continuous parameters (such as proportion of forest cover within a sample unit) can be estimated and interpolated to correct for a range of dates (as a simple proxy for time correction). Linear temporal interpolation has been used for sample units with a size of 10 km × 10 km (Achard *et al.*, 2014) and are used in the PRODES programme for full Landsat scenes (INPE, 2019).

The good practices highlighted in this section are summarized in Table 1.

Table 1 Common issues that can impact the ability to accurately detect, interpret, characterize and precisely estimate area change, along with associated good practices

Type of issue	Description	Good practices
Radiometric anomalies	Clouds and cloud shadows	**Use satellite cloud-free composites.** Prepare composited representative depictions of the land surface for a given time period.
Spatial inconsistencies	Differing pixel sizes or poor image co-registration between the two dates of the assessed period	**Use the same spatial resolution** in Time 1 and Time 2 to construct the change map for stratification. **Check for co-registration issues** and apply adjustments if necessary to avoid false change.
Spectral inconsistencies	Spectral differences caused by seasonality, particularly in areas of pronounced wet and dry seasons (deciduous forest)	**Use imagery for stratification and for reference data from the same season (ideally the wet season).** **Use time series visualizations to understand typical spectral dynamics to aid in interpreting change.**
Temporal inconsistencies	**Temporal differences between the dates of satellite imagery or composites used to create the stratification layers (maps) and the imagery used for the reference data (visually interpreted sample units)**	**Collect the reference data from the same period(s) as the data (map) used to stratify the area.** Alternatively analysing a time series of reference data can help determine the land use at the target dates. For continuous parameters (e.g. proportion of forest cover) linear temporal interpolation can be used to estimate the parameters across time. When a wide range of dates is required to create cloudfree mosaics for stratification, track the dates of the pixel for comparison with the dates of the reference imagery, especially when the reference data detects change in a stable stratum. The stratification map could be further stratified by whether the pixel date is close to or deviates from the target date. **Use all available sensors to increase the likelihood of having reference data close to the target dates.** **Evaluate all images available in the reference period and not only those at the start and end dates to facilitate correct interpretations.**

Source: Authors' own elaboration.

New good practices based on expert knowledge and experience

Use of satellite cloud-free composites

While single date imagery is needed for visual interpretation activities in sample-based monitoring approaches, cloud-free temporal composites are expected to support the production of detailed thematic maps by national institutions, such as tree cover density, annual tree cover change and tree cover disturbance alerting. Satellite cloud-free composites (for example, over an annual period) can be used to produce forest cover maps by national institutions that can be used for stratification in the sampling design.

Two existing Sentinel-2 composite products can be mentioned as examples: the Copernicus Sentinel-2 Global Mosaic product (S2GM) and the JRC Sentinel-2 L1C cloud-free composites product. More recently, Norway has launched an International Climate and Forests Initiative to allow access to Planet's fine-resolution, analysis-ready mosaics of the world's tropics (Norway's International Climate and Forests Initiative Satellite Data Program [NICFI]).

Quality assurance/quality control approaches to improve interpretation quality

Correct interpretation of sample units is critical for obtaining accurate and precise results from any sample-based assessment of area. Therefore, all possible measures should be incorporated to assist interpreters with classification. Dedicated QA/QC measures should be put in place for interpreters, which include: exercises that involve group review and discussion of examples at the end of each data collection session; a clear interpretation hierarchy; independent expert review of a percentage of the sample interpretations; and time allocated for multiple interpreters to independently classify the same location. Consistency in the interpretation across interpreters should also be ensured, and the spatial resolution should be included in the database to analyse the results by taking this information into consideration. A cross-check should also be completed using medium-resolution images also in the case of very high-resolution imagery plots. Such measures will increase the amount of time required for this step of the area estimation process but will help ensure the quality of results and reduce the likelihood of having to revisit this step once the analysis has been completed (see Section 4 for more information).

Identify knowledge gaps that need to be addressed by research and development

Uncertainty: the use of time series information for area estimation will require research and development of statistics suitable to analyse the uncertainty of metrics based on time series.

References

INPE (Brazilian National Institute for Space Research). 2019. Metodologia Utilizada nos Projetos PRODES e DETER (Methodology used in PRODES and DETER Projects). São José dos Campos, Brazil, Brazilian Ministry of Science, Technology and Innovations. http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes/pdfs/Metodologia_Prodes_Deter_revisada.pdf

Shimabukuro, Y.E., Beuchle, R., Grecchi, R.C. and Achard, F. 2014. Assessment of forest degradation in Brazilian Amazon due to selective logging and fires using time series of fraction images derived from Landsat ETM+ images. *Remote Sensing Letters*, 5(9): 773–782. https://doi.org/10.1080/2150704X.2014.967880

1.3 Maintaining consistency as imagery improves

by

German Obando Vargas, Erik Lindquist and Randy Hamilton

It is essential to consider that according to paragraph 11 of Decision 14/CP.19,[8] the reporting entity of the country must maintain "consistency in methodologies, definitions, comprehensiveness, and the information provided between the assessed reference level and the results of the implementation of the [REDD+] activities referred to in decision 1/CP.16, paragraph 70" (UNFCCC, 2014, p. 40).

However, remotely sensed information and the software (algorithms) and hardware used to estimate activity data are continually evolving. Improvements in spatial, temporal, and radiometric resolution mean better and more accurate characterization of the land surface and changes over time. This constant evolution is challenging when meeting operational reporting requirements that specify consistent approaches through time, such as Decision 14/CP.19.

Unfortunately, the use of better spatial and temporal imagery during the reporting periods can negatively affect the accuracy of the country's estimate of emission reductions. In the context of visually interpreted SBAE, if a country assesses more change in recent years because of the increased resolution of available imagery that allows more changes to be observed, the forest emission estimation for the monitoring period is methodologically inconsistent and not comparable with the FREL. The change in imagery makes it impossible to estimate accurate emission reductions for the country.

According to the technical assessment scope in the annex to Decision 13/CP.19, paragraph three,[9] the technical team of experts (TTE) appointed by the UNFCCC Secretariat provides feedback to the countries on areas of further technical improvement: "As part of the technical assessment process, areas for technical improvement [are] identified and these areas and capacity-building needs for the construction of future forest reference emission levels and/or forest reference levels may be noted by the [country]" (UNFCCC, 2014, p. 37). If the technical team of experts identifies technical improvement for activity data estimation, improved imagery must be considered to address activity data issues in the next reference emission levels.

Despite the need to maintain consistency between the reference level and monitoring period, countries can take advantage of imagery improvement in their assessment to understand the actual impact of their REDD+ implementation and ensure the conservativeness of their emission reduction assessment. The following suggestions (summarized in Table 2) are provided for two different types of analysis, based on expert knowledge and experience:

- **Estimating the uncertainty of the LULC maps**. In the case of using maps (and pixel counting) to estimate change, the specialist can use better spatial and temporal resolution imagery to collect reference data to estimate the uncertainty of the LULC maps. Any improvement in the imagery becomes available.[10] For example, analysts can review the portions of the time series interpreted with coarse-resolution images to ensure the time series' consistency with the interpretation of more recent very high-resolution imagery. Especially in those cases where the confidence in the LULC interpretation is not high, the analyst can use improved imagery to revise the class of change assigned in the original analysis. Once the time series analysis is revised, the reporting entity must recalculate the activity data and the FREL to maintain consistency.

[8] See https://unfccc.int/resource/docs/2013/cop19/eng/10a01.pdf#page=39

[9] See https://unfccc.int/resource/docs/2013/cop19/eng/10a01.pdf#page=34

[10] The bias in the interpretation of medium-resolution imagery such as Landsat and Sentinel was consistently greater (8 percent in average) than that observed for very fine-resolution imagery in the Dominican Republic SBAE analysis.

© FAO/Lorenzo Catena

- **Directly estimating areas of change**. In the case of a country using an SBAE approach, practitioners have the following options for using better imagery while still maintaining consistency:

    - **Assess the conservativeness of emission reductions**. The reporting entity must maintain consistency in the minimum map unit and imagery sources used in the reference period through the reporting periods to ensure consistency between the activity data estimates of the reference level and the REDD results. In this case, the practitioner can use better imagery not necessarily to report but to ensure that activity data estimates based on the interpretation of original imagery sources are conservative. For example, the analyst can calculate the bias of the estimate of areas of change with the original imagery using the improved imagery for the monitoring period.

    - **Adjust the time series analysis**. It is considered good practice to track LULCs through time to accurately model carbon fluxes (see Section 2.3). Interpretation of the visually interpreted sample plots through time is therefore essential for the temporal tracking of LULCs. This type of time series analysis can be adjusted when better spatial and temporal resolution becomes available.[11] For example, analysts can review the portions of the time series interpreted with coarse-resolution images to ensure the time series' consistency with the interpretation of more recent very fine-resolution imagery. Especially in those cases where the confidence in the LUCC interpretation is not high, the analyst can use improved imagery to revise the class of change assigned in the original analysis. Once the time-series analysis is revised, the reporting entity must recalculate the activity data and the FREL to maintain consistency. Recalculate the entire time series. When “improved” imagery become available for the whole time series (reference level and monitoring periods), the country can switch to this improved data.

[11] The bias in the interpretation of medium resolution imagery such as Landsat and Sentinel was consistently greater (8 percent in average) than that observed for very fine resolution imagery (VFRI) in the Dominican Republic SBAE analysis.

- **Recalculate the entire time series**. When sources for this purpose would result in a better estimate of the map's uncertainty. However, the same type of imagery should be used to create the maps.

- **Prepare the stratification map for a reporting period**. Practitioners can use better spatial, temporal, and radiometric resolution data to prepare LULC maps to stratify the estimation area.

The following knowledge gaps related to this issue need to be addressed by research:

↘ What are the effects of using different spatial or temporal resolution datasets for stratifying or map-making on sample-based estimates of change?

↘ What are the effects of different spatial, temporal or radiometric resolutions on reference sample information and thus on area estimates and confidence intervals?

Table 2 Use of improved imagery while still maintaining consistency in area estimation

Scenario	Best practices
Sample plots are used as reference data to estimate the uncertainty of wall-to-wall LULC change maps	↘ Use finer spatial and temporal resolution imagery to collect reference data to estimate LULC change maps' uncertainty. Any improvement in the imagery sources would result in a better estimate of the LULC change map's uncertainty.
Sample-based area estimation	↘ Maintain minimum map unit and imagery sources used in the reference period through the reporting periods. ↘ Use better data not necessarily to report but to ensure that estimates based on the interpretation of original imagery sources are consistent and conservative. ↘ When better imagery is available for the entire time series, then a switch to this improved data can be made. ↘ Use improved spatial, temporal, and radiometric resolution data to prepare LULC change maps to stratify the area of estimation for the reporting periods.

Source: Authors' own elaboration.

References

UNFCCC (United Nations Framework Convention on Climate Change). 2014. *Report of the Conference of the Parties on its nineteenth session, held in Warsaw from 11 to 23 November 2013*. UNFCCC. https://unfccc.int/resource/docs/2013/cop19/eng/10a01.pdf#page=39

Section 2

Sampling design

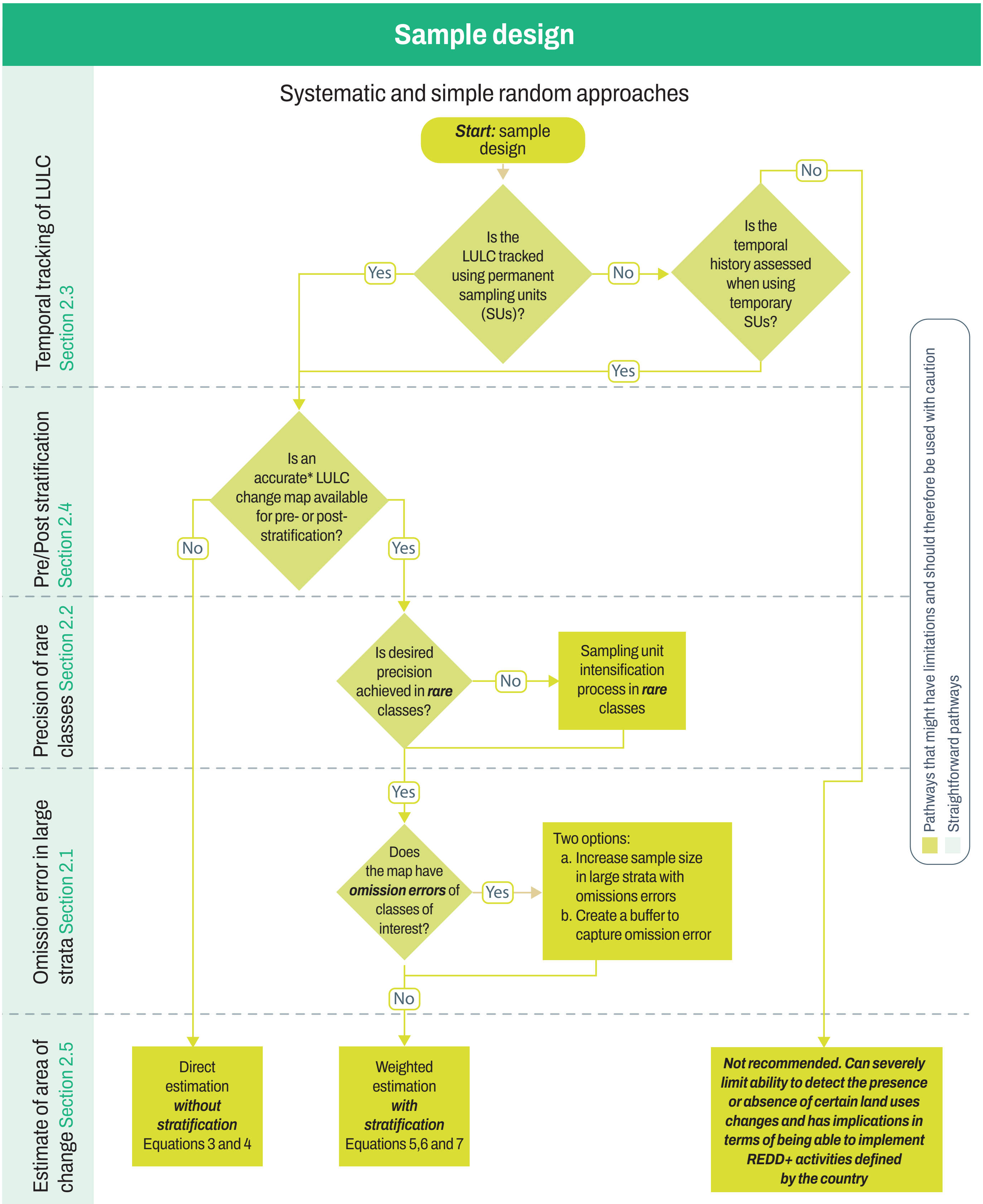
When implementing SBAE, particular attention should be paid to developing a sampling design that achieves the desired outcome in an efficient and cost-effective manner. Over the years, countries have implemented a variety of sampling designs, encountering several challenges along the way. This section attempts to address and provide practical guidance on some of these challenges and associated questions related to sampling design that have arisen from country experiences. In particular, this section presents special considerations and recommendations on:

1. estimation difficulties caused by errors of omission of a rare category of interest, such as deforestation, occurring in large strata (Section 2.1);
2. sampling design of a multipurpose monitoring system (Section 2.2);
3. tracking the temporal history of sample units to correctly identify land use changes (Section 2.3);
4. updating a high-quality base map as source of stratification rather than creating new strata each reporting cycle (Section 2.4); and
5. theoretical and practical implications and differences among two-stage, cluster, and two-step sampling strategies (Section 2.5).

Figure 1 presents a decision tree that shows the interrelationship of these topics and outlines the flow of sampling design decisions.

©FAO/Roberto Cenciarelli

Figure 1 Decision tree of interrelated topics in Section 2 and order of decisions to be made



Notes: Courtesy of German Obando-Vargas.

* An accurate map is considered to be one in which the areas of change are within the confidence intervals of the sample-based estimates from an independent sample. It is recommended that a confusion matrix be constructed and the omission errors evaluated.

Source: Authors' own elaboration.

2.1 Omission errors in large strata in stratified area estimation

by

Steve Stehman and Pontus Olofsson

Description of the problem

Stratified sampling is often implemented for activity data applications, where the strata constructed from a map of the region of interest include the target classes for estimation, such as deforestation or degradation, and one or more strata for the non-target classes. We will focus on the target class of deforestation to illustrate the issues. Map omission errors of deforestation can greatly influence the precision of area estimates produced from the stratified sample data.

Overview of existing good practices to address the issue

Olofsson *et al*. (2020) provide an excellent illustration of this potential impact of omission errors on the variance, $V(\hat{p})$, of the estimated proportion of area, $\hat{p}$ (see key information summarized in Table 3 below). The sample in the example has three strata: deforestation, non-forest, and forest. The impact of deforestation omission errors is clearly demonstrated by stratum 3, the forest stratum, which contributes over 90 percent of the variance of $\hat{p}$ even though the omission error rate of deforestation in the forest stratum is only 0.217 percent (Table 3). Olofsson *et al*. (2020) further noted the strong impact of a single omission error on $\hat{p}$. The estimated percent area of deforestation was 0.42 percent, whereas if the single omission error in stratum 3 had not occurred, the estimated percent area of deforestation would decrease to 0.27 percent – nearly a 33 percent decrease. Even when omission errors occur, the stratified estimator $\hat{p}$ is still unbiased. However, the potential negative impact of omission errors is that the variance of $\hat{p}$ may be severely inflated.

Table 3 Example data illustrating impact of omission errors on the variance of the estimated proportion of area

			Sample size, n_h		Percent of $V(\hat{p})$	
h = stratum	W_h	p_h	Initial	Optimal	Initial	Optimal
1 = deforestation	0.0037	0.72864	199	40	0.6	4.6
2 = non-forest	0.2835	0.00001	211	22	0.2	2.5
1 = forest	0.7128	0.00217	460	808	99.3	92.9

Notes: $V(\hat{p})$ = potential impact of omission errors on the variance. W_h = proportion of area in stratum *h* for the study region, and p_h = proportion of area of stratum *h* that is deforestation based on the reference classification. The "Initial" sample size per stratum n_h is the allocation of the n = 870 sample pixels used by Olofsson *et al.* (2020). In the original data, p_2 = 0, but to avoid having stratum 2 contribute 0 to the variance, p_2 was set to 0.00001 to allow for a small probability that omission errors do exist in this stratum even though none appeared in the sample selected.

Source: Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E. and Sanz, M.J. 2020. Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236: 1–9. https://doi.org/10.1016/j.rse.2019.111492

Stratified sampling will increase precision relative to simple random or systematic sampling (reduce standard errors) to the degree that actual deforestation is concentrated in the map deforestation stratum and nearly absent from the other two strata (the increase in precision is directly related to the accuracy of the stratification map). The effectiveness of the stratification is determined by the proportion of area of deforestation in each stratum, denoted as p_h ($_h$ denotes a stratum). That is, in the deforestation stratum of Table 3, p_1 should be large, and for the other non-forest and forest strata, p_2 and p_3 should be 0. The sample size n_h allocated to each stratum and the proportion of the region of interest covered by each stratum (W_h) are the other factors that determine the variance of $\hat{p}$:

$$V(\hat{p}) = \sum_{h=1}^{H} \frac{W_h^2 p_h (1 - p_h)}{n_h}$$

(Equation 1)

In practice the variance is estimated from the sample data by replacing p_h by the estimated proportion $\hat{p}_h$ and changing the denominator to $n_h - 1$. Omission errors of deforestation would result in $p_h > 0$ for the non-forest and forest strata. Furthermore, because W_h is typically very small for the deforestation stratum and therefore correspondingly large for the other strata, having $p_h > 0$ ($h = 2$ or $h = 3$) when W_h is large will substantially increase the variance.

Two options for diminishing the impact of omission errors are to increase the sample size n_h in the strata with large W_h (modify the sample allocation) or to reduce W_h for those strata that contain omission errors. The second option requires defining an additional "buffer" stratum that is intended to capture omission errors. The specific steps to implement either of these options depend on whether the sample has already been selected or if the choice of sampling design has not been finalized, so both situations will be addressed.

The sample size allocation remedy to mitigate the impact of omission errors is guided by examining the optimal allocation for a set of W_h and p_h values specified for an application, where the optimal allocation is given by Cochran (1977):

$$n_h = \frac{nW_h \; p_h (1 - p_h) C_h^{-0.5}}{\sum_{i=1}^{H} W_i \; p_i (1 - p_i) C_i^{-0.5}}$$

(Equation 2)

In the numerical results presented, the cost of sampling in each stratum (C_h) is assumed equal. If the project is still at the planning stage, it is necessary to hypothesize the values of p_h input to the optimal allocation formula. If the sample has been selected, the estimated p_h from the sample data can be used. In either case, the p_h values can be varied to ascertain the sensitivity of the optimal allocation to the specification of p_h. At the sampling design planning stage, the optimal allocation result could be implemented directly. If the sample has already been selected, then it would be necessary to augment the sample to adjust the allocation closer to the optimal allocation.

New good practices based on expert knowledge and experience

The example data from Olofsson *et al*. (2020) provide a good numerical illustration of the sample allocation decisions. In this example, the sample data have already been collected via the allocation labelled as "Initial" (Table 3). The optimal allocation based on the p_h values estimated from this initial sample is also shown in Table 3. The stratification is reasonably effective as the standard errors for the initial allocation and also a proportionally allocated sample (Table 4) are substantially smaller than the standard error that would have been obtained by simple random sampling (without the strata). Optimal allocation would have achieved an additional 20 percent reduction in standard error relative to the initial allocation. If we sought to improve precision by augmenting the initial sample to more closely reflect optimal allocation, we could increase the sample size in stratum 3, the stratum that dominates the estimated variance. The standard error of estimate for the augmented sample decreases by approximately 0.01 percent for each increment of 100 sample pixels added to stratum 3 (Table 4).

An additional 300 sample pixels added to the initial sample would result in the same standard error of the estimate that would have been achieved by optimal allocation of the original sample, so the 300 added sample pixels would be needed to compensate for the initial suboptimal allocation.

Table 4 Comparison of standard errors of estimates

Design	Standard error (%)
Simple random	0.221
Proportional	0.162
Optimal	0.121
Initial with n_3 = 460	0.155
Initial with n_3 = 560	0.141
Initial with n_3 = 660	0.130
Initial with n_3 = 760	0.121

Note: Comparison of standard errors of estimates obtained using different sampling designs and stratified sample allocations, where the initial sample allocation is augmented by increasing n_3 in increments of 100 starting from the original n_3 = 460 from Olofsson *et al.* (2020) for the forest stratum. The percent estimated area of deforestation is 0.400 percent. Standard errors are computed using values for W_h and p_h shown in Table 3.

Source: Authors' own elaboration.

Augmenting the sample incurs the additional cost of collecting more reference data. Furthermore, the additional sample units must be selected in a manner that maintains the probability sampling feature of the design. If the initial sampling design was stratified randomly, it is straightforward to implement the same stratified protocol to select the additional sample units (Overton and Stehman, 1995). Section 2.2 describes methods for augmenting a stratified systematic sample (systematic sampling within each stratum).

The goal of the option of creating a buffer stratum is to reduce the contribution of variance from the strata containing omission errors. For example, for the strata in Table 3, a new stratum (h = 4) would be created from the elements of forest stratum 3 such that W_4 is small and most of the omission errors of stratum 3 are moved into stratum 4. Because $p_4 > 0$, the new stratum 4 would still contribute variance to $V(\hat{p})$, but the smaller W_4 (compared to the original W_3) should reduce the variance associated with stratum 4 relative to the original stratum 3, and any remaining omission errors in the new stratum 3 would be rare so that p_3 is very small. Since omission errors would more likely be found near areas mapped as deforestation, the new stratum would often be constructed by creating a buffer around the area mapped as the deforestation stratum (thus the name "buffer" stratum).

A buffer stratum could be employed in the sampling design, bringing to bear the issues discussed earlier regarding sample size allocation strata, or in a post-stratified estimator. In the post-stratified application, no new sample units are added, but the stratified estimator would be calculated with the original stratum 3 being split into two strata, the buffer stratum and remaining area of stratum 3. A single buffer stratum cannot be constructed as parts of two or more strata used in the original stratified sample selection. Therefore, it would be necessary to create a buffer stratum within each stratum of the original stratified design for which the concern with omission errors was present. The post stratified approach raises an added caution related to transparency. Often the number of omission errors is very small: perhaps one to three. Therefore, it would be natural to try out a variety of buffer stratum options until one is found that captures all omission errors because this would reduce the sample estimated variance of $\hat{p}$ considerably. However, such an iterative process to define a buffer stratum would lead to bias because the approach is constructed to produce a certain outcome (no omission errors in strata with large W_h). If buffer strata are used in a post-stratification setting, protocols need to be specified and adhered to guarantee that the buffer strata have not been tailored to specific sample data of the application.

Identify knowledge gaps that need to be addressed by research

How best to create buffer strata will be highly dependent on the specific application. Two issues that will inform this process will require additional research. The first issue would focus on creating the maps that identify effective buffer strata (remote sensing issues). The second issue addresses the statistical question of how the trade-offs between area of omission captured by the buffer stratum (p_4 in our illustrative example of Table 3) and total area assigned to the buffer stratum (W_4) impact precision. The buffer stratum is effectively a map classification of deforestation, so users' and producers' accuracy of the buffer classification of deforestation will determine p_4. Therefore, the issue to be resolved is determining the best choice of W_4 to reduce $V(\hat{p})$ given different values of users' (UA) and producers' accuracy (PA) for the buffer stratum.

References

Birigazzi, L., Gamarra, J.G.P. and Gregoire, T.G. 2018. Unbiased emission factor estimators for large-area forest inventories: domain assessment techniques. *Environmental and Ecological Statistics*, 25(2): 199–219. https://doi.org/10.1007/s10651-018-0397-3

Cochran, W.G. 1977. *Sampling techniques, third edition*. Wiley series in probability and mathematical statistics. New York City, USA, Wiley.

Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E. and Sanz, M.J. 2020. Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236: 1–9. https://doi.org/10.1016/j.rse.2019.111492

Overton, W.S. and Stehman, S.V. 1995. Design implications of anticipated data uses for comprehensive environmental monitoring programmes. *Environmental and Ecological Statistics*, 2(4): 287–303. https://doi.org/10.1007/BF00569359

2.2 Sampling designs for multipurpose monitoring

by

Charles Scott, Randy Hamilton, Christophe Sannier and Paul Patterson

Description of the problem

Countries participating in REDD+ must develop a national forest monitoring system (NFMS) to track changes in their forests and the associated greenhouse gas emissions. However, they are increasingly encouraged to recognize the value of developing multipurpose monitoring systems to satisfy other national and international purposes (FAO, 2017, 2018, 2020). Examples of other monitoring needs may include agricultural crop surveys, characterizing trees outside of forests, watershed management and urbanization.

The diverse data requirements of these monitoring needs can affect the timing and type of data collected as well as the sampling design used to collect the data. For example, an urban planner may only need a single-date map, while greenhouse gas reporting requires change estimates (no map required) at a specified frequency. REDD+ reporting focuses on characterizing rare occurrences of change while the agriculture sector may need data on the current extent of a common crop. Because of these differences, it can be difficult to design a monitoring system that is optimal for all purposes. However, there can be significant efficiencies gained and costs saved by developing a monitoring system that satisfies many of the needs. A unified monitoring system can also avoid the frustrations of having multiple systems that produce incompatible or inconsistent results. For example, reporting for REDD+, greenhouse gas inventories, and nationally determined contributions (NDCs) require similar or overlapping data that should be consistent.

There are many types of multipurpose monitoring systems, such as map-based or sample-based using either observations collected in the field or visually on imagery. This section focuses specifically on the use of visually interpreted, sample-based monitoring systems for multiple objectives beyond REDD+; however, such systems may be designed to integrate with other types of monitoring systems or to accommodate the collection of other types of data, such as field-based data, using the same sampling frame.

Here, area and area change estimates are assumed to be sample-based and to be determined based on visual interpretation of fine-resolution imagery, but could also involve ground observation, such as crop surveys. In some cases, maps created using medium-resolution imagery are used as a source of stratification for sample-based estimation. In multipurpose sampling, there is often interest in taking ground-based measurements on a subsample of the imagery to further increase the accuracy of the estimates and/or to measure variables which cannot be observed in imagery, such as species, understory vegetation, crop yields or soil types. When assessing area change, ideally finer-resolution imagery would be used that is of the same dates as was used to create the area change map. To assess map accuracy or estimate the area of classes of interest, a probability sample of the study area (population) that was stratified using a map must be performed by human interpretation using finer spatial resolution data. The following describes several probability sampling methods which are applicable for multipurpose monitoring.

Overview of existing good practices

There are several steps and important considerations to be followed when designing multipurpose monitoring systems. A key first step is to identify intersectoral information needs. Applications such as the Design Tool for Inventory and Monitoring (DTIM) can help identify these needs (Köhl *et al.*, 2009). Beyond identifying information needs, other steps and considerations include developing the sampling design, response design (see Section 3), and determining whether and how to integrate with other types of monitoring systems, such as mapping or NFIs. This section will focus on the pros and cons of different probability sampling methods for the purpose of monitoring beyond REDD+ using an SBAE framework.

These designs are described in some detail in Boschetti *et al*. (2016), GFOI (2018, 2020), and more generally in Cochran (1977). This section will also touch on considerations for integration with other types of monitoring systems.

Stratified random sampling

Stratified random sampling (STR) of a population can be conducted by selecting sample units on a map to draw samples within each stratum (map class) to estimate the true area of each class more accurately, especially in the case of rare classes such as change. The sample can be allocated in several ways, such as proportional (to area), Neyman (proportional to variability), and optimal (also considering cost). For REDD+ and some other applications where interest is in the class estimates and not just the overall mean, Olofsson *et al*. (2014) suggest having a minimum sample size in all strata (such as 10 to 20), then allocating the remaining sample units proportionately to the rest of the strata. This ensures an adequate sample size to meet precision requirements for rare strata, such as areas of change. These sample units (plots) are then distributed either randomly or systematically within each stratum. On each plot, either a single class is assigned or the proportion of each class falling within each sample plot is observed using fine-resolution imagery or field measurements for one, two or more dates of interest. The STR estimators are used to estimate the area of each class and their variances (GFOI, 2020;[12] Cochran, 1977[13]) (see also Section 2.5 of this document).

Simple random sampling versus systematic designs

Stratification often requires the use of a map and is not necessarily the most efficient method for a multipurpose approach. In fact, it may be difficult to: (i) find a stratification that would fulfil all the objectives of the multipurpose survey, and (ii) optimize the sample allocation. As an alternative, plots can be randomly or systematically located with the same intensity across the entire population of interest. The main advantage of this approach is its simplicity because the only requirement is to identify the entire population which, in this case, is the geographical area over which estimates should be produced. The main drawback of both simple random sampling (SRS) and systematic sampling (SYS) designs is that some portions of the population (rare classes) may not be adequately sampled. Systematic approaches overcome potential geographic distribution problems (such as dense clusters or major gaps) that can occur with SRS but can pose a problem if the landscape features cyclical patterns that match the grid interval. To resolve this issue, it is a good practice to combine both approaches by using SRS within grid cells (Sannier *et al.*, 2014).

Sample units can be defined as points or areas (see Section 3.1) for any of the sampling designs. In a systematic approach, sample locations can be determined using a regular grid with a random origin and orientation. Alternatively, the grid typically defines square or hexagonal cells (that better ensure all are of equal area across very large geographic areas). Randomly placing plots within the cells is preferred to using grid corners if the survey is associated with permanent ground samples to avoid the potential of land managers using information about the fixed grid to determine where the plots are on their land and treat them differently, such as not harvesting so they can increase REDD+ payments (Kohl *et al.,* 2015).

For both SYS and SRS, either a single binary class value or the proportion of each plot by area class is observed and the sample mean is normally used for estimating the area of each class and its variance. Systematic variance estimators could be applied and typically result in smaller variance (Strand, 2017) but can be more complicated to apply.

While equal probability sampling methods are robust for measuring a variety of attributes on sample units of interest over time, they are not as efficient as STR for single or a few correlated attributes (especially those that are rare). However, post-stratification can be used to improve the precision for an estimator with a subset of attributes. This is frequently accomplished by overlaying the population with a map of the area classes of interest or of strata which are homogeneous with respect to the attributes. Each plot is assigned to a single map class – a process known as post-

[12] See Equation 6 and Equation 7.

[13] See Equation 5.1 and Equation 5.1.

stratification. Using finer-resolution imagery or field sampling, either the land class is assigned to the entire plot or the proportion of each plot falling into each class is measured. Post-stratification estimators are then applied to improve the precision of the estimates (GFOI, 2020)[14] (see Section 2.5).

Several practitioners have contributed informal tools that correspond with some of the methods outlined in this document. These tools have been assembled into an area sampling toolkit that can be used to plan and analyse monitoring activities.

Intensification of a permanent sample

Each of the aforementioned designs has its advantages. However, especially for rare classes, the desired precision requirements may not be achievable with a simple random or systematic approach. Instead, a combined approach can be used to intensify a systematic or random set of permanent plots.

Starting with the equal probability sample, the first step is to use a suitable stratification map and calculate post-stratified estimates and the precision of the estimates. If the precision of the post-stratified estimates is sufficient, then the task is complete. Otherwise, the sample size needs to be increased.

The following steps describe the process of computing the additional number of sample units required in each map class based on the precision requirements of each land class:

1. Compute the stratum variances for each land class using the base set of plots assigned to the stratum. Variability occurs when a stratum has more than one land class (from finer-resolution imagery or ground data). The approach to stratification used by Olofsson *et al.* (2020) (see Section 2.1) can be applied to further improve precision by reducing omission errors.

2. Using the desired precision level for each land class, calculate the overall sample size needed using Cochran (1977).[15]

3. Allocate the total sample size to each map class using Cochran's Equation 5.26 for each land class.

4. Find the maximum sample size for each map class across all land classes.

5. The intensification sample size is the map class's maximum sample size minus the number of base grid plots that fall in the map class.

6. The intensified plots should be distributed in the same manner as the initial plots (SRS or SYS) within each map class.

A spreadsheet tool was developed to estimate sample sizes in this intensification case where interest is in one or more classes. The tool analyses the initial data to estimate the sample sizes to meet precision requirements for each land class then optimally allocates the plots to the strata. The tool can also be used to evaluate the optimal number of points per plot. This tool and others are available in the area sampling toolkit.

While randomly locating the additional sample points is straightforward (as described in Section 2.1), intensifying a systematic sample requires more thought. One approach is to subdivide the existing grid by the appropriate factor to achieve the desired sample size, such as doubling by cutting each grid cell in half, then choosing points in the unsampled grid cells. Alternatively, a second grid can be randomly placed on the stratum to select the desired number of additional plots. The grid cell size is computed as the stratum area divided by the number of additional samples needed. When done this way, the permanent and intensified plots can be treated the same for estimation purposes.

[14] See Equation 10.

[15] See Equation 5.47.

If the intensification is applied to a certain subset of (rare) categories where the precision needs to be improved, then the efficiency can be increased by first evaluating whether the intensified plot contains information on the rare categories. If not, there is no need for a detailed analysis of the plot and it can be classified as a zero for the class of interest; otherwise, a detailed analysis of the plot is conducted (the proportion of the plot in the class or classes of interest are measured).

Pros, cons and good practices

Stratified random sampling is very efficient for achieving precise estimates by area class as well as achieving overall estimates, but not so much for one attribute (or perhaps a few correlated attributes) for a single date or between two points in time. In addition, the map used for stratification needs to be sufficiently accurate for the stratification to be efficient, as this can be problematic for rare classes such as change classes when there are substantial omission errors in the map used. Olofsson *et al*. (2014, 2020) proposed a method to mitigate this problem (see Section 2.1). Systematic (or simple random) sampling treats all areas equally, so it is robust for observing all attributes of a population which is sampled for a range of resources and for changing objectives over time. Often such plots are permanent for improved precision of net change as well as the ability to estimate gross change (observing both the original and final area class, such as for deforestation) for any two points over time (see Section 2.3). National forest inventories, for example, typically use permanent plots for these reasons. Typically, post-stratification starts with an equal probability sample and is followed by stratification (which is the reverse order of STR). Like STR, post-stratification improves the precision, however it does not allow for different sample size allocation to further improve precision. A different map can be used for different subsets of attributes, such as for estimates of area change (activity data), cropland area, or NFIs. This could prove to be an advantage over STR by providing more precise estimates for a wider range of attributes for multipurpose monitoring. Intensification of a permanent sample combines a base sample grid with a subsequent stratified sample to intensify the sample in specific strata (map classes), thus incorporating the benefits of each of the designs. However, it also comes with added complexity and is still only most efficient for the attributes most correlated with the stratification chosen. Nevertheless, different post-stratifications of the original base grid can easily be tested to see if they result in improved estimates for other attributes.

Unlike ground sampling, area change can be estimated using new sample plots due to the long time series of remotely sensed data available. Thus, any of the sampling designs described can be used effectively for area change (see Section 2.3). However, if any permanent area and ground plots are used, then any new stratification must be performed on them as well.

Considerations for integration with other monitoring systems

Depending on a country's information needs, a truly robust multipurpose monitoring system may require the harmonious integration of several distinct monitoring systems such as a field-based NFI or mapping system. In this case, the monitoring system based on visually interpreted sample units should be designed to integrate with these other systems. Such integration can generate synergistic information that is of greater value than that produced independently by each system. For example, if NFI field plots are co-located with a subset of permanent visually interpreted plots, the visually interpreted data can be used as a source of post-stratification for the field plots to improve the precision of the estimates obtained from the field. Land use and land cover maps could be generated or validated using the visually interpreted data. In this case, common land use and land cover classification systems should be used for both systems. Also, special consideration should be given to the size of the plots and the number of points used within the plots to estimate the class proportions (see Section 3.2). Integration with other types of monitoring will require other considerations and adaptations.

Costa Rica's National Land Use, Land Cover, and Ecosystems Monitoring System (SIMOCUTE) is a good example of an integrated multipurpose monitoring system (Molina-Murillo, 2020).[16] It harmoniously integrates visually interpreted SBAE, using a systematic grid of sample plots, with the country's NFI and mapping programs using common land use and land cover classification systems. Since this system covers all lands, it can generate information for REDD+ as well as for the forestry sector in general, the agricultural sector, and others (Scott *et al.*, 2009).

In Europe, the Land Use and Coverage Area frame Survey (LUCAS) is a multipurpose survey carried out every three years covering crop types as well as other land uses. It is a two-stage survey based on a systematic grid every 2 km over the entire European Union. A stratification is developed based on visual interpretation of aerial photography to select a sample of points in each selected stratum that are then surveyed in the field (Gallego and Delincé, 2010).

Future considerations

Since the wide availability of finer-resolution imagery and of the relevant tools with which to rapidly interpret such imagery is relatively recent, much can still be learned about how best to utilize them. Since maps take time to create, one possibility for rapidly estimating change is to use double sampling for stratification (Cochran, 1977). In the first phase, create two strata by visually assessing many plots, such as from a dense systematic grid, on fine-resolution imagery as to whether or not they have changed. Then draw a small subsample (second phase) of the "no change" stratum (the set of no change plots). Using the same imagery, assign detailed land classes to each of these plots, which may be clusters of points. In the "change" stratum, draw a large or complete sample of plots. In a similar fashion, assign detailed classes to all points within each of these plots. Many of the points will have changed, but some will not. The estimates by class from the two strata can be combined to provide efficient estimates of change, because there are no omission errors of change in the "no change" stratum and the sample size for change will be large to satisfy precision requirements. Conceptually, this should allow for a rapid yet precise approach to estimating the area of all stable and change classes. This approach has not yet been tested.

This approach could also be applied when starting with a base grid of permanent plots. These plots could be visually interpreted into the detailed stable and change classes (as in the second phase above). The variability of each of the key classes of interest could be used to compute the sample sizes needed to meet their precision requirements, such as for deforestation, degradation and afforestation. For each class, the ratio of the required sample size to the number of plots in that class in the base grid is the intensification factor. Compute this for all classes of interest, then apply the maximum of the intensification factors to determine the dense first phase grid described above for the first phase sample. The dense grid should be constructed so that the base grid is a subset (as described in the intensification of permanent plot section above). While each of these two approaches will help ensure that the precision requirements for all classes are met, this approach also provides a means to determine how dense the first phase grid should be.

[16] See https://simocute.go.cr

References

Boschetti, L., Stehman, S.V. and Roy, D.P. 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. *Remote Sensing of Environment*, 186: 465–478. https://doi.org/10.1016/j.rse.2016.09.016

Cochran, W.G. 1977. *Sampling techniques, third edition*. Wiley series in probability and mathematical statistics. New York City, USA, Wiley.

FAO (Food and Agriculture Organization of the United Nations). 2017. *Voluntary guidelines on national forest monitoring*. Rome. https://www.fao.org/3/I6767e/I6767e.pdf

FAO. 2018. *Strengthening national forest monitoring systems for REDD+*. National Forest Monitoring and Assessment Working Paper No. 47. Rome. https://www.fao.org/3/ca0525en/CA0525EN.pdf

FAO. 2020. Global Forest Resources Assessment 2020 (FRA 2020). Rome. https://doi.org/10.4060/ca9825en

Gallego, J. and Delincé, J. 2010. The European land use and cover area-frame statistical survey. In: Benedetti, R., Piersimoni, F., Bee, M. and Espa, G., eds. 2010. *Agricultural survey methods: based on papers presented at the 1998, 2001, 2004 and 2007 International Conferences on Agricultural Statistics*. Chichester, Wiley.

GFOI (Global Forest Observations Initiative). 2018. *Summary of country experiences and critical issues related to estimation of activity data*. GFOI. https://www.fao.org/gfoi/news-events/news/detail/en/c/1159538/

GFOI. 2020. *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, third edition*. Rome, FAO. https://www.fs.usda.gov/rm/pubs_journals/2020/rmrs_2020_espejo_a001.pdf

Köhl, M., Scott, C.T., Lister, A.J., Demon, I. and Plugge, D. 2015. Avoiding treatment bias of REDD+ monitoring by sampling with partial replacement. *Carbon Balance and Management*, 10(1): 11. https://doi.org/10.1186/s13021-015-0020-y

Molina-Murillo, S.A., ed. 2020. Sistema Nacional de Monitoreo de Cobertura y Uso de la Tierra y Ecosistemas [National System for Monitoring of Land cover and Use and Ecosystems]. *Ambientico*, 273. https://www.ambientico.una.ac.cr/wp-content/uploads/tainacan-items/31476/33819/273.pdf

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A. 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148: 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E. and Sanz, M.J. 2020. Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236: 1–9. https://doi.org/10.1016/j.rse.2019.111492

Sannier, C., McRoberts, R.E., Fichet, L.V. and Makaga, E.M.K. 2014. 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. https://doi.org/10.1016/j.rse.2013.09.015

Scott, C.T., Bush, R. and Brewer, K. 2009. Design tool for inventory and monitoring. In: McWilliams, W., Moisen, G. and Czaplewski, R., eds. *Forest Inventory and Analysis (FIA) Symposium 2008*. Fort Collins, USA, USDA Forest Service, Rocky Mountain Research Station. www.researchgate.net/publication/228725906_Design_tool_for_inventory_and_monitoring

Strand, G.H. 2017. A study of variance estimation methods for systematic spatial sampling. *Spatial Statistics*, 21: 226–240. https://doi.org/10.1016/j.spasta.2017.06.008

2.3 Temporal tracking of land use in the context of temporary versus permanent sample units

by

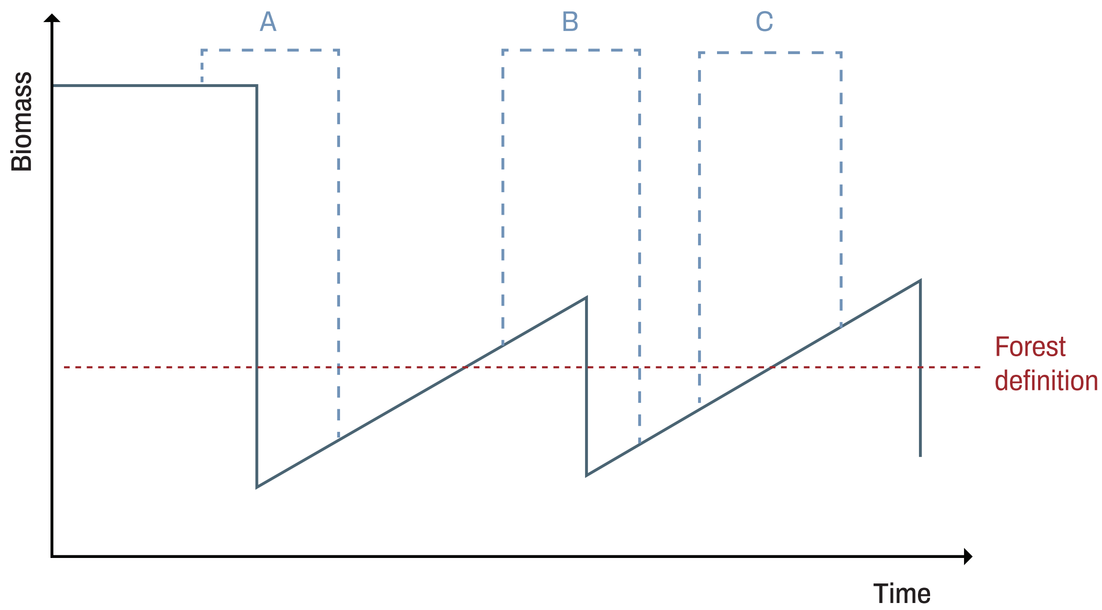
Andres Espejo and Randy Hamilton

Stratified random sampling has frequently been used to estimate the areas of change during a given monitoring period as this approach is a statistically efficient way to estimate the area of change. However, usually STR is implemented in successive monitoring periods, in the form of independent surveys (new sample units are drawn in each survey so they are temporary).

Unfortunately, the use of independent surveys and temporary sample units does not enable the consistent and explicit tracking of land use spatially and temporally (IPCC Approach 3). This might have serious implications as to our ability to detect actual land use change or enable the implementation of the definitions of REDD+ activities defined by the country. For example, trees can be removed or reappear in the landscape without a change in land use, which occurs in the case of forest plantations, slash-and-burn agriculture, or logging in natural forest. If independent surveys and temporary sample units are used, the changes in these systems could be mischaracterized as land use change or any one of these changes could be identified as deforestation, forest degradation or enhancement of carbon stocks, depending on the timing of survey (compare with Figure 2).

The use of independent surveys and temporary sample units could also lead to "double detection" of land use transitions if the timing of assessments happens to coincide with consecutive cutting or regeneration cycles of a forest plantation or the clearing (compare with Figure 1), such as regrowth and clearing cycles that are common in secondary forests in Central Africa.

Figure 2 Characterization of a unit of land subject to conversion from natural forest to a cyclical system



Note: The same unit of land could be classified as deforestation (A, B) or enhancement (C) depending on the part of the system cycle covered by the monitoring period. In this specific case, the same unit of land could be classified as deforested twice, which depending on the country's definition of deforestation, might cause a "double detection".

Source: Author's own elaboration.

The lack of temporal tracking can also make it difficult to accurately model carbon fluxes. For instance, in enhancement of carbon stocks, it is important to know the age of a regenerating forest to be able to apply growth models and correctly estimate carbon fluxes. Two-date temporary sample units lack the ability to determine forest age using imagery.

The challenges presented above can be overcome through three approaches:

- **Temporal history of temporary sample units**: Assessing the history of each sample unit in the years prior to the monitoring period maximizes the likelihood of correctly determining whether a land use change has occurred or not. It also maximizes the probability of and being able to effectively apply the REDD+ definition adopted by the country. For instance, Indonesia defines deforestation as the conversion of natural forest cover to other land cover categories (plantation forest or non-forest lands) that occurred once in an area, meaning that deforested areas that might regenerate and again meet the forest definition were not taken into account a second time in the emission calculation from deforestation.[17] If independent surveys and independent sample units are used in successive surveys, there is the risk of a "double detection" as previously described. In order to resolve this, the strategy followed with maps of masking out areas previously reported as deforested may be replicated by ensuring that the interpreters look into the past history of the sample unit and ensure that no losses have occurred previously.

- **Permanent sample units**: An alternative to the historical review of temporary sample units is to use an SYS or SRS of permanent sample units. This ensures the temporal tracking of the units through time. However, systematic sampling is not as efficient for characterizing small areas of change and could require a very large sample to achieve the desired precision for each variable of interest, such as area of deforestation. Alternatively, the precision could be improved by post-stratifying a base systematic sample and, if necessary, adding additional temporary sample units into areas of change or other areas of interest (see Section 2.1 and Section 2.2). The temporal history of these new sample units would need to be assessed as previously described.

- **Permanent sample units with stratification**: Another possible solution, which is a combined approach of using a systematic grid of permanent sample units and stratified estimation with temporary sample units, is to implement a variation on the approach that Suriname has implemented (Suriname, 2018). Suriname created a high-quality base map that was then updated with cumulative change (including all areas that have been restored or deforested progressively since the beginning of the reference period). The initial stratified sample units would become permanent and new sample units (that could be interpreted for the whole time series and become permanent) would be added to the initial sample in new areas of change that appear. The use of cumulative stratification and permanent sample units would enable the temporal tracking of units through time with a lower sampling intensity. This task could become complex over time, as the number of strata would increase progressively, which would complicate the estimation process (see Section 2.4).

One last consideration regarding the use of temporary and permanent sample units is the ability to characterize the age of regenerated forests which might be required to estimate removals from enhancements of carbon stocks. The age of regenerated forests can be determined by looking at the history of the sample units and assigning an age class to the sample unit (first approach above); however, in order to obtain reasonably precise estimates of forest in the different age classes, a sufficient number of sample units would be required for each age class, which might be complex. The second approach would present this same limitation in terms of efficiency, but the third approach could be used efficiently to estimate areas of different age classes.

[17] See https://redd.unfccc.int/files/frel_submission_by__indonesia_final.pdf#page=14

References

Suriname. 2018. *Forest Reference Emission Level for Suriname's REDD+ Programme*. Paramaribo, REDD+ Suriname, National Institute for Environment and Development in Suriname. https://redd.unfccc.int/files/2018_frel_submission_suriname.pdf

© UN-REDD/Cory Wright

2.4 New stratification versus updating a base map for stratified area estimation

by

Inge Jonckheere, Randy Hamilton and Rémi d'Annunzio

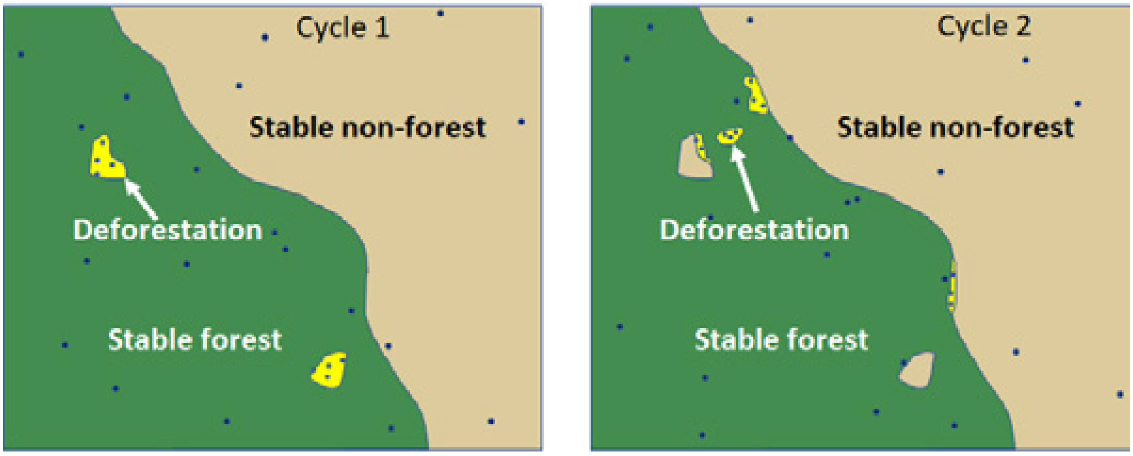
Some countries create a new stratification map for each reporting interval (with a new set of sample units), while others have created a very high-quality base map that is updated through time. In the latter case, new sample units are added in areas of change and existing sample units are reevaluated. Both options are possible: here we will discuss the pros and cons of these two approaches and the use of temporary versus permanent sample units.

Although this is a high-priority topic for countries, there is neither much available guidance nor much existent scientific literature. Clearly both complementary and new guidance is needed based on more recent research in this field.

Based on current literature, it appears that either approach could be effective. However, the cost of creating an updated map versus the cost of creating a new stratification map is a relevant consideration. The temporary versus permanent sample unit question depends on the desired objectives (see Section 1.2). If the objectives require a time series of observations at the same locations, then it is necessary to use permanent plots. On the other hand, if the objectives are a series of separate estimates for multiple change periods, such as a series of biannual estimates, then having new strata for each change period is probably desirable.

Both options are graphically explained in Figure 3 and Figure 4.

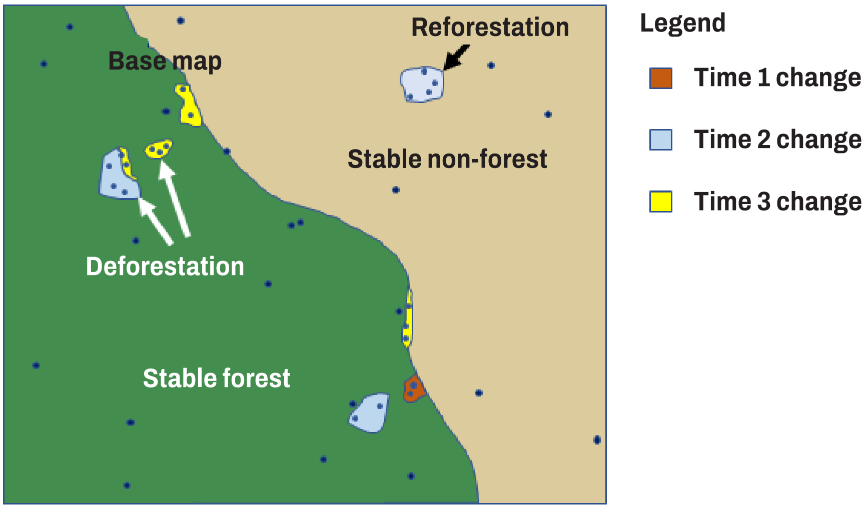
Figure 3 Illustration of a traditional stratified random sample with temporary sample units



Note: New sample units are drawn from new change maps each reporting cycle. Sample intensities typically vary between strata to optimize sampling efficiency.

Source: Authors' own elaboration.

Figure 4 Illustration of a stratified random sample with base map updated through time and sampl using permanent plots



Note: The original plots are reinterpreted as well as all new plots added to change areas.
Source: Authors' own elaboration.

The example of the Suriname FREL 2018 shows the update of a very high-quality map in practice (Suriname, 2018).

There are other possible alternatives for the future, also including deep learning and upcoming artificial intelligence technologies (Lang, 2019; Araza *et al.*, 2020). For example, the use of Sentinel-2 (or other high-resolution data) in conjunction with airborne LIDAR or GEDI (Dubayah *et al.*, 2020; Bruening *et al.*, 2023) may make automated algorithms more trustworthy than visual interpretation (Dubayah *et al.*, 2020). If normal optimal or other base maps are still intended to be used, one could also use these algorithms as the reference samples.

References

Araza, A., De Bruin, S., Herold, M., Quegan, S., Labriere, N., Rodriguez-Veiga, P., Avitabile, V. *et al.* 2022. A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps. *Remote Sensing of Environment*, 272: 112917. https://doi.org/10.1016/j.rse.2022.112917

Bruening, J. M., May, P. B., Armston, J. D., & Dubayah, R. O. 2023. Precise and Unbiased Biomass Estimation From GEDI Data and the US Forest Inventory. *Frontiers in Forests and Global Change*, 6, 1149153.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M. *et al.* 2020. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing*, 1: 100002. https://doi.org/10.1016/j.srs.2020.100002

Lang, N. 2019. *Deep learning reconstructs forests in 3D using only satellite images*. In: *EcoVisionETH*. Cited 17 September 2020. https://medium.com/ecovisioneth/mapping-forest-structures-with-deep-learning-c1d3c1b41e4e

Suriname. 2018. *Forest Reference Emission Level for Suriname's REDD+ Programme*. Paramaribo, REDD+ Suriname, National Institute for Environment and Development in Suriname. https://redd.unfccc.int/files/2018_frel_submission_suriname.pdf

2.5 Estimators for sample-based area estimation in finite and infinite populations

by

Paul Patterson, Andrew Lister and German Obando-Vargas

In SBAE, sample units are typically selected using an SYS, SRS, or STR design. When the sample units represent a physical area (plot), another SYS or SRS of points (or pixels) within each of the sample units is often interpreted to characterize the proportion of land use, land cover, or change within the plot. The reader should not confuse sample units with population units; population units in this situation are the pixels or points within the sample unit (plot). Area estimation using this type of subsampled sample unit (plot) is the focus of this section. Typically, a probability-based survey provides a sampling strategy; the sampling strategy includes the population paradigm (such as finite or infinite), a probability-based sampling design (such as SYS, SRS, or a more complex sampling design), and the choice of an appropriate estimator (an equation) of the attribute of interest that aligns with the population paradigm and sampling design (Särndal *et al.*, 1992). The statistical properties of the estimators follow from the sampling strategy.

For area estimation where there are sample units (plots) and a subsample within each of the sample units, there are three theoretically different sampling strategies that are used in literature: two-stage, two-step, and cluster. The two-stage and cluster sampling strategies can be found in any text on survey sampling, while two-step was proposed recently (see Annex I or Patterson [2012]). The differences in the sampling strategies are not in the sampling of the sample units (plots) or in the subsample of the sample units, but in the population paradigm and how the plot is defined. Interestingly, all three strategies have the same estimator (mean proportion of the attribute of interest) despite their theoretical differences. The variance estimators (equations) are also the same. Note the variance estimator for SRS is also commonly used as the variance estimator for systematic sampling (Bechtold and Patterson, 2005).

This section will present the shared estimator of the proportion and variance estimator for the three sampling strategies when either there is an SRS, SYS or STR of the sample units. The presentation includes an intuitive explanation of the estimators (equations). Key considerations arising from the theoretical differences between the three sampling strategies will then be discussed (for more information on theoretical differences and why they matter, the three sampling strategies are described and compared in more detail in Annex I).

Estimator and variance estimator

The estimator of the proportion and variance estimator are provided first for either SRS or SYS of the sample units (plots). This is followed by the estimator of the proportion and the variance estimator for STR of the sample units and post-stratified estimation (Patterson, 2012). Note that general terminology familiar to practitioners is used to describe the estimators here. However, each strategy has its own unique terminology that is presented in Annex I.

There is a region, R, and an attribute of interest, such as: a land use such as forest or cropland; a land cover such as a tree, shrub, or grass; a change such as conversion from forest to cropland; or cropland to secondary forest. The objective is to estimate the proportion of the attribute of interest within the R and denote the proportion by P_R. There is an SYS or SRS of sample units from R. The sample size is n. For each sample unit, a systematic or simple random subsample of points (also called population units) is drawn; for this explanation, it is assumed that the number of points, m, is constant. For s_{ij}, the jth point within the ith sample unit, the variable y_{ij} should have the value of 1 if s_{ij} intersects the attribute of interest and zero otherwise. The estimator of P_R is denoted $\hat{P}_R$ and the equation is:

$$\hat{P}_R = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} y_{ij}\right)$$

(Equation 3)

The subequation, $\frac{1}{m}\sum_{j=1}^{m} y_{ij}$, is the proportion of points within the ith sample unit that intersect the attribute of interest (an estimate of the proportion of the attribute of interest in the ith sample unit). The proportions are then averaged over the n sample units to obtain an estimate over R of the proportion of the attribute of interest.

The variance estimator is denoted by $v(\hat{P}_R)$ and the equation is:

$$v(\hat{P}_R) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{P}_i - \hat{P}_R\right)^2$$

(Equation 4)

$\hat{P}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$ is the estimate of the proportion of the attribute of interest within the ith sample unit. While this variance estimator is unbiased for SRS, it is also commonly used as the variance estimator for SYS (Bechtold and Patterson, 2005).

For both STR and post-stratified estimation, the estimator of the proportion is the same; however, the variance estimators are not. The variance estimators are presented later. If there are H strata, where the weight (proportion) of each stratum is W_h and the number of sample units is n_h, then the estimate of the proportion of the attribute of interest is:

$$\hat{P}_{RS} = \sum_{h=1}^{H} W_h\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\left(\frac{1}{m}\sum_{j=1}^{m} y_{hij}\right)\right)$$

(Equation 5)

This is simply a weighted sum of the estimated proportion within each stratum; a subscript S has been added to denote either a stratified or post-stratified estimator, with the estimator within each stratum being the simple random sample or systematic estimator previously discussed.

For STR, where each stratum is sampled with either a simple random sample or systematic sample of the sample units followed by a subsample of points within the sample units (plots), the variance estimator is again a weighted sum of the variance estimator within each stratum. The variance estimator is denoted by $v_S(\hat{P}_{RS})$, with the subscript S for stratified. The equation is:

$$v_S(\hat{P}_{RS}) = \sum_{h=1}^{H} W_h^2\left(\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(\hat{P}_{hi} - \hat{P}_h\right)^2\right)$$

(Equation 6)

$\hat{P}_{hi} = \frac{1}{m}\sum_{j=1}^{m} y_{hij}$ and $\hat{P}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\left(\frac{1}{m}\sum_{j=1}^{m} y_{hij}\right)$

For post-stratified estimation, where the sample of the population is either a simple random sample or a systematic sample of sample units (plots) followed by a subsample of the points within the sample units, the variance estimator is again a weighted sum of the estimated variance within the strata. However, the weights are adjusted to account for the sample size within strata being unknown until the sample is drawn. The variance estimator is denoted by $v_{PS}(\hat{P}_R)$ with the PS for post-stratified.

The equation is:

$$v_{PS}(\hat{P}_{RS}) = \sum_{h=1}^{H} \left(\frac{W_h}{n} + \frac{1 - W_h}{n^2}\right)\left(\frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (\hat{P}_{hi} - \hat{P}_h)^2\right)$$

(Equation 7)

The $\frac{W_h^2}{n_h}$ from the previous equation $\left(\frac{W_h}{n} + \frac{1-W_h}{n^2}\right)$ to account for the sample size within strata being unknown until the population sample is drawn; $n = \sum_{i=1}^{H} n_h$ is the total sample size (see Section 5.9 of Cochran [1977]).

The above equations use a fixed size, m, for the subsample of the points within the sample units. However, this does not always have to be the case. For example, to reduce costs, the United States Forest Service Image-based change estimation project (a forest monitoring project that relies on sample-based interpretation of fine-resolution imagery) uses the protocol that if a sample unit contains change, then 49 points are photo-interpreted; if there is no change, then five points are photo-interpreted (Megown *et al*., 2015; for the current version of the protocol see USDA [2019]). The estimators given above can also be extended to the case where m = 5 or 49. The extension is valid because the factor that determines m is a property of the sample unit (whether change occurs) and not of the sample.

Practical considerations

As previously mentioned, there are three theoretically different sampling strategies (two-stage, two-step, and cluster) that ultimately all use the same form of the estimator of the proportion and variance estimator. Ultimately, the choice of a sampling strategy rests with the user's preferences. However, the theoretical differences and unique assumptions among the three strategies have some practical implications that should be considered. Key implications are briefly reviewed in this subsection and more details can be found in Annex I. Since many forest monitoring projects are constructed to provide information for international reporting, carbon market participation, or emissions reductions incentives programs, it is important to understand and clearly articulate the selected sampling strategy. The more detailed discussion on the sampling strategies in Annex I is provided as a reference to which the interested practitioner can refer when documenting their work and justifying their choice of sampling strategy. Table 5 presents an overview of terminology and a summary of key differences among the three strategies that are explained in Annex I; the associated estimator previously described in this section are also consolidated into this table for the practitioner's convenience.

As mentioned, the theoretical differences and unique assumptions among the three sampling strategies have some practical implications. One of the theoretical differences that merits particular attention is the issue of sample unit crossing strata boundaries. When either stratified or post-stratified sampling is used with any of the three sampling strategies, each sample unit is assumed to be in one and only one stratum. However, it is easy to find examples where sample units straddle stratum boundaries. For two-stage and cluster sampling strategies, this problem cannot be easily overcome without ignoring this assumption. However, for a two-step sampling strategy, which assumes an infinite rather than a finite population, the sample unit value is collapsed (assigned) to the centre point of the sample unit and hence is in one and only one stratum. Therefore, the straddling of stratum boundaries is not an issue for the two-step strategy (see Annex I). Practitioners who use a two-stage or a cluster sampling strategy often assign the stratum encountered at an arbitrary point within the sample unit, such as the centre point, to the sample unit in both ground-based and image-based forest inventories. Although this violates a fundamental assumption, the impact is assumed to be minimal. The authors are unaware of any studies that evaluate this.

Table 5 Summary of the terminology, concepts and estimator equations for the two-stage, two-step and cluster sampling strategies

Parameter		Sampling strategy		
		Two-stage sampling (finite population)	Single stage or cluster sampling (finite population)	Two-step sampling (infinite population)
Region	R	Region of interest tessellated by a finite number of sample units	Region of interest tessellated by a finite number of sample units (clusters)	Region of interest consisting of an infinite population of points
Number of sample units	N	$N = \text{area}(R)/a_{SU}$, where a_{SU} = the area of each sample unit; there is a finite population of sample units	$N = \text{area}(R)/a_{SU}$, where a_{SU} = the area of each sample unit; there is a finite population of sample units (clusters)	There is an infinite number of points, with a potential *sample unit or support region* centred at any point
Number of population units* (points) within a sample unit	M	$M = a_{SU}/a_M$, where a_{SU} = the area of an sample unit, and a_M = the area of a population unit; each sample unit is tessellated into M units, with a sample of size m drawn	$M = a_{SU}/a_M$, where a_{SU} = the area of an sample unit, and a_M = the area of a population unit; each sample unit is tessellated into M units, with each of the M units measured	An infinite number of points (population units) exists within each sample unit or support region, with a sample size of m drawn
Estimator of the proportion of the attribute of interest	$\hat{P}_R$	$\hat{P}_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i$		(Equation 3)
Stratified or post-stratified sampling estimator	$\hat{P}_{RS}$	$\hat{P}_{RS} = \sum_{h=1}^{H} W_h \left(\frac{1}{n_h}\sum_{i=1}^{n_h}\left(\frac{1}{m}\sum_{j=1}^{m} y_{hij}\right)\right)$		(Equation 5)
Variance estimator under **simple random sampling**	$v(\hat{P}_R)$	$v(\hat{P}_R) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\left(\frac{1}{m}\sum_{j=1}^{m} y_{ij}\right)_i - \hat{P}_R\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{P}_i - \hat{P}_R)^2$		(Equation 4)
Variance estimator for the **stratified** estimator	$v_s(\hat{P}_{RS})$	$v_s(\hat{P}_{RS}) = \sum_{h=1}^{H} W_h^2 v(\hat{P}_h) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\hat{P}_{hi} - \hat{P}_h)^2\right)$		(Equation 6)
Variance estimator for the **post-stratified** estimator	$v_{PS}(\hat{P}_{RS})$	$v_{PS}(\hat{P}_{RS}) = \sum_{h=1}^{H}\left(\frac{W_h}{n} + \frac{1-W_h}{n^2}\right)\left(\frac{1}{(n_h-1)}\sum_{i=1}^{n_h}(\hat{P}_{hi} - \hat{P}_R)^2\right)$		(Equation 7)

Note: * Population units refer to the units within the sample units, which are referred to as "points" in the body of this section. However, in two-stage and cluster strategies, these population units are two-dimensional in nature (i.e. have area); nevertheless, they are still often referred to as points. In a two-step strategy, the population units are dimensionless points.

Source: Authors' own elaboration.

Another difference among the three strategies has to do with the shared assumption that the sample units (plots) should tessellate the population. Using two-stage or cluster strategies, it is not always possible to do this in either ground-based or image-based forest inventories. While it may be impossible to tessellate the population using these two strategies, the assumption is that the impact on the estimator, variance and variance estimator is minimal. The authors are unaware of any studies that evaluate this assumption. By comparison, a two-step strategy will always perfectly tessellate the region and avoid this theoretical complication.

Another benefit of the two-step strategy is that it can be used to give guidance (as described in Section 3.1) on issues that are not as easily addressed using the other strategies. Although there are several benefits to the two-step sampling strategy, it can seem complex, at first, to many practitioners since most have been formally trained in finite sampling theory (two-stage and/or cluster) and delving into the infinite population theory (two-step) may require additional study.

One final theoretical consideration relates to the distribution of the points within the sample units (plots). In both the two-stage and two-step sampling strategies, the second stage or second step sample of the points is assumed to be a simple random sample from the sample unit. In many applications of SBAE, a fixed systematic sample is used for the sample of the second stage or second step. This is also a common practice in forest inventory. In fact, the location of the centre points of the sample units is usually based on a systematic sample. The usual argument is that even though the systematic sample is fixed, it is one of many possible systematic samples so it should be treated as a random systematic sample as opposed to a fixed sample. Additionally, a simple random sample variance estimator is generally used and is considered to be a conservative estimate of the variance. The cluster sample strategy treats the points as a fixed systematic sample. The authors recommend a simulation study be conducted to evaluate the validity of this argument for the case of considering the fixed systematic sample of points to be a random systematic sample.

References

Bechtold, W.A. and Patterson, P.L. 2005. *The Enhanced Forest Inventory and Analysis Program: National Sampling Design and Estimation Procedures*. Asheville, USA, USDA, Forest Service, Southern Research Station. https://doi.org/10.2737/SRS-GTR-80

Cochran, W.G. 1977. *Sampling techniques, third edition*. Wiley series in probability and mathematical statistics. New York City, USA, Wiley.

Megown, K., Lister, A., Patterson, P., Frescino, T., Jacobs, D., Webb, J., Daniels, N. and Finco, M. 2015. Image-based change estimation (ICE): monitoring land use, land cover and agent of change information for all lands. In: Stanton, S.M. and Christensen, G.A. 2015. *Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) Symposium 2015*. Portland, USA, USDA, Forest Service, Pacific Northwest Research Station. https://www.fs.usda.gov/research/treesearch/50336

Patterson, P.L. 2012. Photo-based estimators for the Nevada photo-based inventory. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station. https://www.fs.usda.gov/rm/pubs/rmrs_rp092.pdf

Särndal, C., Swensson, B. and Wretman, J. 1992. *Model assisted survey sampling*. New York City, USA, Springer-Verlag.

USDA (**Department of Agriculture of the United States of America**). 2019. *Forest inventory and analysis national image change estimation, version 3.0*. Internal report. Salt Lake City, USA, Forest Service.

©FAO/Roberto Cenciarelli

Summary of Section 2

by

German Obando-Vargas and Randy Hamilton

Section 2 addresses a number of sampling design questions that various countries around the world have raised in the context of implementing SBAE. Table 6 summarizes key considerations and recommendations highlighted throughout the section.

Table 6 Summary of key considerations and recommendations related to sampling design for sample-based area estimation

Consideration	**Sampling design**		
	Systematic sampling (SYS)	Simple random sampling (SRS)	Stratified random sampling (STR)
Precision obtained *(Section 2.1 and Section 2.2)*	SYS and SRS sample all areas equally (i.e. equal probability of selection) and are not optimized for any particular category. As a result, the precision obtained from SYS and SRS (particularly for rare classes such as deforestation) is frequently lower than that obtained from STR. Post-stratification of SYS or SRS can improve the precision of the estimates; however, it does not allow for different sampling intensities to further improve precision. Combining the original sample with a new STR sample, using a high-quality map as source of stratification, can significantly improve the precision by intensifying the sample where needed.		STR involves pre-stratifying the sample and is a very efficient way to achieve precise estimates of area by class and overall. However, the map used for stratification needs to be sufficiently accurate to achieve this efficiency. Poor map accuracy can be particularly problematic for rare classes (e.g. change classes) when there are substantial omission errors in the map.
Suitability for multipurpose monitoring *(Section 2.2)*	Because SYS and SRS treat all areas equally, these designs are robust for sampling a variety of attributes for a range of resources and accommodate changing objectives over time, which can be ideal for multipurpose monitoring. Because all areas have equal probability of selection, statistically valid estimates can be derived for any particular area of interest (e.g. provinces). In addition, any number of different maps can be used for post-stratification to improve the precision of the estimates according to any particular user's interests (e.g. to estimate cropland, forest, wetland or other areas). SYS and SRS are limited in their ability to precisely estimate areas of rare classes due to low sample size in these classes. This can be overcome if the overall sample size is *very* large or if the original SYS or SRS sample is combined with a new STR sample that together with the original sample achieve the required sample sizes in all strata. In terms of multipurpose monitoring, SYS and SRS designs could prove to be advantageous over STR because they can provide estimates for a wider range of attributes.		Typically, in STR, the sample is pre-stratified and allocated unequally among strata to optimize sampling efficiency for particular categories of interest. Therefore, STR would not necessarily be efficient for a multipurpose approach because it would be very difficult to: (i) find a single stratification that would fulfil all the objectives of the multipurpose survey; (ii) use any other sources of stratification for other purposes; and (iii) optimize the sample allocation for different objectives.

Consideration	Sampling design		
	Systematic sampling (SYS)	Simple random sampling (SRS)	Stratified random sampling (STR)
Ability to track land use and land cover between monitoring periods *(Section 2.3)*	SYS and SRS sample units are often permanent, which facilitates reinterpretation through time and easy temporal tracking of LULCs. However, if surveys are independent and use temporary sample units, the land use and land cover cannot be tracked through time at the plot level, making it difficult to accurately model carbon fluxes. This can be addressed through special accommodations such as: (i) assessing the history of each temporary sample unit in the years prior to the monitoring period, or (ii) converting the temporary sample units to permanent units and assessing them through time.		STR typically involves an independent stratified sample for each monitoring period. Because of this, the land use and land cover cannot be tracked through time at the plot level, making it difficult to accurately model carbon fluxes. To compensate for this lack of temporal history, the history of each temporary sample unit in the years prior to the monitoring period should be assessed.
Possibility to apply special adjustments to obtain desired precision in rare classes *(Section 2.2)*	***Yes***. SYS and SRS are not optimized to achieve high precision in rare classes. However, the precision can be improved. One option is to increase the overall intensity of the sample to increase the sample size in the rare classes, but this is often very costly. A high-quality map can be used to post-stratify the data to improve the precision of the estimates; however, this is generally inadequate. The most efficient option is to increase the sample size where needed by combining the original sample with a new STR sample, using the high-quality map for pre-stratification.		***Yes.*** Additional sample units can be added (sample unit intensification) in key strata**.**
Special adjustments needed to address omission error in large strata *(Section 2.1)*	Errors of omission are a characteristic of STR samples and do not apply to SYS or SRS unless some type of stratification or post-stratification is applied as a followup step, such as for intensification.		***Yes***. There are two options: (i) increase sample size in large strata with omission errors and/or in other strata as required to achieve desired precision; or (ii) create a buffer to capture omission error.
Area estimator: The area of a land cover type, land use, or forest-to-non-forest transition in region R (A_R) is generally known as the area of domain D (A_D). It is possible to estimate the area of domain "A_D" using point sampling, with the following expression.	$\hat{A}_D = A_R \times \hat{P}_R$ (using definitions from Equation 3)		$\hat{A}_{DS} = A_R \times \hat{P}_{RS}$ (using definitions from Equation 5)
Standard error estimator for A_D; $v(\hat{P}_R)$ is the variance of the mean of the attribute of interest in the domain of interest in the region.	$S(\hat{A}_D) = A_R \times \sqrt{v(\hat{P}_R)}$ (using definitions from Equation 4)		$S(\hat{A}_{DS}) = A_R \times \sqrt{v_s(\hat{P}_{RS})}$ (using definitions from Equation 6 or Equation 7 for post-stratification)
Confidence intervals*	$CI_D = \hat{A}_D \pm \left(t_{\alpha/2} \times S(\hat{A}_D)\right)$ (using definitions from Equation 12)		$CI_D = \hat{A}_{DS} \pm \left(t_{\alpha/2} \times S(\hat{A}_{DS})\right)$

Note: * *See* p. 27 of Thompson, S.K. 2012. *Sampling, third edition*. Hoboken, USA, John Wiley and Son.
Source: Authors' own elaboration.

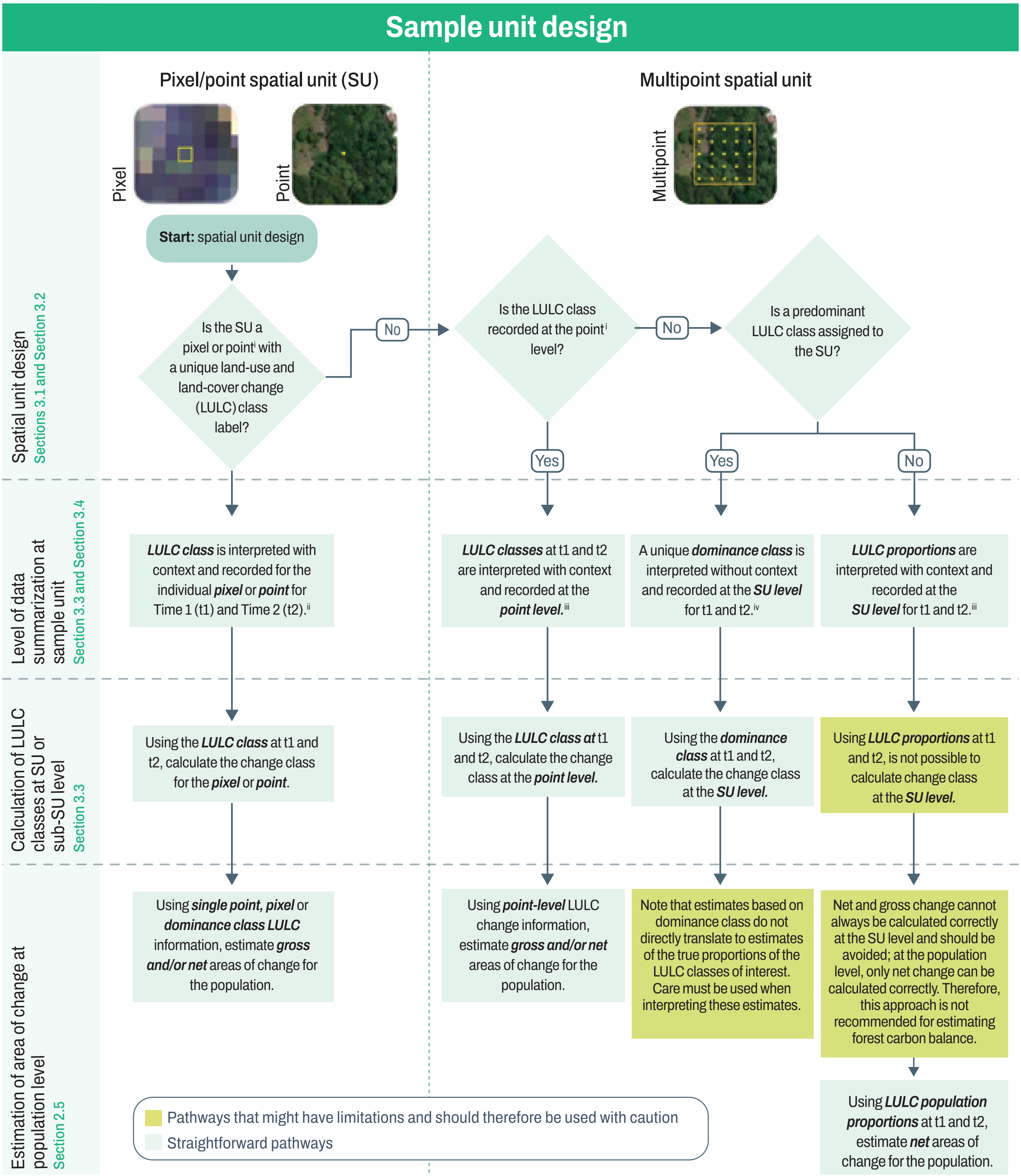
Section 3

Response design

The most effective designs for SBAE or other types of inventories are constructed using a logical sequence of steps. These steps, variations of which have been described elsewhere (Lister *et al*., 2015) include assessing information needs, identifying constraints, determining allowable error for key attributes (commonly expressed as a given sampling error at a specified confidence level, such as a sampling error of 10 percent of the estimate with a 95 percent confidence level), designing the sample unit (or plot), and determining the sampling design.

This section seeks to provide practical guidance on the plot design step by exploring: the implications of using point-based or pixel-based sample units versus multipoint area-based plots (Section 3.1); considerations related to the size of an area-based plot and the number of points with which to subsample the plots (Section 3.2); the impacts of different labelling protocols for area-based plots (Section 3.3); and the effects of interpreting land use with or without context (Section 3.4). Figure 5 presents a decision tree that shows the interrelationship of these topics and outlines flow of plot design decisions.

Figure 5 Decision tree of interrelated topics in Section 3 and order of decisions to be made



Notes: Courtesy of German ObandoVargas.

i The term “point” in the pixel/point sample unit column refers to a sample unit that is a single dimensionless point. In the multipoint sample unit column, “points” refers to the points within a plot that are evaluated to determine the land use and land cover composition of the plot.

ii **Interpretation with context-individual point/pixel scenario:** the land use context around the point (or pixel) is evaluated to determine the corresponding land use. Specifically, the interpreter mentally delineates boundaries around the land use intersected by the point/pixel while applying the appropriate definitions, such as minimum area, canopy cover (in the case of forest), and potentially minimum widths. The corresponding land use is assigned to the point/pixel.

iii **Interpretation with context:** the land uses are interpreted by applying the land use definitions to the landscape patterns within and without the multipoint sample unit. The interpreter mentally delineates boundaries around the different land uses while applying their respective definitions, such as minimum area, canopy cover (in case of the forest), and potentially minimum widths. The corresponding land use is assigned to the points within the sample unit. These may be recorded at the point level or converted to proportions and recorded at the sample unit level depending on the pathway.

iv **Interpretation without context:** for forest land use, the forest definition is applied with the sample unit as the frame of reference for applying the forest definition rather than the irregular landscape patterns (inside and outside the sample unit) as the frame of reference. If the minimum canopy cover threshold within the sample unit is met, the entire sample unit is considered to be forest.

Source: Authors’ own elaboration.

© FAO/Enric Català Contreras

References

Lister, A.J., Scott, C.T., Brewer, C.K. and Stein, B.R. 2015. *Forest inventory design principles – challenges and solutions*. Proceedings of the XIV World Forestry Congress, September 7–11, 2015, Durban, South Africa. FAO. http://foris.fao.org/wfc2015/api/file/55e721863a4836e65ca6b509/contents/d287e4b5-6b9a-4b65-8112-13a5d0b5d073.pdf

3.1 Point, pixel, or multipoint area-based sample units: practical considerations

by

Paul Patterson, Randy Hamilton and Andrew Lister

Different types of sample units have been used for SBAE. A 2018 GFOI white paper briefly reviews the main types that are commonly used (GFOI, 2018). This section compares three specific types of sample units: points, pixels, and larger area-based sample units.

Points are infinitely small whereas pixels are associated with an area, such as a 30 m × 30 m Landsat pixel). When using either points or pixels as sample units, a unique land use and land cover is visually interpreted and assigned to each unit. In other words, the data at the level of the sample unit are binary in nature, meaning that for any class of interest, it is either present or absent in the sample unit; each unit receives a 1 or a 0, and the set of 1s and 0s are summarized to create estimates (Arévalo *et al*., 2020; Frescino *et al*., 2016). Good practices for SBAE using pixels as sample units have been documented elsewhere (Olofsson *et al*., 2014, 2020). It is worth noting that the forest minimum map unit will always (in the case of points) or generally (in the case of pixels) be larger than the sample unit. Therefore, of necessity, when using point-based or pixel-based sample units, the context within which the sample unit falls will need to be assessed to determine the land use in accordance with the forest definition and minimum map unit (for more information on this topic, see Section 3.4).

With larger area-based sample units (or plots), the land use and land cover composition is frequently characterized by overlaying a systematic grid of points and interpreting from fine-resolution imagery the land use and land cover intersected by the points to estimate the percent composition of the different classes within the plot (Frescino *et al*., 2009, 2016; Lister *et al*., 2019; Patterson, 2012; Tzamtzis *et al*., 2019). In this case, the plot-level data can be considered continuous in nature (for example, 80 percent forest remaining forest and 20 percent deforestation from forest to grassland).

Area estimates and the associated confidence intervals produced from point-based or pixel-based sample units are handled statistically in the same way. However, it can be shown that for the same number of sample units, area-based multipoint units will yield higher-precision (smaller variance) estimates than those calculated using binary data from single point-based or pixel-based sample units (for a more detailed explanation, see Why multipoint sample units will have smaller variance than single-point sample units in Section 3.1). To obtain a similar precision using single point-based or pixel-based sample units, the sample size would need to be increased. However, there is a trade-off between the increased costs of using more sample units versus the increased costs of using a multipoint design with a smaller sample size that takes longer to evaluate. The number of additional point or pixel sample units needed to achieve the same precision will depend on the distribution of the various classes in the landscape. Ideally, a study should be carried out to compare the time it takes to interpret point or pixel sample units versus the larger multipoint plots in relationship to the differences in precision achieved to determine which is most efficient for a particular landscape (Section 3.2 describes how to conduct these types of studies).

Due to the increased precision obtainable using multipoint, area-based sample units, they are sometimes preferred over point or pixel sample units. However, point or pixel sample units have some practical implementation advantages over larger area-based units. For example, a larger area-based sample unit falling near a population boundary could feasibly extend beyond the population boundary, violating the assumption of complete tessellation of the population with sample units (see Section 2.5). In order to not violate statistical assumptions, the physical area of the sample unit that extends beyond the boundary could be reflected back into the study area using a geometric transformation approach (Patterson, 2012) or some other partial plot correction method. Single point-based sample units will never have this problem. In the case of pixel-based sample units, population boundaries are frequently defined to coincide with the pixelated edges of an image, eliminating the problem as well.

A similar problem to that of sample units straddling a population boundary can also occur along strata boundaries in stratified area estimation when using a finite-population sampling strategy such as two-stage or cluster. These strategies assume the sample units occur in one and only one stratum. In practice, the stratum of the centre point of the plot is sometimes assumed to be the stratum of the entire plot even when it spans strata. Although this violates a fundamental assumption, the impact is generally assumed to be minimal. An infinite-population sampling strategy such as two-step on the other hand, avoids the problem theoretically by collapsing the information to the centre point of the sample unit (see Section 2.5). As at population boundaries, point or pixel sample units also avoid the problem.

Why multipoint sample units will have smaller variance than single-point sample units

In Section 2.5 and Annex I, the two-step estimator, $\hat{\hat{P}}_R$, was introduced and its variance $V\left(\hat{\hat{P}}_R\right)$ was given. See those sections for any undefined notation that is used in this box. To show, for the same number of sample units, area-based, multipoint sample units will yield higher precision (lower variance) estimates than those calculated using binary data from point or pixel sample units, the alternative form of the variance, $V\left(\hat{\hat{P}}_R\right)$, presented in Annex I is useful. The alternate form is:

$$V\left(\hat{\hat{P}}_R\right) = \frac{1}{n}[P_R(1-P_R)] - \left(\frac{1}{n} - \frac{1}{nm}\right)\frac{1}{\|R\|}\int_R P(s)\big(1-P(s)\big)ds$$

(Equation 8)

P_R is the proportion of the attribute of interest over the region R and $P(s)$ is the proportion of the attribute of interest over the sample unit centred at s.

In the case of using binary data from a point or pixel sample unit, the proportion over the sample unit centred at s is equal to 0 or 1, where the point is considered as the sample unit. $P(s)$ is equal to 0 or 1 and the second term of the alternative form of $V\left(\hat{\hat{P}}_R\right)$ is equal to 0.

Hence, $V\left(\hat{\hat{P}}_R\right) = \frac{1}{n}[P_R(1-P_R)]$ in the case of binary data from a point or pixel sample unit. In the case of area-based, multiplepoint units: $0 \leq P(s) \leq 1$ for any s in the region R. If $0 \leq P(s) \leq 1$, then $P(s)\big(1-P(s)\big) \geq 0$.. For some s in R, there will be a mix of the attribute of interest within the sample unit centred at s, which implies that $0 < P(s) < 1$, implying that $\int_R P(s)\big(1-P(s)\big)ds > 0$. Since $\left(\frac{1}{n} - \frac{1}{nm}\right) > 0$, the second term will be greater than zero. This implies that the variance, $V\left(\hat{\hat{P}}_R\right)$, for the area-based two-step estimator computed with multiple points in the support region, for the same number of sample units, will be smaller than the variance calculated using binary data from point or pixel sample units.

If the interpreted data are to be used for purposes beyond estimating the areas of the different land use and land cover classes, such as for calibrating or validating models to produce maps, then other special considerations may apply (see Section 3.2).

References

Arévalo, P., Olofsson, P. and 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. https://doi.org/10.1016/j.rse.2019.01.013

Frescino, T.S., Moisen, G.G., Megown, K.A., Nelson, V.J., Freeman, E.A., Patterson, P.L., Finco, M., Brewer, K. and Menlove, J. 2009. *Nevada Photo-based Inventory Pilot (NPIP) photo sampling procedures*. RMRSGTR222. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station. https://doi.org/10.2737/RMRS-GTR-222

Frescino, T.S., Moisen, G.G., Patterson, P.L., Freeman, E.A. and Menlove, J. 2016. *Nevada Photo-Based Inventory Pilot (NPIP) resource estimates (20042005)*. RMRSGTR344. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station.

GFOI (Global Forest Observations Initiative). 2018. *Summary of country experiences and critical issues related to estimation of activity data*. GFOI. https://www.fao.org/gfoi/news-events/news/detail/en/c/1159538

Lister, A., Lister, T. and Weber, T. 2019. Semi-Automated Sample-Based Forest Degradation Monitoring with Photointerpretation of High-Resolution Imagery. *Forests*, 10(10): 896. https://doi.org/10.3390/f10100896

Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E. and Sanz, M.J. 2020. Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236: 1–9. https://doi.org/10.1016/j.rse.2019.111492

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A. 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148: 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Patterson, P.L. 2012 Photo-based estimators for the Nevada photo-based inventory. RMRSRP92. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station. https://www.fs.usda.gov/rm/pubs/rmrs_rp092.pdf

Tzamtzis, I., Federici, S. and Hanle, L. 2019. A Methodological Approach for a Consistent and Accurate Land Representation Using the FAO Open Foris Collect Earth Tool for GHG Inventories. *Carbon Management*, 10(4): 437–450. https://doi.org/10.1080/17583004.2019.1634934

3.2 Area-based sample unit design: number of points and sample unit size

by

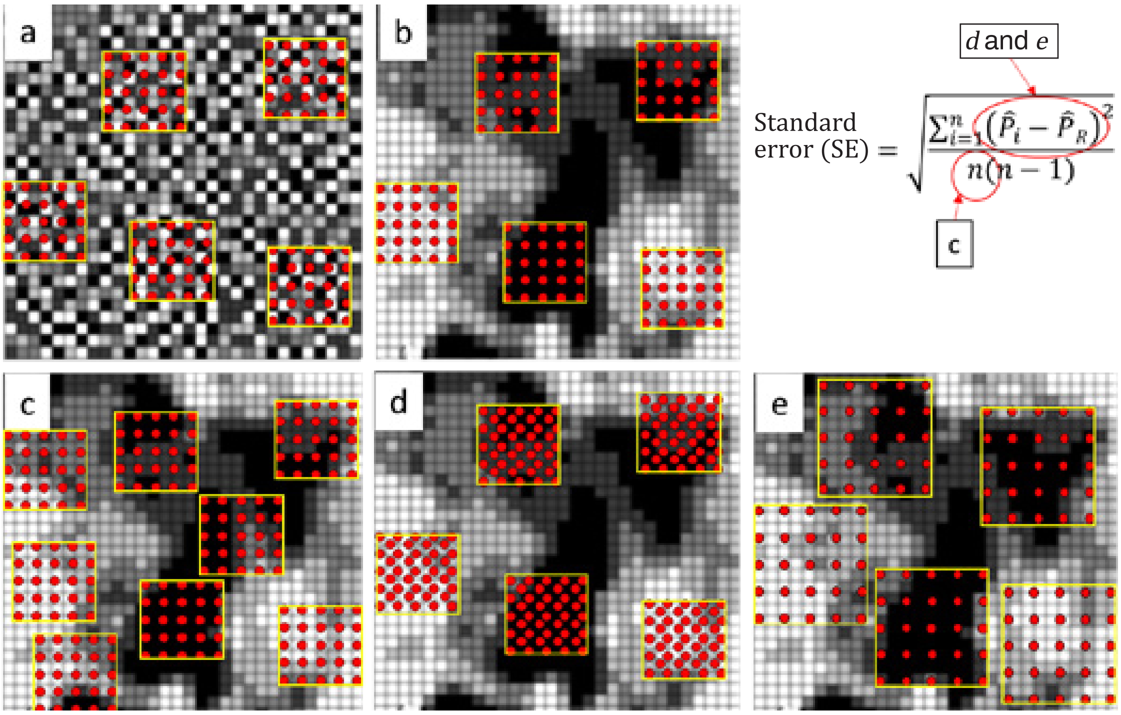
Andrew Lister, Paul Patterson and Randy Hamilton

Common questions posed by countries implementing SBAE using larger area-based sample units (or plots) include: how many plots are needed; how large should they be and how many points should be placed within each one? The optimal design of area-based sample units can be informed from basic principles of statistics. Figure 6 empirically illustrates several important principles. Standard error (SE in Figure 6) is a commonly used measure of uncertainty and is used in the construction of confidence intervals. Standard error is the square root of the variance of the estimate from Section 2.5. Figure 6a shows a landscape where the plots contain a mix of class values – about the same mix as occurs on the landscape as a whole. What this means is that each plot-level proportion ($\hat{P}_i$) will be very much like the sample mean proportion ($\hat{P}_R$), and standard error will be low because the numerator of the equation will be small. Figure 6b, however, shows a landscape where each plot only has one or possibly two class values. In this example, some plots are very different from the sample mean proportion of each class, leading to a larger numerator and a higher standard error. Thus, we see that the landscape configuration can have a significant impact on the standard error.

To lower the standard error in the landscape of Figure 6b, one can increase n (put out more plots), increasing the denominator of the standard error equation, as shown in Figure 6c. Another option is to increase the number of points within the plot (Figure 6d), which will tend to lower the standard error. The next section on number of points per sample unit shows explicitly how an increase in the number of points per plot affects the variance; a reduction in the variance causes a reduction in the standard error. The relationship between the variance and the standard error is complicated and not obvious, as well as beyond the scope of this paper. Finally, one can increase the size of the plot to capture more of the variability on the landscape on each plot, making each plot-level proportion more like the sample mean proportion, lowering the numerator and standard error (Figure 6e).

In summary, the standard error can be reduced by increasing the sample size, n, increasing the number of points within the plot or by increasing the size of the sample unit. The trick of optimal design is choosing which combination of options (Figure 6c, Figure 6d and Figure 6e) leads to the lowest cost evaluation that will meet the needs of users.

Figure 6 Illustration of two landscapes with different spatial distributions



Notes: Illustration of two landscapes with different spatial distributions of the landscape elements (a) and (b), along with different sample unit designs that affect the calculation of the standard error (SE): (c) increased sample size, *n*, (d) increased point density per plot and (*e*) increased plot size. In the equation, $\hat{P}_i$ is the proportion of points in the class of interest on plot *i*, $\hat{P}_R$ is the average proportion across all plots, and *n* is the number of plots in the sample (see Section 2.5 for details on calculating the proportions).

Source: Authors' own elaboration.

Number of points per sample unit

Countries implementing SBAE using larger area-based sample units frequently ask how many points, m, should be evaluated in a sample unit. If the m points are assumed to be chosen using a simple random or systematic sample, the variance of the estimator can be used to give direction on the size of m. To illustrate, the variance from the two-step sampling strategy (see Section 2.5) will be used. Note that the concept is the same using the other strategies described in Section 2.5. The formula involves two terms:

$$Var = \frac{1}{n}\frac{1}{\|R\|}\int_R (P(s) - P_R)^2 ds + \frac{1}{nm}\frac{1}{\|R\|}\int_R P(s)\big(1 - P(s)\big) ds$$

(Equation 9)

The first term calculates the variance among sample units while the second term calculates the variance within sample units. If we let $C_1 = \frac{1}{\|R\|}\int_R (P(s) - P_R)^2 ds$ and $C_2 = \frac{1}{\|R\|}\int_R P(s)\big(1 - P(s)\big) ds$ then variance can be expressed as:

$$Var = \frac{1}{n}C_1 + \frac{1}{nm}C_2$$

(Equation 10)

First, observe that increasing m only affects the second term, by the factor of $\frac{1}{m}$. Therefore, evaluating 5 points instead of 1 point would decrease the second term by a factor of one-fifth (0.2), while increasing the number of points to 25 would decrease the second term by an additional factor of one-fifth (0.2). The next logical step would be 49 points, which would only decrease the second term by approximately a factor of one-half (approximately 0.5), while increasing the required interpretation effort significantly.

The second observation is that the decrease in the total variance (combining the first and second terms) achieved by increasing m is dictated by the relative sizes of C_1 and C_2. For example, if C_2 is one-fifth the size of C_1 (it is expected that C_2 would generally be smaller than C_1), then the factor by which the total variance decreases as m increases from m_1 to m_2 is:

$$Factor = \frac{\frac{1}{n}C_1 + \frac{1}{nm_2}\frac{1}{5}C_1}{\frac{1}{n}C_1 + \frac{1}{nm_1}\frac{1}{5}C_1} = \frac{1+\frac{1}{5m_2}}{1+\frac{1}{5m_1}}$$

(Equation 11)

If we use 5 points instead of 1 point, the total variance decreases by a factor of 0.87. While increasing the number of points from 5 to 25 would decrease the variance by an additional factor of 0.97, increasing the number of points from 25 to 49 would decrease the variance by an additional factor of 0.99. In this scenario, the variance reduction becomes minimal increasing m beyond 5. In the other extreme, if C_2 is five times C_1, then increasing the number of points from 1 to 5 points would decrease the variance by a factor of 0.33, while increasing the number of points from 5 to 25 would decrease the variance by an additional factor of 0.60. Increasing the number of points from 25 to 49 would decrease the variance by an additional factor of 0.98. In this scenario, the variance reduction becomes minimal beyond an m of 25. In agreement with these results, a study conducted in Costa Rica showed that reducing the number of points from 49 to 25 points did not significantly decrease the precision of the estimate (Ortiz-Malavasi, 2019).

It can be shown that the sizes of C_1 and C_2, and hence their relative sizes, are dictated by the distribution of the category being estimated. It is recommended to conduct a simulation study to compare the relative sizes of C_1 and C_2 for various distributions of the category being estimated. Apart from the precision considerations associated with the number of points within the sample unit, interpretation time and the associated costs are also important considerations. The following section presents a workflow for a simple optimization experiment that could be performed to determine the optimal number of points per plot as a function of interpretation time.

There are some additional considerations that may impact the choice of m. For example, Frescino and Patterson (2017) showed that with too few points in the sample unit, some rare categories may be missed. If rare categories are known to exist but are not detected by the inventory, the recommendation is to note in the results that there is insufficient data to report on that category, and to consider additional studies with a different survey design to focus on that category. Patterson and Finco (2011) present a method to derive an upper-bound for the proportion of an undetected rare category.

If the survey is multipurpose and map building is one of the objectives, then the precision of the estimates for individual plots may be important and a larger number of points per plot may be needed to adequately characterize the plot composition. For example, this would be particularly important if the objective were to use plot-level values as training data to create a percent canopy cover map.

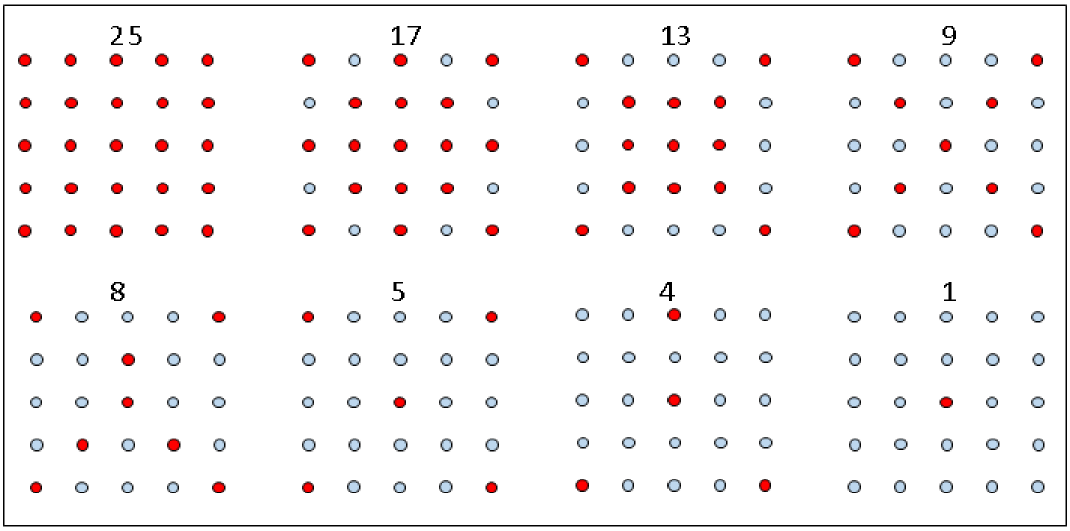
Workflow for a simple plot design optimization study

This section describes a study that can be conducted to optimize the sample size and number of points per plot while taking into account the time costs of interpretation. Several practitioners have contributed informal tools that correspond with some of the methods presented in this box. These tools have been assembled into an area sampling toolkit that can be used to plan and analyse monitoring activities.

For a study, assign at least 30 plots to each relevant subpopulation found on a map. For example, assign 30 plots each to each stable and change stratum in a study area for REDD+ reporting.

- For each plot, generate a grid of 25 points over a sample unit of interest (commonly 1 hectare or larger).
- Interpret each plot, carefully recording: the class values for each point; the time it takes to interpret and enter data on each point and transition to the next point (on plot costs); and the time it takes to load and begin the first point of the next plot after finishing the last point of the previous plot (between plot costs). Calculate and store the mean time per point and the mean time between plots per subpopulation. Ideally, use a data entry interface, like that found in the Open Foris Collect Earth Online tool (Saah *et al*., 2019), which minimizes data entry and plot transition times. Other tools are described by Schepaschenko *et al*. (2019).
- Calculate means, standard deviations, and coefficients of variation (CV) ($CV = \frac{standard\ deviation}{mean}$) from the sample data for classes of key variables from the land use or cover class data collected in the second step.
- Subset the dataset: calculate means, standard deviations, and coefficients of variation with plots constructed from subsets of the points (e.g. with subplots in the following configurations, as seen in Figure 7).

Figure 7 Grid of 25 points spread over an area of 1 hectare



Note: Initially, the full number of points will be interpreted (25 red points), and coefficients of variation (CV) will be calculated. Coefficients of variation will then be calculated from plots constructed with subsets of the points arrayed in regular patterns (examples of configurations of points in red indicate possible subsets to try).
Source: Authors' own elaboration.

- Calculate the required sample size (n) to achieve the desired precision for each subpopulation using the following equation:

$$required\ n = \left(\frac{CV \cdot t}{E}\right)^2$$

(Equation 12)

t = the $(1 - \alpha/2)^{th}$ percentile of the t distribution with n-1 degrees of freedom, $1- \alpha$ is the confidence level associated with the desired precision, CV is the coefficient of variation expressed as a percent, and E is the desired precision standard error of the mean expressed as a percentage of the mean. For example, if the coefficient of variation of the sample of plot-level values for one of the configurations is 50 percent, and the desired precision of an estimate is a 95 percent confidence interval that is 10 percent of the estimate, the calculation would be (using a conservative value of 2 for t) as follows:

$$n_{required} = \left(\frac{50 \cdot 2}{10}\right)^2 = 100\ plots$$

(Equation 13)

- For each configuration (for each subpopulation) calculate the total time required to complete the inventory and achieve the desired precision of the estimate using the following equation:

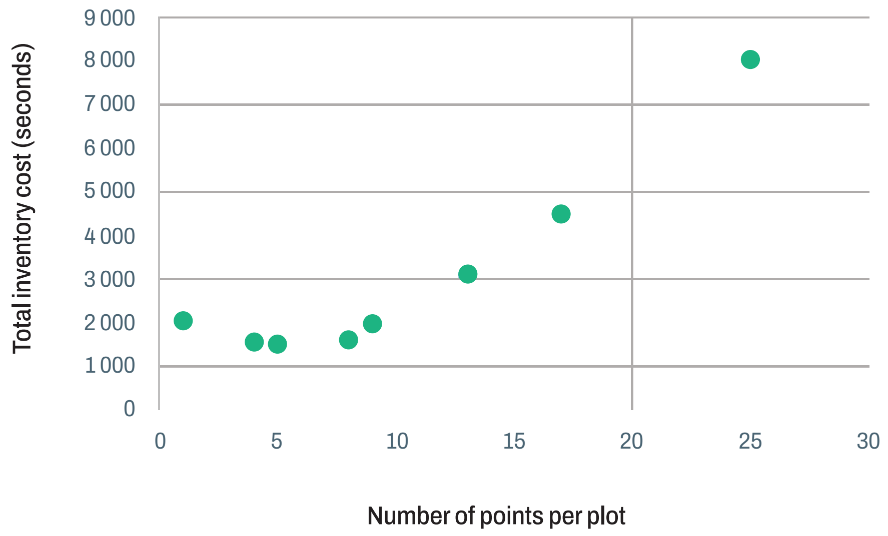
$$Total\ time = n_{required} \times m \times\ t_{on} + \left(n_{required} - 1\right) \times t_{between}$$

(Equation 14)

t_{on} and $t_{between}$ are the mean per point time and the mean between plot time, respectively, calculated as described in the second step; m is the number of points used in the design being evaluated.

- Create the following graphic, which should show a similar pattern, for the inventory as a whole (summing costs across subpopulations, or giving the larger strata more weight if the range of stratum size is large) (see Figure 8).

Figure 8 Example of the relationship between the number of points used in a plot design and the total time cost required to achieve a chosen desired precision



Source: Authors' own elaboration.

In the example of Figure 8, there is a clear minimum of the function that would define the relationship between the required number of points to achieve the desired precision and the total time cost – somewhere around 5 points per plot (note: data is hypothetical). While the exact minimum could be found through calculation of derivatives, it is probably sufficient to choose a subplot count through inspection of a graphic like Figure 8 as a guide. An example of this approach is described in Lister *et al*. (2014).

In the example presented in this section, stable and change strata are used as subpopulations, and it is important to understand total inventory costs because the functional relationships like that shown in Figure 8 can differ between subpopulations that have different landscape-scale forest and non-forest patterns.

Sample unit size and point distribution

Experience has shown that best results with multipoint designs are achieved by separating the points as much as possible within the plot area (Lister *et al*., 2014). This is in agreement with traditional forest inventory ground-based cluster plot theory, which shows that this minimizes wasted effort from collecting redundant information from the same patch on a single plot (Thompson, 2012; Cochran, 1977). Ideally, plots should be designed such that points are outside the range of spatial autocorrelation of the phenomenon under study; this approach, using semivariance analysis, was used in the design of the plot in the NFI of Tanzania (Tomppo *et al*., 2014).

There is likely a theoretical maximum separation distance among subplots that becomes either illogical or impractical. For example, designs with points spread over a 100 km^2 area might yield very precise estimates, but they might be so time-consuming to implement that it would have been more cost-effective to use more, but smaller, plots. Another problem arises with large plots: they are more likely to cross stratum or population boundaries than smaller plots, necessitating some sort of correction to deal with partial plots. A similar experiment to that described in the workflow for a simple plot design optimization study could be performed to test the impact of plot size and time considerations.

In addition to those already mentioned, there can be other practical implementation considerations as well. For example, it can be difficult to interpret land use from small plots if very fine-resolution imagery is not available, which may necessitate the use of larger plots. For countries interpreting land use without context (see Section 3.3 and Section 3.4 for important considerations), the plot size may be determined by the minimum area component of the forest definition. It is also important to keep in mind that if the data are to be used for purposes beyond estimating areas, such as for training and validation of maps, other restrictions on the plot size may apply.

In summary, increasing plot size and the spacing between points within the plots can increase the precision of the estimates; however, increasing plot size must be balanced against practical implementation considerations, the requirements of other users of the data (such as map training and validation purposes), and other national or international requirements.

Conclusion

In conclusion, there are a number of different factors that impact the standard error and precision of an estimate derived from a sampling design that uses area-based sample units. These factors include the distribution of the classes of interest in the landscape, sample size, size of the plot, number of points within the plot, and the distribution of the points. Adjusting any of these can potentially increase the precision of the final estimates.

Plot design choices can also have important impacts on overall assessment costs; pilot studies (as described in the workflow for a simple plot design optimization study) can help guide these decisions. It is important to conduct a pilot study in a manner consistent with how data will be reported. Whether or not a plot design optimization study is conducted often depends upon the level of interest or capacity of those designing the assessment.

Apart from statistical considerations, other practical or even political considerations may impact plot design decisions such as: costs versus monetary benefits; the availability of imagery with adequate resolution for a particular plot size; forest minimum area constraints; and other use requirements such as using the data as training or validation data for mapping. Regardless of a country's particular situation, it is instructive for inventory designers and practitioners to understand the theoretical (Figure 6) and practical implications of different plot design decisions when designing an SBAE monitoring system.

References

Cochran, W.G. 1977. *Sampling techniques, third edition*. Wiley series in probability and mathematical statistics. New York City, USA, Wiley.

Frescino, T.S. and Patterson, P.L. 2017. Number of Points Needed for Image-Based Change Estimation. In: Healey, S.P. and Berrett, V.M. 2017. *Doing more with the core: Proceedings of the 2017 Forest Inventory and Analysis (FIA) Science Stakeholder Meeting, October 24– 26, 2017*. RMRSP75. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station.

Lister, T.W., Lister, A.J. and Alexander, E. 2014. Land use change monitoring in Maryland using a probabilistic sample and rapid photointerpretation. *Applied Geography*, 51: 1–7. https://doi.org/10.1016/j.apgeog.2014.03.002

Ortiz-Malavasi E. 2019. Reestimación de datos de actividad para el periodo de referencia del ERPD de Costa Rica 1998 al 2011 mediante Evaluación Visual Multitemporal: Documento I – Metodología de trabajo, definición de tamaño de parcela y protocolo para preparación de base de datos [Re-estimation of activity data for the reference period of the ERPD of Costa Rica 1998 to 2011 through Multitemporal Visual Assessment: Document I - Working methodology, plot size definition and protocol for database preparation]. Unpublished manuscript. San José, REDD+ Secretariat.

Patterson, P.L. and Finco, M. 2011. Calculation of upper confidence bounds on proportion of area containing not-sampled vegetation types: an application to map unit definition for existing vegetation maps. *Mathematical and Computational Forestry & Natural-Resource Sciences* 3(2): 98–101

Saah, D., Johnson, G., Ashmall, B., Tondapu, G., Tenneson, K., Patterson, M., Poortinga, A. *et al.* 2019. Collect Earth: An online tool for systematic reference data collection in land cover and use applications. *Environmental Modelling & Software*, 118: 166–171. https://doi.org/10.1016/j.envsoft.2019.05.004

Schepaschenko, D., See, L., Lesiv, M., Bastin, J.F., Mollicone, D., Tsendbazar, N.E., Bastin, L. *et al.* 2019. Recent Advances in Forest Observation with Visual Interpretation of Very High-Resolution Imagery. *Surveys in Geophysics*, 40(4): 839–862. https://doi.org/10.1007/s10712-019-09533-z

Thompson, S.K. 2012. *Sampling, third edition*. Hoboken, USA, John Wiley and Son.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H.M., Chamuya, N., Zahabu, E. and Otieno, J. 2014. A sampling design for a large area forest inventory: case Tanzania. *Canadian Journal of Forest Research*, 44(8): 931–948. https://doi.org/10.1139/cjfr-2013-0490

3.3 Labelling protocols and sample unit data summarization

by

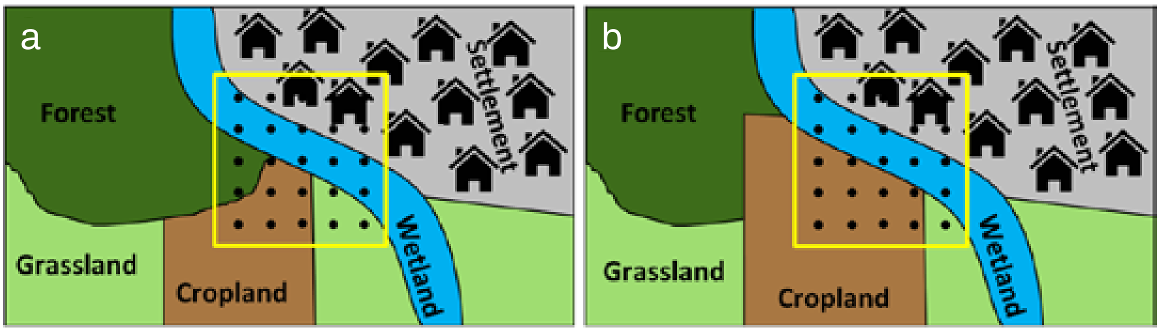
Paul Patterson, Frédéric Achard, Andrew Lister and Randy Hamilton

This section discusses the pros and cons of three protocols sometimes used to label the land use and land cover of area-based sample units (or plots) by interpreting a set of points within the sample unit. These include: i) assigning a single dominance class to the sample unit; ii) recording the land use proportions for each sample unit; or iii) recording and maintaining the point-level land use and land cover labels, such as using the Open Foris Collect Earth Online (CEO) tool, which allows point-level tracking through time.

Dominance class

In the case of assigning a dominance class, a single land use class is assigned to each plot based on a set of predefined rules, and the land use classes are often interpreted without context (see Section 3.4). For example, Bastin *et al*. (2017) promoted determining dominance class (predominant land use) using a hierarchy rule. "Predominance" is understood here to refer to a given land use category exceeding a predetermined threshold of the plot area; additional hierarchy rules also apply. For example, the percent occupancy threshold could be 30 percent with the following hierarchical order to apply in the case of ties: forest, cropland, grassland, settlement, wetlands and other lands. In Figure 9a, forest, cropland, grassland, and settlement occupy less than 30 percent of the plot, and wetlands are greater than 30 percent; therefore, this plot would be classified as dominance class wetlands. Countries using a dominance class approach may use different thresholds as well as different interpretation rules than those illustrated in Bastin *et al*. (2017).

Figure 9 Comparison of a land use mosaic in Time 1 (a) and Time 2 (b) showing cropland expansion



Source: Authors' own elaboration.

It is important to note that the estimates of the area or percent composition of each class in the population, based on a dominance approach, have different meanings than those based on recording land use proportions or retaining point-level land use calls within the plots. For example, an estimate of the proportion of forest calculated from dominance class plots must be interpreted as the proportion of locations on the landscape where a plot centred at that location is covered by at least 30 percent forest – while the estimate of cropland, for example, would be the proportion of the locations on the landscape for which a plot centred at that location would be covered by less than 30 percent forest and greater than 30 percent cropland. The estimate of forest (or any other land use) calculated in this fashion does not translate to an estimate of the true percent forest (or other land use) cover in the landscape, which is what would be expected from the other two labelling protocols. It is critical to understand and correctly interpret this difference in meaning of the estimates.

One drawback of using the dominance class paradigm is that a small shift in the land classes on the plot can lead to a radical change in the dominance class of the plot. For example, Figure 9 shows cropland increasing between Time 1 and Time 2 at the expense of forest and grassland. As a result, the dominance class, due to the hierarchy rules, changes from wetland to cropland; however, the same proportion of the plot as before is wetland and in reality, no change in wetland occurred. In addition, a small amount of deforestation occurred, which was not detected.

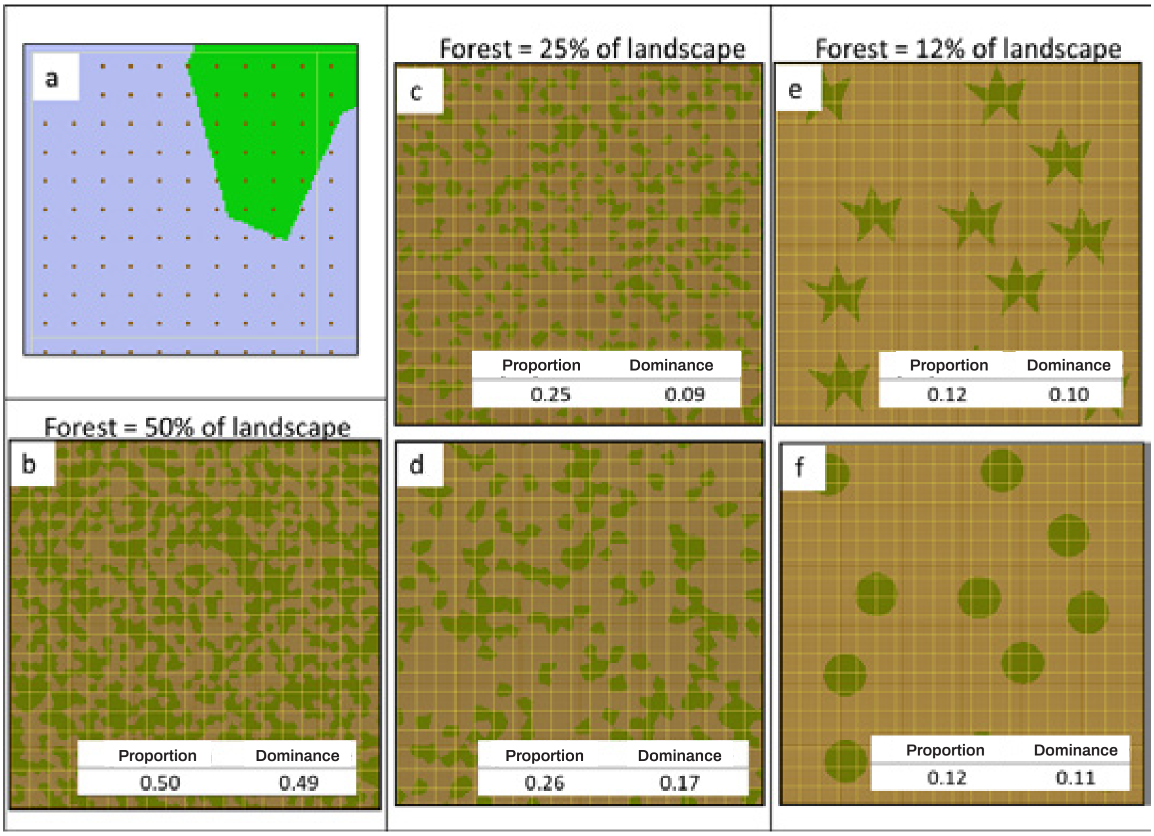
There are other possible drawbacks to assigning dominance classes. First, subtle changes in the land use composition in the plot, such as from 35 percent forest coverage to 29 percent coverage, can cause the dominance class to change, and depending on the number of such changes, could cause inflated estimates of change. Second, and in a contrary manner, some large changes (for example, from 100 percent forest coverage to 31 percent coverage on the sample unit) will not lead to a change in dominance class, and thus will be undetected.

It can be shown that in certain landscape configurations, considerable differences can exist between estimates based on dominance class versus those calculated from the points within the plots and plot-level proportions. To illustrate, consider the results of the experiment depicted in Figure 10. In it, a population consisting of 361 hectares was constructed. A grid of plots, each 100 m × 100 m, was overlain on the population. Within each plot, a grid of 100 points was superimposed as shown in Figure 10a, and each point was labelled with a forest or non-forest class based on the type of patch it intersected. Several landscape types with different proportions of forest and forest patch configurations were generated within the population boundary (Figure 10b–f). Population estimates of forest proportion were calculated using both the dominance class and proportion approaches. In the case of dominance class, each plot received a 1 (present) or a 0 (absent) for the class of interest, with dominance based on > 50 percent occupancy. Using the proportion approach, each plot received a proportion of each class based on the proportions calculated from the point counts.

The results of this study, which are only based on one realization of the experiment, suggest that the proportion approach more closely reflects the true proportion of the forest class than the dominance approach, and that the differences are related to some combination of the proportion of the landscape occupied by the class of interest and the patch size and configuration. Proportions of classes that occur in smaller (relative to the plot size), more isolated patches appear to be prone to greater differences than classes occurring in larger, contiguous patches. Therefore, the results suggest that more research is needed on this important topic, and that considerable differences in the estimates or variances from the two approaches can occur.

© UN-REDD/Cory Wright

Figure 10 Design and results of a simulation experiment for population estimates of the proportion of the forest



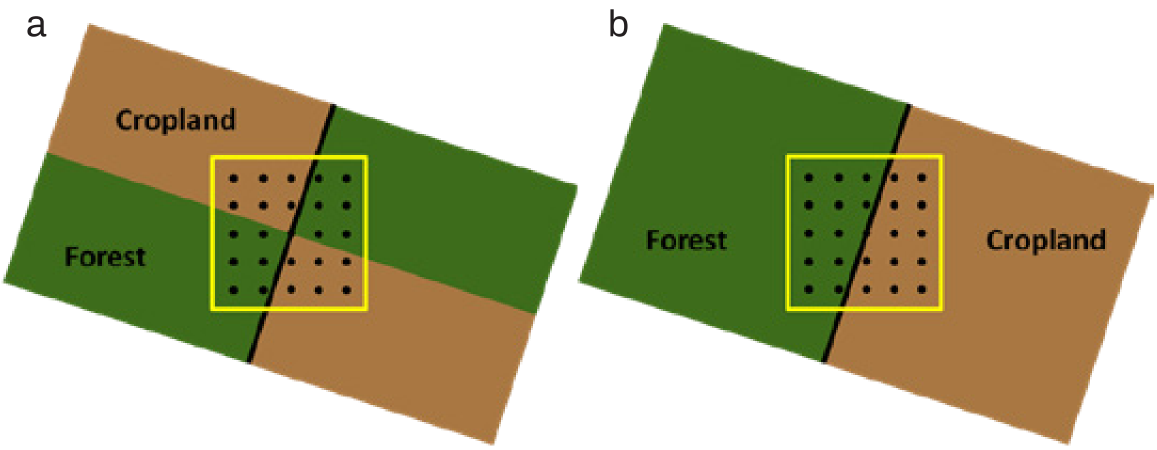
Notes: Design and results of a simulation experiment in which a grid of 100 points is superimposed on each of a set of 361 100 m × 100 m plots. Population estimates of the proportion of the forest (green) class were calculated using two methods: i) the proportion approach, in which each of the 100 points was labeled with a cover class, and the proportion forest was assigned to each plot; and ii) the dominant cover class on each plot was determined based on point counts, and for estimates of forest, 1 (present) or 0 (absent) was assigned to each plot. Means from the set of 361 plots were calculated and are presented in the tables at the bottom of each landscape image. The figure is divided into the following parts: a) a depiction of the 100 points superimposed on one plot, with the class of interest (green) indicated in the upper right; b) a simulated landscape that is 50 percent forest; c) a simulated landscape that is 25 percent forest, with small, isolated patches; d) a 25 percent forest landscape with larger, more contiguous patches; e) a 12 percent forest landscape with star-shaped patches to assess the effects of elongated patches; and f) a 12 percent forest landscape with round patches, assessing effects of compact patches.
Source: Authors' own elaboration.

It is also important to reiterate, as mentioned previously, that estimates based on dominance are not directly comparable to those based on proportions from point-level data. In addition, dominance class estimates cannot be directly interpreted as estimates of the true proportion of the attribute of interest, but rather as the proportion of locations on the landscape where a plot centred at that location is covered by (in this case) more than 50 percent of the attribute of interest. In cases where the dominance threshold is significantly greater or less than 50 percent, the difference between the approaches would be expected to be even greater.

Land use proportions

Recording the land use proportions within the plots significantly increases the amount of detail about the composition of the landscape. Unlike the dominance class approach, estimates of forest (or other land use) can be interpreted as the proportion of that land use occurring in the landscape. However, with only proportions recorded at the plot level, some types of change can go unnoticed. For example, there is a possibility of missing reciprocal changes that occur within sample units. In Figure 11, the proportion of cropland (50 percent) and secondary forest (50 percent) in the sample unit is the same in both Time 1 and Time 2. If only proportions are recorded, it would appear that no change occurred. However, in reality, 25 percent of the secondary forest remained secondary forest, 25 percent of the cropland remained as cropland, 25 percent of the secondary forest transitioned to cropland, and 25 percent of the cropland transitioned to secondary forest, which yields a far different carbon accounting result. In other plots with even more land uses present, with some or all of them changing, it can be impossible to know what land use transitioned to what with only the Time 1 and Time 2 proportions. Because of this, only net population-level changes can be accurately calculated from these data. Therefore, this approach is not recommended for estimating forest emissions and removals, which should ideally be based on gross changes.

Figure 11 Hypothetical scenario showing two properties



Note: In Time 1 (a), each property is half cropland and half secondary forest. After a number of years, in Time 2 (b), the cropland in the left property has regenerated to secondary forest, while the secondary forest in the property to the right has been cut to become cropland. In the sample unit, the proportions of the two land uses remains the same.
Source: Authors' own elaboration.

Point-level land use information

The third option of recording and maintaining point-level land use and cover labels provides the most exact information and allows reciprocal and other within-plot land use transitions to be tracked with exactness. In Figure 11, for example, this approach would allow point-level changes to be tracked to correctly identify that 25 percent of the sample unit was deforested and 25 percent regenerated, and that the remaining land uses remained stable. Therefore, when feasible, it is recommended to record the land use calls for each point within each plot. If the point values are recorded at Time 1 and Time 2, the gross proportions of each change category can be calculated per plot and summarized to generate gross population-level estimates for each change category. The Open Foris Collect Earth Online tool[18] permits point-level calls to be recorded. Although this method is recommended, one challenge of recording point-level information is that image shift from Time 1 to Time 2 can make it appear that change has occurred when it has not. Ideally, interpreters should view the Time 1 and Time 2 imagery simultaneously to mentally compensate for any image shift that occurs to ensure that the correct calls are made for each point.

[18] See http://www.openforis.org/tools/collect-earth-online.html

References

Bastin, J.F., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C. *et al.* 2017. The extent of forest in dryland biomes. *Science*, 356(6338): 635–638. https://doi.org/10.1126/science.aam6527

3.4 Interpretation paradigms: Interpretation without or with context

by

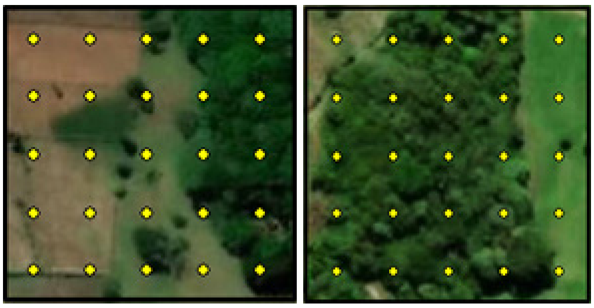
Randy Hamilton and Paul Patterson

Sample-based area estimation has been implemented using a variety of sample unit designs, labelling protocols and interpretation rules. Unfortunately, these varied approaches can sometimes lead to very different results. Several sample unit designs and the implications of different labelling protocols were discussed in a previous GFOI white paper (GFOI, 2018). In this section we will focus on the interpretation rules (including the application of the forest minimum map unit component of the forest definition) associated with the particular case of interpreting points within a plot to determine the land use. Two basic types of interpretation rules have been implemented, which will be referred to as interpreting the land use without or with context.

Interpretation without context

In the case of interpretation without context, the sample unit would typically be the size of the minimum area or minimum map unit of the forest definition (or in some cases larger). The interpreter determines whether the sample unit is forested by assessing whether the amount of tree cover within it meets or exceeds the forest canopy cover definition and ideally that the other components of the forest definition are met (Figure 12). This determination is made without considering the land uses and patterns outside of the sample unit. In other words, the forest definition is applied with the sample unit as the frame of reference for the minimum map unit rather than with the irregular landscape patterns as the frame of reference. If the minimum canopy cover threshold within the sample unit is met, the entire sample unit is considered to be forest. Simple classification rules, like those of Bastin *et al*. (2017), can be used to label the sample unit with a dominance class (see Section 3.3 for a discussion of this approach).

Figure 12 Assessing interpretation without context



Notes: When interpreting without context, the interpreter only considers the landscape within the sample unit. In this example, the sample units are 2 ha in size and the forest definition requires canopy cover greater than or equal to 60 percent with a minimum map unit of 2 ha. In the upper figure, 36 percent (9 of 25 points) of the sample unit has tree cover; therefore, this sample unit would be considered non-forest In the lower figure, 76 percent (19 of 25 points) of the sample unit has tree cover; therefore, this sample unit would be considered forest.

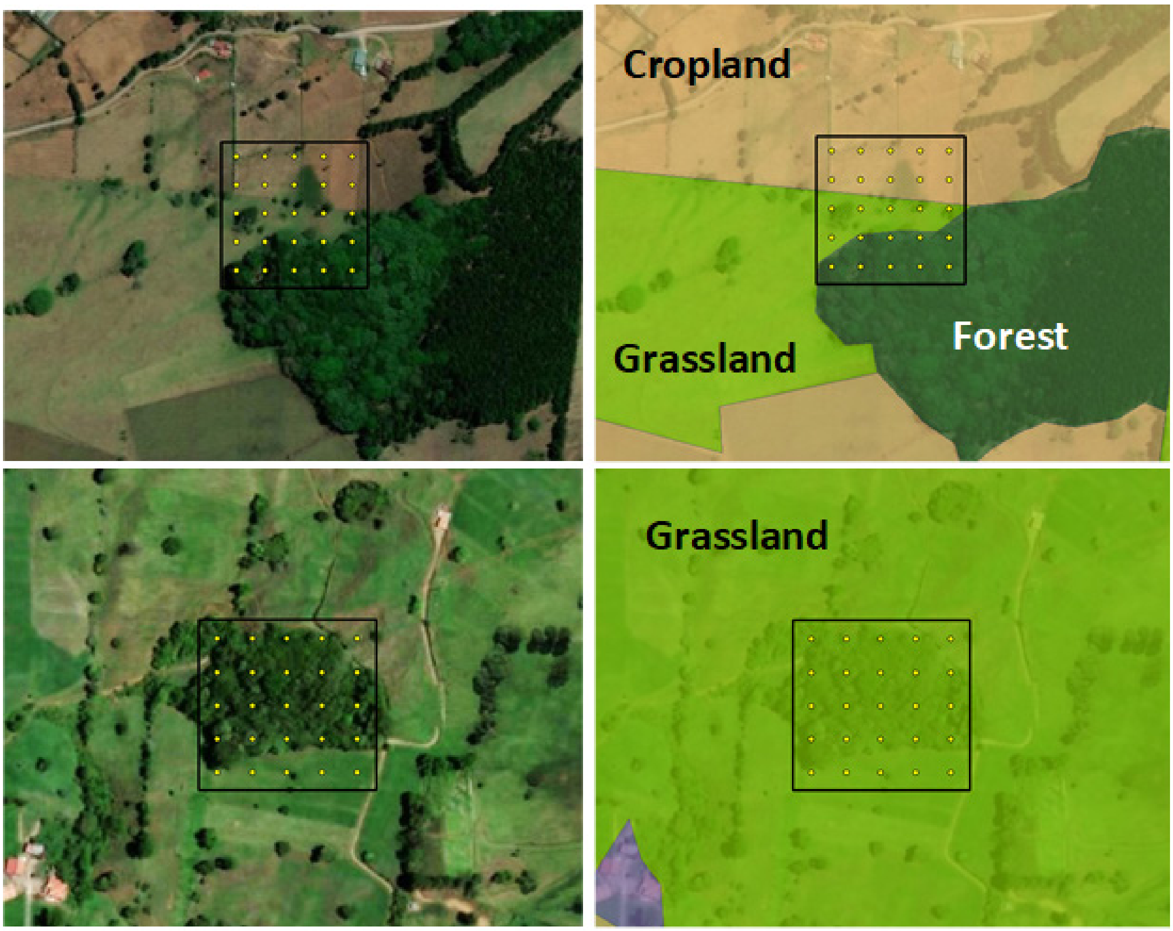
Source: Authors' own elaboration.

Interpretation without context could be difficult to apply in multipurpose monitoring systems where other land uses, with their own cover and minimum map unit definitions, must also be applied and reconciled with each other. Under certain conditions, this approach could also lead to results considerably different from those obtained from interpreting each point within a plot with context (these differences are explained in more detail in Section 3.3). For example, consider a scenario with a forest definition of 30 percent canopy cover in a highly fragmented landscape that only contains small patches of trees, most of which are smaller than the forest definition minimum map unit size. In this scenario, applying the minimum map unit definition to the individual patches would show very little forest, while applying the minimum map unit definition to the sample unit could result in a much higher estimate of forest (see Section 3.3 on dominance class for a description of the subtle differences on how these two estimates differ and should be interpreted).

Interpretation with context

In the case of interpretation with context, the interpreter examines the patterns of land use both within and without the sample unit, and mentally delineates (or delineates in a Geographical Information System) boundaries around the different land uses, while applying their respective definitions such as minimum area, canopy cover (in the case of forest), and potentially minimum widths (Figure 13). In this case, the land use definitions are applied with the landscape patterns as the frame of reference. The corresponding land use is then assigned to the points within the sample unit to determine the proportion of each land use within the sample unit. In general, interpretation with context will provide a more exact estimate of the areas of the different land uses than interpretation without context. Furthermore, using the paradigm with context aligns with the intuitive, traditional concept that land uses occur in patches, or functional units, rather than the elemental landscape unit being an arbitrary, square-shaped unit (or a unit of any shape).

Figure 13 Assessing interpretation with context



Note: When interpreting with context, the interpreter considers the landscape within and without the sample unit and applies the land use definitions to the patterns in the landscape. The plot locations in this figure are the same as in Figure 12. Just as in Figure 12, the sample unit is 2 ha in size and the forest definition requires canopy cover greater than or equal to 60 percent with a minimum map unit of 2 ha. In the upper figures, the land use within the sample unit is a mix of grassland (24 percent), cropland (40 percent), and forest (36 percent). In the lower figures, the patch of trees does not reach 2 ha; therefore, the land use is considered grassland. Note how different the land use calls are for interpretation with context versus without context (compare to Figure 12).

Source: Authors' own elaboration.

References

Bastin, J.F., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C. *et al.* 2017. The extent of forest in dryland biomes. *Science*, 356(6338): 635–638. https://doi.org/10.1126/science.aam6527

GFOI (Global Forest Observations Initiative). 2018. *Summary of country experiences and critical issues related to estimation of activity data*. GFOI. https://www.reddcompass.org/mgd/resources/GFOI-MGD-3.1-en.pdf

Summary of Section 3

by

German Obando-Vargas and Randy Hamilton

Section 3 addresses a variety of response design questions that various countries around the world have raised in the context of implementing SBAE. Table 7 summarizes key considerations and recommendations highlighted in the section.

Table 7 Summary of key considerations and recommendations related to plot and response design for sample-based area estimation

Consideration	**Type of sample unit and level of data summarization at sample unit**			
	Pixel/point *(Section 3.1)*	Multipoint area-based *(Section 3.1)*		
	Single land use and land cover class recorded for each sample unit *(Section 3.1)*	Land use and land cover classes recorded at point-level within sample unit *(Section 3.3)*	Land use and land cover class proportions recorded for each sample unit *(Section 3.3)*	Land use and land cover dominance class recorded for each sample unit *(Section 3.3)*
Data type at sample unit level	Binary	Continuous	Continuous	Binary
Precision obtained for same number of sample units *(Section 3.1)*	Lower than or equal to continuous data	Higher than or equal to binary data	Higher than or equal to binary data if LULC class proportions are directly recorded by the analyst	Not comparable (since estimating different characteristics)
Are special adjustments needed if sample unit spans population boundaries? *(Section 3.1)*	No (n/a)	Yes	Yes	Yes
Are special adjustments or accommodations needed if sample unit spans stratum boundary? *(Section 2.5, Section 3.1)*	No (n/a)	No, if using two-step sampling strategy Yes, if using two-stage or cluster strategy	No, if using two-step sampling strategy Yes, if using two-stage or cluster strategy	No, if using two-step sampling strategy Yes, if using two-stage or cluster strategy

Consideration	Type of sample unit and level of data summarization at sample unit			
	Pixel/point *(Section 3.1)*		Multipoint area-based *(Section 3.1)*	
	Single land use and land cover class recorded for each sample unit *(Section 3.1)*	Land use and land cover classes recorded at point-level within sample unit *(Section 3.3)*	Land use and land cover class proportions recorded for each sample unit *(Section 3.3)*	Land use and land cover dominance class recorded for each sample unit *(Section 3.3)*
Are there any limitations on labelling and level of data summarization at the sample unit data level? *(Section 3.3)*	No. Maintaining the pixel or point-level land use and cover labels provides the most exact information and allows reciprocal and other within-plot land use transitions to be tracked with exactness Yes. Co-registry problems between fine-resolution images from Time 1 to Time 2 can make it appear that change has occurred when it has not		Yes. This approach is not recommended for estimating forest emissions and removals since net and gross change cannot always be calculated correctly at the sample unit level	Yes. Estimates based on dominance class do not directly translate to estimates of the true proportions of the land use and land cover classes of interest. Also, small changes in the plot's land use and land cover classes can lead to radical changes in the plot's dominance class. In addition, some large changes will not lead to a change in dominance class, and will thus be undetected
Can the data be used as training/validation for maps? *(Section 3.2)*	Yes	Yes, but special considerations apply	Yes, but special considerations apply	Yes, but special considerations apply
Are there limitations for estimating change at the population level? *(Section 3.3)*	No	No	Yes. Gross changes generally cannot be calculated correctly. Only net changes at the population level should be calculated	Yes. In certain landscapes, the results may be very different from those obtained using point-level information and do not represent true land use and land cover areas
Are there special considerations related to interpretation of land use with or without context? *(Section 3.4)*	Must be interpreted with context unless the pixel is the size of or smaller than the forest minimum map unit	Must be interpreted with context	Must be interpreted with context	Interpreted without context. The plot becomes the frame of reference for the forest minimum map unit. Landscape patterns are ignored

Note: When interpreting with context, the interpreter considers the landscape within and without the sample unit and applies the land use definitions to the patterns in the landscape. The plot locations in this figure are the same as in Figure 12. Just as in Figure 12, the sample unit is 2 ha in size and the forest definition requires canopy cover greater than or equal to 60 percent with a minimum map unit of 2 ha. In the upper figures, the land use within the sample unit is a mix of grassland (24 percent), cropland (40 percent), and forest (36 percent). In the lower figures, the patch of trees does not reach 2 ha; therefore, the land use is considered grassland. Note how different the land use calls are for interpretation with versus without context (compare to Figure 12).
Source: Authors' own elaboration.

Section 4

Quality assurance/ quality control

4

4.1 Assessing and reporting quality of reference data

by

Steve Stehman and Ron McRoberts

Detailed description along with graphic examples

Visual interpretation of remotely sensed data as a means of producing reference data is susceptible to error (a label is assigned that does not correspond to reality) or there may be inconsistencies among multiple interpreters that induce bias into statistical estimators (Foody, 2009).

Some countries that implement SBAE attempt to implement good practices using a QA/QC process, and some countries try to estimate the effects of errors associated with the visual interpretation of plots by cross-interpreting or by using independent third parties to interpret part of the plots. However, the results of these efforts are not necessarily integrated into overall estimates of uncertainty.

In addition, countries have no incentive to integrate interpreter error, because doing so could increase the uncertainty of their estimates with the possible impact of reducing their results-based payments. Also, QA/QC procedures such as reinterpretation of sample units can increase monitoring costs.

Overview of existing good practices

Interpreter error and interpreter variability can have substantial adverse effects on area estimates as well as estimation of uncertainty, potentially introducing bias and thereby reducing compliance with the IPCC good practice guidelines. To know whether uncertainties can be reduced you must first correctly estimate them. At the very least, the effects of interpreter variability and error should be estimated and documented.

We define reference data error as an incorrect labelling of the ground condition for a sample unit, and reference data variability as differences among replicate interpretations of the same sample unit. Reference data error can contribute to bias in the estimators of activity data, whereas reference data variability can inflate the uncertainty (standard errors) of the area estimates. Assessment of reference data error requires availability of "gold standard" data for assessing interpreter error. Reference data variability can be estimated using repeat interpretations for at least a subsample of the full sample selected to estimate activity data.

Multiple QA/QC procedures have been proposed to minimize systematic and random interpreter error during the label assignment process. In addition, good practice guidelines for characterizing the quality of visual interpretations that comprise reference data and for integrating the effects of interpreter uncertainty into overall uncertainty estimates have been proposed.

Summarized good practices

Improving consistency

The following actions can be taken to improve consistency:

- decrease bias of activity data estimators by using more interpreters, perhaps as many as five to seven (MGD 3.0), in order to resolve interpretations that have reduced confidence;
- develop labelling protocols and/or standard operating procedures (SOPs);
- implement common training regimes prior to the start of data collection and interpretation;

- implement a double interpretation process and calibration at the start of the interpretation process, which can be gradually reduced as the differences among interpreters decrease to the point they can be considered similar (MGD 3.0); and
- conduct consistency checks and discuss problematic cases throughout the process.

Resolving different interpretations (when there is not unanimous agreement)

The following approaches can be used to resolve different interpretations:

- final "expert" decision (MGD 3.0), which may be used when there is a person who is more familiar with the area or has more experience than others;
- consensus decision (MGD 3.0), which refers to seeking consensus among interpreters when there are different labels assigned to the same plot;
- majority interpretation, which is when there are at least 3 interpreters for the same plot and two (or the majority) of them assign the same label; and
- proportion of reference class assignments (for example, 0.5 deforestation).

It is necessary to provide guidance on when to use which approach. For example, the first approach assumes an "expert" exists, whereas other approaches assume sets of interpreters with roughly equal ability. There may be some cases (sample units) for which there is not a majority interpretation, so it may still require consensus or an expert to decide those cases. Using a proportion of class assignments would require estimation formulas different from those used when a single reference class label is applied (not an insurmountable problem).

Estimating interpreter variability

The following approaches can be taken to estimate interpreter variability:

- multiple interpreters for *all* sample units (McRoberts *et al*., 2018); and
- random *subsample* selected for duplicate interpretation (Pengra *et al*., 2020).

Incorporating protocols to estimate interpreter variability accomplishes several objectives. First, the data can be used to describe agreement among interpreters, information that conveys transparency to the reference data collection process. An agreement matrix similar in format to the error matrix typically used for accuracy assessment would suffice to describe agreement. Second, the data can be used to provide feedback to interpreters during the course of reference data collection, identifying common sources of confusion and possibly identifying interpreters who are performing differently from the norm. Third, these data may be used to incorporate reference data variability into the overall assessment of uncertainty of the area estimators. In applications for which four or more interpreters examine every sample unit, there is likely no need to collect additional data for assessing interpreter variability. Agreement among interpreters can be readily characterized from such data and providing summary analyses of agreement would be sufficient to meet the transparency objective of describing agreement. Therefore, the following text applies primarily to applications in which three or fewer interpreters interpret each sample unit.

Interpreter variability and agreement can also be estimated by selecting a subsample of the full sample selected for estimating activity data and obtaining duplicate interpretations for the subsample. In applications for which two or three interpreters are used for each sample unit, the duplicate interpretation should be obtained by a different set of interpreters. The fact that the sample unit is receiving a duplicate interpretation should be unknown to the interpreters. Ideally, the subsample should be a probability sample from the main activity sample to facilitate incorporation of interpreter variability into the estimator of variance of the activity data. Selecting the subsample using simple random sampling would be easiest to implement and simplify variance estimators that incorporate reference data variability.

Interpreter variability would contribute an additional component of variability to the uncertainty of activity data estimators. To date, only variability associated with the sample selection has been incorporated into variance estimation (the variability defined by design-based inference). McRoberts *et al*. (2018) provide a method to estimate variance that includes variability associated with interpreters along with sampling variability (the hybrid inference approach). In the McRoberts *et al*. (2018) application, each sample unit has multiple interpreters. A different approach to estimating variance would be required when only a subsample of repeat interpretations is available. Särndal *et al*. (1992) provide a theoretical framework for this purpose (similar to the idea of hybrid inference) and variance estimator formulas that are applicable to when a subsample of repeated interpretations has been obtained. This methodology has not yet been applied to activity data estimation.

The size of the subsample of duplicate interpretations is also an issue that requires further research. Assuming a fixed total budget for reference data interpretations, clearly resources diverted to duplicate interpretations reduce the sample size available for estimating activity. We propose an initial guideline that the subsample constitutes somewhere in the range of 5–10 percent of the full activity data sample. A subsample of 100 duplicates would likely be sufficient to provide a reasonably precise estimate of agreement, but a larger subsample may be needed to estimate agreement for multiple time intervals of the reference data collection process. A larger subsample would also produce a more precise estimate of the contribution of interpreter variability to the total variance of the area estimator.

Identify knowledge gaps that need to be addressed by research and development

Even when there is not a template used for reporting interpreter error, two possibilities for reporting include:

- providing an agreement matrix for interpreters based on data from repeat interpretations; and
- obtaining and reporting interpreter agreement data for the time period during which reference data are obtained (early, middle and late).

A topic that would benefit from research is a comparison of the efficacies of the four proposed approaches for resolving interpreter disagreements (in particular, information on the trade-offs between costs and uncertainty for the four approaches):

- final "expert" decision (MGD 3.0)
- consensus decision (MGD 3.0)
- majority interpretation
- proportion of reference class assignments

There could be more guidance on the training and capacity development around the visual interpretation QA/QC process.

There should be more practical development to estimate the interpreter variability and integrate it into the overall variance estimate.

References

Foody, G.M. 2009. The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, 30(12): 3275–3281. https://doi.org/10.1080/01431160902755346

McRoberts, R.E., Stehman, S.V., Liknes, G.C., Næsset, E., Sannier, C. and Walters, B.F. 2018. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 142: 292–300. https://doi.org/10.1016/j.isprsjprs.2018.06.002

Pengra, B.W., Stehman, S.V., Horton, J.A., Dockter, D.J., Schroeder, T.A., Yang, Z., Cohen, W.B., Healey, S.P. and Loveland, T.R. 2020. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. *Remote Sensing of Environment*, 238: 111261. https://doi.org/10.1016/j.rse.2019.111261

Section 5

Way forward

by

Rémi d'Annunzio, Andreas Vollrath and Erik Lindquist

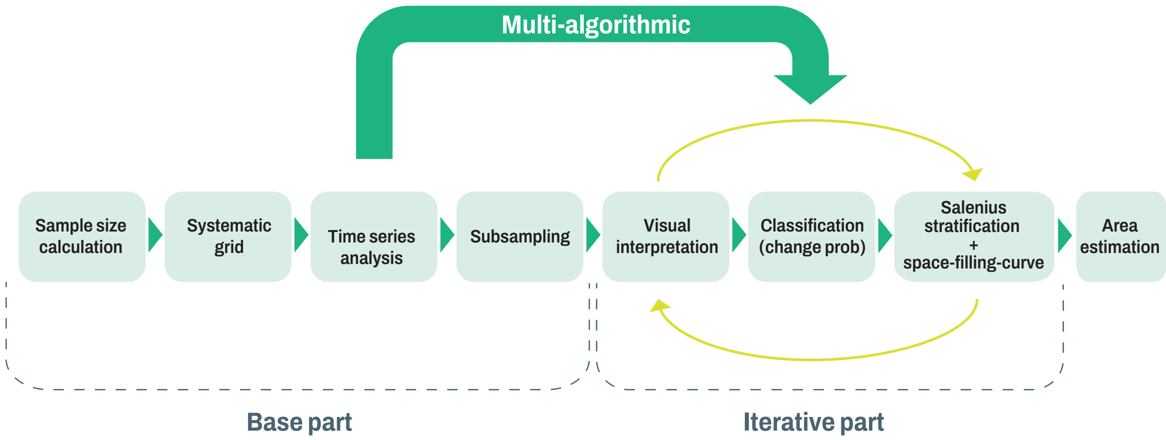

© FAO/Tsigie Befekadu

While stratified area estimation (SAE), or the practice of using a classified map to design a reference sample, has been widely recognized as a standard for producing results in accordance with best practices (Stehman, 1997; Olofsson *et al.*, 2013, 2014), using a random distribution of optimally allocated sample units can generate misleading results due to errors of omission of mapped deforestation (Olofsson *et al.*, 2020).

Practical solutions to the problems caused by errors of omission of change in a stratified area estimation approach imply exceptionally large sample sizes that are time consuming to analyse manually and thus make visual interpretation of each sample unit often not feasible given the tight time limits normally required for reporting. Time-consuming though it may be, results obtained with a large reference sample will be more precise than with a small sample.

To ease the burden of reference sample interpretation and decrease the time required to produce results, ensemble sample-based area estimation (eSBAE) has been developed and is tested in different pilot countries. It consists of a hybrid approach for area estimation, combining visual interpretation and machine learning. The proposed hybrid approach incorporates improvements to various aspects throughout its full workflow, aimed at optimizing the sample size.

Figure 14 Workflow to implement ensemble sample-based area estimation



Source: Authors' own elaboration.

Capturing unbiased area estimates of rare events, such as deforestation and forest degradation, with low levels of uncertainty, requires large sample sizes regardless of the sampling design (Pagliarella *et al.*, 2018) and the area covered. However, manually interpreting the samples is time-consuming and often not possible within reporting deadlines. To address this, FAO suggests a set of tools that optimize sample selection for visual interpretation. Throughout the last year, this procedure has been piloted in several countries, resulting in a significant reduction in effort to achieve acceptable levels of uncertainty without introducing further bias.

The starting point of this approach is a very dense systematic grid (1 km × 1 km or 2 km × 2 km spacing). In a subsequent step, a unique probability of forest change is assigned to each sample using auxiliary data sources from global products in combination with data-driven information extraction routines (such as time-series analysis), which can potentially include national maps and information layers. This process, known as ensemble classification or stacking, improves the distinction between stable and changing forest areas (Healey *et al.*, 2018) – hence ensemble SBAE.

Unlike the discrete classification of stable and unstable forest classes used in stratification for stratified area estimation (Stehman, 1997; Olofsson *et al.*, 2013, 2014), the continuous variable of forest change probability from the ensemble classification process is utilized, allowing for a gradual representation of change likelihood over the entire area.

It turns out that the statistical distribution of forest change probability for all samples is heavily skewed toward low probabilities, as most samples are in core forest areas or areas outside forests. An optimized framework for subsampling such distributions is known as the Dalenius type of stratification, followed by Neymann allocation for sample selection (Hidiroglou and Kozak, 2018). The samples are selected in a spatially balanced way, using the concept of a space-filling curve (Lister and Scott, 2009).

One advantage of this workflow is that the final stratum of a high-likelihood of change is usually larger compared to maps of change and no-change making omissions in the large no-change stratum highly unlikely. Omission errors have been a major concern for stratified area estimation (Olofsson *et al.*, 2020), as even a few of them result in elevated levels of uncertainty due to their huge weight. In contrast, the eSBAE workflow targets a clean stable stratum free of omissions. This may lead to higher uncertainties in the change strata, as sample weights are initially higher. However, reducing uncertainty now can be achieved more rapidly by intensifying on change stratum, where less points are present.

The FAO Forestry Division is developing a series of notebooks from the Open Foris initiative to implement this approach using Jupyter notebooks on the System for Earth Observation Data Access, Processing and Analysis for Land Monitoring (SEPAL) platform (SEPAL, 2023). Parts of the process can also be used to prioritize sample selection for QA/QC or to support intensified sampling in stable strata using stratified area estimation.

References

Healey, S.P., Cohen, W.B., Yang, Z., Brewer, C.K., Brooks, E.B., Gorelick, N., Hernandez, A.J. *et al.* 2018. Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment*, 204: 717–728. https://doi.org/10.1016/j.rse.2017.09.029

Hidiroglou, M.A. and Kozak, M. 2018. Stratification of Skewed Populations: A Comparison of Optimisation-based versus Approximate Methods. *International Statistical Review*, 86(1): 87–105. https://doi.org/10.1111/insr.12230

Lister, A.J. and Scott, C.T. 2009. Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment*, 149(1–4): 71–80. https://doi.org/10.1007/s10661-008-0184-y

Olofsson, P., Foody, G.M., Stehman, S.V. and Woodcock, C.E. 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129: 122–131. https://doi.org/10.1016/j.rse.2012.10.031

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A. 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148: 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Arévalo, P., Espejo, A.B., Green, C., Lindquist, E., McRoberts, R.E. and Sanz, M.J. 2020. Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236: 1–9. https://doi.org/10.1016/j.rse.2019.111492

Pagliarella, M.C., Corona, P. and Fattorini, L. 2018. Spatially-balanced sampling versus unbalanced stratified sampling for assessing forest change: evidences in favour of spatial balance. *Environmental and Ecological Statistics*, 25(1): 111–123. https://doi.org/10.1007/s10651-017-0378-y

SEPAL (System for Earth Observation Data Access, Processing and Analysis for Land Monitoring). 2023. eSBAE notebooks. In: *SEPAL documentation*. Cited 5 September 2023. https://github.com/sepal-contrib/eSBAE_notebooks

Stehman, S.V. 1997. Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, 60(3): 258–269. https://doi.org/10.1016/S0034-4257(96)00176-9

Annex

Annex I
Sampling strategies

by

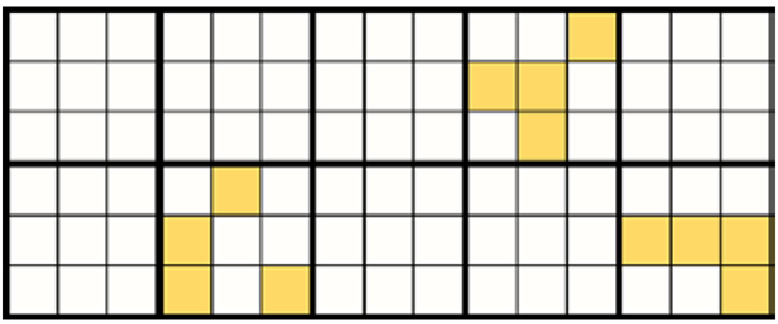
Paul Patterson and Andrew Lister

This annex discusses the theoretical differences among the two-stage two-step and cluster sampling strategies presented in Section 2.5 for those who wish to understand the theoretical underpinnings in greater depth; however, the information presented in Section 2.5 will be sufficient for many practitioners. The discussion presented here is designed as an overview and tries to provide enough details without being a complete theoretical discussion; some results will be stated, with references cited for those who wish to see the details of the derivations. A subsection is dedicated to each of the strategies. Each of the subsections will cover simple random sampling and systematic sampling first, followed by stratification and post-stratification. Any notation defined in Section 2.5 will not be redefined in this part of the paper.

Two-stage sampling from a finite population

In two-stage sampling, the sample units are considered a first-stage sample of primary sampling units (PSUs). The PSUs are supposed to be disjointed and cover the region (tessellate the region). The number of PSUs that tessellate the region is denoted N; typically, this is the size of R, measured in appropriate units, divided by the size of the plot in the same units. The second stage is a sample of the population units that make up the PSU. The PSU contains M population units, and there are m population units sampled within each PSU (Figure A1.1).

Figure A1.1 Two-stage sampling from a finite population



Note: Example of a population that has been tessellated by $N = 10$ PSUs, denoted by the thicker lines. Each PSU contains $M = 9$ population units. The sample shown here contains $n = 3$ PSUs and within each PSU the sample contains $m = 4$ population units.

Source: Authors' own elaboration.

As already mentioned, the sample units are considered the PSUs and the points within a sample unit are considered the population units. For this to be a finite population the "points" must be two-dimensional instead of one-dimensional and these two-dimensional objects should tessellate the sample unit (in the literature these objects are referred to as secondary sampling units; here, we will continue to refer to them as points). For a derivation of the following results, see Sections 10.1–10.4 of Cochran (1977).

The two-stage estimator of the proportion of the attribute of interest within the region R, denoted by $\hat{P}_R$, is:

$$\hat{P}_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i$$

(Equation 15)

$\hat{P}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$ is an estimate of the proportion of the attribute of interest in the ith PSU (sample unit) of the sample. Next is the equation of the variance of the estimator. Before giving the equation, a couple of items need to be defined. The PSUs are supposed to tessellate the region R. Let $i = 1 \ldots N$, denote the index of the N PSUs that tessellate the region R and let P_i equal the proportion of the attribute of interest within the ith PSU. If N is much larger than n and M is much larger than m then the two finite population correction factors, $\left(1-\frac{n}{N}\right)$ and $\left(1-\frac{m}{M}\right)$, are ignored and the variance of $\hat{P}_R$ is:

$$V(\hat{P}_R) = \frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}(P_i - P_R)^2 + \frac{1}{nm}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - P_i\right)^2\right)$$

(Equation 16)

The variance estimator, also ignoring the finite population correction factors, is:

$$v(\hat{P}_R) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{P}_i - \hat{P}_R\right)^2 + \frac{1}{Nn(m-1)}\sum_{i=1}^{n}\hat{P}_i\left(1-\hat{P}_i\right)$$

(Equation 17)

Since $0 \leq \hat{P}_i \leq 1$, the product $\hat{P}_i\left(1-\hat{P}_i\right)$ is less than or equal to 0.25 and hence $\frac{1}{n}\sum_{i=1}^{n}\hat{P}_i\left(1-\hat{P}_i\right)$ is less than or equal to 0.25. Considering this, if the number of PSUs, N, is large, the second term can be ignored and the following is used as the reduced variance estimator:

$$v(\hat{P}_R) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{P}_i - \hat{P}_R\right)^2$$

(Equation 18)

In stratified two-stage sampling each stratum is sampled using an independent two-stage sample, which means that each stratum is tessellated by PSUs and each PSU is in one and only one stratum. In post-stratified two-stage sampling, the PSUs in the sample are assigned to a stratum after the two-stage sample is drawn. Each stratum is tessellated by PSUs in it and each PSU is in one and only one stratum. It is important to note that in practice it is not uncommon to find examples of PSUs crossing strata boundaries, which is a violation of the assumption. Practitioners who use a two-stage sampling strategy and stratification often assign the stratum encountered at an arbitrary point within the sample unit (such as the centre point) to the sample unit in both ground-based and image-based forest inventories. Although this violates a fundamental assumption, the impact is assumed to be minimal. The authors are unaware of any studies that evaluate this.

To simplify the formulas, the proportion of the attribute of interest in the hth stratum will be denoted by P_h and its estimate denoted by $\hat{P}_h$; $h = 1, H$. Also, for the hth stratum, let W_h equal the stratum weight, Nh equal the total number of PSUs, n_h the number of sample units (PSUs) in the sample, and m the number of points sampled in each of the sample units. Then the estimator for both stratified two-stage sampling and post-stratified two-stage sampling is given by:

$$\hat{P}_{RS} = \sum_{h=1}^{H} W_h \hat{P}_h = \sum_{h=1}^{H} W_h \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \left(\frac{1}{m} \sum_{j=1}^{m} y_{hij} \right) \right)$$

(Equation 19)

The variance of stratified two-stage sampling is denoted by $V_S(\hat{P}_{RS})$, and is given by:

$$V_S(\hat{P}_{RS}) = \sum_{h=1}^{H} W_h^2 V(\hat{P}_h)$$

(Equation 20)

$V(\hat{P}_h)$ is given by $V(\hat{P}_h) = \frac{1}{n_h} \frac{1}{N_h - 1} \sum_{i=1}^{N} (P_{hi} - P_h)^2 + \frac{1}{n_h m} \frac{1}{N_h} \sum_{i=1}^{N} \left(\frac{1}{M-1} \sum_{j=1}^{M} (y_{hij} - P_{hi})^2 \right)$ with the finite population correction factors ignored. A variance estimator is given by:

$$v_s(\hat{P}_{RS}) = \sum_{h=1}^{H} W_h^2 v(\hat{P}_h) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (\hat{P}_{hi} - \hat{P}_h)^2 \right)$$

(Equation 21)

The finite population correction factors are ignored and the second term of the traditional variance estimator is ignored. The derivation of a variance estimator for the post-stratified estimator is based on the derivation in Section 5A.9 of Cochran (1977). Let $n = \sum_{i=1}^{H} n_h$ be the total sample size. The variance estimator based on the derivation is:

$$v_{PS}(\hat{P}_{RS}) = \sum_{h=1}^{H} \left(\frac{W_h}{n} + \frac{1-W_h}{n^2} \right) \left(\frac{1}{(n_h-1)} \sum_{i=1}^{n_h} (\hat{P}_{hi} - \hat{P}_R)^2 \right)$$

(Equation 22)

The second term of the traditional two-stage variance estimator is ignored.

Two-step sampling from an infinite population

The two-step sampling strategy is based on an extension of the Horvitz-Thompson estimator to design-based infinite population sampling (Cordy, 1993) combined with Stevens and Urquhart's (2000) results on support regions. For details of the derivations of the stated results for the two-step sampling, see Patterson (2012). In two-step sampling, the region R is considered a continuous population of points; R is also referred to as an infinite population, since there are an infinite number of points in the region R. It is worth mentioning that in sampling from an infinite population, the resolution of the imagery must be fine enough relative to the attributes of interest that we can confidently interpret what the attribute of interest would be at the point level.

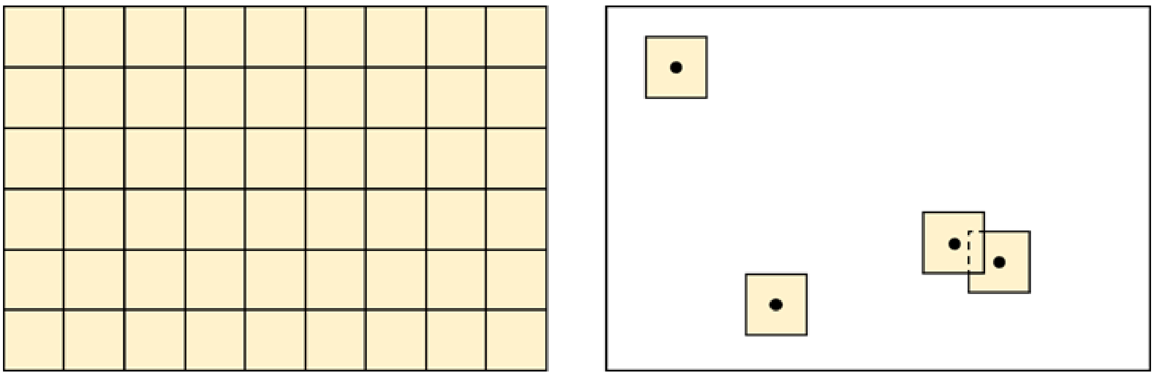
In the two-step strategy, what we have referred to as the "sample units" in the body of the document are defined as support regions for the points. The points are the true sample units in this strategy and the proportion of the attribute of interest within the support region is assigned to the centre point of the support region. The centre point, by its very nature, can occur in only one stratum – a key key assumption of all three strategies. The point nature of the two-step strategy, therefore, overcomes the theoretical issue faced by the two-stage and cluster strategies of having sample units of a finite size that may span strata boundaries.

If s is a point in the region R, let $P(s)$ equal the proportion of the attribute of interest within the support region centered at s. This corresponds, in two-stage sampling, to P_i, which is equal to the proportion of the attribute of interest within the ith PSU. In two-stage sampling, there is a finite number of sample units (PSUs) that tessellate R. In two-step sampling, there is a potential support region centred at any point – it is "potential" because a support region does not exist until the sample is drawn (Figure A1.2). For the sample, $s_1, \dots, s_i, \dots, s_n$ of the support region centres (using an infinite population), an unbiased single-step estimator of P_R would be $\frac{1}{n}\sum_{i=1}^{n} P(s_i)$ the average of the support region proportions. This is "similar" to the single stage (or cluster) sample of the PSUs. In two-step sampling, the $P(s_i)$ is estimated by a point sample of size m from the support region, that is $s_{i1}, \dots, s_{ij}, \dots, s_{im}$. As with the two-stage sample, define the variable y_{ij} to have the value 1 if s_{ij} intersects the attribute of interest and zero otherwise. Then, the two-step estimator, $\hat{\hat{P}}_R$, of P_R is:

$$\hat{\hat{P}}_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\hat{P}(s_i)$$

(Equation 23)

Figure A1.2 Two-stage sampling versus two-step sampling



Note: In two-stage sampling (left) a finite number of sample units (PSUs) tessellate the region *R*; in two-step sampling (right), a potential support region is centred at any point (it is "potential" because a support region does not exist until the sample is drawn). In this particular case, four points with their support regions are shown. As seen in the figure, support regions could overlap. If drawn systematically, the support regions would not overlap.
Source: Authors' own elaboration.

$\hat{P}(s_i) = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$ is the estimator of $P(s_i)$ described previously. Note that the estimator for two-stage sampling has the same algebraic form as the estimator for two-step sampling; it is the sampling design that differs between two-stage sampling and two-step sampling. The variance in the two-stage sampling strategy consists of two terms; each term involves a sum over all the N PSUs which tessellate R. In two-step sampling, there is an infinite number of points; infinite addition is accomplished using the technique of integration over R, denoted by the symbol $\int_R$. The variance of $\hat{\hat{P}}_R$ is:

$$V\left(\hat{\hat{P}}_R\right) = \frac{1}{n}\frac{1}{\|R\|}\int_R \left(P(s) - P_R\right)^2 ds + \frac{1}{nm}\frac{1}{\|R\|}\int_R P(s)\big(1 - P(s)\big)ds$$

(Equation 24)

$||R||$ is the size of R which is a measure of number of points in R; in two-stage sampling, this is equivalent to N, the number of PSUs which tessellate R. The variance of the two-step sampling strategy, $V\left(\hat{\hat{P}}_R\right)$, can be presented in an alternate form, which is used in Section 3.1. The alternative form is:

$$V\left(\hat{\hat{P}}_R\right) = \frac{1}{n}[P_R(1 - P_R)] - \left(\frac{1}{n} - \frac{1}{nm}\right)\frac{1}{||R||}\int_R P(s)\big(1 - P(s)\big)ds$$

(Equation 25)

The derivation of the alternate form assumes the support region has a reflection property for support regions near the boundary, meaning that if a support region extends beyond the region boundary, the proportion outside the region is reflected back into the region.

An unbiased variance estimator for $\hat{\hat{P}}_R$ is as follows (which is equivalent to the reduced variance estimator of the two-stage estimator):

$$v\left(\hat{\hat{P}}_R\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{P}(s_i) - \hat{\hat{P}}_R\right)^2$$

(Equation 26)

Stratification and post-stratification follow the same construction process as used for the two-stage estimator.

Cluster sampling from a finite population

In this application of cluster sampling, the sample unit is the set (or cluster) of points; it is a systematic array of secondary units. When conducting SBAE under cluster sampling, the "points" must be two-dimensional instead of one-dimensional and the clusters of these two-dimensional objects should tessellate the population. For the derivation of the following results, see Chapter 12 of Thompson (2012). The estimator, as stated previously, is the same as for the two-stage and two-step strategies, namely:

$$\hat{P}_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{j=1}^{M} y_{ij}$$

(Equation 27)

M is the number of points (population units) in each cluster; a capital M is used since there is a census of all secondary units in the cluster. Note the inner sum, $P_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$, is not an estimate of the proportion within the ith sample unit, but rather the actual proportion based on a census of the ith sample unit (in two-stage sampling, the proportion would be estimated).

Ignoring the finite population correction factor, the variance of the cluster estimator is:

$$V\left(\hat{P}_R\right) = \frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}(P_i - P_R)^2$$

(Equation 28)

Since we sample the elements within the cluster, there is no component in the variance to measure the variability within the cluster. The variance estimator, $v(\hat{P}_R)$, is given by the following (which is algebraically the same as the reduced variance estimator for the two-stage estimator and variance estimator for two-step estimator):

$$v(\hat{P}_R) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (P_i - \hat{P}_R)^2$$

(Equation 29)

Stratification and post-stratification follow the same construction process as used for the two-stage estimator, and the assumption that the clusters will be in one and only one stratum applies. However, in practice, this is often not the case. As with the two-stage sampling strategy, practitioners who use a cluster sampling strategy and stratification often ignore this assumption, but the impact is assumed to be minimal. The authors are unaware of any studies that evaluate this.

References

Cochran, W.G. 1977. *Sampling techniques, third edition*. Wiley series in probability and mathematical statistics. New York City, USA, Wiley.

Cordy, C.B. 1993. An extension of the Horvitz—Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18(5): 353–362. https://doi.org/10.1016/0167-7152(93)90028-H

Patterson, P.L. 2012 *Photo-based estimators for the Nevada photo-based inventory*. RMRSRP92. Fort Collins, USA, USDA, Forest Service, Rocky Mountain Research Station. https://www.fs.usda.gov/rm/pubs/rmrs_rp092.pdf

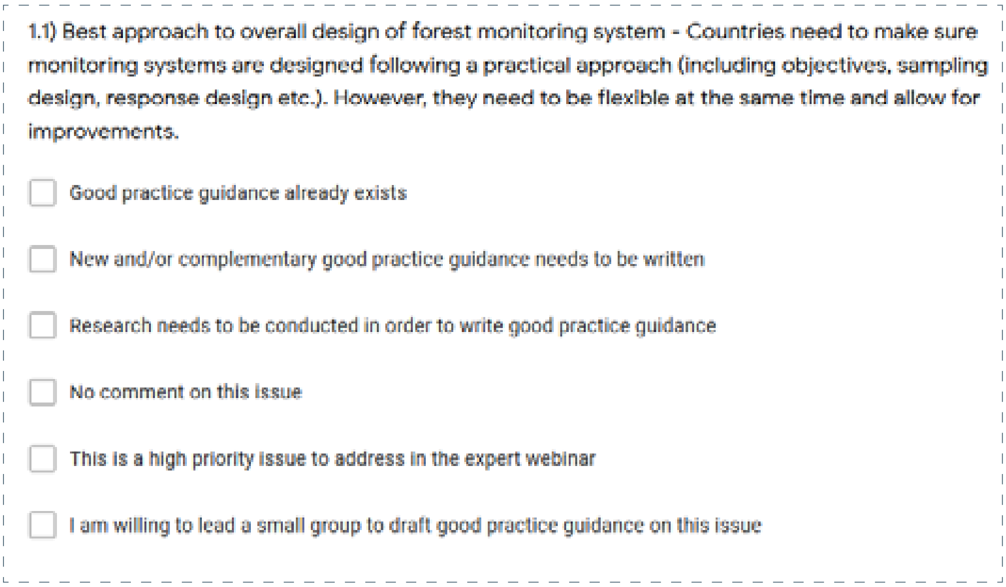
Stevens, D.L. and Urquhart, S. 2000. Response designs and support regions in sampling continuous domains. *Environmetrics*, 11: 13–41.

Thompson, S.K. 2012. *Sampling, third edition*. Hoboken, USA, John Wiley and Son.

Annex II
Baseline surveys and results

This paper was written based on: i) three surveys of experts to prioritize issues; ii) online meetings with experts to discuss issues; and iii) expert input to form sections. This annex presents the full list of topics included in the surveys, to give a fuller picture of the subject of area estimation as it stands and serves as a background to document how this paper was written. The three surveys were structured as follows: each issue that had been identified as of potential interest or importance was briefly described and respondents were asked to agree or disagree with six statements in relation to the issue (see Figure A2.1), which enabled prioritization of issues for inclusion in this paper.

Figure A2.1 Screenshot of survey question indicating how the issue of interest was presented



1.1) Best approach to overall design of forest monitoring system - Countries need to make sure monitoring systems are designed following a practical approach (including objectives, sampling design, response design etc.). However, they need to be flexible at the same time and allow for improvements.

- [ ] Good practice guidance already exists
- [ ] New and/or complementary good practice guidance needs to be written
- [ ] Research needs to be conducted in order to write good practice guidance
- [ ] No comment on this issue
- [ ] This is a high priority issue to address in the expert webinar
- [ ] I am willing to lead a small group to draft good practice guidance on this issue

Source: Authors' own elaboration.

General design issues

General design of monitoring system

- Best approach to overall design of forest monitoring system – Countries need to make sure natural monitoring systems are designed following a practical approach (including objectives, sampling design, response design, etc.). However, they need to be flexible at the same time and allow for improvements.

Considerations for multipurpose monitoring

- Considerations for multipurpose monitoring – Sampling should be useful for other purposes beyond area estimation. What are the considerations? Are there good practice recommendations?

- Integration of monitoring components – Ideally all monitoring components should be integrated (for example, NFI with area estimation).

- Monitoring for degradation, afforestation, etc. – Development has typically focused on deforestation and less on degradation, afforestation, reforestation and carbon enhancement (SMF and conservation). What adjustments need to be made to adequately monitor these other activities?
- Value of maps – Don't devalue maps too much. They are important for planning and other activities, not just for area estimation. How can mapping classification uncertainty be used for reporting or planning?

Issues related to varying dates and qualities of imagery for mapping and/or visual interpretation

- Lack of or gaps in high-resolution imagery – High-resolution imagery is frequently not available for all parts of a country and may have a cost implication. What can countries do if there are gaps in the fine-resolution coverage or if it is not available at all?
- Missing data – Missing data (for example, due to clouds, shadows, lack of recent imagery)
- Mixed-resolution imagery – When full-coverage fine-resolution is not available, countries will interpret plots from both fine- and moderate-resolution imagery. What are the implications of this practice? What are good practice recommendations?
- Impacts of widening date range of imagery – Countries frequently widen the date range for interpreting the imagery to +/- 1 year, +/- 2 years, etc. to ensure all plots are interpreted. What are the impacts and recommendations?
- Date mismatches: Strata versus reference data – Mismatch of the dates of imagery used to create a stratification map and the dates of imagery of the reference data frequently occur. May this be a source of omission errors? What guidance can be recommended?
- Impacts of improved temporal resolution – Improvements in temporal resolution may lead to increased omission errors. What can be the incentive for countries to improve in this regard?
- Impacts of improved spatial and temporal resolution – Improvements in spatial and temporal resolution may lead to increased detection of deforestation and higher deforestation rates. Therefore, there may be no incentive for countries to improve in this regard, especially if the increased spatial and temporal resolution is not available for both the reference and monitoring period.
- Data from NFI versus from remote sensing – Even when it is not very common, some countries already have permanent NFI plots, that are revisited at regular intervals, and can those provide better information on activity data than remote sensing?
- Maintaining consistency as imagery improves – Imagery improves through time. Should countries change image sources as new/better image sources become available? How can this be done while maintaining data consistency? Would it be worthwhile to run a Landsat-based and fine-resolution imagery-based system in parallel to calibrate against, providing a consistent basis to compare against?

Lack of capacities

- Lack of statistical expertise – Countries need to involve experts in statistics, but they are not always involved in the design; usually good expertise in remote sensing but not in sampling design or statistical analysis.
- Need for local capacities – Local capacities are indispensable to make monitoring systems sustainable in the long term. Is there more that can be done to ensure the proper capacities are developed?
- Lack of documentation standards – Not all countries have proper documentation and data archiving of processes allowing reconstruction of results. What can be done to encourage improvement in this regard?
- Lack of SBAE expertise – In some regions it can be very challenging to find experts who can correctly implement SBAE. There may be a trade-off between complexity and transparency. For example, they may implement a complex approach, but depend on external consultants to explain it. What can countries do that lack adequate local expertise?
- Developing monitoring capacity in other sectors – Forest monitoring is moving forward and improving, but other sectors are not. Is there more that should be done to develop consistent methodologies for these other sectors?

Choosing the most appropriate tool for local needs and conditions

- Choosing the most appropriate tool – There are different tools available, each one has its pros and cons. Are those pros and cons well understood by countries? How can countries make informed decisions?

- Lack of or poor-quality internet – Limited internet availability and bandwidth is an issue for some countries, which can significantly impact the use of image streaming services such as Planet data and also the use of tools such as Collect Earth or Collect Earth Online. What can be recommended to countries having these limitations? Guidance and/or comments:

Quality assurance/quality control

Good practices for quality assurance/quality control

- Incomplete implementation of QA/QC – Countries sometimes implement only some parts of the QA/QC process and usually not from the beginning. What can be done to help countries better implement QA/QC?

- Lack of good practice guidance – There is not detailed good practice guidance on QA/QC for all the possible processes in a monitoring system. Can additional good practice guidance be developed?

Interpreter error

- Assessing and using interpreter error – There is somewhat limited guidance on how to characterize interpreter error and how to use the resulting data. Additional guidance is needed.

- Integrating interpreter error into overall uncertainty estimates – Some countries are quantifying interpreter error (such as using within and cross-interpreter interpretations), but not integrating it into the overall error. How can the interpreter error be integrated in the uncertainty analysis? Countries have no incentive to integrate interpreter error because it will increase the uncertainty of the estimates and may reduce their payments.

Reporting

Better communicating and addressing reporting inconsistencies

- Reconciling differences between SBAE and maps – Many countries that use SBAE also create maps. Area estimates from the two products will differ and may create confusion, because countries need to publish shape files or raster due to transparency issues. Inconsistencies in areas reported by maps versus samples can also be difficult to explain and difficult for decision-makers to accept.

- Value of maps – Maps are used for many purposes other than stratification.

- Addressing inconsistencies due to improvements – Making improvements while still maintaining consistency can be challenging. For example, how can finer-resolution imagery for recent periods be combined in a compatible way with coarser-resolution historical imagery? When the improvement enhances the accuracy, countries must make decisions on whether to prioritize consistency or accuracy.

- Incomparable data due to different definitions – Currently countries use vastly different definitions that make the data incomparable between countries.

Incentivizing countries to produce more complete and transparent reports

- Integrating more sources of error – Is reporting more sources of error disincentivized by donors? Are policy changes needed?

- Transparency and data sharing – Transparency and data sharing can sometimes be an issue (such as coordinates of random plots). How can these be improved?

Difficulties of characterizing land use change over short periods of time and variable reporting requirements

- Higher uncertainty due to reduced deforestation – If a country reduces deforestation, this means it becomes a rarer feature over the results period and therefore the error of the estimates increases.
- Difficulty of measuring land use change over short periods – There is a mismatch in mandates (for example, UNFCCC and Forest Carbon Partnership Facility [FCPF] ask for reports each year or every two years, respectively) and guidelines (for example, IPCC asks for land use change). It is difficult to characterize land use change over short periods; therefore, land cover change is generally reported. Adequately characterizing land use change requires the analysis of longer time series.

Sampling design issues

Good practices for stratified area estimation

- Omission errors in stable classes – What are recommended approaches to deal with omission errors occurring in large stable classes, which significantly increase the error of the estimates?
- Buffering approaches – Countries are using different approaches to develop buffers to capture omission errors (for example, Peru is using Morphological Spatial Pattern Analysis and Colombia is using deforestation risk). What are recommended approaches?
- New stratification versus updating a base map – Some countries create a new stratification map for each reporting interval (with a new set of sample units) while others have created a very high-quality base map that is updated through time. In the latter case, new sample units are added in areas of change and existing sample units are reevaluated. What are the pros and cons of the two approaches and using temporary versus permanent sample units?
- Sources of stratification – Some countries lack the capacity to develop their own change maps but need a reliable source of stratification for area estimation. What are recommended alternatives?
- Challenges of global maps for stratification – In some countries such as areas with dry forest, global products have very low accuracy because they have been calibrated using training data from humid and dense forests. What can be done in these cases?
- Recreating strata until desired results achieved – Some countries have created new stratification maps when omission errors are detected by stratified sampling to avoid the associated penalization. Is this valid? What are the implications of this practice?
- Applying new strata to existing pre-stratified samples – Some countries are creating new stratification maps and applying them to a sample that was stratified using a different map. What are the implications of this practice?
- Number of sample units required in stable strata (issue 1 of 2) – How many sample units are needed in stable classes to verify that there are no errors of omission?
- Number of sample units required in stable strata (issue 2 of 2) – How many sample units are needed in the stable classes to reduce the weight of errors of omission of rare classes (especially since deforestation may be rare)? Is there a minimum sample size in the stable classes to be safe?
- Strata labels versus reference levels – Do map classes and labels for stratification have to match the phenomena that are being measured? For example, if deforestation is being estimated, do the strata have to be derived from wall-to-wall information on deforestation?
- Sample units crossing strata boundaries – Countries have encountered situations in which a sample unit falls near a boundary and overlaps into one or more strata. How should they handle the situation? This is especially relevant when using a rare class as a stratum.

- Sample allocation to strata – What is the best way to allocate sample units to strata (equal, optimal, etc.)?
- Arbitrary stratification – Some countries are selecting somewhat arbitrarily the areas in which to intensify their sample. What are the implications of this practice?
- Overlapping samples – occasionally two random sample units may overlap. How should this be handled? Keep both or throw one away?
- Rule-of-thumb for stratification map accuracy – Sometimes accuracy of a change map appears to be extremely low after a stratified area estimate is obtained. Is there a rule-of-thumb for minimum user accuracy/producer's accuracy of the change class for a map to be an efficient stratifier?

Stratified versus systematic area estimation

- Stratified versus systematic area estimation – Some countries are unsure about whether stratified or systematic area sampling is better. What are the pros and cons of the two approaches?
- Preferred design for REDD+ and greenhouse gas only – If a country is interested in designing a monitoring system for only REDD+ and greenhouse gas reporting, is there a preferred design?
- Preferred design for multipurpose monitoring – What is the recommended sample design for a multipurpose monitoring system? For example, what is the best sample design?
    - A base reinterpreted systematic grid coupled with a stratified intensified sample of temporary systematic sample units.
    - A base reinterpreted systematic grid coupled with a nested, wall-to-wall intensified grid in which every sample unit is visually reviewed, but only those containing change are recorded.
    - A stratified sample of a very high-quality base map in which the base map is updated in each reporting cycle with change areas. The original sample units are reinterpreted, and new sample units are added to the change areas, which also are reinterpreted through time.
    - Post-stratifying the base systematic grid using a change map.
    - New samples are drawn from new change maps each reporting cycle (temporary sample units in each cycle).
    - Other?

Temporary versus permanent reference samples

- Temporary versus permanent interpreted sample units – When is it best to use temporary sample units (interpreted for a single reporting period) versus permanent sample units (reinterpreted through time)? What are the considerations related to stratified versus systematic sampling? For REDD+ and greenhouse gas reporting, how important is it to use permanent image plots (sample units) to facilitate comparisons among all monitoring events? What needs beyond REDD+ require data from permanent sample units?
- How to use permanent sample units in stratified area estimation – Permanent sample units can be challenging to work with in stratified area estimation. If permanent sample units are required, what is the best design and approach to use them efficiently?
- Temporary sample units and double counting – How can double counting of change (deforestation, reforestation, degradation, and enhancements) be avoided when the monitoring system uses temporary sample units for visual interpretation?

What are good practices for systematic area estimation?

- Sampling intensity to achieve desired precision – Countries frequently find that few samples fall in areas of change (such as deforestation) and sometimes use a grid density based on resources available (time, capacities, budget). What is the minimum density needed to capture the phenomena of interest (deforestation, degradation) with acceptable precision?
- Adjustments for REDD+ and greenhouse gas reporting – If a country wishes to use systematic sampling to satisfy other needs, what adjustments are needed to obtain the precision needed for REDD+ and greenhouse gas reporting?

Other sample design issues

- Finite versus infinite sampling paradigm – Some countries have used a finite and others an infinite sampling paradigm. Is there a preferred approach? What are the pros and cons of the two approaches?
- Subnational assessments – Some countries are interested in doing subnational assessments. What good practices apply to these assessments (such as nesting approaches)?
- Assessing carbon losses from deforestation – Considering that post-deforestation land use has much larger uncertainties, when countries consider post-deforestation carbon contents in their emission factors, the emission estimates are affected by this assessment. The question then arises: should countries assess post-deforestation once and deduct a fixed quantity from the forest carbon stock to avoid overestimating emissions (and emission reductions) from deforestation, or should they assess this annually where this information may affect the quantity of emissions reductions?
- Other sample designs – Area estimation assumes normal distributions; should other distributions be considered (such as binomial)? Is there a need for more experimental design?
- Country size considerations – Countries vary vastly in size. Are some practices more applicable for larger or smaller countries? Is it easier to make a good map for a small country? Can mapping in a large heterogeneous area cause omission errors to be more likely?
- Combining different sized sample units – If sample unit size changes in the sample design, can a design with multiple sample unit sizes be used?
- Impacts of sample unit size on sample design – Does the sample unit size have any effect on the sample design? Can the sampling size be reduced if the sample unit size is very large?

Response design issues

Designing the land use and land cover classification system(s) for reference data

- Number of reference classes – Some countries use only forest/non-forest classes, others use the IPCC classes, and others use additional, more detailed classes. What are the considerations and good practice recommendations?
- Land use versus land use and land cover – Some countries use only a land use classification system, others use both land use and land cover, and others use a single mixed land use and land cover system. What are the implications and the pros/cons? When would a country want to interpret both land use and land cover? For example, some countries have expressed interest in characterizing tree cover within pasture and agricultural land, which requires both land use and land cover.
- Impact of imagery used for interpretation – How do the spatial, temporal, spectral resolutions of the imagery available impact the design of the classification system?
- Difficult land uses to interpret – Some land use classes are difficult to identify and characterize, such as shifting cultivation systems that rotate between agriculture and forest. What are good practices related to class definitions and identifying these difficult land uses?
- Number of forest classes – How can a country find balance between the number of forest classes and the precision of the emission factors to optimize overall precision and minimize effort? For example, if a country has emission factors for various forest types, is it better to estimate the areas of deforestation for all forest types and apply the refined emission factors for each (this reduces the sample size in each class and increases sampling error), or lump the forest type classes to increase the sample size and improve the precision of the area estimates, but requiring the use of a less precise emission factor? How can a country find the optimal balance?
- Assessing deforestation by forest type – Many countries use different emission factors per forest type requiring activity data to be produced per forest type. What is the recommended approach to get deforestation estimates by forest type, by considering forest type in the sampling design as a stratification or can this information be assigned through labelling of the reference data?

Defining the strata for stratified area estimation

- Creating effective strata – what are good practice recommendations for creating an effective stratification layer? How many strata?

Sample unit (plot) design

- Point samples versus larger sample units or support regions – Is it more efficient to use many single point (or pixel) sample units or fewer larger sample units (for example, 0.5 ha, 1 ha, 2 ha, 10 ha) containing multiple points per sample unit? This question encapsulates the question of generating binary versus continuous data at the level of the sample unit. What are the pros and cons? Is it worth the time and resources for a country to do comparison studies; would it ultimately improve their results?
- Optimal sample unit size – What considerations should be taken into account when determining sample unit size? What is the optimal sample unit size (for example, Landsat pixel, 1 ha, 2 ha, 5 ha)? How can optimal sample unit size be determined? What are the trade-offs? Should the sample unit size correspond to the minimal area in the forest definition?
- Sample unit shape – Does sample unit shape matter (for example, circular versus square versus hexagonal)? What factors should a country consider?
- Image segments and polygons as sample units – A few countries create maps for stratified area estimation using image segmentation. Can the segments be used as sampling units? What are the considerations and implications versus traditional point-type and plot-type designs? Can this approach be recommended?
- Number of points per sample unit – Countries using support region-based or plot-based samples use different numbers of points within the sample units. What are the pros and cons of higher or lower numbers of points? Is there an optimal number of points, how can it be determined and is it worth the effort?
- Two-stage cluster design – Can the points within sample units be used as a two-stage cluster design?

Interpretation paradigms

- Degree of data summarization assigned to sample unit – Countries use different sample unit interpretation paradigms when using sample units larger than a single pixel, such as: assigning only a dominance class to the sample unit; recording the proportions of all land uses and covers within the sample unit; or recording point-level land use and cover data (such as using the Open Foris Collect Earth Online tool). What are the key considerations, as well as pros and cons? Is there a recommended approach?
- Interpretation with or without context – Considering land use minimum area definitions, some countries interpret land use considering the context outside of the sample unit, while others do not. For example, if forest is defined as > 2 ha, and a small patch of forest falls in a sample unit, the sample unit is considered partially forested; the sample unit would be 2 ha in size and if the tree cover within the sample unit reaches the cover definition, the sample unit is considered forested whether or not the area of the patch of trees reaches 2 ha. What are the pros and cons? What are the recommended good practices? Implications of land use class definitions?
- Assessing change not visible from imagery – How can change that is not visible in most imagery be assessed, such as change occurring underneath a tree canopy?
- Approaches to interpreting change across long intervals – When interpreting change within long time periods (such as for a FREL), is it better to interpret only first and last dates and then record dates of change for sample units that change or interpret each time period within the range independently? For example, some countries first interpret the sample units at the beginning and end of the period and second, for sample units that have changed, they review a Landsat time series graph and/or other ancillary data to identify the year(s) in which change occurred. Other countries interpret all sample units for each date range of interest within the full range of dates. What are the pros and cons of the two approaches? Is one preferred over the other?

Definitions of forest, deforestation and degradation

- Forest, deforestation, and degradation definitions – Countries are applying widely varying definitions that are not necessarily comparable and may or may not be clearly defined or may not be defined such that they can be reliably identified. Also, some countries are using multiple definitions for different initiatives. What good practice guidance can be provided?

Annex III
Participants of webinars and respondents of surveys

Frédéric Achard
Naikoa Aguilar-Amuchastegui
Eric Bullock
Oswaldo Carrillo
Laura D'aietti
Rémi d'Annunzio
Emily Donegan
Andres Espejo
Javier Garcia Perez (Gamarra)
Sandro Federici
Yelena Finegold
Randy Hamilton
Martin Herold
Inge Jonckheere
Kay Kallweit
Erik Lindquist
Andrew Lister
Luis Martinez Angel
Ron McRoberts
Carmen Lourdes Meneses Tovar
José Maria Michel
Erik Næsset
German Obando-Vargas
Adrián Ochoa Sanchez
Yong Pang
Paul Patterson
Marieke Sandker
Christophe Sannier
Muri Soares
Steve Stehman
Naila Yasmin

This paper addresses the urgent technical issues encountered in sample-based area estimation (SBAE) for international reporting purposes and greenhouse gas data within the agriculture, forestry, and other land use sector, including Reducing Emissions from Deforestation and Forest Degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries (REDD+). It offers guidance on monitoring forest dynamics, accounting for interpreter variability, defining sample units, and determining the number of assessments required.

By drawing on country experiences and expert consultations, the paper provides practical recommendations, consolidates established practices, and identifies areas requiring further research. It serves as a valuable resource for donors, academia, and countries utilizing or considering SBAE for REDD+ or other international reporting purposes. This comprehensive guide offers insights into current practices and limitations while promoting a deeper understanding of the field.

For more information, please contact:

Forestry Division – Natural Resources and Sustainable Production

NFO-Publications@fao.org

www.fao.org/forestry/en

Food and Agriculture Organization of the United Nations
Rome, Italy

With the financial support of

UK Government

THE WORLD BANK

GFOI Global Forest Observations Initiative

With the technical support of

SilvaCarbon

ISBN 978-92-5-138531-9

CC9276EN/1/01.24