Food and Agriculture Organization of the United Nations



White paper

Good practices in sample-based area estimation

Second edition

White paper

Good practices in sample-based area estimation

Second edition

Andreas Vollrath
Inge Jonckheere
Randy Hamilton
José Maria Michel
Emily Donegan



Food and Agriculture Organization of the United Nations

Rome, 2026

REQUIRED CITATION:
Vollrath, A., Jonckheere, I., Hamilton, R., Michel, J.M. and Donegan, E., eds. 2026. *Good practices in sample-based area estimation.* White paper. Second edition. Rome, FAO. https://doi.org/10.4060/ce2179en.

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations (FAO) concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO in preference to others of a similar nature that are not mentioned.

The views expressed in this information product are those of the author(s) and do not necessarily reflect the views or policies of FAO.
ISBN 978-92-5-141096-7
© FAO, 2026

Second edition 2026
First edition 2024



Some rights reserved. This work is made available under the Creative Commons Attribution- 4.0 International licence (CC BY 4.0: https://creativecommons.org/licenses/by/4.0/legalcode.en).

Under the terms of this licence, this work may be copied, redistributed and adapted, provided that the work is appropriately cited. In any use of this work, there should be no suggestion that FAO endorses any specific organization, products or services. The use of the FAO logo is not permitted. If a translation or adaptation of this work is created, it must include the following disclaimer along with the required citation: "This translation [or adaptation] was not created by the Food and Agriculture Organization of the United Nations (FAO). FAO is not responsible for the content or accuracy of this translation [or adaptation]. The original English edition shall be the authoritative edition."

Any dispute arising under this licence that cannot be settled amicably shall be referred to arbitration in accordance with the Arbitration Rules of the United Nations Commission on International Trade Law (UNCITRAL). The parties shall be bound by any arbitration award rendered as a result of such arbitration as the final adjudication of such a dispute.

Maps. The boundaries and names shown and the designations used on these map(s) do not imply the expression of any opinion whatsoever on the part of FAO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers and boundaries. Dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Third-party materials. This Creative Commons licence CC BY 4.0 does not apply to non-FAO copyright materials included in this publication. Users wishing to reuse material from this work that is attributed to a third party, such as tables, figures or images, are responsible for determining whether permission is needed for that reuse and for obtaining permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

FAO photographs. FAO photographs that may appear in this work are not subject to the above-mentioned Creative Commons licence. Queries for the use of any FAO photographs should be submitted to: photo-library@fao.org.

Sales, rights and licensing. FAO information products are available on the FAO website (www.fao.org/publications) and print copies can be purchased through the distributors listed there. For general enquiries about FAO publications please contact: publications@fao.org. Queries regarding rights and licensing of publications should be submitted to: copyright@fao.org

In preparing this publication, the authors used Claude to support the literature review and text revisions, while Adobe Firefly was used to generate the initial cover illustration. All content has been thoroughly reviewed to ensure accuracy, reliability, compliance with data protection standards, and adherence to ethical principles.

Cover illustration:
© FAO/Massimiliano Martino

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 publication 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. Additionally, the publication will give non-experts an overview of area estimation, as well as its applications and limitations.


© FAO/Guillermo Carlos

Contents

	Acknowledgements	x
	Foreword to the second edition	xi
	Abbreviations	xii
	Introduction	**1**
1	**General design of monitoring system**	**3**
1.1	Role of sample-based area estimation in monitoring degradation, reforestation and afforestation	4
	References	7
1.2	Issues related to varying dates and qualities of imagery in sample-based area estimation	8
	References	11
1.3	Maintaining consistency as imagery improves	12
	References	14
2	**Sampling design**	**17**
2.1	Omission errors in large strata in stratified area estimation	22
	References	25
2.2	Sampling designs for multipurpose monitoring	26
	References	31
2.3	Temporal tracking of land use in the context of temporary versus permanent sample units	33
	References	46
2.4	Sampling considerations for operational land use change monitoring	47
	References	55
2.5	Estimators for sample-based area estimation in finite and infinite populations	56
	References	60
2.6	Improving sampling efficiency with continuous auxiliary variables	62
	References	74
2.7	Assessment for small areas	76
	References	80
2.8	Summary of Section 2	83

3	**Response design**	**87**
	References	90
3.1	Point, pixel, or multipoint areabased sample units: practical considerations	91
	References	93
3.2	Area-based sample unit design: number of points and sample unit size	94
	References	100
3.3	Labelling protocols and sample unit data summarization	101
	References	105
3.4	Interpretation paradigms: Interpretation without or with context	107
	References	109
3.5	Updating reference labels for operational monitoring	110
	References	114
3.6	Summary of Section 3	117
4	**Quality assurance/quality control**	**121**
4.1	Assessing and reporting quality of reference data	122
	References	125
	Annex	**127**
	Annex I Sampling strategies	128
	References	133
	Annex II Baseline surveys and results	134
	Annex III Participants of webinars and respondents of surveys	142

Figures

1	Design and results of a simulation experiment for population estimates of the proportion of the forest	19
2	Hypothetical scenario showing two properties	20
3	Relative margin of error, expressed as the half-width of the 90 percent confidence interval relative to the estimated proportion of the target class, as a function of sample size, for different proportions of the target class	35
4	Relative efficiency of stratified random sampling compared with simple random sampling, as a function of commission and omission error, for equal allocation (left column) and proportional allocation (right column), as well as for target class proportions of 1 percent, 10 percent and 50 percent of the area of interest	39
5	Relative efficiency of stratified random sampling compared with simple random sampling, as a function of commission and omission error, for midway allocation (left column) and optimal allocation (right column), as well as for target class proportions of 1 percent, 10 percent and 50 percent of the area of interest	41
6	Risk of undersampling, expressed as the probability that the sample misses the change completely, as a function of sample size and within-stratum change proportion (Equation 7)	43
7	Illustrations of intensification of simple random and systematic sampling designs for repeated monitoring	49
8	Illustrations of different options for adjusting stratified sampling design for repeated monitoring	51
9	Illustrations of double sampling involving initial dense sampling grid and subsequent stratified sampling	52
10	Example of a continuous layer of change probability based on the Probability Map Subtraction approach (left) and a subsequent K-means stratification (right) for an area around Porto Velho, Brazil	63
11	Continuous layers for deforestation in Salonga, central Congolese lowland forests	66
12	Strata boundary determination methods (with three strata as an example)	67
13	Options for change area assessment within a small area of interest	77
14	Decision tree of interrelated topics in Section 3 and order of decisions to be made	89
15	Illustration of two landscapes with different spatial distributions	95
16	Grid of 25 points spread over an area of 1 hectare	97
17	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	97
18	Comparison of a land use mosaic in Time 1 (a) and Time 2 (b) showing cropland expansion	101
19	Design and results of a simulation experiment for population estimates of the proportion of the forest	103
20	Hypothetical scenario showing two properties	104
21	Assessing interpretation without context	108
22	Characterization of a unit of land subject to conversion from natural forest to a cyclical system.	115
23	A potential approach to update permanent sample labels for operational monitoring of forest change using sample-based area estimation	116

Tables

1	Estimated deforestation area, standard error, and relative efficiency of the regression and difference estimators under two stratification schemes and two sample allocation strategies, based on a total sample size of 5 000 units	10
2	Use of improved imagery while still maintaining consistency in area estimation	14
3	Example data illustrating impact of omission errors on the variance of the estimated proportion of area	22
4	Comparison of standard errors of estimates	24
5	Overview of potential sampling approaches for monitoring updates of reducing emissions from deforestation and forest degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks (REDD+)	53
6	Summary of the terminology, concepts and estimator equations for the two-stage, two-step and cluster sampling strategies	59
7	Different stratification methods and sample allocation for three strata and a sample size of 5 000 units	67
8	Precision of estimated forest change area, for various combinations of stratification and allocation of a sample of 5 000 units	68
9	Model-assisted estimation approaches and their estimators	69
10	Estimated deforestation area, standard error, and relative efficiency of the regression and difference estimators under two stratification schemes and two sample allocation strategies, based on a total sample size of 5 000 units	71
11	Summary of key considerations and recommendations related to sampling design for sample-based area estimation	83
12	Summary of key considerations and recommendations related to plot and response design for sample-based area estimation	111

Box

1	Estimated deforestation area, standard error, and relative efficiency of the regression and difference estimators under two stratification schemes and two sample allocation strategies, based on a total sample size of 5 000 units	10


© FAO/Cristian Palacios Felte

Acknowledgements

This paper, across its first and second editions, was written by the Forest Monitoring and Data Platforms team members at the Food and Agriculture Organization of the United Nations (FAO). The first edition was edited by Inge Jonckheere, Randy Hamilton, José Maria Michel and Emily Donegan. The second edition was edited by Andreas Vollrath.

Specific sections were also written by FAO and external experts. First edition contributions were made by José Maria Michel, Emily Donegan, Frédéric Achard, Erik Lindquist, German Obando-Vargas, Randy Hamilton, Charles T. Scott, Andrew Lister, Ron McRoberts, Steve Stehman, Pontus Olofsson, Remi d'Annunzio, Christophe Sannier, Andres Espejo, Paul Patterson and Inge Jonckheere. Second edition contributions were made by Andreas Vollrath, Nandika Tsendbazar, Javier Garcia-Perez and Sytze de Bruin.

The authors are grateful to the following experts who volunteered to review the content of the second edition: Charles T. Scott, Steve Stehman, Ron McRoberts, Randy Hamilton, Andrew Lister, Erik Næsset and Christophe Sannier. Their valuable feedback improved the clarity and quality of the revised text.

The first edition text was copyedited by Alex Gregor and Rachel Golder, with graphic design by Roberto Cenciarelli and Lorenzo Catena. The second edition text was copyedited by Alex Gregor, with graphic design by Massimiliano Martino. For the first edition, Sara Maulo and Vanessa Vertiz supported the publication process. Vanessa Vertiz supported the publication process for the second edition.

The first edition was produced by FAO thanks to financial support 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. The second edition was also produced by FAO, thanks to financial support from the United Kingdom, as well as Norway's International Climate and Forest Initiative (NICFI).

Foreword to the second edition

by

Andreas Vollrath

When the first edition of *Good practices in sample-based area estimation* appeared in 2024, it set out to answer a practical need. Countries generating activity data for the estimation of greenhouse gas emissions and emission reductions from land use and land use transition had adopted sample-based area estimation (SBAE) widely; however, they kept meeting the same unresolved questions and had little consolidated guidance to turn to. The first edition drew its agenda directly from country experience and the expert consultations that preceded it, and it has since become a reference for practitioners, reviewers and partners working on forest area estimation.

This second edition is a product of the same collaboration, extended over the years that followed. Working alongside national teams, technical partners and reviewers on real measuring, reporting and verification (MRV) systems have shown us where the first edition was most useful, and where practitioners still needed more. The revised and new sections gathered here respond to that feedback. They share a common thread: a deeper and more coherent treatment of sampling design, and a shift in how the map that supports it is understood.

That shift is most explicit in the new section on the use of maps representing a continuous variable related to the likelihood of forest change (Section 2.6). Here the map is treated less as a source of precise cartographic truth and more as a helper whose purpose is to make sampling more efficient. Seen this way, a layer that is quicker and cheaper to produce – and that flags where change is uncertain rather than asserting exactly where it occurred – can be just as valuable for area estimation as a polished categorical map. This is not an argument against mapping. Detailed categorical products retain their own cartographic value, but they are demanding to produce and reporting deadlines are often tight. A probability layer can be generated quickly enough to meet those deadlines and still support a sound estimate, while also serving as a stepping stone towards the more laborious discrete maps when time and resources allow.

Four further contributions complete the picture. A new dedicated section on sample size and allocation (Section 2.3) gives practitioners concrete guidance on how many sample units to collect and how to distribute them among strata. It introduces a midway allocation that, through its simplicity and robustness, offers a simple yet strong way to allocate a sample efficiently, even in the case when multiple categories are of interest. It further shows that how the sample is allocated can matter just as much as the stratification layer itself.

Section 2.4 turns to the multi-temporal dimension of sampling design, showing how permanent samples can be established in stratified designs, as well as ease updating and yield more consistent estimates across the different parameters a country must report.

Section 2.7 addresses small area estimation techniques for when national estimates need to be broken down into smaller regions. Section 3.5 sets out strategies for linking the outputs of change-detection algorithms to visual interpretation, so that permanent sample units can be updated efficiently in operational monitoring, complementing the design considerations of Section 2.4.

Taken together, these additions reflect what we have learned by doing this work with partners rather than in the abstract. As in the first edition, the aim is not to prescribe a single correct method, but to help countries and their partners choose well among defensible options, and to make robust area estimation a little more attainable in practice.

Abbreviations

AI	artificial intelligence
AOI	area of interest
ART	Architecture for REDD+ Transactions
BFAST	Break detection For Additive Season and Trend
BUR	Biennial Update Report
CCDC	Continuous Change Detection and Classification
eSBAE	ensemble sample-based area estimation
FAO	Food and Agriculture Organization of the United Nations
FREL	Forest Reference Emission Level
GFC	Global Forest Change
GFOI	Global Forest Observations Initiative
GIS	geographic information system
GREG	generalized regression estimator
IPCC	Intergovernmental Panel on Climate Change
LULC	land use and land cover
MGD	Methods and Guidance Document
MoE	margin of error
NFI	national forest inventory
PROMS	probability map subtraction
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
SBAE	sample-based area estimation
SRS	simple random sampling
STR	stratified random sampling
SYS	systematic sampling
TMF	Tropical Moist Forests
TREES	The REDD+ Environmental Excellence Standard
UNFCCC	United Nations Framework Convention on Climate Change
VLM	vision language model

Units and formulae

CV	coefficient of variation
h	stratum
ha	hectare(s)
k	kilometre(s)
k^2	square kilometre(s)
m	metre(s)
m^2	square metre(s)
M	number of population units (points) within a sample unit
N	number of sample units
$\hat{p}$	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


© FAO/Enric Català Conteras

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). Staff of the Food and Agriculture Organization of the United Nations (FAO) 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 the second edition 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).



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 light detection and ranging [LIDAR] sensors 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 the Global Ecosystem Dynamics Investigation (GEDI) instrument on the International Space Station and BIOMASS Earth Explorer mission of the European Space Agency (ESA) (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

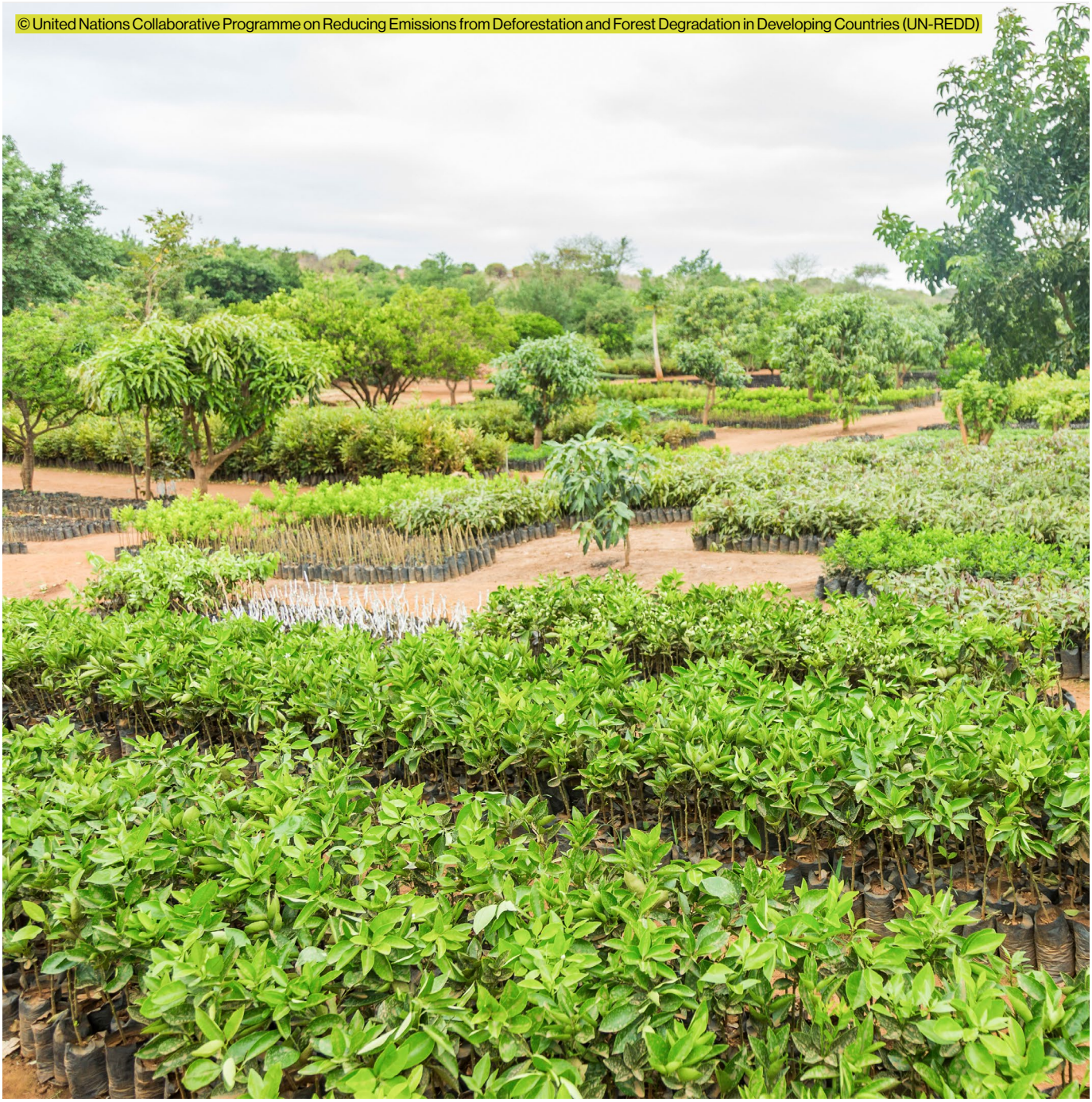
[5] See https://www.forestcarbonpartnership.org/system/files/documents/Guatemala_ERPD_11_05_2019.pdf

[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.


© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)

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). 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). 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). https://www.forestcarbonpartnership.org/system/files/documents/Guatemala_ERPD_11_05_2019.pdf

FCPF. 2019b. Dominican Republic. Emission Reductions Programme Document (ERPD). 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 datasets 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 multi-temporal 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 detection 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 Programa de Cálculo do Desmatamento da Amazônia (Satellite-based Monitoring of Deforestation in the Brazilian Amazon Forest [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 Sentinel-2 L1C cloud-free composites product of the European Commission's Joint Research Centre (JRC). 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 3,[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.



↘ **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 3.5). 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 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. 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*. https://unfccc.int/resource/docs/2013/cop19/eng/10a01.pdf#page=39

© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)/Alice Van der Elstraeten





Section 2

Sampling design



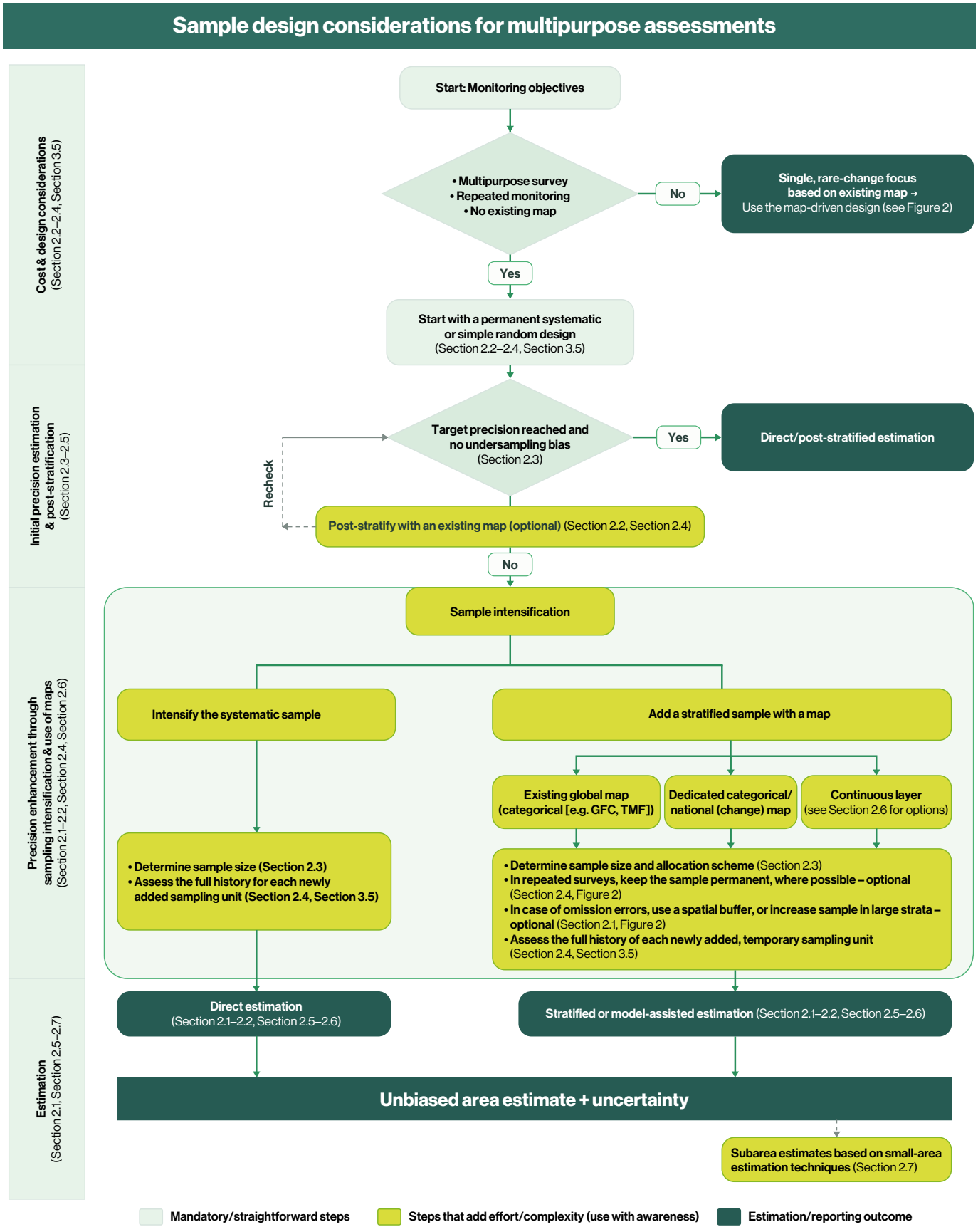
When implementing SBAE, particular attention should be paid to developing a sampling design that achieves the desired outcome in an efficient and cost-effective manner. Over the years, countries have implemented a variety of sampling designs, encountering several challenges along the way. This section addresses these challenges and provides practical guidance on the questions related to sampling design that have arisen from country experience. Since the first edition, this guidance has been considerably expanded. Alongside the original considerations, it now gives dedicated treatment to the sizing and allocation of the reference sample, the demands of operational monitoring across successive reporting periods, the use of maps that represent forest change as a continuous variable, and the estimation of change for small areas within the overall area of interest.

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. the sampling design of a multipurpose monitoring system (Section 2.2);
3. practical strategies for reducing survey costs by efficiently determining the size of the reference sample and its allocation among strata (Section 2.3);
4. sampling considerations for the operational monitoring of land-use change across successive reporting periods, in particular the trade-off between the temporal consistency of permanent samples and the efficiency of temporary ones (Section 2.4);
5. the theoretical and practical implications and differences among two-stage, cluster and two-step sampling strategies (Section 2.5);
6. improving sampling efficiency by using maps that represent forest change as a continuous variable, whether to define strata or support model-assisted estimation (Section 2.6); and
7. the assessment of change for small areas within the overall area of interest for situations in which design-based estimation for those areas is not feasible (Section 2.7).

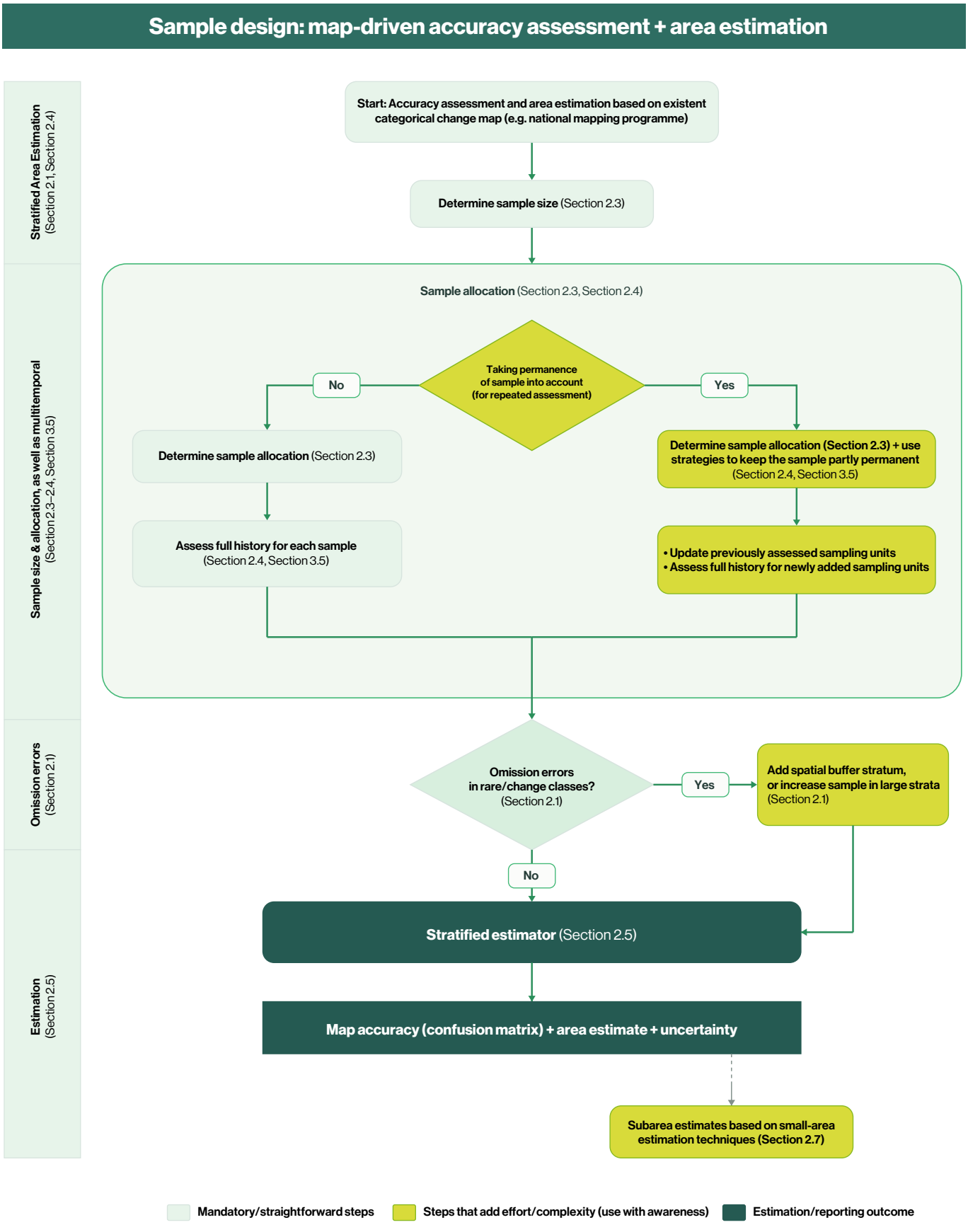
While each topic stands on its own, the decisions behind them are interrelated. In particular, the choice between permanent and temporary samples and the use of a map for stratification carry implications that go beyond sampling design and shape aspects of the response design covered in Section 3. Figure 1 and Figure 2 present decision trees that show the interrelationship of these topics and outline the flow of sampling design decisions in a broader context of a multipurpose, repeated assessment that may need to fulfill various monitoring objectives (Figure 1) or targets the validation and bias-adjustment of a map that is central to the reporting of area estimates of REDD+ categories and beyond (Figure 1).

Figure 1 Decision tree of interrelated topics in Section 2 and the order in which sampling design decisions are made for a repeated, multipurpose oriented assessment



Notes: GFC – Global Forest Change; TMF – Tropical Moist Forest. The post-stratification loop reuses the existing sample without new interpretation; move onto intensification only if precision is still insufficient. Continuous-layer options are listed in Section 2.6. Guard against undersampling (rule of three) throughout (Section 2.3); update permanent-sample labels across reporting periods (Section 3.5).
Source: Authors' own elaboration.

Figure 2 Decision tree of interrelated topics in Section 2 and the order in which sampling design decisions are made for map-driven area estimation and map accuracy assessment



Source: Authors' own elaboration.


© FAO/Roberto Cenciarelli
NFI Project
Papua New Guinea

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). 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
3 = 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 sampling (SRS) or systematic sampling (SYS) (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):

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

(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 SRS (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: 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.

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 (SYS 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' accuracy (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 SRS and 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

[12] See Equation 6 and Equation 7.

[13] See Equation 5.1 and 5.3.

with respect to the attributes. Each plot is assigned to a single map class – a process known as post-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].

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

[14] See Equation 20 and Section 2.5.

[15] See Equation 5.47.

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.

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 sampling (SYS) (or SRS) 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 (Section 2.4 and Section 3.5). 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 (Section 2.4). 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 LULC 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 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 described previously). 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 previously 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). 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*. 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


© FAO/Lorenzo Catena

2.3 Practical strategies for reducing survey costs by efficiently determining sample size and allocation

by

Andreas Vollrath, Sytze de Bruin, Nandika Tsendbazar and Javier Garcia-Perez

Description of the problem

Determining an appropriate sampling design includes the process of deciding how many sampling units are placed where – that is, determining the size of the sample and how the total reference sample is distributed across the sampling frame (the sample allocation). These decisions are an integral part of the sampling design (see Section 2.2) and closely linked to the choices made before: the sampling strategy (systematic or stratified) and, for stratified designs, the stratification layer (for example, a forest change map).

If the sample collection costs do not differ between the strata, which is likely the case when the reference labels are obtained by visual interpretation, the total reference sample size may be determined based on a given available budget – that is, the reference sample size (n) equals the budget divided by the cost per sampling unit. If, instead, achieving a target level of precision is paramount, it is possible to obtain an educated guess on the necessary sample size before the actual data collection starts. For SYS and SRS, this process is straightforward (see Equation 3) – as long as a rough estimate for the proportion of the target class exists (for example, from previous surveys or global datasets).

Stratified random sampling approaches are typically more efficient for targeting relatively rare events such as forest change and necessitate a smaller sample size. This involves an extra step of how that total reference sample should be divided among the strata and is referred to as sample allocation. Alongside the strategy for stratification itself, this decision can have a substantial effect on the precision of the resulting estimates. Two sampling surveys may use the same stratification and the same sample size, and thus the same interpretation effort and cost, yet produce area estimates varying degrees of precision, simply because the distribution of sample units was different among the strata.

Since each sampling unit must be visually interpreted, efficiency translates directly into cost: a design twice as efficient delivers the same precision for half the budget. In theory, an optimal sample allocation approach (Neyman allocation) exists. In practice, however, determining it requires a priori knowledge of the variability within each stratum, knowledge that might not be available, or knowledge that comes at the additional cost of collecting a preliminary sample. As a consequence, no single allocation scheme can be recommended for all situations, and the choice involves a trade-off between theoretical efficiency and the reliability of the assumptions it rests on.

It should be noted that neither the choice of stratification nor the choice of allocation scheme introduces systematic errors (bias) into the area estimates. Statistical rigour requires a probability-based sample and an estimator that matches the design (Stehman and Foody, 2019). Once both conditions are met, the estimates remain unbiased regardless of how unsuitable the map or how inefficient the allocation may be. These decisions instead determine the efficiency of the survey, and thus the cost at which a given degree of precision is achieved. A poor stratification or allocation does not produce wrong estimates; it produces costly ones.

The one exception arises from undersampling. When a stratum receives too few sampling units, the rare events within it may not be captured by the sample at all, and its variance can no longer be estimated reliably. In practice, every stratum should therefore be assigned a minimum number of sampling units, large enough to capture the rare events it may contain. Once this minimum is secured, all remaining design choices are questions of efficiency, not of correctness, a distinction that matters for practitioners planning a survey as much as for verifiers assessing one.

Overview of existing good practices

Systematic sampling

Under a SYS or SRS design, all sampling units are selected with equal probability across the sampling frame, and the only decision to be made is the total sample size. The underlying concept is intuitive: we want to estimate the proportion of a certain class – for example, the fraction of the area affected by deforestation – by interpreting a limited number of sampling units rather than the entire territory. Because only a subset of the population is observed, the resulting estimate carries an uncertainty, expressed as a margin of error around the estimated proportion (the half-width of its confidence interval). The larger the sample, the smaller this margin of error becomes. Cochran (1977) formalized this relationship in a widely used formula that determines the sample size (n) required to reach a desired precision, based on three quantities: the expected proportion of the class of interest (p), the acceptable margin of error (e), and the confidence level at which that margin should hold, represented by the critical value (Z) from the standard normal distribution:

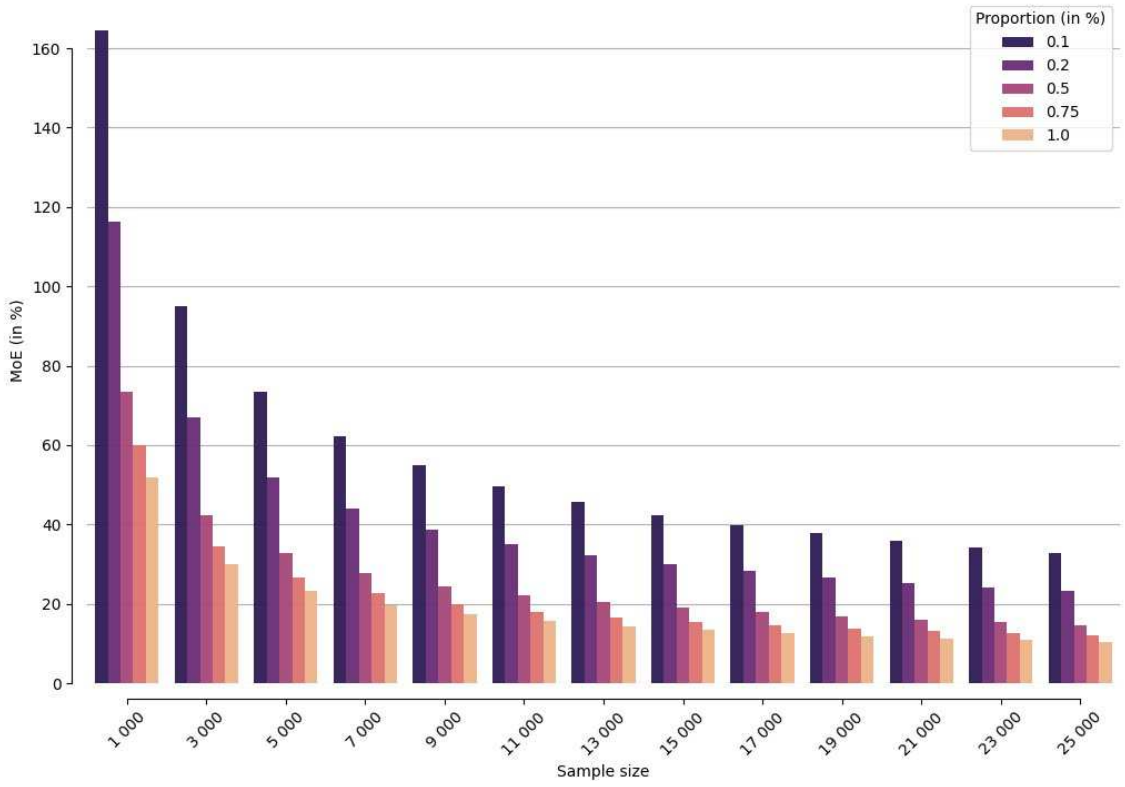
$$n = \frac{Z^2 p(1-p)}{(ep)^2}$$

(Equation 3)

Two properties of this relationship deserve attention. First, precision comes at a disproportionate cost: reducing the margin of error by half requires quadrupling the sample size. Consequently, as sample size increases, each additional sampling unit yields progressively smaller gains in precision (see Figure 3). Second, the expected proportion or area of the target class itself strongly influences the required sample size. Rare classes such as forest change often cover well below 1 percent of a country's area. This necessitates considerably larger sample sizes to be estimated with the same level of precision relative to the target's variable proportion. This is illustrated in Figure 3, which shows the relative margin of error for different combinations of sample size and class proportion: for a given sample size, the relative error increases sharply as the proportion becomes smaller. This is a direct consequence of equal-probability sampling. When the class of interest is rare, only very few sampling units will actually fall into it, and it is these few units that carry the information on which the estimate rests.

It is important to note that Figure 3 expresses uncertainty as a relative margin of error (as a percentage of the estimated proportion of the target class within the sampling frame). Relative errors are naturally higher for smaller proportions, but this does not necessarily mean the estimate is less informative in practical terms: a 40 percent relative error on a change proportion of 0.25 percent corresponds to an uncertainty of only ±0.1 percentage points of the total area, while a 10 percent relative error on a proportion of 20 percent corresponds to ±2 percentage points – 20 times larger in absolute terms. The required sample size with a targeted standard error that is expressed as a proportion (relative errors) can also be calculated following a slightly different version of Equation 3 (see Equation 3.2 in Tyukavina *et al.* [2025]). Both perspectives should be kept in mind when laying out the sampling design and calculating the required sample size, as well as when interpreting the precision of the resulting area estimates.

Figure 3 Relative margin of error, expressed as the half-width of the 90 percent confidence interval relative to the estimated proportion of the target class, as a function of sample size, for different proportions of the target class



Notes: MoE – margin of error.
Source: Computed from a sample size formula (Equation 3) for an equal-probability (SRS or SYS) sample from Cochran, W.G. 1977. *Sampling techniques, Third edition*. New York, USA, John Wiley and Sons.

Stratified sampling

When a map is used to stratify the sampling frame, the allocation problem consists of two decisions:

1. choosing the total sample size; and
2. dividing this total among the strata.

The two are treated separately in the following subsection, although they are not independent: the total that is needed depends on how it will be divided, and vice versa.

Sample size calculation for stratified sampling

Four distinct approaches are highlighted here to compute the reference sample size required to achieve a target level of precision using an STR design. They differ in how the quality of the stratification map is factored into the calculation. In all cases, it is important to note that these sample size calculations should be treated as a "rough" guide because they strongly depend on the a priori information used in the calculations (such as expected or guessed change fractions, relative efficiency and assumed stratification map quality). In practice, those are not always well known, as they will only be revealed by the sampling exercise itself.

The first approach is a deliberate back-of-the-envelope calculation that is useful for a quick pre-assessment in the absence of any information on the map's quality. It does not replace the formal calculations that follow, but it provides a rough first estimate within minutes, based on a single assumption and – if Figure 3 is used as a look-up table – without any formula at all. The

starting point is the sample size for the target class under an unstratified design, obtained from Equation 3 or read directly from Figure 3. To account for the typical gain in efficiency that stratification provides, this sample size is then divided by an assumed relative efficiency that an STR design might bring (see Box 1). Stratified designs built on high-quality maps typically reach relative efficiencies of around 3 to 5 for rare classes (Arévalo *et al.*, 2020) (Figure 4 and Figure 5). If Equation 3 yields a sample size of 15 000 for a given target error, a stratified design with an assumed relative efficiency of 3 would accordingly require about 5 000 units.

A second approach, presented as Equation 12 in Olofsson *et al.* (2014) determines the sample size from an assumed overall accuracy of the map used for stratification and the desired width of the confidence interval around it. Similar to the first approach, its appeal lies in its simplicity, but relying on overall accuracy is dangerous in the context of activity data. The classes of greatest interest are typically the change categories (which are rare); overall accuracy, dominated by the large stable classes, is a very poor predictor of how well those rare classes are estimated, likely resulting in a too small sample size (with an elevated risk of undersampling bias, as described in the following text).

Equation 13 in Olofsson *et al.* (2014) presents another sample size calculation that builds on the class-specific user's accuracies. Its drawback is that it requires user's accuracies for all map classes, quantities that have to be guessed from previous maps or approximated by an initial sampling exercise (added cost).

Olofsson *et al.* (2014) are explicit about this limitation, describing sample size planning as an "inexact science" and recommending that the calculation be repeated for a range of plausible accuracies before a final decision is taken. It is also worth noting that the resulting sample sizes tend to be modest and may prove too small in practice for the precise estimation of rare change classes. The recent land cover validation guideline also details similar approaches for calculating total sample size for stratified sampling and sample size estimation for multi-estimation purposes (Section 3.5 in Tyukavina *et al.* [2025].

The fourth approach takes advantage of the fact that the total sample size is mostly determined by the largest stratum, and therefore only needs a priori knowledge of the proportion of the target variable therein. As this approach is not largely documented, it is laid out in more detail here:

1. compute the sample size for the largest stratum, using Equation 3 based on the guessed change fraction in the largest stratum;

2. compute the sample fractions, given the chosen allocation scheme; and

3. compute the total sample size as that of Step 1, divided by the sample fraction for the largest stratum from Step 2.

For example, a discrete map depicting deforestation and forest degradation over 5 percent of the area of interest is used for stratification. The other 95 percent of the area is depicted as stable forest. Our initial guess is that 1 percent of the mapped stable forest is also subject to deforestation and forest degradation, so $p = 0.01$ for the large stable stratum. Our aim is to assess the change fraction in that stratum with the probability of a maximum 25 percent relative margin of error at a confidence level of 0.9:

- For a 0.9 confidence level, the significance level, α = 0.10, and the corresponding critical value from the standard normal distribution, is $z\left(1-\frac{\alpha}{2}\right) = z\left(0.95\right) = 1.645$.
- Hence, the sample size for the largest stratum (Equation 3) is $1.645^2 \times 0.01 \times 0.99 \ / \ (0.25 \times 0.01)^2$, which results in 4 286 points to interpret.
- Assuming a proportional sample allocation, 95 percent of the total sample size would be allocated to the stable stratum.

Hence, the total sample size is then 4 286 / 0.95 = 4 512.

Its economy of assumptions comes with a sensitivity that should not be overlooked: the required sample size rises quickly as the expected proportion decreases so that a seemingly small change in the guessed proportion can double the total sample size. If we change *p* to 0.005 for the given example, the total sample size balloons to 9 069 samples. On the other hand, this approach is more robust in avoiding undersampling bias, as described in the following subsection.

Strategies of sample allocation for stratified sampling

For STR, the question of how to divide the total sample among the strata has traditionally been answered pragmatically. The good practice guidance of Olofsson *et al*. (2014) discusses the range of options between optimal (Neyman) and proportional allocation, recognizing that no single scheme serves all estimation objectives. They propose midway-type compromises – such as allocating a fixed 50 units to 100 units to each rare change stratum and distributing the remainder proportionally – for their example calculation.

What this guidance does not provide, however, is a quantitative link between the chosen allocation and the efficiency of the resulting design. Skakun (2025) demonstrates how maps with identical overall accuracy can differ substantially in how efficiently they support area estimation. Benchmarking stratification maps on overall accuracy or F-score should therefore be discouraged when area estimation is the objective. Instead, a map should be evaluated by what it is used for: when the objective is area estimation, the relevant question is not how accurate the map looks, but how much sampling effort it saves when used for stratification – the comparison of the statistical efficiency between two sampling designs (see Box 1 for further explanation).

An interesting emphasis of Skakun's remarks is that for rare classes, the producer's accuracy (the inverse of omission error) matters considerably more than the user's accuracy (the inverse of commission error) with regard to the statistical efficiency of area estimation. This finding links directly to Section 2.1, where the problematic influence of omission errors on the uncertainty of area estimates is discussed. A general rule, independent of the selected allocation, is therefore to avoid maps with large omission errors for stratification purposes. This has a practical implication for the choice of a stratification map. Certain ready-made products may have been designed to avoid overpredicting change, thereby accepting omission errors in exchange for a high user's accuracy. This does not make them the best option for stratification.

Box 1. Relative efficiency – *the value* of a good stratification

When areas are estimated from a probability sample, the map used for stratification does not determine whether the estimate is correct. Design-based estimators remain unbiased even when the underlying map is poor, as errors in the map are corrected by the sample data[i]. What the map does determine is how precise the estimate is for a given effort: a good map concentrates the interpretation effort where the changes actually happen, so that fewer sampling units are wasted on areas that hold little information. This gain can be quantified by the relative efficiency (RE), which compares the variance (i.e. the squared uncertainty) of the estimate under simple random sampling (SRS) with the variance under the stratified (STR) design, for the same total sample size n:

$$\mathrm{RE} = \frac{\mathrm{var(SRS)}}{\mathrm{var(STR)}}$$

A relative efficiency of 3 means the stratified design achieves the same precision as a simple random sample 3 times its size – or, put in terms of budget, that the same precision can be reached with only a third of the sampling units:

$$\mathrm{n_{STR}} = \frac{\mathrm{n_{SRS}}}{\mathrm{RE}}$$

The practical consequences are substantial. In a worked example by Skakun,[i] estimating a class covering 15 percent of the area with a target precision of 5 percent requires 2 267 sampling units under SRS, but only 693 units when stratifying with a map of moderate quality (producer's accuracy of 0.9, user's accuracy of 0.7) – a reduction of about 70 percent in interpretation cost. The gains are largest exactly where they are needed most: for rare classes such as forest change, high-quality maps typically reach relative efficiencies of 3 to 5, whereas for classes covering large area proportions the advantage of stratification largely disappears.

Source:
[i] Skakun, 2025
[ii] Arévalo *et al.*, 2020

The most straightforward approaches to sample allocation are equal and proportional allocation (Cochran, 1977; Stehman, 2012). Their appeal lies in their simplicity, as they require nothing more than the number of strata (H), the total sample size (n), and, in the proportional case, the class weights (W_h) (the area proportions of the map classes):

$$n_{h,\ equal} = \frac{n}{H}$$

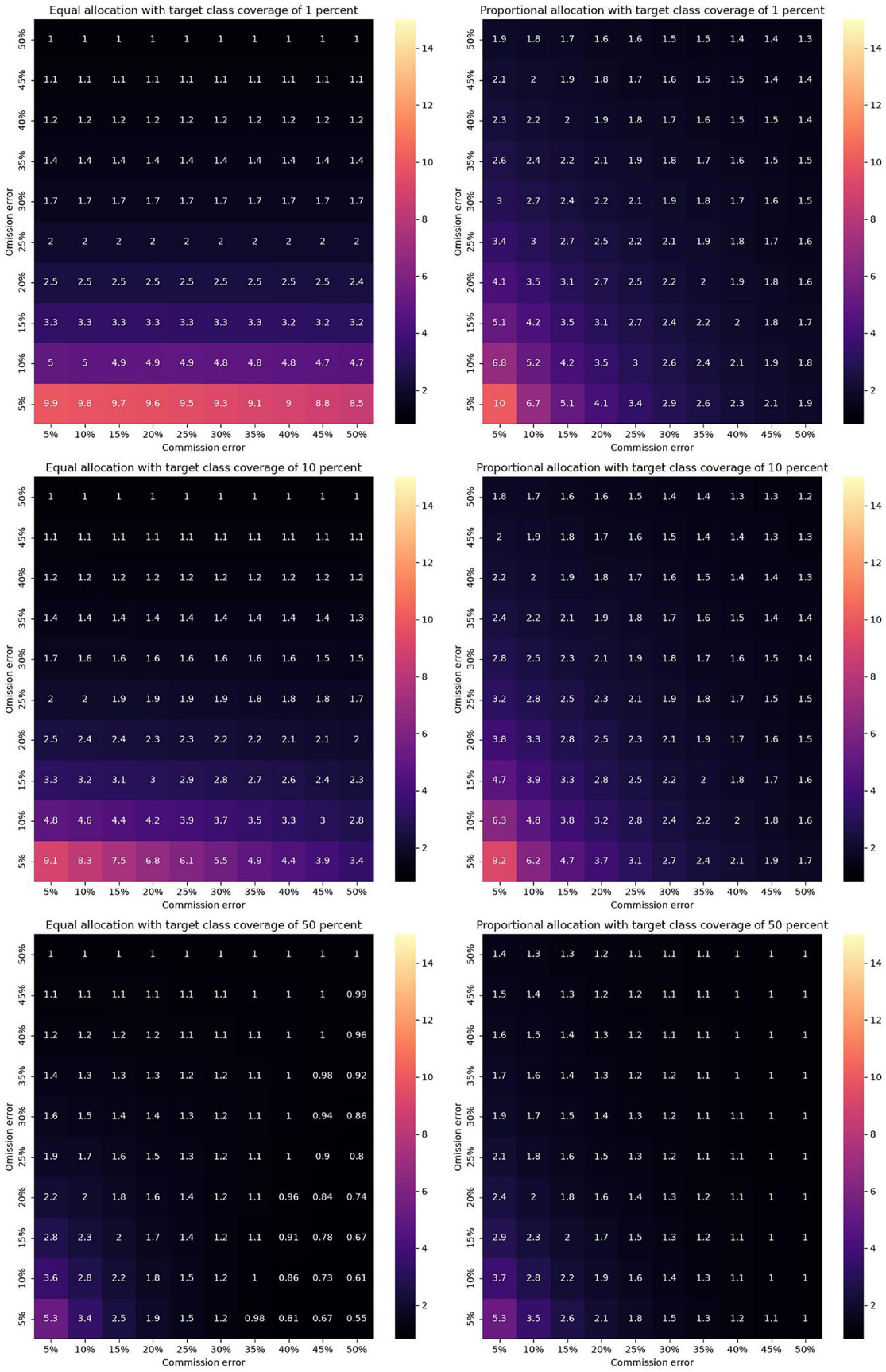
(Equation 4)

$$n_{h,\ proportional} = W_h \times n$$

(Equation 5)

Their efficiency depends on the error distribution of the stratification map (Figure 4). Equal allocation tends to be more efficient when commission errors prevail over the omission error for the target class. Proportional allocation, in turn, retains a higher relative efficiency when omission errors dominate or errors are balanced.

Figure 4 Relative efficiency of stratified random sampling compared with simple random sampling, as a function of commission and omission error, for equal allocation (left column) and proportional allocation (right column), as well as for target class proportions of 1 percent, 10 percent and 50 percent of the area of interest



Note: AOI – area of interest.

The optimal Neyman allocation is the most efficient choice among all accuracy combinations (Figure 5); it should be used whenever the within-stratum variability can be credibly estimated. One qualification is essential; however, the optimal allocation, as shown in Figure 5, presumes perfect knowledge of the within-stratum variances. In reality, this amounts to a chicken-and-egg problem, as the variance within a class is a direct function of the very change proportion the survey is meant to estimate. If it were known beforehand, no sample would be needed in the first place. In practice, it is derived either from previous assessments or from a first, preliminary sampling exercise. The optimal allocation can therefore only be approximated from assumed variances, and its real-world efficiency is bounded by the quality of those assumptions. Moreover, the implications and costs differ depending on whether a stratum receives too few sampling units (undersampling) or too many (oversampling). The contribution of a stratum to the overall uncertainty is inversely related to the number of units it receives, meaning that the fewer units a stratum holds, the more weight each single unit carries. Undersampling a stratum that turns out to be more variable than expected therefore sharply inflates the uncertainty of the survey. In its extreme form, undersampling does not only reduce precision but risks missing rare events altogether (see undersampling bias in the text that follows). Oversampling, in turn, is far more forgiving. The surplus units in a stratum that turns out to be less variable than expected are not lost – they merely contribute less than they would have elsewhere. Since the true variability is unknown at the design stage, it is safer to guard against the costly mistake than to gamble on having guessed the variances correctly.

An allocation scheme that approaches the optimum without requiring any prior knowledge is therefore introduced here as the midway allocation. Following the intuition of Olofsson *et al*. (2014), the compromise in distributing sampling units between the large stable strata and the small change strata can be cast into an explicit and simplistic rule, in which the sample size of each stratum (*h*) is the average of its equal and proportional shares, as follows:

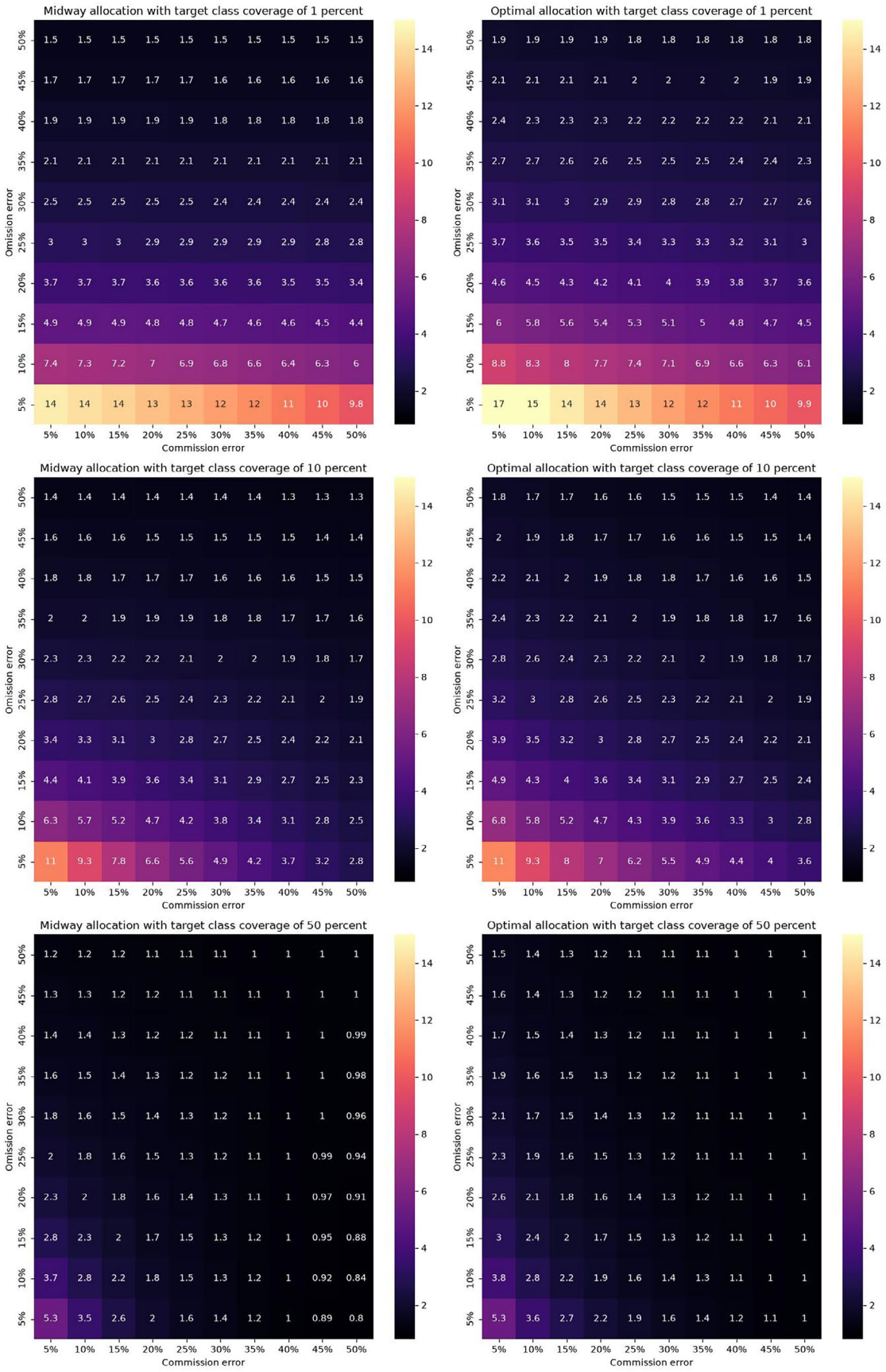
$$n_{h,midway} = \frac{n_{h,equal} + n_{h,proportional}}{2} = \frac{\frac{n}{H} + W_h \times n}{2}$$

(Equation 6)

Given a map of two classes, with a 98 percent stable and a 2 percent change stratum, the equal allocation would distribute 500 samples in both strata, independent of their size. The proportional allocation, in turn, would only allocate 20 samples in the change class, and the remaining 980 samples in the stable one. The midway allocation, in this specific case, results in 260 sampling units in the rare change strata, and 740 sampling units in the stable one.

From a theoretical point of view, Equation 6 can also be interpreted as a proportional allocation in which the mapped stratum weights are pulled halfway towards equality: each stratum is allocated as if its weight were the average of its mapped proportion (W_h) and the equal share (1 / H). Statisticians refer to this as shrinkage: when a piece of information is not fully reliable, its influence is deliberately softened rather than trusted in full. If the within-stratum variabilities are unknown, the mapped areas alone should not be allowed to determine the entire allocation.

Figure 5 Relative efficiency of stratified random sampling compared with simple random sampling, as a function of commission and omission error, for midway allocation (left column) and optimal allocation (right column), as well as for target class proportions of 1 percent, 10 percent and 50 percent of the area of interest



Notes: AOI – area of interest.

As illustrated in Figure 5, this simple allocation formula comes remarkably close to the optimal allocation in terms of relative efficiency, and it performs better than equal or proportional allocation under almost all error combinations. The only case in which proportional allocation is slightly superior is the combination of higher omission error rates with very low commission errors. Given that a truly optimal allocation is unattainable in practice, the midway allocation is almost as good as the optimum for practical purposes.

Another, more practical advantage of the midway allocation arises when multiple target variables are to be covered, such as deforestation and forest degradation. Neyman allocation can only be optimized for one of them at a time, since the within-stratum variances entering the formula are specific to a single target variable. As a result, an allocation that is optimal for one parameter is generally not optimal for another. The midway allocation, by contrast, uses no variable-specific information at all: it is parameter-independent by construction and therefore serves all target variables reasonably well. For multiparameter surveys, it is thus not merely an approximation to the optimum, but often the most sensible choice altogether.

For completeness, it should be noted that further allocation schemes exist beyond those presented here. Hidiroglou and Kozak (2018) provide a comprehensive overview, including power and X-proportional allocation, which position themselves between the extremes in a similar spirit to the midway allocation. Like the Neyman allocation, however, these schemes depend on a priori knowledge of the within-stratum variability or of an auxiliary variable correlated with the target class; situations in which they hold a true advantage over the schemes presented here are rare. Readers interested in the full range of options and their formal properties are referred to the original article.

A cautionary note on stratification and the efficiencies of different allocation comes from Pagliarella *et al*. (2018), who compared spatially balanced (systematic-like) sampling with various allocation schemes in stratified designs. In their study, stratified sampling was never better than spatially balanced sampling, with proportional allocation performing slightly worse and strongly unbalanced allocations performing markedly worse in comparison. This does not align with general sampling theory: a stratified design outperforms an unstratified one whenever the strata separate change from no-change reasonably well, and the size of that gain depends directly on the user's and producer's accuracies of the stratification map. Their finding is rather specific to the test case, which uses a single artificial population with a particular map error structure and known within-stratum variances, and should not be generalized.

Undersampling bias: the risk of missing rare events

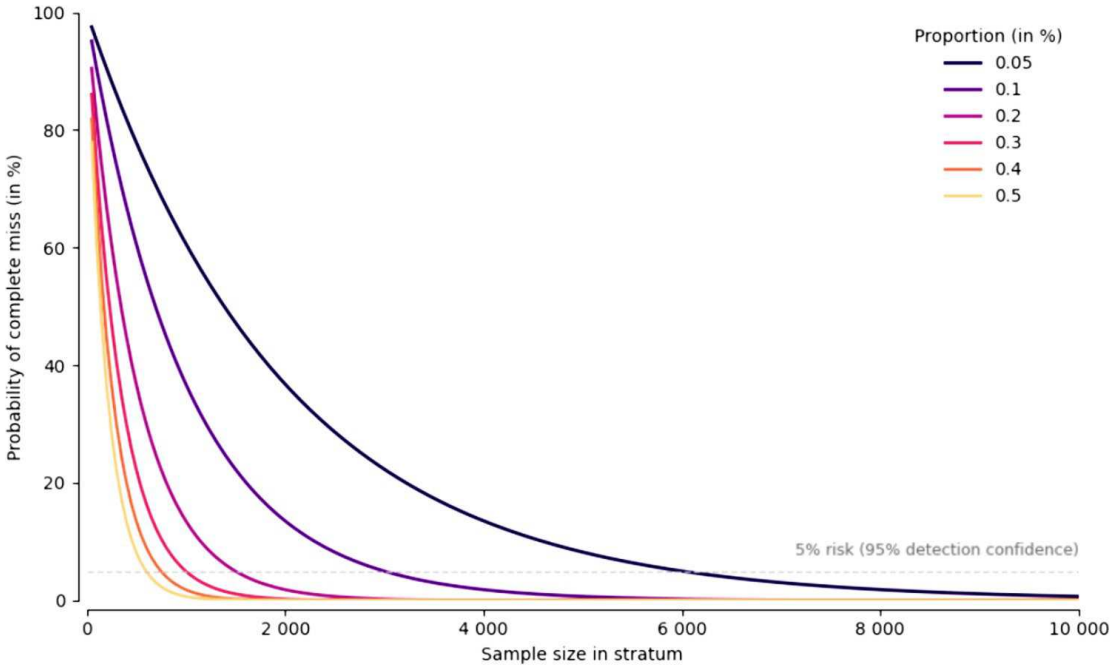
A biased estimate would result from a stratum receiving so few sampling units that a rare event occurring within it is likely to be missed entirely. The typical case in activity data is a change that the map has omitted and that therefore sits in the large stable stratum. Whether the sample detects it is a matter of simple probability. If the undetected change covers a proportion (p) of the stratum and n_h sampling units are allocated to it, the probability (P_0) that not a single unit falls into the change is:

$$P_0 = (1 - p)^{n_h}$$

(Equation 7)

In this equation, (p) = 0.005 and 100 sampling units, $P_0 = 0.995^{100} \approx 0.61$. In other words, there is a 61 percent chance that the survey reports zero change for that stratum, with an estimated standard error of zero, suggesting certainty where there is none. Although the estimator remains unbiased in the formal, design-based framework, any individual survey will then most likely underestimate the change. This is the practical meaning of undersampling bias. Equation 7 has been used to depict the risk of undersampling as a function of sample size and various within-stratum change proportions (Figure 6).

Figure 6 Risk of undersampling, expressed as the probability that the sample misses the change completely, as a function of sample size and within-stratum change proportion (Equation 7)



Notes: The dashed line marks the 5 percent risk (95 percent detection confidence).

The same relationship can be read in reverse: rather than asking how likely a given change proportion is to be missed, one can ask the minimum proportion that a given sample size can reliably detect.

$$p_{min} = 1 - \alpha^{\frac{1}{n_h}} = \frac{-\ln(\alpha)}{n_h}$$

(Equation 8)

At 95 percent confidence, $p_{min} \approx \frac{3}{n_h}$, which is the well-known rule of three (Hanley and Lippman-Hand, 1983). A stable stratum with 100 interpreted units can therefore only reliably reveal change covering more than about 3 percent of its area at 95 percent confidence level. With 1 000 units, this bound tightens to 0.3 percent. Reporting it is a simple and transparent way for both practitioners and verifiers to quantify the maximum plausible extent of undetected change. Expressed as a share of the total mapped area rather than of the stratum alone, a bound of 3 percent within a stable stratum covering 98 percent of the map corresponds to roughly 2.9 percent of the total area, a figure that can exceed the size of the mapped change class itself.

This concept of detection minimum also shows why the fourth approach of the sample size calculation is the safest guard against undersampling bias. By dimensioning the largest stratum for a target precision of the change proportion expected within it, it automatically allocates far more units than the bare detection minimum of Equation 8 requires. In the example in the preceding text, the target of a 25 percent relative error for a change proportion of 1 percent led to 4 286 units in the stable stratum, whereas merely detecting that change with 95 percent confidence would require about 300. A sample sized for precision therefore clears the detection threshold by a wide margin, provided the guessed change proportion is not grossly overestimated.

A final remark concerns the calculation of the confidence interval. The standard procedure, known as the Wald interval, underestimates the uncertainty for small proportions, and it is the reason why a stratum with zero detections reports an uncertainty of zero. Stehman and Xing (2022) show that this undercoverage already occurs when the target class covers roughly 10 percent or less of the area of interest, a situation that applies to virtually all activity data.

They recommend alternatives based on the Wilson interval instead. Since these alternatives factor in the sample size of the affected stratum, they yield a non-zero uncertainty even when no change is detected, offering a simple way to report estimates with a realistic uncertainty at low sample sizes and thus at reduced cost. Further elaborations on this issue, as well as practical solutions, are presented in Section 3.5 of Tyukavina *et al*. (2025).

Way forward

Sample size and allocation are not isolated decisions. The target error and the available budget provide the immediate guidance, but both are embedded in a broader context of monitoring objectives.

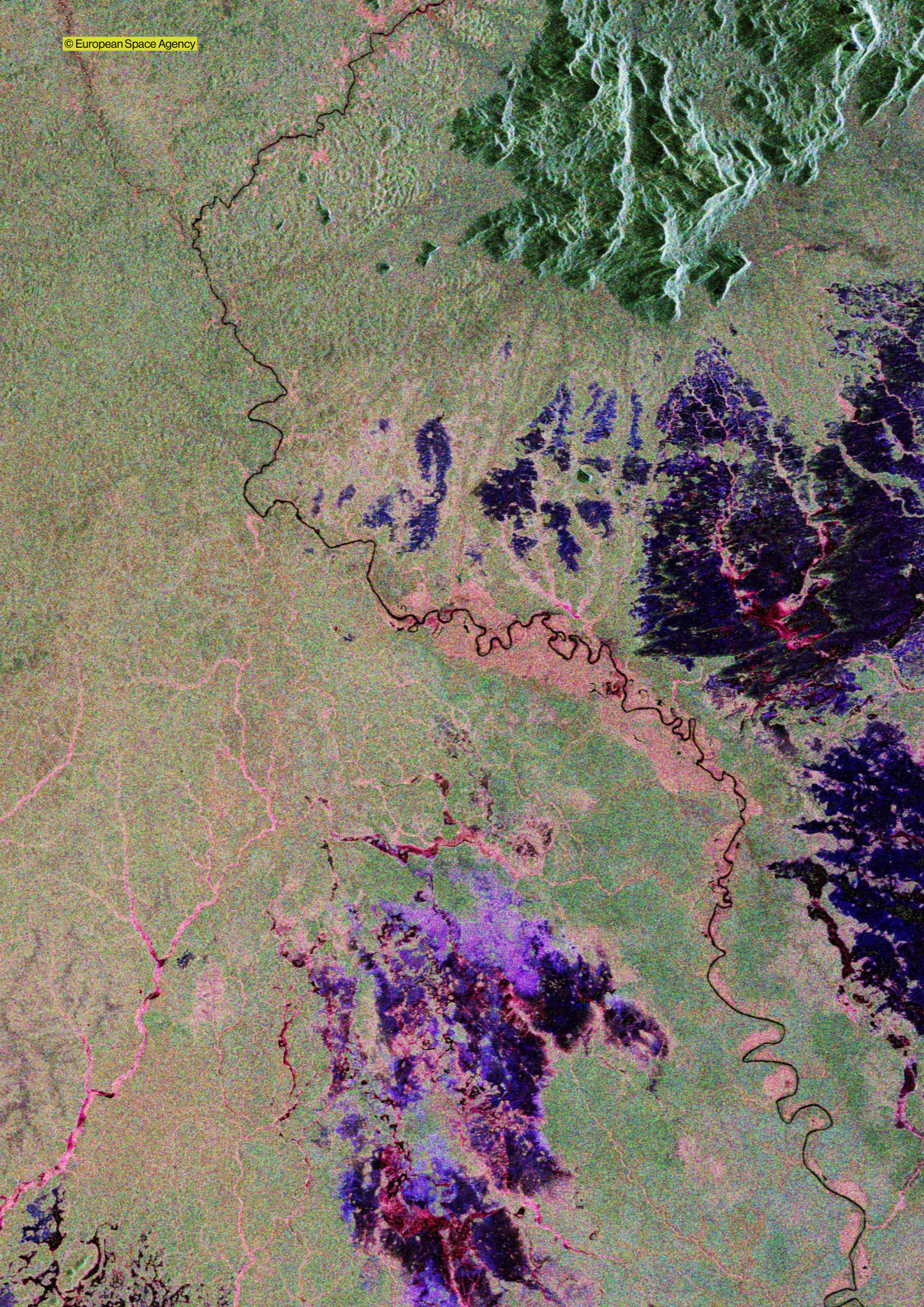
An unstratified sampling design is the simplest point of departure. The sample size required for an expected target error is straightforward to calculate (Equation 3), and the design offers cost advantages that go beyond the single consideration of sample size: the sampling units are permanent (Section 2.4), which facilitates regular updating (Section 2.4 and Section 3.5), and the same grid can serve multiple monitoring purposes (Section 2.2). Still, its main drawback is the typically large sample size that is required to capture rare change classes with sufficient precision.

Stratification, instead, might be the preferred option where the budget does not allow for such sample sizes, or where an existing map is meant to be integrated into the survey, with its accuracy assessment as a secondary objective. In that case, the efficiency of the survey rests on two elements: the quality of the map used for stratification, and the way the sampling units are distributed among the strata. The second is easily overlooked, yet it can be just as decisive for the precision that is ultimately achieved; therefore, it is a major factor in the cost of a survey.

The two designs can also be combined. If a survey is intended to serve multiple purposes and to be updated regularly, a systematic grid of permanent sampling units can act as a consistent backbone of the sampling design. In each monitoring period, the grid can be further intensified with additional temporary sampling units to better capture rare events (Section 2.4). Such a setting has an important advantage for allocation: within-stratum variances can already be calculated from previously collected samples, so that an optimal allocation of the additional units becomes feasible where a single parameter is of primary interest. If the survey is meant to deliver estimates of several parameters, such as deforestation and forest degradation, a midway allocation may be the simplest way to serve multiple target variables simultaneously, since an allocation that is optimal for one parameter is rarely optimal for another.

The midway allocation is also the recommended starting point in situations where no prior information exists on the quality of the stratification. If a single parameter is of primary interest, a practical refinement is to hold back a share of the sampling units in the first iteration. If deemed necessary, a second iteration can then be added in which the within-stratum variances are estimated from the results of the first, making an optimal allocation of the remaining units straightforward. Since the held-back units would have been interpreted in any case, this two-step procedure approaches the efficiency of an optimal design at no additional cost.

While most of these considerations are purely about efficiency, practitioners would do well to check their designs for undersampling bias and, in the case of zero detections, report the maximum change that may have gone undetected (Equation 8), or apply alternative ways for calculating the uncertainty.


© European Space Agency

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

Cochran, W.G. 1977. *Sampling techniques, Third edition*. New York, USA, John Wiley and Sons.

Hanley, J.A. and Lippman-Hand, A. 1983. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA*, 249(13): 1743–1745.

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.

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

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): 111–129. https://doi.org/10.1007/s10651-017-0378-y

Skakun, S. 2025. The impact of map accuracy on area estimation with remotely sensed data within the stratified random sampling design. *Remote Sensing of Environment*, 326: 114805. https://doi.org/10.1016/j.rse.2025.114805

Stehman, S.V. 2012. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sensing Letters*, 3(2): 111–120. https://doi.org/10.1080/01431161.2011.541950

Stehman, S.V. and Foody, G.M. 2019. Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 231: 111199. https://doi.org/10.1016/j.rse.2019.05.018

Stehman, S.V. and Xing, D. 2022. Confidence intervals for proportion of area estimated from a stratified random sample. *Remote Sensing of Environment*, 280: 113193. https://doi.org/10.1016/j.rse.2022.113193

Tyukavina, A., Stehman, S.V., Foody, G.M., Bontemps, S., See, L., Olofsson, P., Tsendbazar, N.E., Radoux, J., Komarova, A., Serre, B.M., Song, X.P., d'Andrimont, R., Koren, G., Potapov, P., Bullock, E.L., Campbell, P., de Bruin, S., Defourny, P., Friedl, M.A., Fritz, S., Hansen, M.C., Herold, M., Lamarche, C., Lesiv, M., Mané, L., Meroni, M., Nickeson, J.E., Pelletier, F., Pickens, A., Reiche, J., Schepaschenko, D., Tarrio, K., Verhegghen, A., Woodcock, C. and Xiao, X. 2025. Land cover and change map accuracy assessment and area estimation good practices protocol, Version 1.1. In: Tyukavina, A., Stehman, S.V., Foody, G., Bontemps, S., Komarova, A., Tsendbazar, N.E. and Nickeson, J., eds. *Good Practices for Satellite Derived Land Product Validation*. Land Product Validation Subgroup (WGCV/CEOS). https://doi.org/10.5067/doc/ceoswgcv/lpv/lc.001

2.4 Sampling considerations for operational land use change monitoring

by

Nandika Tsendbazar, Sytze de Bruin and Andreas Vollrath
(with adopted content from the first edition written by Inge Jonckheere, Randy Hamilton and Remi d'Annunzio)

Description of the problem

Consistent data on national land use, land use change, and associated emissions and removals are crucial for transparent reporting on climate mitigation actions. Regular updates are needed to track progress, necessitating an efficient operational monitoring system capable of precisely capturing rare occurrences of land use change. While many researchers have focused on sampling design and representation issues in this context, their assessments were retrospective analysis – estimating changes in land use that occurred in the past (Arévalo *et al*., 2020; Tang *et al*., 2021) – but without expanding much on implications for future assessments.

Operational monitoring of land use over time differs from a one-time retrospective change assessment because the sampling design needs to be flexible to account for the spatial variations of land use change occurrences in the future. Adaptation of the sampling has implications for the precision of land use change assessments from successive surveys. The rigour of statistical estimates requires proper consideration of the estimation uncertainty. However, in many official communications or reporting, the concept of estimation uncertainty is often omitted for simplicity. This omission makes it difficult to interpret apparent discrepancies in estimates over time that are, in fact, within the expected range of statistical uncertainty.

When designing a periodically repeated sample assessment, the SBAE practitioner must find ways to minimize temporal inconsistencies from sampling (avoiding changing the design at each survey to be "optimal" for that point in time), while keeping the sampling effort manageable (avoiding designs so inefficient that they require prohibitively large sample sizes). Simultaneously, optimizing both is inherently challenging, representing a fundamental trade-off between approaches. Depending on the priority given to consistency versus efficiency, two main concepts exist: permanent and temporary sampling designs.

A permanent sampling design involves establishing fixed sampling units that are repeatedly measured without changing the sample units' locations. Typically, this is the case when using systematic or simple random sample designs or fixed stratified designs (for example, based on ecozones). The use of permanent sample units is straightforward, and their reuse ensures a consistent tracking of land use in time. It also allows for optimized survey designs that can build on previous interpretations. However, while a permanent design offers temporal consistency, it can be challenging to adequately represent rare phenomena like forest area change for all monitoring periods. Its reduced precision in assessing small areas of change can only be compensated by a larger sample size (often by a factor of 3 to 5 compared to stratified sampling).

In contrast, a temporary sampling design involves selecting a new set of sampling units for each monitoring period, treating each survey as temporally independent. This approach is often coupled with a stratified sampling design informed by where change is more likely to occur. Stratified random sampling can be a powerful tool for reducing sampling effort, specifically when targeting rare phenomena like forest area change (see Section 2.1 and Section 2.6). The efficiency depends on the quality of the map used for stratification, as well as the allocation scheme, and can be calculated through the relative efficiency, as described in Section 2.3.

In a retrospective evaluation, the approach of employing a new stratification (hence, a new sample each time) for biennial evaluation was found to produce more precise estimates (lower uncertainty around each biennial estimate) of deforested areas in the Colombian Amazon, compared to employing one single stratification through 2001 and 2016. However, it also required an increased reference data collection effort (Arévalo *et al*., 2020). One example of employing biennial updates in stratification in the context of international reporting is the Biennial Update Report submission of Paraguay, which followed this type of monitoring using temporary sampling (Paraguay, 2018).

Drawing a fresh set of sample locations for every survey, however, carries a drawback regarding the interpretation of estimates between two successive surveys. Every area estimate produced from a sample is accompanied by uncertainty, expressed as the standard error or confidence interval reported alongside it. That uncertainty is the direct expression of the sampling variability (the variation in a result that arises solely from which locations happen to be selected). Even in the absence of any real change on the ground, two surveys built on different sets of locations will return somewhat different area estimates, because each captures a different subset of the landscape.

This behaviour has direct consequences for area estimation of multiple classes across successive surveys using temporary samples, which is typically the case for operational monitoring systems. Such surveys typically capture information from several classes that describe the same reality – most notably stable forest area and forest area change, which are logically linked and, within a single survey, estimated from the same sample and therefore mutually consistent. For a permanent sample, this consistency holds over time, as a deforested sampling unit adds to the deforestation estimate while, at the same time, reducing the stable forest-area estimate. The combined variability coming from the two independent samples of a temporary design will affect both classes in different ways and nothing constrains the estimates to remain mutually consistent, as each class carries its own, independent sampling variability for the new set of locations.

This divergence explains an outcome that can appear contradictory. Under temporary sampling, the estimated area of stable forest may show an increase from one period to the next, while the net forest-change estimate for the same interval points to a loss. The two figures are simply different, as they are independently sampled views of the same underlying reality, each surrounded by its own uncertainty. As the divergence falls well within that uncertainty, it is statistically plausible. For a country reporting officially, however, presenting two headline figures that at first sight disagree is difficult to communicate and defend, which is itself a reason to weigh a design's temporal consistency alongside its precision within any single period.

An additional practical consideration that impacts the overall workload is the effort required to interpret new samples, compared to revisiting existing samples. Potentially, to limit cost, sample units from no-change stratum of a previous reporting period can be migrated to the current reporting period, provided that the reference label is updated (see Section 3.5). When employing a temporary sampling design, it is crucial to assess the history of each sample unit in the years preceding the monitoring period. Although it comes with increased workload, this historical analysis increases the likelihood of correctly determining whether a land use change has occurred. Section 3.5 delves more into the considerations around the response design of repeated assessments.

One might consider another strategy, where each new assessment would use an updated stratification that covers both the current monitoring and the previously covered historical period (always starting from the same year in the past). The implication of such a strategy is that already-reported estimates for the historical period would be constantly updated and changed. Under the UNFCCC technical assessment process, this strategy would be problematic, as the assessments need to be methodologically consistent, and no changes can be made to the historical data used in the reference level. As such, countries face a challenge when seeking to adopt new and improved methodologies while maintaining consistent estimates over time.

Overview of existing and potential practices

Despite the importance of having robust strategies for continuous and updated monitoring, there is still limited guidance and research on cost-effective and scientifically sound approaches that countries can use. Existing practices are mainly based on a sampling design that is purely permanent or purely temporary. To a lesser extent, more advanced strategies are used that seek to alleviate the drawbacks of both approaches, while benefiting from their advantages.

Dealing with simple random and systematic sampling for repeated monitoring

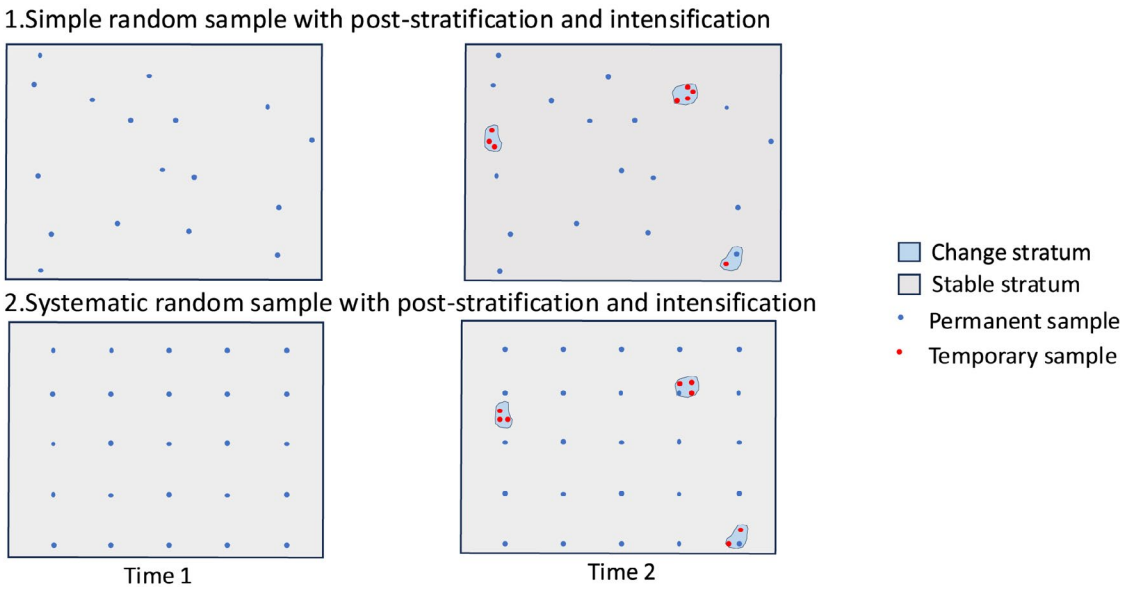
Post-stratification

A relatively simple and straightforward approach to reduce the sampling uncertainty of an unstratified sample without increasing the sample size is the employment of **post-stratification**. Here, a permanent simple random or systematic sample is regrouped using a map, and a post-stratified estimator is applied for area estimation. Accordingly, the weights for each sample unit in the successive assessments vary. However, it should be noted that post-stratification does not solve issues for capturing rare phenomena, such as forest change, if the initial sample is not large enough (it does not correct for an undersampling bias). A sufficiently dense sample is needed to capture the rare occurrences of forest change. Compared to using a permanent sample without stratification, post-stratification adds the temporally changing strata for increased precision. An example of using post-stratification in international reporting is the Cambodian biennial update report (Cambodia, 2020), where an existing dense SYS (6 km grids) was post-stratified into forest change and no change. If the precision requirement is met, this design is simple to implement and maintain for future updates. There is also the advantage that post-stratification can be done for various categories, making it more suitable for multipurpose assessments (see Section 2.2).

Intensification of a permanent sample

Another strategy is the augmentation of a permanent sample. Suppose that after analysing the original (unstratified) sample, the estimation uncertainty around key target classes is too large so that an increased sample size is needed. This can be achieved in at least two ways. One is to intensify the SYS or SRS design over the full area of interest. This would only affect the sample size, but not the sampling design, and give the possibility to reuse the added sample units in future assessments, ensuring temporal consistency. However, this approach comes with high implementation costs as a very large sample size may be needed to precisely capture the rare classes (Section 2.3).

Figure 7 Illustrations of intensification of simple random and systematic sampling designs for repeated monitoring



Source: Authors' own elaboration.

A second possibility is to use a change map for stratification and employ sample intensification over one or more strata of interest (Figure 7). The advantage is that the intensification can be targeted at the specific classes/strata of interest, which reduces the effort in sample collection as compared to the former strategy. It is important to note that in this case, a stratified sampling strategy is used, where the sampling intensities vary among the strata. The added sampling units are temporary, although only for the target change strata, which adds a source of variability leading to potential inconsistencies between subsequent assessments, as previously explained. This way of temporary sampling intensification using a permanent sample as the basis aims to alleviate the main issues with temporary and permanent sampling designs by increasing consistency over time (using permanent sampling units), while keeping the sample size manageable for a desired precision level (through temporary sampling units).

One example where this temporary sample intensification approach was implemented was to assess the accuracy of operational land cover maps for national monitoring (Pengra *et al.*, 2020). Starting with a permanent simple random sample, intensification was applied to estimate accuracy and area, targeting specific classes of change (Stehman *et al.*, 2021).

If the sampling design to be intensified is systematic, there are three slightly different strategies to place new sample units:

1. dividing existing grid spacing within the mapped change stratum (for example, by two and selecting new sample units in the thus far unsampled grids);

2. applying a separate, denser systematic grid in the change stratum to be intensified; and

3. intensifying the sample by additional random sampling within one or more strata.

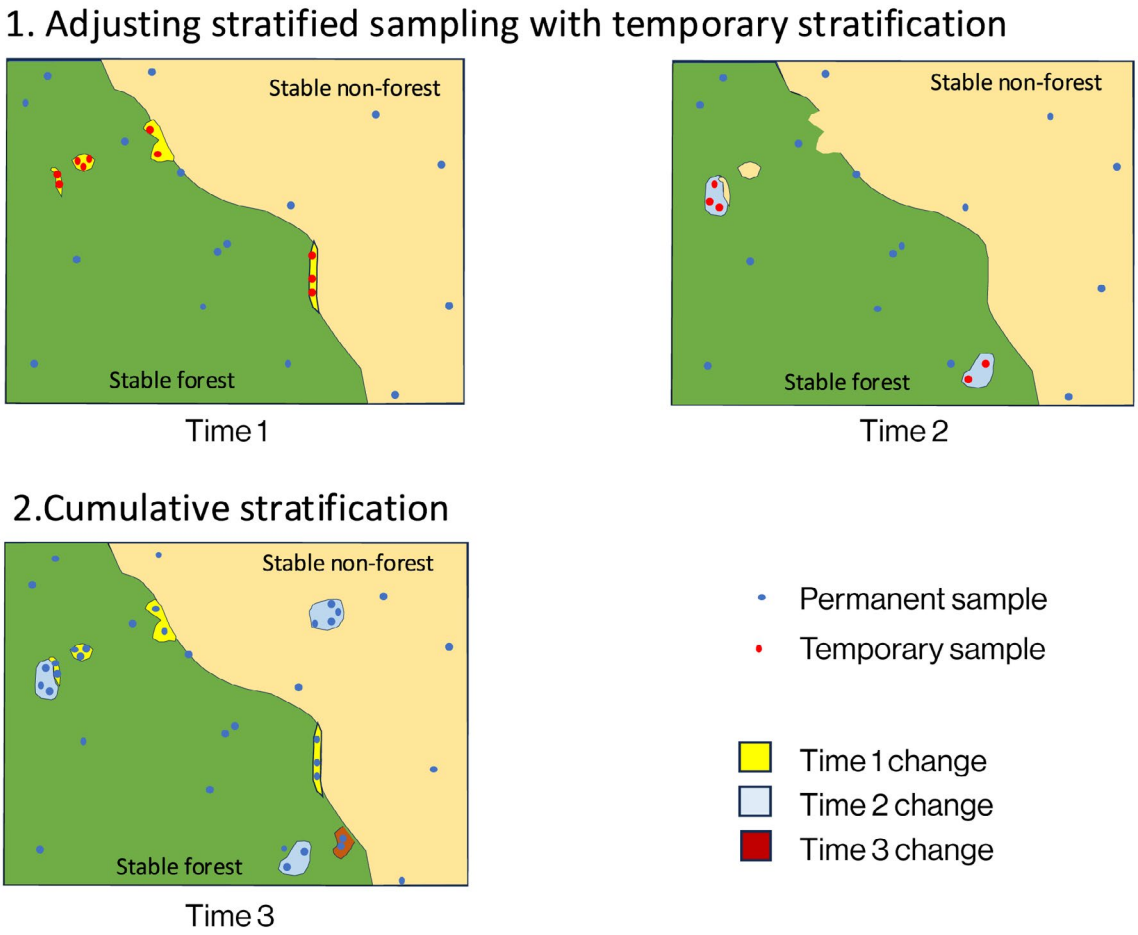
The first case makes use of the previously selected sample in the existing grid, post-stratifies the sample, and adds new sample units within the change stratum. The sample size in the mapped change stratum is increased by a factor of 4, 8, 16, etc., according to the grid subdivision. For the second case, a denser grid is applied to the change stratum with increased sampling intensity; for the third case, a number of additional sample units is randomly selected and added as new sets of sample units for one or more strata. Approaches used in Eva *et al.* (2010) can serve as an example for the first case. They intensified a continental-level 1-degree systematic design by 0.25 degree systematic design for deforestation area estimation in French Guiana.

Dealing with stratified random sampling designs for repeated monitoring

A stratified random sample with partially permanent sample units can be a useful strategy for SBAE practitioners whose starting point is an STR design at Time 1. This approach aims to keep the sampling effort manageable while increasing consistency over time, compared to an entirely independent survey for each assessment. Here, two options are available: partially updating an existing design with a temporary stratification; or cumulating change stratification over time (Figure 8).

An additional feature of updating a stratification – either through added temporary or permanent strata – is that the required sample sizes for the stable strata can be calculated based on the previous assessments. The population parameters of the target class have been previously estimated for the permanent strata, which allows targeting a specific standard error by using Neyman's formula (Section 2.3). This contributes to sample size optimization and hence, improved cost efficiency of the survey over time.

Figure 8 Illustrations of different options for adjusting stratified sampling design for repeated monitoring



Source: Authors' own elaboration.

Partial update of an existing stratification through temporary strata

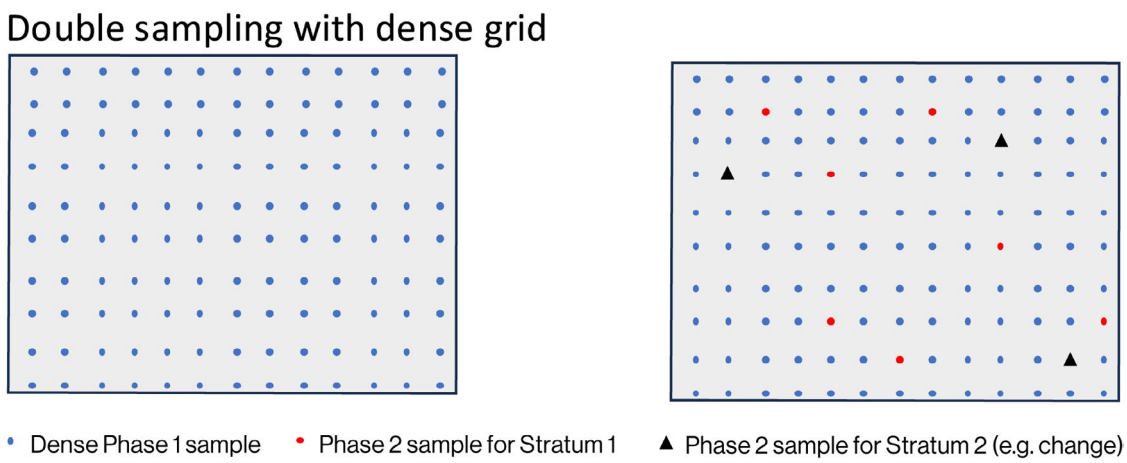
Stratified random sampling allows the possibility to adjust the strata – for example, by adding a new stratum where changes are expected to occur, according to a map (Stehman *et al*., 2012). For operational validation of annual land cover maps, Tsendbazar *et al*. (2021) implemented an STR approach with temporary strata. For each update period, new land cover change strata are added to the original strata (initially used to select the permanent sample). New sample units are then added to those in the change strata to ensure sufficient sample representation for the rare class (such as land cover change). Since the new strata also reduce the area of the original strata, the inclusion probabilities for the sample are recalculated for all strata, including the original and the temporary stratifications. A stratified estimator is then used with the recomputed inclusion probabilities. For the next monitoring period, these steps are repeated, but without utilizing the temporary strata and the temporary sample units of the previous survey. This approach has the advantage of keeping the stratification relatively simple for repeated monitoring while targeting precise estimation. However, this comes at the cost of not reusing the temporary sample units and increased variability in the change area estimates over time, due to the spatial misalignment of multiple temporary samples, although this is expected to be limited, as it combines a permanent sample with a temporary sample.

Partial update of an existing stratification through cumulative stratification

Instead of adding a temporary stratification to the original stratification for a single reporting period, these strata become part of the permanent stratification. This means newly added sample units for the first monitoring update period are kept and used for the second monitoring update with additional strata targeting change for the second monitoring period. Suriname, in 2018, implemented this approach. The use of cumulative stratification can be efficient in change area estimation for all the monitoring updates. However, it leads to stratification becoming highly complex over time. In addition, this requires an increasingly larger sample size to maintain, consequently, increasing costs. A potential way to alleviate this is to simplify the stratification after a cycle of 5–10 years and create new base strata, although with the drawback of affecting historical reports.

Dealing with double sampling for repeated monitoring

Figure 9 Illustrations of double sampling involving initial dense sampling grid and subsequent stratified sampling



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.

Double sampling is a common approach where the cost of stratifying the entire area is significant. For example, it may be too expensive to produce a national-scale forest change map for the sole purpose of supporting stratified sampling. In double sampling, a large first-phase sample is taken and the sampling units are labelled using a low-cost automated procedure – for example, employing a combination of dense time-series forest change algorithms only on the first-phase sample instead of wall-to-wall mapping. It should be noted that the sample should still be multiple thousands or more to accurately estimate the stratum weights (see Section 2.6). A second sample is next drawn from the first sample following an STR approach. The second sample is labelled using a more costly yet more accurate procedure, such as visual interpretation. This design can be adapted for repeated monitoring. For example, a time-series change-detection analysis can be conducted on first-phase sample locations (such as dense grids) to stratify this first sample. A second sample is then subsampled within each stratum for actual field data collection or visual interpretation.

For each new reporting period, new (temporary) second sample units can be subsampled from the first sample (such as dense grids). As with the previous stratified sampling options, a decision can be made on whether to keep the second sample temporary or permanent (such as using cumulative stratification).

For ease of comparison, the different approaches discussed in this section are summarized in Table 5.

Table 5 Overview of potential sampling approaches for monitoring updates of reducing emissions from deforestation and forest degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks (REDD+)

Sampling approach	Complexity	Statistical efficiency	Consistency (temporal tracking)	Cost and operational effort	Key considerations and mitigations	Examples
Temporary sampling only						
New stratification each update period	Straightforward implementation; no legacy maintenance.	Efficient for the current period, but increased variability in change estimates of land use.	Previous samples discarded; need to track specific land use trajectories over time.	New interpretation required for every period; cannot reuse previous labels.	Reuse old samples in "no-change" areas if labels are updated; interpret land use history for all units.	Forest change monitoring, retrospective analysis;[i] REDD+ monitoring.[ii]
Permanent and/or temporary sampling						
SRS/SYS + post-stratification and intensification	Straightforward to implement; stratified estimators needed to account for unequal inclusion probability.	Post-stratification improves precision; intensification specifically targets change areas.	Base permanent sample ensures consistency; intensified (temporary) units require long-term temporal tracking. For SYS, the systematic grid needs adjustment for intensification.	Intensified samples are temporary (not reused); requires extra effort for temporary units.	For temporary units, interpret also the land use history.	SRS: operational land cover validation.[iii, iv] SYS: forest change monitoring.[v, vi]
Adjusting stratification with temporary strata	Stratification adjustment is simple, but inclusion probabilities must be recalculated.	Efficient for precise change estimation within the specific reporting period.	Base sample ensures consistency, but added temporary units require long-term tracking.	Temporary sample units are not reused; extra effort required for the temporary units.	For temporary units, interpret also the land use history.	Operational land cover validation.[vii]
Cumulative stratification	Stratification becomes increasingly complex as new change strata accumulate.	Efficient for precise change estimation across all reporting periods.	The permanent sample is expanded with each reporting period. It allows long-term tracking if land use history is added for each new location.	Larger sample size required to keep all strata up-to-date; high interpretation costs.	Simplify stratification after a 5–10-year cycle to reduce complexity and costs.	REDD+ monitoring.[viii]
Double sampling for stratification for dense SRS or SYS	Moderate to high: requires large initial sample to reduce strata weight uncertainty. Map for stratification not necessary; point-based assessment.	Moderate: a very dense sample is needed to ensure precise estimation for the change class.	Moderate: very dense permanent sample offers a unified, consistent system. The second sample (subsample) is varied per period, so it requires temporal tracking of sample units in the change stratum.	Variable: gain in effort for point-based analysis should outweigh the cost of creating a stratification map.	The first sample is used to calculate stratum weights, so this option is beneficial if the second sample varies per reporting period (temporary sample).	National forest inventory.[ix]

Source: Authors' own elaboration.

[i] 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

[ii] Paraguay. 2018. *Biennial Update Report: REDD+ Technical Annex to the UNFCCC*. Resultados alcanzados por la República del Paraguay en la Reducción de Emisiones por Deforestación, para pago por los resultados de REDD+ [Results achieved by the Republic of Paraguay in the reduction of emissions from deforestation for REDD+ results-based payment]. Asunción.

[iii] 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

[iv] Stehman, S.V., Pengra, B.W., Horton, J.A. and Wellington, D.F. 2021. Validation of the U.S. Geological Survey's Land Change Monitoring, Assessment and Projection (LCMAP) Collection 1.0 annual land cover products 1985–2017. *Remote Sensing of Environment*, 265: 112646. https://doi.org/10.1016/j.rse.2021.112646

[v] Cambodia. 2020. *Biennial Update Report: REDD+ Technical Annex to the UNFCCC*. Phnom Penh.

[vi] Eva, H., Carboni, S., Achard, F., Stach, N., Durieux, L., Faure, J.-F. and Mollicone, D. 2010. Monitoring forest areas from continental to territorial levels using a sample of medium spatial resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(2): 191–197. https://doi.org/10.1016/j.isprsjprs.2009.10.008

[vii] Tsendbazar, N., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R. and Duerauer, M. 2021. Towards operational validation of annual global land cover maps. *Remote Sensing of Environment*, 266: 112686. https://doi.org/10.1016/j.rse.2021.112686

[viii] Government of Suriname. 2018. *Forest Reference Emission Level for Suriname's REDD+ Programme*.

[ix] 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.

New approaches and knowledge gaps

The main challenge of operational monitoring is to implement a forest monitoring system that is feasible without compromising statistical rigor and reducing the uncertainty of the forest change area estimations. Despite the direct need for multiple, frequent assessments (such as REDD+ annual updates), continuous sample-based monitoring of forest change lacks substantial examples from which lessons can be learned. Due to the limited examples and literature, there is currently no systematic assessment of best practices.

Based on the potential of different sampling schemes in the previous section, here we identify key aspects to consider when implementing operational forest change monitoring.

- Sampling scheme *flexibility*: Flexibility of sampling design is vital to facilitate adjustments in the sampling scheme to improve sampling representativeness in newly changed areas. Known for its flexibility to target specific areas of interest, stratified sampling (Stehman *et al*., 2012) is an obvious choice. However, post-stratifying SRS and SYS in combination with sample intensification within new change areas has also shown to be valuable for updated monitoring (Eva *et al*., 2010; Stehman *et al*., 2021). These approaches are also relatively straightforward to implement and hence attractive when setting up an operational forest change monitoring system. They also show advantages with regard to multipurpose monitoring objectives.

- Balancing sampling *precision* and the complexity of the sampling design: Intensifying the existing sample for newly changed areas ensures precise estimation, but can increase sample size and complexity in sample design over time, leading to higher operational costs. The challenge is to keep sampling design simple yet effective. Leveraging remote sensing technology is crucial, as it allows for more focused sampling efforts by accurately identifying change areas. This reduces the need for unnecessary sample size expansion while maintaining statistical rigor in forest change estimations.

- *Reusing previous sample units*: In scenarios where temporary strata are used to target specific change areas, the challenge is deciding whether to retain or discard sample units from previous periods. While the use of temporary sample units enhances precision, their reuse for long-term monitoring leads to increased complexity (for example, for [post-] cumulative stratification). On the other hand, discarding them, and using new sampling locations for each monitoring period, adds a source of variability that may interfere with the change estimates. Developing methodologies to effectively integrate past and new sample units is an important knowledge gap that requires further research.

Ensuring precision while keeping operational burdens manageable is a key challenge that should be addressed by current and future efforts for operational forest change assessments. This calls for comparative assessments of precision efficiency in different operational monitoring contexts.

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

Cambodia. 2020. *Biennial Update Report: REDD+ Technical Annex to the UNFCCC*. Phnom Penh.

Eva, H., Carboni, S., Achard, F., Stach, N., Durieux, L., Faure, J.-F. and Mollicone, D. 2010. Monitoring forest areas from continental to territorial levels using a sample of medium spatial resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(2): 191–197. https://doi.org/10.1016/j.isprsjprs.2009.10.008

Government of Suriname. 2018. *Forest Reference Emission Level for Suriname's REDD+ Programme*.

Paraguay. 2018. *Biennial Update Report: REDD+ Technical Annex to the UNFCCC*. Resultados alcanzados por la República del Paraguay en la Reducción de Emisiones por Deforestación, para pago por los resultados de REDD+ [Results achieved by the Republic of Paraguay in the reduction of emissions from deforestation for REDD+ results-based payment]. Asunción.

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

Stehman, S.V., Olofsson, P., Woodcock, C.E., Herold, M. and Friedl, M.A. 2012. A global land-cover validation dataset, II: augmenting a stratified sampling design to estimate accuracy by region and land-cover class. *International Journal of Remote Sensing*, 33(22): 6975–6993. https://doi.org/10.1080/01431161.2012.695092

Stehman, S.V., Pengra, B.W., Horton, J.A. and Wellington, D.F. 2021. Validation of the U.S. Geological Survey's Land Change Monitoring, Assessment and Projection (LCMAP) Collection 1.0 annual land cover products 1985–2017. *Remote Sensing of Environment*, 265: 112646. https://doi.org/10.1016/j.rse.2021.112646

Tang, X., Woodcock, C.E., Olofsson, P. and Hutyra, L.R. 2021. Spatiotemporal assessment of land use/land cover change and associated carbon emissions and uptake in the Mekong River Basin. *Remote Sensing of Environment*, 256: 112336. https://doi.org/10.1016/j.rse.2021.112336

Tsendbazar, N., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R. and Duerauer, M. 2021. Towards operational validation of annual global land cover maps. *Remote Sensing of Environment*, 266: 112686. https://doi.org/10.1016/j.rse.2021.112686

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 SYS (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 9)

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:

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

(Equation 10)

$\hat{P}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$ 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 11)

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:

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

(Equation 12)

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

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 13)

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 Equation 11 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 previously 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 6 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 6 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 9)
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 11)
Variance estimator under **SRS**	$v(\hat{P}_R)$	$v(\hat{P}_R) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\left(\frac{1}{m}\sum_{j=1}^{m} y_{ij}\right)_i - \hat{P}_R\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{P}_i - \hat{P}_R)^2$		(Equation 10)
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 12)
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 13)

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.


© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)/Pablo Cambronero

2.6 Improving sampling efficiency with continuous auxiliary variables

by

Nandika Tsendbazar, Sytze de Bruin and Andreas Vollrath

Description of the problem

Obtaining acceptable precision when estimating the area of forest change through visual interpretation of samples is difficult because the changes of interest are rare. Deforestation and forest degradation typically affect only a small fraction of the sampling frame in any given reporting period. Under SRS or SYS, resolving such a rare class to a useful level of uncertainty requires a very large number of reference observations, and the cost of collecting them can quickly become prohibitive.

Stratification is the principal means of controlling this cost by providing a more efficient sampling design (see Section 2.2). In practice, STR for area estimation is most often carried out alongside an existing map, frequently one produced by a national mapping programme. In its absence, practitioners can be constrained from two directions at once. Producing a categorical map to an assured minimum quality – and repeating it for every reporting cycle – can be prohibitively demanding. On the other hand, relying on a large systematic sample can be equally impractical, for the reasons described previously.

Where area estimation is the primary objective and stratification has been identified as the means of reducing the sample size, an auxiliary variable associated with forest disturbance or change in continuous form becomes a viable third option. Beyond discrete spatial buffers (Section 2.1), this strategy equally has the potential to mitigate effects of omission errors, as continuous values of forest change probability can effectively serve as a proxy for "potentially uncertain areas", capturing nuances within the underlying data that discrete maps might miss.

This section sets out how such variables can be put to use. It begins with the ways in which forest change can be represented in continuous form – for example, as a modelled probability or a magnitude of change rather than a discrete class. It proceeds with the two forms that this auxiliary information can take. One is the well-known wall-to-wall map, while the other depicts a double-sampling approach, where a sample of the variable on a dense grid is conducted (when full mapping is not warranted). Attention then turns to how a continuous variable is used to define strata – and in particular, to the choice of stratum boundaries for the strongly skewed distributions that change probabilities typically exhibit. The section closes with estimators that exploit the same auxiliary information in a different way – most notably, model-assisted estimation, which calibrates the auxiliary layer against a reference sample rather than using it only to stratify.

Overview of existing good practices

Representation of continuous variables for forest change from remote sensing

Several approaches to representing the probability or magnitude of change have been documented in the literature and are presented in the following text. From a practical perspective, producing a continuous layer can be easier than compiling an accurate categorical map. The advantage of a probability-based layer is precisely that it can flag areas of uncertainty, which can then be targeted through probability sampling to improve the precision of the area estimate. The idea is similar to the spatial buffer in Section 2.1, but operates on the statistical distribution (histogram) of the probability layer. A minimum level of quality nonetheless remains essential, as only a sufficiently accurate layer will yield genuine gains in the precision of the final forest-change estimates.

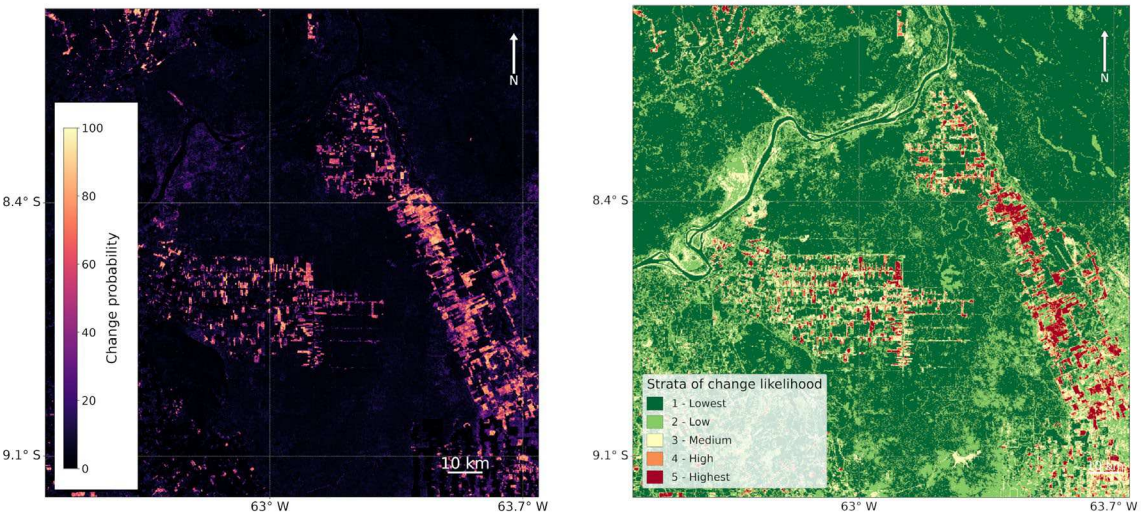
Full resolution approaches

Multi-temporal change analysis of maps that exhibit a continuous representation of tree cover can also represent the magnitude of changes in tree cover or forest as a continuous variable at full resolution. Examples are the global percent tree cover map at 30 m resolution by Hansen *et al*. (2013) and European tree cover density – a Copernicus high resolution layer at 10 m resolution, tropical tree cover percent layer at 0.5 ha level (Brandt *et al*., 2023), and global tree cover fraction at 100 m resolution by Copernicus Global Land Service (CGLS) (Buchhorn *et al*., 2020).

Research has also focused on directly mapping forest change in a continuous format such as the forest disturbance index by Healey *et al*. (2005) and forest change probability based on change vector analysis of three indices by Francini *et al*. (2021). Using a hybrid approach of training, a machine learning algorithm based on input features produced by time-series algorithms (Xu *et al*., 2022) is another way to map potential forest change probabilities.

Another option is to employ supervised classification and directly use the classifier's probability output, instead of its discrete classes. This can be done in two ways: by either directly classifying change (Probability Change Detection [PROCD]), or by subtracting the classifier's probability from two maps depicting tree or forest cover at the start and end of the monitoring period (Probability Map Subtraction [PROMS]). The latter has the advantage of not relying on change training data and was employed for Zimbabwe's Forest Reference Level submission to the UNFCCC (Zimbabwe, 2024). Based on annual layers of tree cover probability, the method was shown to be less prone to seasonal effects by the prevalent dry forests, which usually get mixed up with real change in dense time-series analysis.

Figure 10 Example of a continuous layer of change probability based on the Probability Map Subtraction approach (left) and a subsequent K-means stratification (right) for an area around Porto Velho, Brazil



Notes:
AOI – area of interest. The "Increasing model dependence" arrow bar on the left refers to the position of the conditions represented by diamonds and denotes increasing risk of bias. Refer to the disclaimer on p. ii for the names and boundaries used in this map.
Sources: Authors' own elaboration using SEPAL (System for Earth Observation Data Access, Processing and Analysis for Land Monitoring). 2026. *SEPAL*. https://sepal.io

Aggregation to a coarser resolution

Both continuous and discrete layers representing forest or forest change can be aggregated to a coarser resolution to create continuous wall-to-wall layers and used to estimate areas. Pickering *et al*. (2019) used the Global Forest Cover (GFC) 30 m resolution map to create a wall-to-wall map of forest loss and degradation areas over 24 km × 24 km blocks and used it as a basis for stratified sampling. Similarly, Francini *et al*. (2021) and Sannier *et al*. (2014) aggregated forest change maps at 30 m pixel spacing to 10 km × 10 km and 2 km × 2 km blocks, respectively, and estimated forest change and harvested area using a model-assisted estimator. As the size of the population unit (the grid raster or pixel size) changes in this case, the response design needs to be adapted accordingly.

Double sampling

A continuous auxiliary variable (x) can take two forms, and the choice affects which estimator can be used. Where wall-to-wall mapping of the sampling frame is feasible, the variable takes the form of an exhaustive map. Every population unit then carries a value of the auxiliary variable (x). If mapping the whole frame is too expensive, or cannot be produced at an acceptable accuracy, x is instead determined on a dense but non-exhaustive sample – that is, a grid of first-phase units. The map form supports the map-based estimators further described in the following text as well as in Section 2.5. The dense-sample form leads to the double-sampling approach.

A practical example of the dense-sample form is the ensemble sample-based area estimation (eSBAE) approach developed by FAO, based on the idea of Healey *et al*. (2018). It combines the outputs of several change-detection algorithms, such as BFAST (Verbesselt *et al.,* 2010), Cumulative Sum (CuSUM) and Continuous Change Detection and Classification (CCDC) (Zhu and Woodcock, 2014), with the spectral time-series data and a set of relevant global data layers, to map the probability of change. Constructing a wall-to-wall map with this approach is computationally prohibitive. Processing is therefore limited to a very dense systematic grid, which effectively serves as a first-phase sample. As in a map-based stratified design, the continuous layer is then subdivided into strata, and a probability sample is drawn from them for visual interpretation. If the first-phase sample is large enough (several tens of thousands of sampling units), the standard STR estimator can be used to obtain the final area estimates. This double-sampling approach was applied in Côte d'Ivoire's Forest Reference Level submission (see Section 2.4) to the UNFCCC in 2024 (Côte d'Ivoire, 2024).

Stratification methods for maps depicting a continuous variable

The purpose of employing stratification is to improve the precision of estimates by dividing the population into parts that minimize variability within each stratum while maximizing variability between strata. Therefore, for stratification based on continuous variables, the strata boundaries need to be identified to satisfy this requirement. One particular advantage of using continuous variables over discrete ones is that the within-stratum variability can be approximated from the auxiliary variable itself, without the need for a pre-assessment. This information can be used both in defining the strata boundaries and in the subsequent sample allocation (see Section 2.3).

Popular stratification methods that work well on large datasets are K-means clustering the geometric stratification (Kozak, 2011) and the cumulative root frequency methods (Cumrootf) (Dalenius and Hodges, 1957). The Cumrootf method aims to reduce the estimated variance using the cumulative root frequency distribution of an auxiliary variable. This method has been used to create strata boundaries to estimate the area of forest loss and degradation in Guyana and Brazil (Broich *et al*., 2009; Pickering *et al*., 2019). Suitable for non-symmetric distributions, the geometric method equalizes the coefficient of variation of the auxiliary variable within each stratum based on a geometric progression (Hidiroglou and Kozak, 2018). However, it does not work well for variables with a minimum value close to zero, which is the case when relying on mapped deforestation probabilities. K-means clustering, in comparison, performed better, and has also been used when the aim is to minimize the variance estimator with a fixed total sample size; it has been shown to provide good results for skewed populations (Kozak, 2011). A main advantage of using the K-means clustering method is its suitability for geospatial data (Ballin and Barcaroli, 2020), as it is implemented in various geographic information systems (GIS) and remote-sensing software packages.

In this chapter, the K-means, as well as the Dalenius and Hodges methods for stratification, are demonstrated (see Box 1). It is often recommended to have two to six strata when continuous data is used for stratification as it has been found that there is limited gain in terms of reduction in variance with more than six strata (Cochran, 1977, pp. 132–134).

Optimizing the strata boundaries through random or local search has also been proposed (Brito *et al*., 2010; Kozak, 2004). These optimization-based approaches were developed mainly in other fields and can find the best strata for a given criterion and set of constraints. However, they are computationally intensive and typically impractical for geographic data, where continuous maps cover very large areas at fine spatial resolution. Such methods are therefore considered beyond the scope of this section.

The sample size calculation, as well as the sample allocation, are the natural subsequent steps in stratification and are described in detail (see Section 2.3). For Neyman allocation, estimates of within-stratum variances are required. Because these variances are typically unknown prior to sampling, a common practice is to use variance estimates derived from the mapped auxiliary variable that is assumed to be strongly correlated with the target variable.

In the estimation process, the well-known and already documented estimators for the population mean and the variances are applied for an STR design, as in the case with categorical maps. For completeness, those are repeated here. The estimator of the population mean for STR is as follows:

$$\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h ,$$

(Equation 14)

where $W_h = \frac{N_h}{N}$ or equivalently $W_h = \frac{A_h}{A}$ and $\bar{y}_h$ is the estimated mean of y whithin stratum h,

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

(Equation 15)

The estimator of the population total for STR is:

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h$$

(Equation 16)

The estimator of the variance of the population mean for STR is (Equation 3.5 in Lohr [2019]):

$$\widehat{Var}(\bar{y}_{str}) = \frac{1}{N^2} \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$$

(Equation 17)

where the term $\left(1 - \frac{n_h}{N_h}\right)$ is a correction term for the finite population within a stratum and s_h^2 is the estimated within stratum population variance. In other words:

$$s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

(Equation 18)

The estimator of the variance of the population total for STR is (Equation 3.4 in Lohr [2019]):

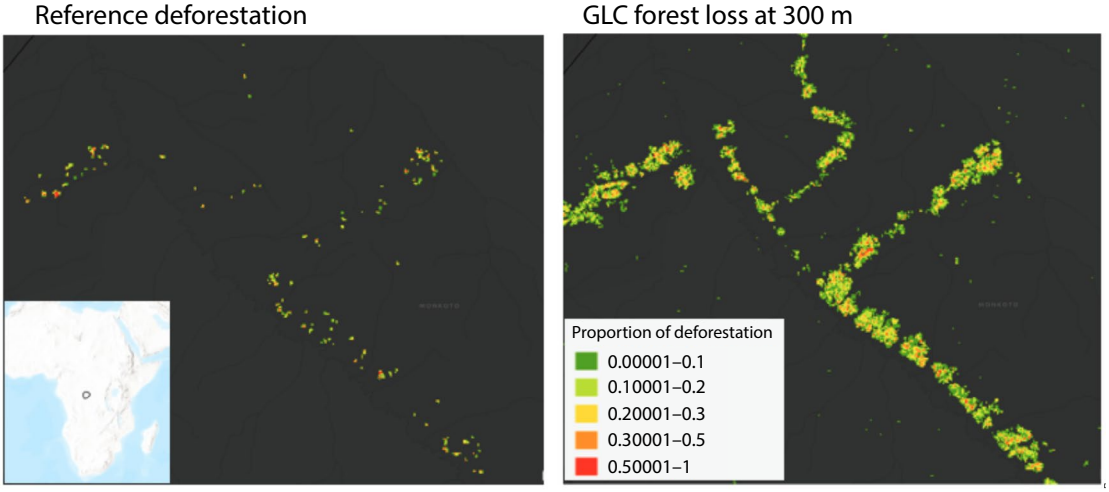
$$\widehat{Var}(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$$

(Equation 19)

Demonstration of forest-loss area estimation using stratification over continuous layers

Using continuous change layers, different methods for stratification boundary and sample allocation are demonstrated in this subsection as an example of SBAE of deforestation using stratified sampling.

Figure 11 Continuous layers for deforestation in Salonga, central Congolese lowland forests

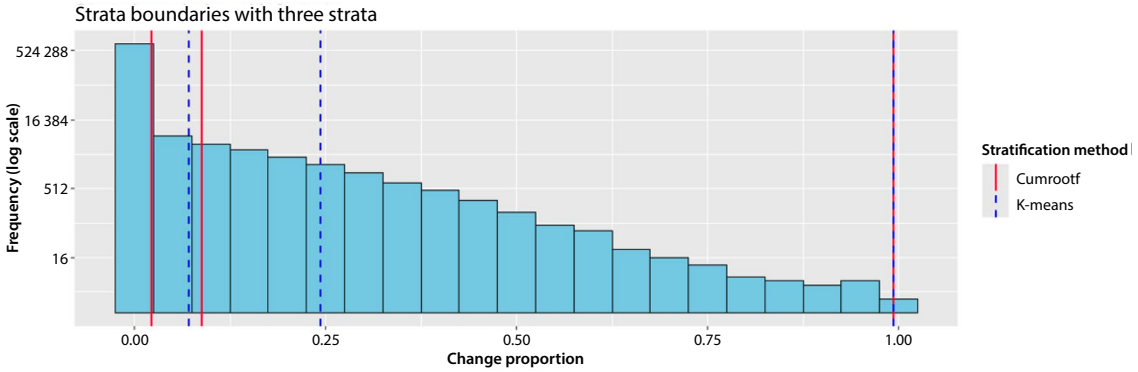


Notes: GLC – Global Land Cover. Left: Visually inspected reference deforested-area proportion at 300 m resolution, adapted from Copernicus HotSpot Monitoring. Right: Global Forest Change (GFC) forest loss (2016–2019), aggregated to 300 m (9 ha) pixel resolution. Refer to the disclaimer on p. ii for the names and boundaries used in these maps.

Sources: Reference deforested-area proportion adapted from Szantoi, Z., Brink, A., Lupi, A., Mammone, C. and Jaffrain, G. 2021. An update and beyond: key landscapes for conservation land cover and change monitoring, thematic and validation datasets for the African, Caribbean and Pacific regions. *Earth System Science Data*, 13: 3767–3789. https://doi.org/10.5194/essd-13-3767-2021 Forest loss adapted from Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D. *et al*. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160): 850–853. https://doi.org/10.1126/science.1244693
Authors' own elaboration using Esri (Environmental Systems Research Institute). 2026. *ArcGIS Pro*. https://www.esri.com/it-it/arcgis/products/arcgis-pro/overview

Based on the GFC product, a continuous forest change map was created at 300 m resolution by totalling up the forest loss pixels for 2016–2019 (Figure 11). Simple random and stratified sampling was simulated to estimate forest loss using a design-based approach. Sample sites were selected based on this stratification layer. To define stratification boundaries, K-means – as well as Dalenius and Hodge's Cumrootf methods – were used with strata numbers varying from 3 to 6.

Figure 12 Strata boundary determination methods (with three strata as an example)



Sources: Authors' own elaboration.

With a fixed total sample size of 5 000, sample sites were then allocated to the three strata using the area proportional allocation, Neyman allocation, equal allocation, and allocation based on the midway allocation, as described in Section 2.3 (see Table 7).

Table 7 presents the precision of the estimated forest-change area (SE), calculated as the square root of the value derived from Equation 17 and multiplied by the area of a sample unit (9 ha). It should be noted that, in this example, the strata differ substantially in size. Consequently, proportional sample allocation assigns relatively few sample units to the smaller strata, resulting in only marginal improvements in estimator efficiency. In contrast, equal allocation among strata markedly improved precision. Neyman allocation produced the greatest gain in efficiency for both stratification approaches. However, achieving this advantage requires reliable prior information on within-stratum variance, which is not always available in practice. When such information is lacking, a compromise "midway" allocation based on the average of equal and proportional allocation may be considered. Depending on the stratification scheme, this approach can provide satisfactory results. As shown in Table 7, for the K-means stratification, the precision obtained with the midway allocation was nearly identical to that achieved with Neyman allocation.

Table 7 Different stratification methods and sample allocation for three strata and a sample size of 5 000 units

Stratification method	**Cumrootf stratification**			**K-means stratification**		
Stratum area proportion (%)	96.9	1.87	1.22	97.9	1.59	0.53
Neyman sample allocation	2 373 (47.5%)	2 081 (41.6%)	545 (10.9%)	3 456 (69.1%)	938 (18.8%)	605 (12.1%)
Proportional sample allocation	4 845 (96.9%)	94 (1.88%)	61 (1.22%)	4 894 (97.9%)	79 (1.58%)	27 (0.54%)
Equal sample allocation	1 667 (33.3%)	1 667 (33.3%)	1 667 (33.3%)	1 667 (33.3%)	1 667 (33.3%)	1 667 (33.3%)
Midway sample allocation	3 256 (65.1%)	880 (17.6%)	864 (17.3%)	3 280 (65.6%)	873 (17.5%)	847 (16.9%)

Note: Compromise sample allocation uses the averages of proportional and equal allocation.
Source: Authors' own elaboration.

Table 8 Precision of estimated forest change area, for various combinations of stratification and allocation of a sample of 5 000 units

Stratification	Allocation	SE (ha)	RE
Cumrootf	Neyman	549	12.4
	proportional	1 833	1.1
	equal	614	9.9
	midway	657	8.7
K-means	Neyman	764	6.4
	proportional	1 806	1.1
	equal	967	4
	midway	770	6.3

Note: SE – standard error. The relative efficiency (RE) is relative to SRS.
Source: Authors' own elaboration.

Model-assisted estimation using a map depicting a continuous variable

Model-assisted estimation offers the possibility of improving the precision and reliability of area estimates by making use of auxiliary data information for all or a large sample of population units, as well as a probability sample of reference observations from sample units (Särndal *et al*., 1992). For example, to increase the precision of estimating forest (or land cover) change areas (y), a remote-sensing-based auxiliary variable (x) that is correlated to y can be used in combination with a probability sample of forest change observations (GFOI, 2020). As model-assisted estimation builds on design-based inference, the use of a probability sample is a must.

Model-assisted estimators are increasingly used with forest inventory and remotely sensed data to increase the precision of estimates of inventory parameters (McRoberts *et al*., 2022). The estimator of the population mean consists of the sum of a predictions-based term (the model part) and a residuals-based adjustment term, which is based on the mean of differences between sample unit observations and their corresponding predictions (Equation 20). The predictions-based term – also referred to as the synthetic estimator – uses an exhaustively sampled mapped auxiliary variable. Typically, the auxiliary mapped variable is first calibrated using some form of regression to a reference sample to predict the target variable.

Depending on how reference sample observations are used in the synthetic estimate, different estimators are used for model-assisted inference (GFOI, 2020). The main distinguishing characteristic between these two is that, for the difference estimator, the reference data used for adjusting the map bias is not used for calibrating the auxiliary variable independent from the synthetic estimate based on map values.

Table 9 Model-assisted estimation approaches and their estimators

Model-assisted approaches	Estimators	Estimator unbiasedness
Map unit values directly derived from the map (without calibration) using a probability sample of reference of an **existing** map; synthetic estimate from the map is independent of the reference data	Model-assisted difference estimator	Unbiased for any sample size (*n*)
Map unit values **calibrated** using a probability sample of reference observations	Model-assisted generalized regression estimator (GREG)	Asymptotically unbiased: a bias of order (1/*n*)
Model predictions of response variables based on the relationship with auxiliary variables		

Source: Authors' own elaboration based on McRoberts *et al.* 2022.

If the same reference dataset is used for both calibrating the map and assessing the residuals between sample unit observations and their corresponding predictions, the regression estimator is used. In this chapter, generalized regression estimator (GREG) refers to using any regression model for calibrating the input map (McRoberts *et al.*, 2022); for example, this can be a linear model, a random forest model or a logistic regression. The GREG estimator is asymptotically unbiased for both variance and mean estimators, which means it approaches unbiasedness only with a large sample size (McRoberts *et al.*, 2022). The regression estimator of the population mean within stratum h is:

$$\bar{y}_{hreg} = \frac{1}{N_h} \sum_{i=1}^{N_h} \hat{y}_{hi} + \frac{1}{n_h} \sum_{i=1}^{n_h} \varepsilon_{hi}$$

(Equation 20)

where $\hat{y}_{hi}$ denotes the predicted value of the target variable for map unit i within stratum h and is the ε_{hi} difference (residual) between the observation at sample unit i within stratum h and the prediction at that location: $\varepsilon_{hi} = y_{hi} - \hat{y}_{hi}$. (Equation 21)

The estimated mean for the entire population is computed by aggregation, as before:

$$\bar{y}_{reg} = \sum_{h=1}^{H} W_h \bar{y}_{h,reg}$$

(Equation 22)

Särndal *et al*. (1992) recognized the effects of sampling variability on the variability of the calibration model parameter estimates, and accordingly, on the variability of the population parameter estimates. They account for this effect by weighting individual sample units by so-called g-weights in the variance estimator for the model-assisted population total. However, for large samples and linear regression, the g-weights are all similar and close to 1 (McRoberts *et al.*, 2022); hence, they are omitted here. If a non-linear regression approach such as random forests is used, the g-weights are not readily computed. In a somewhat different context, Wadoux and Heuvelink (2023) proposed using residuals from cross-validation to assess the residual variability. In either case, the estimator of the population variance in stratum h becomes:

$$s^2_{h,reg} = \sum_{i=1}^{n_h} \frac{\left(\varepsilon_{hi} - \bar{\varepsilon}_h\right)^2}{n_h - 1}$$

(Equation 23)

where $\bar{\varepsilon}_h$ is the mean of the sample residuals within stratum h:

$$\bar{\varepsilon}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \varepsilon_{hi}.$$

(Equation 24)

The regression estimator of the variance of the population mean and total is obtained by replacing $s^2_{h,reg}$ in Equation 17 and Equation 19, respectively.

To achieve improved precision on forest change area estimation using the model-assisted approach, a common method is to use forest change maps in continuous forms (for example, as a proportion of forest change categories over coarser blocks). Pickering *et al*. (2019) produced a wall-to-wall map of forest loss area in blocks of 24 km × 24 km, aggregated from global and national maps, and then used a model-assisted regression and difference estimators to increase the precision of forest loss area estimation. Their results showed that model-assisted estimators incorporating global and national-scale maps improved the precision compared to sample-based estimates. These findings are consistent with Sannier *et al*. (2014, 2016). Francini *et al*. (2021) also created a map of the total area of clearcuts over 10 km × 10 km blocks, and using a model-assisted GREG approach, they corrected the map bias estimated from a probability sampling. Further numerical examples on this can also be found in Box 33 of the *Method and Guidance Document* (GFOI, 2020).

At finer resolution, such as 30 m pixel level, McRoberts *et al*. (2016) used the tree canopy cover proportion layer created from the GFC dataset, as well as a logistic regression model, to predict the forest cover probability. They estimated forest area using a model-assisted GREG estimator, in which the synthetic estimates derived from the map were bias-adjusted by a design-based correction term. The model-assisted estimator yielded more precise estimates than the direct estimator, which relied solely on the sampled forest presence and did not use any auxiliary map information. The GREG estimator has also been used to estimate the proportion of tree cover areas in Europe using *in situ* reference data and Copernicus High Resolution Layer on Tree Cover Density (Sannier *et al*., 2023), as well as forest area estimation using the global forest/non-forest map from the Japanese Aerospace and Exploration Agency (JAXA) Kyoto & Carbon program as spatially explicit auxiliaries and model-assisted estimators with logistic regression modelling (probability of forest) in African miombo forests (Næsset *et al.*, 2016).

Demonstration of forest-loss area estimation using model-assisted approaches

Using the same data used for the example in the previous subsection, we demonstrate area estimation using different model-assisted approaches with various stratification options in this section.

The continuous forest change map at 300 m resolution aggregated from the GFC (Figure 11) forest loss pixels for 2016–2019 were used for stratification to select reference sample pixels. The reference sample forest loss area proportion was then retrieved from the reference map (see Figure 11). When using the GREG estimator, the map-based estimate (GFC continuous forest change at 300 m resolution) was then calibrated using linear regression. The header, “No-calibration”, refers to the use of the difference estimator without calibration.

To define stratification boundaries, K-means – as well as Dalenius and Hodge's Cumrootf methods – were used to create three strata. For sample allocation, proportional and Neyman methods were used. For comparison, an SRS was employed as well. Table 10 lists the estimated forest change area with its precision with varying options for model-assisted approaches and different stratification methods, as a result of 1 000 iterations.

On average, GREG estimates based on using linear regression and sample data allocated to K-means and Cumrootf strata using Neyman allocation yielded the lowest standard errors, and consequently, the highest relative efficiencies. Compared with the results obtained through stratification alone (Table 7), the efficiency gains achieved by applying the GREG estimator were only modest in this case. In contrast, the difference estimator performed worse than the stratified estimators reported in Table 7, underscoring the importance of proper calibration.

Table 10 Estimated deforestation area, standard error, and relative efficiency of the regression and difference estimators under two stratification schemes and two sample allocation strategies, based on a total sample size of 5 000 units

Allocation	Linear model			No calibration		
	Estimated area (ha)	Standard error (ha)	Relative efficiency (compared to SRS)	Estimated area (ha)	Standard error (ha)	Relative efficiency (compared to SRS)
SRS						
–	7 039	1 739	1.2	7 017	2 932	0.4
Stratified Cumrootf (number of strata: 3)						
Neyman allocation	7 092	489	14.9	7 090	548	11.9
Proportional allocation	6 973	1 713	1.2	6 984	2 012	0.9
Stratified K-means (number of strata: 3)						
Neyman allocation	7 076	734	6.6	7 063	863	4.8
Proportional allocation	6 964	1 705	1.2	6 946	1 881	1.0
Reference area						7 066.8 ha

Note: Relative efficiency was calculated using the variance of the SRS estimator as the reference.
Source: Authors' own elaboration.

Estimation process in the case of a double sampling design

A double-sampling approach, such as eSBAE, has been introduced before and is motivated by the situation where a reliable wall-to-wall map cannot be produced, or would be too costly to produce on a regular basis. Instead of mapping the whole sampling frame, the continuous auxiliary variable is determined only on a large first-phase sample. This first-phase sample is a dense systematic grid, and it replaces the map as the source of stratifying information. Double sampling for stratification is effective if the value of increased precision resulting from stratification exceeds the cost of collecting the large preliminary sample (Shaw *et al.*, 2008).

The estimation then proceeds in two phases. In the first phase, the continuous layer is divided into strata using one of the boundary methods previously described. Since there is no map, the stratum weights are not known in advance. They are estimated from the first-phase sample as the proportion of first-phase units that fall in each stratum. In the second phase, a probability sample is drawn within the strata and interpreted visually to obtain the reference observations of change. These provide the mean of the target variable within each stratum (Equation 25).

The area estimate is then formed in the same way as for STR. The estimated stratum weights from the first phase are combined with the stratum means from the second phase. The corresponding estimator and its variance are given in Equation 25 and Equation 26. The variance has two components: the first is the sampling variance of the second-phase observations within strata; the second reflects the uncertainty in the stratum weights estimated from the first phase.

This second component is the price of not having a map, and its size depends on the first-phase sample. If the first-phase sample is large, the stratum weights are estimated precisely and this component becomes small. The double-sampling estimator then approaches the standard stratified estimator, and the two give practically the same result. This is why a very dense first-phase grid – of several tens of thousands of units – is used in approaches such as eSBAE.

Suppose a population is stratified into H classes. Stratum weight, or proportion of population falling in stratum *h*, is $Wh = N_h/N$. Set *n'* as the sample size for the first phase and *n* as that for the second phase. Then, $w_h = nh/n'$ is the proportion of the first sample falling in stratum h for the second-phase sampling.

The population mean is then calculated as follows:

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$$

(Equation 25)

First focus on stratification approach.

If *n'* and *N* are large:

$$\widehat{Var}(\bar{y}_{st}) = \sum_{h}^{L} \frac{w_h^2 s_h^2}{n_h} - \sum_{h}^{L} \frac{w_h s_h^2}{N} + \frac{g'}{n'} \sum_{h}^{L} w_h (\bar{y}_h - \bar{y}_{st})^2$$

(Equation 26)

$$g' = \frac{N - n'}{N - 1}$$

(Equation 27)

If the first phase sample is large, the effect of stratum area uncertainty on the estimates is small and the difference with results computed by Equation 17 is negligible.

Double sampling has been used to estimate detailed forest inventory variables such as forest biomass, volume of growing stock, and deadwood inventory (Ritter and Saborowski, 2014; Westfall *et al.*, 2019). Here, in the first phase, sampling units are stratified based on an auxiliary variable commonly sourced from aerial or satellite imagery. Regular grid intervals (such as 5 km) are used for the systematic first sample (Tomppo *et al.*, 2014), but SRS can also be an option. In the second phase, within-strata subsamples of the first-phase units are measured, usually by field sampling (Ritter and Saborowski, 2014). For forest canopy change assessments, the current remote-sensing technology allows spatially explicit (wall-to-wall) maps using automated processing capabilities. However, the timely collection of reference data on forest change using such maps as stratification is challenged by acceptable accuracy and processing time for wall-to-wall stratification layers. Double sampling for stratification can be a viable and efficient option for sample-based forest change area estimation if precise estimates are achieved while avoiding wall-to-wall map generation for stratification.

Good practice guidance and way forward

The approaches presented in this section are all based on probability samples and design-based estimators; there is not one that is universally preferable. The appropriate choice depends on the landscape characteristics, the data that is available, the capacity of the operator, and the overall sampling strategy. The first consideration is the form of the auxiliary variable. Where a wall-to-wall map can be produced at acceptable cost and accuracy, both stratification and model-assisted estimation are viable choices – the latter adding a calibration step that can extract further precision from the same map. Where a map is lacking or too costly, double sampling for stratification obtains the stratifying information from a large first-phase sample – at the cost of some additional variance from estimating the strata weights that is negligible in the case of large sample sizes.

The demonstrations in this section illustrate the gains that can be expected. In the stratified case, stratification on the continuous layer reduced the standard error of the forest-loss area estimate substantially relative to SRS and post-stratification; among the boundary methods, the Cumrootf method gave a lower standard error than K-means on average, and Neyman allocation outperformed proportional allocation. In the model-assisted case, GREG estimators – using both linear and logistic regression, calibrated on samples drawn under K-means and Cumrootf stratification with Neyman allocation – returned the lowest standard errors and the highest relative efficiency.

These results should be read as indicative rather than prescriptive because the design elements are interdependent: the form of the auxiliary variable constrains the estimator, the estimator constrains the allocation, and all of them depend on the quality of the underlying layer. Map quality therefore remains critical – a continuous layer that omits change will limit the achievable precision as surely as a categorical one. Several questions still require consolidated guidance – including variance estimation for model-assisted estimators based on non-linear models (where the g-weights are not readily available), the use of dense first-phase samples with regression rather than stratification, and standardized reporting of relative efficiency – so that approaches can be compared across countries.

References

Ballin, M. and Barcaroli, G. 2020. *R package SamplingStrata: new developments and extension to Spatial Sampling*. https://doi.org/10.48550/arXiv.2004.09366

Brandt, J., Ertel, J., Spore, J. and Stolle, F. 2023. Wall-to-wall mapping of tree extent in the tropics with Sentinel-1 and Sentinel-2. *Remote Sensing of Environment*, 292: 113574. https://doi.org/10.1016/j.rse.2023.113574

Brito, J., Ochi, L., Montenegro, F. and Maculan, N. 2010. An iterative local search approach applied to the optimal stratification problem. *International Transactions in Operational Research*, 17(6): 753–764.

Broich, M., Stehman, S.V., Hansen, M.C., Potapov, P. and Shimabukuro, Y.E. 2009. A comparison of sampling designs for estimating deforestation from Landsat imagery: a case study of the Brazilian Legal Amazon. *Remote Sensing of Environment*, 113(11): 2448–2454. https://doi.org/10.1016/j.rse.2009.07.011

Buchhorn, M., Bertels, L., Smets, B., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Masiliunas, D. and Li, L. 2020. *Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015–2019*: *algorithm theoretical basis document*.

Cochran, W.G. 1977. *Sampling techniques*. John Wiley and Sons.

Côte d'Ivoire. 2024. *Forest Reference Level submission to the UNFCCC*.

Dalenius, T. and Hodges, J.L. Jr. 1957. The choice of stratification points. *Scandinavian Actuarial Journal*, 1957(3–4): 198–203.

Francini, S., McRoberts, R.E., Giannetti, F., Marchetti, M., Scarascia Mugnozza, G. and Chirici, G. 2021. The Three Indices Three Dimensions (3I3D) algorithm: a new method for forest disturbance mapping and area estimation based on optical remotely sensed imagery. *International Journal of Remote Sensing*, 42(12): 4693–4711. https://doi.org/10.1080/01431161.2021.1899334

GFOI. 2020. *Methods and guidance from the Global Forest Observations Initiative: integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests.*

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D. *et al*. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160): 850–853. https://doi.org/10.1126/science.1244693

Healey, S.P., Cohen, W.B., Yang, Z. and Krankina, O.N. 2005. Comparison of Tasseled Cap-based Landsat data structures for use in forest disturbance detection. *Remote Sensing of Environment*, 97(3): 301–310. https://doi.org/10.1016/j.rse.2005.05.009

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.

Kozak, M. 2004. Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5): 797–806.

Kozak, M. 2011. Comparison of efficiency of geometric stratification and K-means algorithm in univariate stratification of skewed populations. *International Journal of Agricultural and Statistical Sciences*, 7(1): 341–344.

Lohr, S.L. 2019. *Sampling design and analysis*, second edition. CRC Press. https://doi.org/10.1201/9780429296284

McRoberts, R.E., Vibrans, A.C., Sannier, C., Næsset, E., Hansen, M.C., Walters, B.F. and Lingner, D.V. 2016. Methods for evaluating the utilities of local and global maps for increasing the precision of estimates of subtropical forest area. *Canadian Journal of Forest Research*, 46(7): 924–932. https://doi.org/10.1139/cjfr-2016-0064

McRoberts, R.E., Næsset, E., Heikkinen, J., Chen, Q., Strimbu, V., Esteban, J., Hou, Z., Giannetti, F., Mohammadi, J. and Chirici, G. 2022. On the model-assisted regression estimators using remotely sensed auxiliary data. *Remote Sensing of Environment*, 281: 113168. https://doi.org/10.1016/j.rse.2022.113168

Næsset, E., Ørka, H.O., Solberg, S., Bollandsås, O.M., Hansen, E.H., Mauya, E., Zahabu, E. *et al*. 2016. Mapping and estimating forest area and aboveground biomass in miombo woodlands in Tanzania using data from airborne laser scanning, TanDEM-X, RapidEye, and global forest maps: a comparison of estimated precision. *Remote Sensing of Environment*, 175: 282–300. https://doi.org/10.1016/j.rse.2016.01.006

Pickering, J., Stehman, S.V., Tyukavina, A., Potapov, P., Watt, P., Jantz, S.M., Bholanath, P. and Hansen, M.C. 2019. Quantifying the trade-off between cost and precision in estimating area of forest loss and degradation using probability sampling in Guyana. *Remote Sensing of Environment*, 221: 122–135. https://doi.org/10.1016/j.rse.2018.11.018

Ritter, T. and Saborowski, J. 2014. Efficient integration of a deadwood inventory into an existing forest inventory carried out as two-phase sampling for stratification. Forestry: *An International Journal of Forest Research*, 87(4): 571–581. https://doi.org/10.1093/forestry/cpu016

Sannier, C., McRoberts, R.E. and Fichet, L.-V. 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

Sannier, C., McRoberts, R.E. and Fichet, L.-V. 2016. Suitability of Global Forest Change data to report forest cover estimates at national level in Gabon. *Remote Sensing of Environment*, 173: 326–338. https://doi.org/10.1016/j.rse.2015.10.032

Sannier, C., Gallego, J., Langanke, T., Donezar, U. and Pennec, A. 2023. Tree cover area estimation in Europe based on the combination of in situ reference data and the Copernicus high resolution layer on tree cover density. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48: 277–284.

Särndal, C.-E., Swensson, B. and Wretman, J. 1992. *Model assisted survey sampling*. Springer-Verlag.

Shaw, C.H., Boyle, J.R. and Omule, A.Y. 2008. Estimating forest soil carbon and nitrogen stocks with double sampling for stratification. *Soil Science Society of America Journal*, 72(6): 1611–1620. https://doi.org/10.2136/sssaj2007.0219

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.

Verbesselt, J., Hyndman, R., Newnham, G. and Culvenor, D. 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1): 106–115. https://doi.org/10.1016/j.rse.2009.08.014

Wadoux, A.M.J.-C. and Heuvelink, G.B.M. 2023. Uncertainty of spatial averages and totals of natural resource maps. *Methods in Ecology and Evolution*, 14(5): 1320–1332. https://doi.org/10.1111/2041-210X.14106

Westfall, J.A., Lister, A.J., Scott, C.T. and Weber, T.A. 2019. Double sampling for post-stratification in forest inventory. *European Journal of Forest Research*, 138(3): 375–382. https://doi.org/10.1007/s10342-019-01171-9

Xu, L., Herold, M., Tsendbazar, N.-E., Masiliūnas, D., Li, L., Lesiv, M., Fritz, S. and Verbesselt, J. 2022. Time series analysis for global land cover change monitoring: a comparison across sensors. *Remote Sensing of Environment*, 271: 112905. https://doi.org/10.1016/j.rse.2022.112905

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

Zimbabwe. 2024. *Zimbabwe's Forest Reference Emission Level submission to the UNFCCC*. https://redd.unfccc.int/media/zimbabwe_frel_modifiedversion_2024.pdf

2.7 Assessment for small areas

by

Sytze de Bruin

Description of the problem

There may be situations where countries need spatially disaggregated estimates for subareas within their overall area of interest. For example, Ghana has proposed a performance-based benefit-sharing plan[17] for carbon credits assessed through the Architecture for REDD+ Transactions (ART) programme. To assess performance, the country intensifies sampling in so-called hotspot intervention areas to assess whether deforestation was reduced in those areas. The intensified sample information is not used in the overall area estimate, but it will affect the share of benefits to be received.

The Ghana example illustrates one approach to disaggregate an overall area estimate by augmenting the initially small sample within the subarea of interest by a targeted additional probability sample. The augmented probability sample then supports design-based estimation, as described elsewhere in this publication. However, additional sampling may not always be feasible. In such cases, model-based estimation offers alternatives for assessing the change area within the subarea(s) of interest. In model-based estimation of change area, a model is used to predict the categorical change or the probability of change at each location; the total area is obtained by integration over the area of interest. A straightforward example is reliance on an existing categorical change map, potentially supported by an overall map confusion matrix, which is used to account for expected misclassification errors.

A major disadvantage of all model-based approaches is that they inevitably introduce subjectivity, as any modelling exercise requires making assumptions and selecting among alternative model specifications, structures, and identification strategies. However, model-based prediction may be the best available option, and several studies (Brus and de Gruijter, 1997; Emick *et al.*, 2023; McRoberts, 2006; McRoberts *et al.*, 2022; Wadoux and Heuvelink, 2023) describe viable approaches and their characteristics.

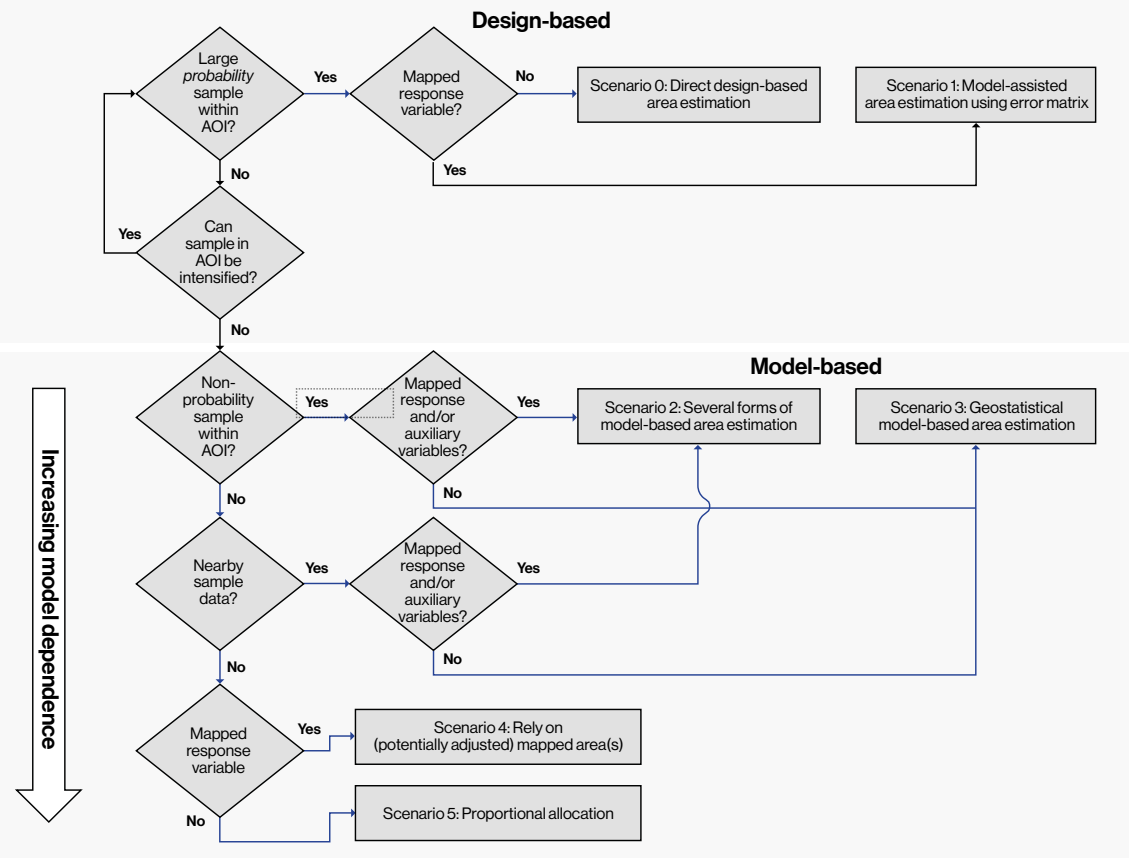
Overview of model-based approaches

Depending on the availability of nearby sample data and on whether a map of the target variable is provided (such as a discrete per-pixel change map or pixel-level change probabilities), a range of simple to more advanced model-based approaches can be used to estimate the change area within the area of interest. These are summarized in Figure 13. In this figure, the term “nearby sample data” refers to observational data collected in adjacent or otherwise spatially contiguous regions whose underlying data-generating processes are assumed to be comparable to those within the area of interest. For completeness, Figure 13 also includes design-based area estimation in situations where a large probability sample of data points is available or can be obtained within the area of interest.

When design-based estimation is not feasible, Figure 13 highlights a set of methods that have been widely demonstrated in the scientific literature for continuous variables such as above-ground biomass (Saarela *et al.*, 2016; Saarela *et al.*, 2022; Araza *et al.*, 2022; Emick *et al.*, 2023; Wadoux and Heuvelink, 2023) and, to a lesser extent, for estimating land-cover class areas or proportions (de Bruin, 2000; McRoberts, 2006, 2010; Sales *et al.*, 2022). These approaches are collectively referred to as model-based estimation.

[17] See https://fc.gov.gh/the-benefit-sharing-plan-bsp

Figure 13 Options for change area assessment within a small area of interest



Note: AOI – area of interest. The "Increasing model dependence" arrow bar on the left refers to the position of the conditions represented by diamonds and denotes increasing risk of bias.

Scenario 0 and Scenario 1: Design-based area estimation

Design-based estimation is model-free and therefore avoids the subjective choices inherent in model specification and identification. Stehman (2013) introduces the direct estimator, which fully relies on the probability sample without using a map of the response variable; it corresponds to Scenario 0 in Figure 13. Considering Scenario 1 in Figure 13, Stehman (2013) provides a comprehensive treatment of design-based estimators for class-area estimation using data derived from an error matrix constructed from simple random or stratified probability samples; in Stehman (2014), this framework is extended to cases in which the strata do not correspond to the map classes. Næsset *et al*. (2016) provide model-assisted area estimators using continuous covariates. A detailed discussion of design-based estimation is also presented in previous subsections of this section.

Scenario 2: Model-based area estimation

Model-based inference necessarily depends on subjective choices regarding model specification and identification, and these choices directly influence the reliability of the resulting inference. Several characteristics of model-based prediction affect this reliability. Incorrect model specification – for example, assuming linearity for inherently non-linear processes – typically introduces model bias (Ståhl *et al.*, 2024). Ståhl *et al*. (2024) refer to this as conventional model bias, defined as the difference between the expected value of the model-based predictor and the expected value of the property being predicted. Even when conventional model bias is absent, model-based predictors may still exhibit bias due to what the authors term "design bias", which manifests as a tendency to overestimate small true values and underestimate large true values. In addition, preferential sampling can distort model-parameter estimation and thereby introduce systematic error in predicted population means or totals. Such systematic errors are often accompanied by an underestimation of prediction uncertainty.

Scenario 2 employs wall-to-wall mapped response variables or auxiliary variables to predict, for every location, the probability of presence of the target class. A range of modelling approaches has been used for this purpose, including indicator kriging (de Bruin, 2000), logistic regression (McRoberts, 2006), generalized linear geostatistical modelling (Diggle and Ribeiro Jr, 2007), and machine learning (Sales *et al*., 2022). By summing the class probabilities within the target area and multiplying this total by the pixel area, the expected class area is obtained. De Bruin (2000) applied stochastic simulation to quantify uncertainty in class-area estimates, whereas McRoberts (2006) modelled covariances among pairs of pixel-level predictions within the area of interest using a variogram of standardized residuals. An alternative proposal to use block kriging for quantifying uncertainty about area estimates remains under investigation. The resulting predictions and their associated uncertainty depend strongly on the adequacy of the inferred variogram model and the spatial representativeness of the sample. Consequently, the arrow in Figure 13 illustrates that model dependence increases when no probability sample is available within the area of interest.

Scenario 3: Geostatistical model-based area estimation

In the absence of wall-to-wall mapped covariates, geostatistical methods (de Bruin, 2000; Diggle and Ribeiro Jr, 2007) can still be used to predict the probability of presence of the target class. In this setting, spatial dependence is inferred directly from the sample data, and predictions are obtained through indicator kriging or generalized linear geostatistical modelling. These approaches allow the construction of spatially explicit probability surfaces when auxiliary variables are unavailable, and, similar to Scenario 2, the expected class area is obtained by summing the class membership probabilities over the target area and multiplying the resulting total by the pixel area. Again, the reliability of the resulting predictions depends critically on the quality of the estimated spatial correlation structure and the spatial coverage of the sample.

Scenario 4: Rely on potentially adjusted mapped areas

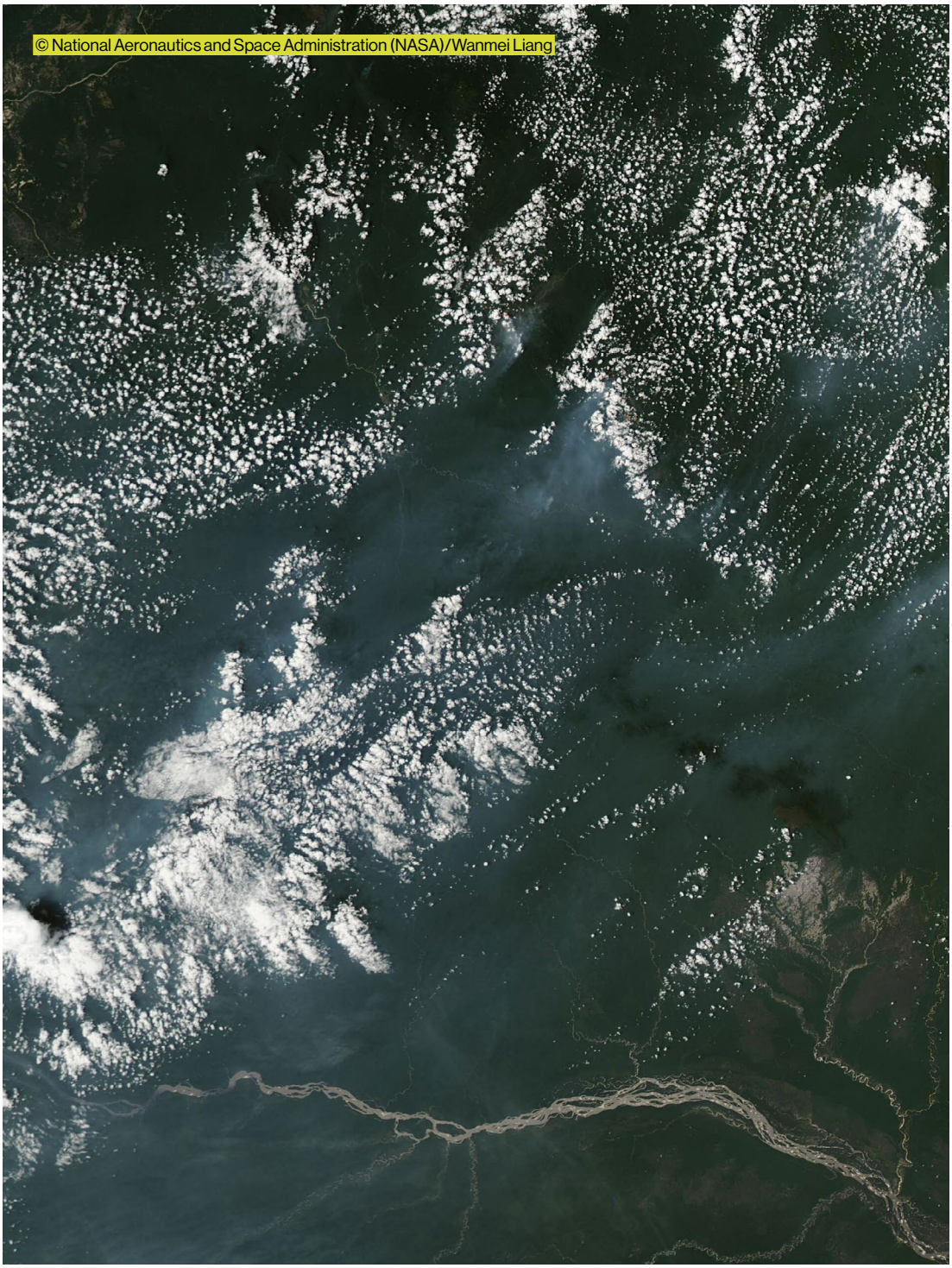
In the absence of nearby reference data, directly using an uncalibrated map is conceptually straightforward but highly error-prone. Moreover, without additional information, such an approach does not support uncertainty estimation. When the map provides discrete per-pixel labels indicating whether each pixel belongs to the target class, the class area is obtained by counting the pixels labelled as that class and multiplying this count by the pixel area. However, this procedure is known to yield measurement-biased area estimates because classification errors translate directly into erroneous pixel counts. This issue is especially pronounced for rare classes (Stehman, 2005; Witjes *et al*., 2024), as low-probability classes, such as the change class, tend to be systematically underrepresented in the mapped output.

When an error matrix is available and is assumed to hold for the small area of interest, it can be used to correct the pixel-count-derived area estimates by accounting for misclassification rates. This straightforward approach presumes that the systematic map error is uniform across the entire area, which is typically not the case. Even so, it is generally considered better than using the pixel count without any correction. An example is the forest reference emission level (FREL) submission to the UNFCCC of Ethiopia, where this method has been used to achieve bias-corrected estimates of forest loss and gain across various biomes (Ethiopia, 2017).

An alternative procedure can be used if the map depicts change probabilities. These are then summed over the area of interest, and the resulting sum is multiplied by the pixel area to obtain the expected change area. However, without further information, the latter approach does not allow an assessment of confidence intervals for the class area (de Bruin, 2000). To obtain such intervals, one requires the residual covariances between pairs of predicted class probabilities within the area of interest, which can be estimated through variogram modelling (McRoberts, 2006; Araza *et al*., 2022; Wadoux and Heuvelink, 2023).

Scenario 5: Proportional allocation

Scenario 5 applies when there is insufficient sample data within or near the target small area to support a reliable estimate, and no suitable map or ancillary spatial information is available. In such circumstances, the range of defensible estimation methods is extremely limited. A pragmatic approach is to allocate the overall change area to the subarea in proportion to its share of the total area. In the small area estimation literature, this corresponds to the broad area ratio estimator (Asian Development Bank, 2020). It borrows strength from the broader area by applying the overall change rate to the subarea, and is a recognized option precisely when no other information is available to support a more specific estimate. However, this method assumes that change is uniformly distributed across the landscape, and therefore completely ignores any spatial variation in change patterns.



References

Araza, A., de Bruin, S., Herold, M., Quegan, S., Labriere, N., Rodriguez-Veiga, P., Avitabile, V., Santoro, M., Mitchard, E.T.A., Ryan, C.M., Phillips, O.L., Willcock, S., Verbeeck, H., Carreiras, J., Hein, L., Schelhaas, M.-J., Pacheco-Pascagaza, A.M., da Conceição Bispo, P., Laurin, G.V. and Lucas, R. 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

ART (Architecture for REDD+ Transactions). 2021. *The REDD+ Environmental Excellence Standard (TREES)*, Version 2.0. https://www.artredd.org/trees/standard-and-templates

Asian Development Bank. 2020. *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices.* Manila. https://doi.org/10.22617/TIM200160-2

Brus, D.J. and de Gruijter, J.J. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1): 1–44. https://doi.org/10.1016/S0016-7061(97)00072-4

de Bruin, S. 2000. Predicting the areal extent of land-cover types using classified imagery and geostatistics. *Remote Sensing of Environment*, 74(3): 387–396. https://doi.org/10.1016/S0034-4257(00)00132-2

Diggle, P.J. and Ribeiro Jr, P.J. 2007. *Model-based geostatistics*. New York, USA, Springer. https://doi.org/10.1007/978-0-387-48536-2

Emick, E., Babcock, C., White, G.W., Hudak, A.T., Domke, G.M. and Finley, A.O. 2023. An approach to estimating forest biomass while quantifying estimate uncertainty and correcting bias in machine learning maps. *Remote Sensing of Environment*, 295: 113678. https://doi.org/10.1016/j.rse.2023.113678

Ethiopia. 2017. *Ethiopia's forest reference level submission to the UNFCCC.* https://redd.unfccc.int/media/ethiopia_frel_3.2_final_modified_submission.pdf

McRoberts, R.E. 2006. A model-based approach to estimating forest area. *Remote Sensing of Environment*, 103(1): 56–66. https://doi.org/10.1016/j.rse.2006.03.005

McRoberts, R.E. 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment*, 114(5): 1017–1025. https://doi.org/10.1016/j.rse.2009.12.013

McRoberts, R.E., Næsset, E., Saatchi, S. and Quegan, S. 2022. Statistically rigorous, model-based inferences from maps. *Remote Sensing of Environment*, 279: 113028. https://doi.org/10.1016/j.rse.2022.113028

Næsset, E., Ørka, H.O., Solberg, S., Bollandsås, O.M., Hansen, E.H., Mauya, E., Zahabu, E., Malimbwi, R., Chamuya, N., Olsson, H. and Gobakken, T. 2016. Mapping and estimating forest area and aboveground biomass in miombo woodlands in Tanzania using data from airborne laser scanning, TanDEM-X, RapidEye, and global forest maps: a comparison of estimated precision. *Remote Sensing of Environment*, 175: 282–300. https://doi.org/10.1016/j.rse.2016.01.006

Saarela, S., Holm, S., Grafström, A., Schnell, S., Næsset, E., Gregoire, T.G., Nelson, R.F. and Ståhl, G. 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. *Annals of Forest Science*, 73(4): 895–910. https://doi.org/10.1007/s13595-016-0590-1

Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Yang, Z., Andersen, H.-E., Dubayah, R.O., Qi, W., Duncanson, L.I., Armston, J.D., Gobakken, T., Næsset, E., Ekström, M. and Ståhl, G. 2022. Comparing frameworks for biomass prediction for the Global Ecosystem Dynamics Investigation. *Remote Sensing of Environment*, 278: 113074. https://doi.org/10.1016/j.rse.2022.113074

Sales, M.H.R., de Bruin, S., Souza, C. and Herold, M. 2022. Land use and land cover area estimates from class membership probability of a random forest classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 4402711. https://doi.org/10.1109/TGRS.2021.3080083

Ståhl, G., Gobakken, T., Saarela, S., Persson, H.J., Ekström, M., Healey, S.P., Yang, Z., Holmgren, J., Lindberg, E., Nyström, K., Papucci, E., Ulvdal, P., Ørka, H.O., Næsset, E., Hou, Z., Olsson, H. and McRoberts, R.E. 2024. Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time, and how this affects applications. *Forest Ecosystems*, 11: 100164. https://doi.org/10.1016/j.fecs.2023.100164

Stehman, S.V. 2005. Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, 96(3): 466–474. https://doi.org/10.1016/j.rse.2005.04.002

Stehman, S.V. 2013. Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, 132: 202–211. https://doi.org/10.1016/j.rse.2013.01.016

Stehman, S.V. 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing*, 35(13): 4923–4939. https://doi.org/10.1080/01431161.2014.930207

Wadoux, A.M.J.-C. and Heuvelink, G.B.M. 2023. Uncertainty of spatial averages and totals of natural resource maps. *Methods in Ecology and Evolution*, 14(5): 1320–1332. https://doi.org/10.1111/2041-210X.14106

Witjes, M., Herold, M. and de Bruin, S. 2024. Iterative mapping of probabilities: a data fusion framework for generating accurate land cover maps that match area statistics. *International Journal of Applied Earth Observation and Geoinformation*, 131: 103932. https://doi.org/10.1016/j.jag.2024.103932


© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)/Cory Wright


© FAO/Roberto Cenciarelli
SAMSUNG
PLOT [3]
LAND FEATURES [1]
Navigation
PLOT
LAND FEATURES
LAND FEATURE DETAILS
OBJECT DESCRIPTION
TREE DETAILS
SEEDLING DETAILS
FINALIZATION
Land feature ID
1
Recommended land class name
Forest Plantation
Land feature notes
Evidence of planted trees
Current page:
LAND FEATURES
1
Delete this
Show as table
Show names
PLOT
LAND FEATURE DETAILS

2.8 Summary of Section 2

by

German Obando-Vargas and Randy Hamilton (with second-edition additions from Andreas Vollrath)

Section 2 addresses a number of sampling design questions that various countries around the world have raised in the context of implementing SBAE. Compared with the first edition, this section has been expanded with dedicated guidance on determining the sample size and its allocation among strata (Section 2.3), using maps that represent forest change as a continuous variable (Section 2.6), and estimating change for small areas within the overall area of interest (Section 2.7). Table 11 summarizes the key considerations and recommendations highlighted throughout the section.

Table 11 Summary of key considerations and recommendations related to sampling design for sample-based area estimation

Consideration	SYS and SRS	STR
Precision obtained *(Section 2.1–2.3)*	Both 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). Both 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	SYS and SRS	STR
Ability to track LULC between monitoring periods *(Section 2.4 and Section 3.5)*	Both 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 LULC 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 LULC 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. Partly permanent designs, cumulative stratification, and double sampling (Section 2.4) allow part of the stratified sample to be kept permanent across periods.
Possibility to apply special adjustments to obtain desired precision in rare classes *(Section 2.3–2.4)*	Neither SYS nor SRS is 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.	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 follow-up step, such as for intensification.	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.
Determining the sample size and its allocation *(Section 2.3)*	The only quantity to determine is the total sample size, which follows from the target margin of error and the expected proportion of the class of interest.[i] Because every unit has an equal probability of selection, rare classes such as forest change require large sample sizes to be estimated with useful precision; there is no allocation decision to make.	In addition to the total sample size, the sample must be divided among the strata. The allocation scheme (equal, proportional, midway or Neyman) can influence the precision as much as the stratification itself, so that a design twice as efficient delivers the same precision for half the budget. A minimum number of units should be secured in every stratum to avoid undersampling bias (“rule of three”). Where within-stratum variances are unknown, a midway allocation (the average of equal and proportional) approximates the optimum without requiring prior information and serves multiple target variables at once.

Consideration	SYS and SRS	STR
Use of a map representing a continuous variable *(Section 2.6)*	A continuous layer of forest-change probability can be used to post-stratify the equal-probability sample, or as an auxiliary variable in a model-assisted estimator (difference or generalized regression [GREG]) to improve precision without changing the design.	A continuous layer can define the strata directly (stratification on a continuous variable [e.g. K-means or cumulative-root-frequency boundaries]), which avoids the need for an accurate categorical map. Where a wall-to-wall layer is too costly to produce, a dense first-phase sample (double sampling) supplies the stratifying information; model-assisted estimation can further calibrate the layer against the reference sample. Section 2.6 lists the available options.
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). The area of domain A_D can be estimated using point sampling with the following expression. *(Section 2.5–2.6)*	$\hat{A}_D = A_R \times \hat{P}_R$ (using definitions from Equation 9)	$\hat{A}_{DS} = A_R \times \hat{P}_{RS}$ (using definitions from Equation 11)
Standard error estimator for A_D, where $v(\hat{P}_R)$ 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 10)	$S(\hat{A}_{DS}) = A_R \times \sqrt{v_s(\hat{P}_{RS})}$ (Equation 12 and Equation 13)
Confidence intervals	$CI_D = \hat{A}_D \pm \left(t_{\alpha/2} \times S(\hat{A}_D)\right)$	$CI_D = \hat{A}_{DS} \pm \left(t_{\alpha/2} \times S(\hat{A}_{DS})\right)$

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

[i] Cochran, W.G. 1977. *Sampling techniques, Third edition*. New York, USA, John Wiley and Sons.



Section 3

Response design

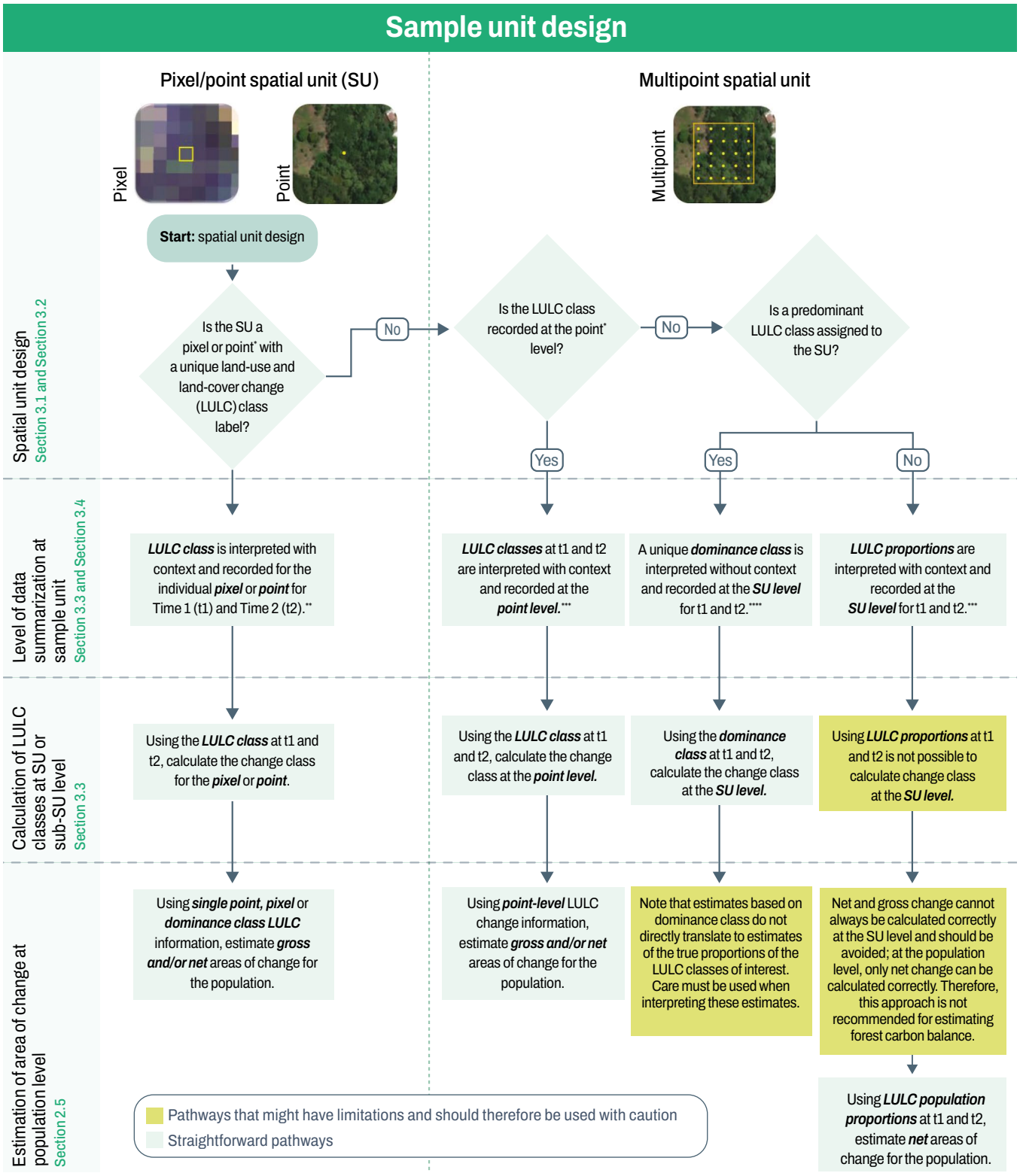


The most effective designs for SBAE or other types of inventories are constructed using a logical sequence of steps. These steps, variations of which have been described elsewhere (Lister *et al.*, 2015) include assessing information needs, identifying constraints, determining allowable error for key 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 14 presents a decision tree that shows the interrelationship of these topics and outlines flow of plot design decisions.



Figure 14 Decision tree of interrelated topics in Section 3 and order of decisions to be made



Notes: Courtesy of German Obando-Vargas.

* 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" refer to the points within a plot that are evaluated to determine the LULC composition of the plot.

** **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.

*** **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.

**** **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.



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 LULC 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 LULC composition is frequently characterized by overlaying a systematic grid of points and interpreting from fine-resolution imagery the LULC 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 $V(\hat{\hat{P}}_R)$ was given. See those sections for any undefined notation. 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, $V(\hat{\hat{P}}_R)$, 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 28)

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 $V(\hat{\hat{P}}_R)$ is equal to 0.

Hence, $V(\hat{\hat{P}}_R) = \frac{1}{n}[P_R(1-P_R)]$ in the case of binary data from a point or pixel sample unit. In the case of area-based, multiplepoint units: $0 \le P(s) \le 1$ for any s in the region R. If $0 \le P(s) \le 1$, then $P(s)\big(1-P(s)\big) \ge 0$.. 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, $V(\hat{\hat{P}}_R)$, 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 LULC 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*. 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://www.mdpi.com/1999-4907/10/10/896

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*,

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

by

Andrew Lister, Paul Patterson and Randy Hamilton

10(4): 437–450. https://doi.org/10.1080/17583004.2019.1634934

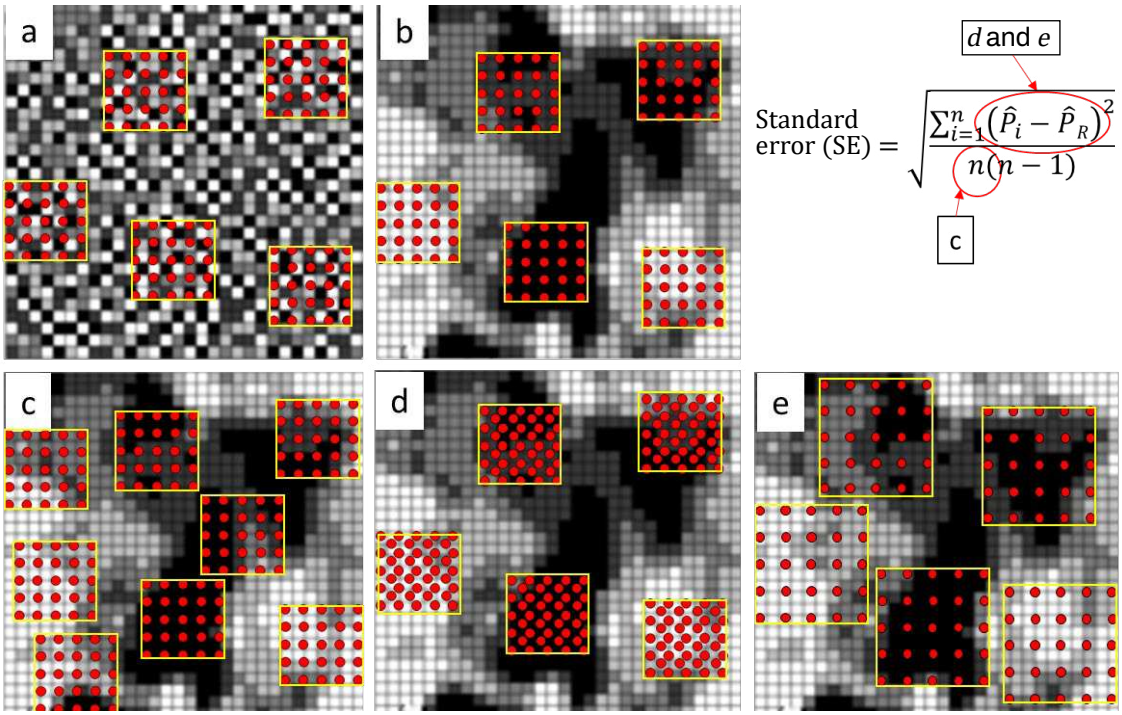
Common questions posed by countries implementing SBAE using larger area-based sample units (or plots) include: how many plots are needed; how large should they be and how many points should be placed within each one? The optimal design of area-based sample units can be informed from basic principles of statistics. Figure 15 empirically illustrates several important principles. Standard error (SE in Figure 15) 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 15a 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 15b, 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 15b, one can increase n (put out more plots), increasing the denominator of the standard error equation, as shown in Figure 15c. Another option is to increase the number of points within the plot (Figure 15d), 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 15e).

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 15c, Figure 15d and Figure 15e) leads to the lowest cost evaluation that will meet the needs of users.



Figure 15 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).
Refer to the disclaimer on p. ii for the names and boundaries used in this map.

Source: Authors' own elaboration using Esri (Environmental Systems Research Institute). 2026. *ArcGIS Pro*. https://www.esri.com/it-it/arcgis/products/arcgis-pro/overview

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 29)

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 30)

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 31)

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.

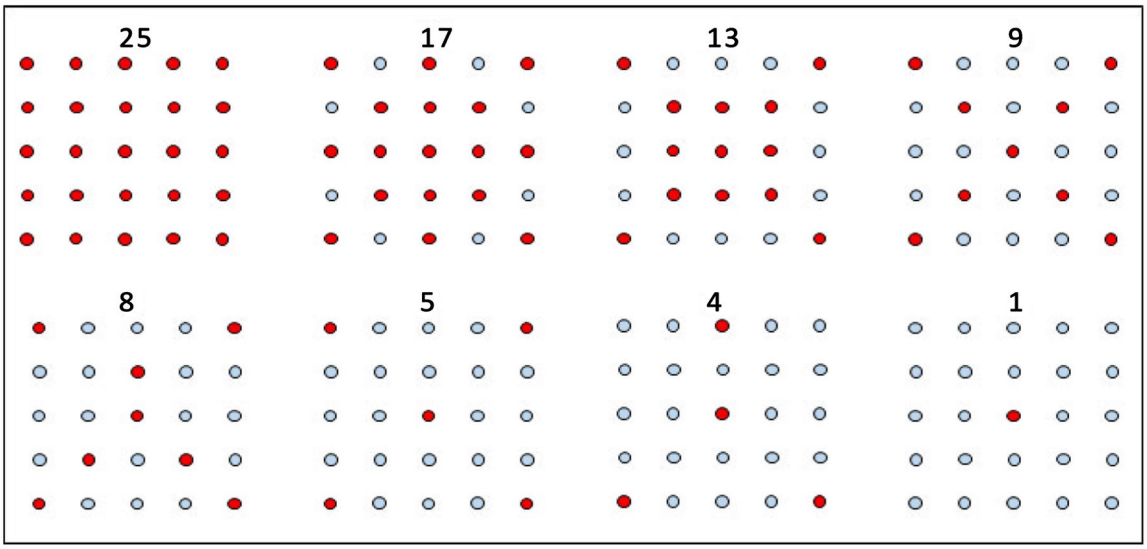
Workflow for a simple plot design optimization study

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. 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 16).

Figure 16 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 32)

$t = the\ (1\ -\alpha/2)^{th}$ 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 33)

- 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:

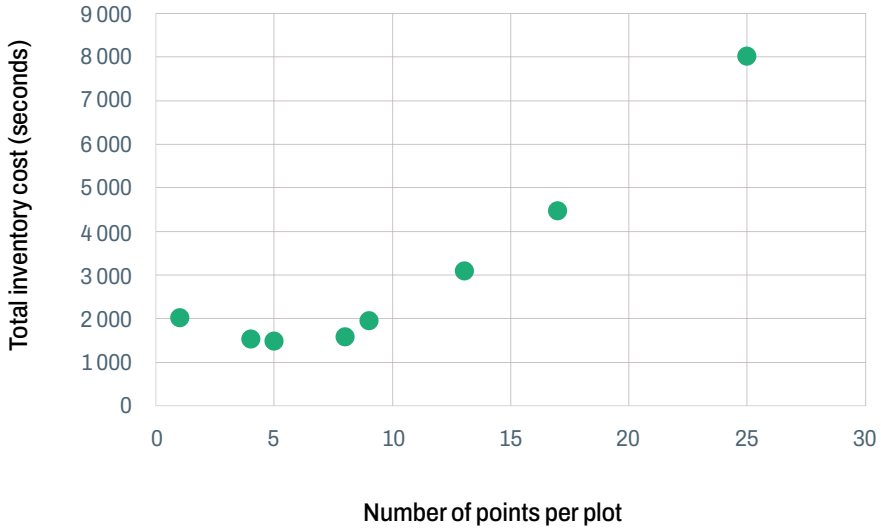
$$Total\ time = n_{required} \times m \times\ t_{on} + \left(n_{required} - 1\right) \times t_{between}$$

(Equation 34)

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 16).

Figure 17 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 17, 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 17 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 17 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 the United Republic 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 14) 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

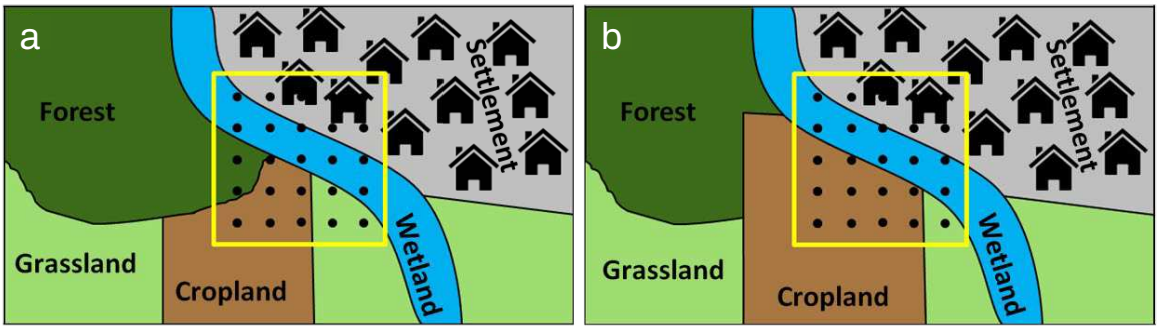
Paul Patterson, Frédéric Achard, Andrew Lister and Randy Hamilton

This section discusses the pros and cons of three protocols sometimes used to label the LULC 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 LULC 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 18, 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 18 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 18 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.

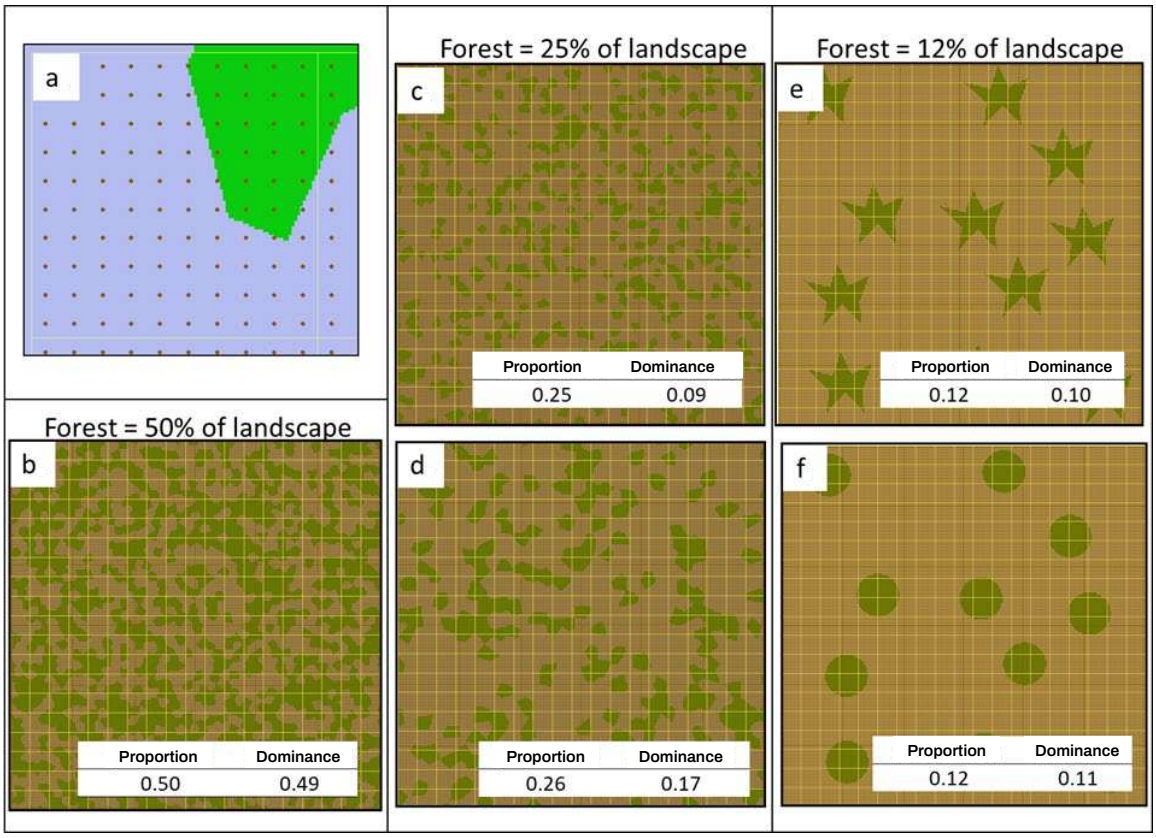
There are other possible drawbacks to assigning dominance classes. First, subtle changes in the land use composition in the plot, such as from 35 percent forest coverage to 29 percent coverage, can cause the dominance class to change, and depending on the number of such changes, could cause inflated 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 19. 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 19a, 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 19b). 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.


© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)/Cory Wright

Figure 19 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.

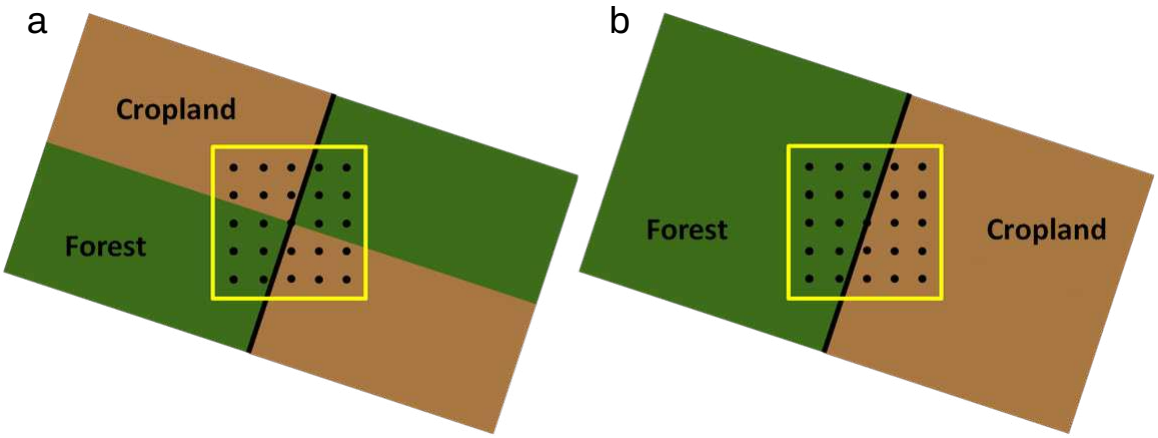
It is also important to reiterate, as mentioned previously, that estimates based on dominance are not directly comparable to those based on proportions from point-level data. In addition, dominance class estimates cannot be directly interpreted as estimates of the true proportion of the attribute of 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 20, 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 20 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.

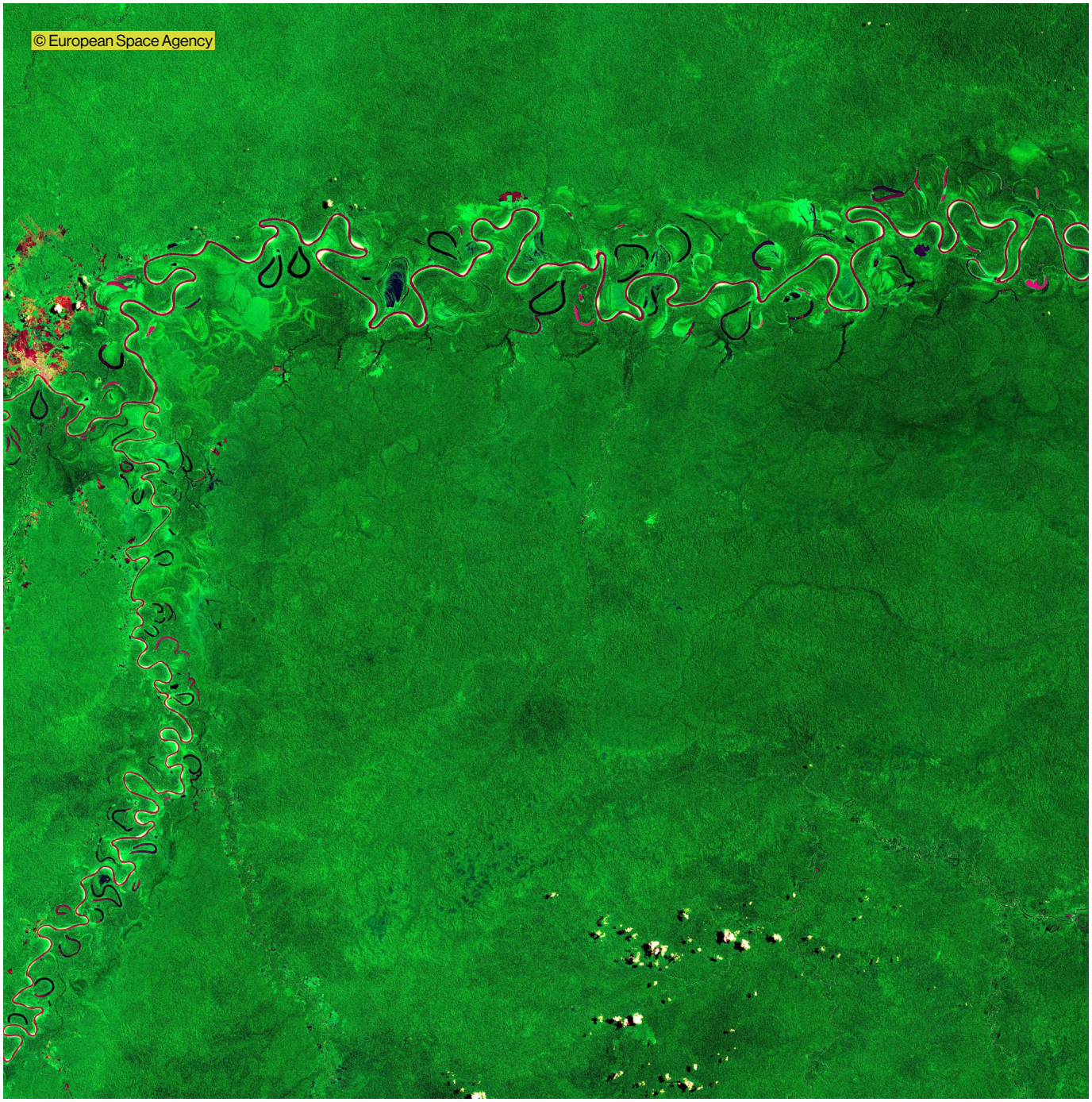
Point-level land use information

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 20, 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




© FAO

3.4 Interpretation paradigms: Interpretation without or with context

by

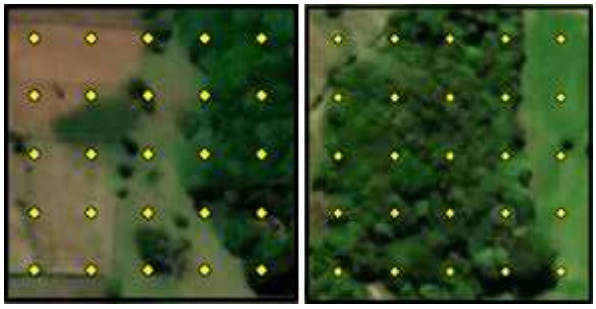
Randy Hamilton and Paul Patterson

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 21). 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 21 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.

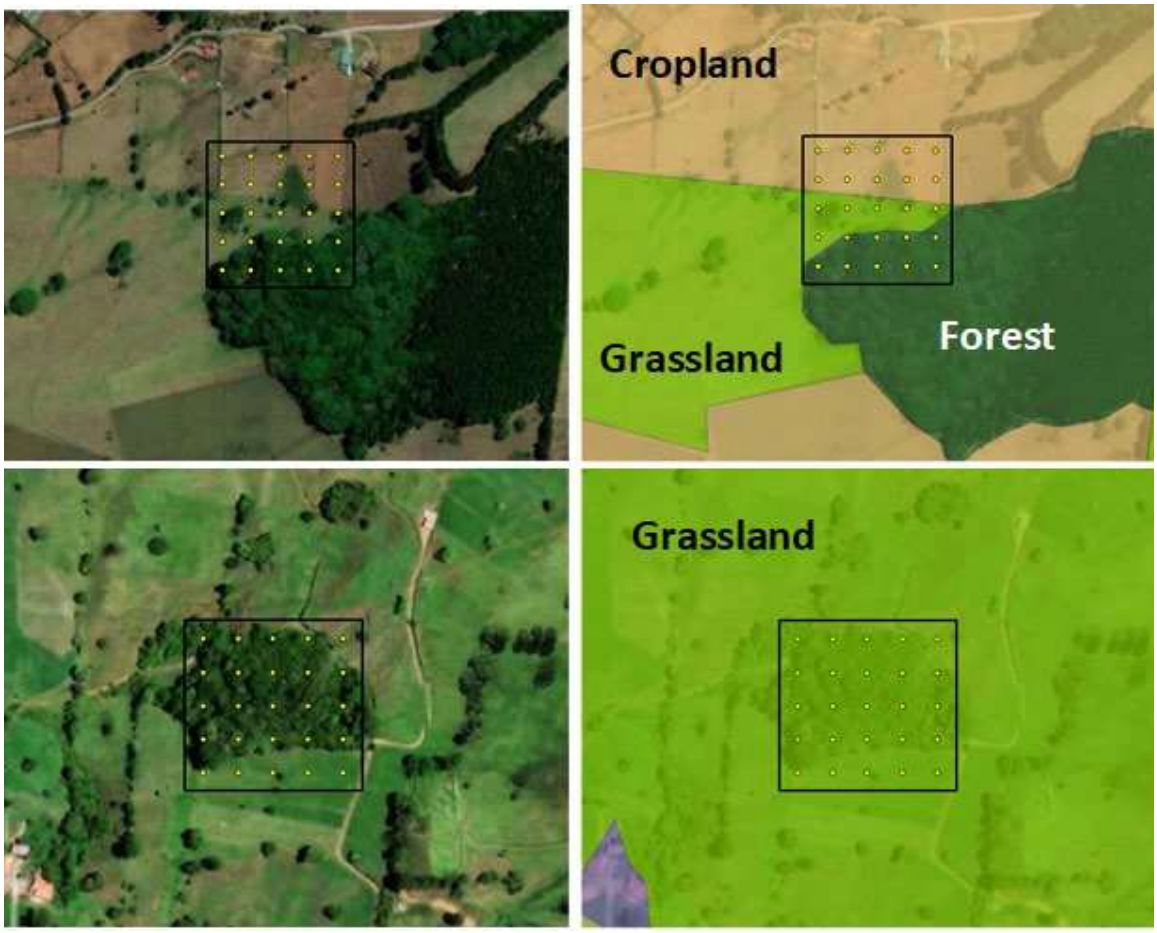
Interpretation without context could be difficult to apply in multipurpose monitoring systems where other land uses, with their own cover and minimum map unit definitions, must also be applied and reconciled with each other. Under certain conditions, this approach could also lead to results considerably 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 [GIS]) 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 22). 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 22 Assessing interpretation with context



Notes:
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 21. Just as in Figure 21, 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 21).
Refer to the disclaimer on p. ii for the names and boundaries used in this map.
Source: Authors' own elaboration using Esri (Environmental Systems Research Institute). 2026. *ArcGIS Pro*. https://www.esri.com/it-it/arcgis/products/arcgis-pro/overview

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*. https://www.reddcompass.org/mgd/resources/GFOI-MGD-3.1-en.pdf



© United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)/Cory Wright

3.5 Updating reference labels for operational monitoring

by

Nandika Tsendbazar (with existing and adopted content from the first edition written by Andres Espejo and Randy Hamilton)

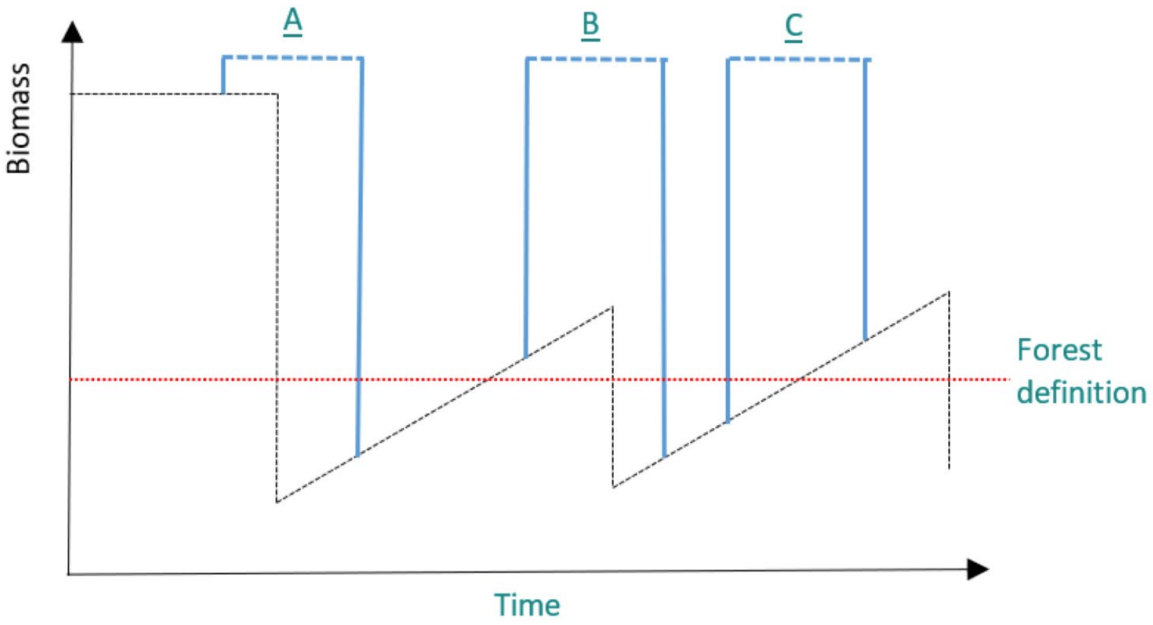
For operational assessments of forest change area, accurate and up-to-date reference labels are crucial. The quality of the reference dataset can adversely affect area estimates and their estimated uncertainty for SBAE. Guidance on ensuring high-quality reference datasets and assessing reference data quality is included in Section 4.1. The reference datasets are often created through visual interpretation of high spatial and temporal resolution images (Tarko *et al*., 2021).

As outlined in Section 2.4, for operational monitoring, a distinction is drawn between a permanent sample and a temporary sample. When monitoring using a permanent sample, the reference labels are updated through time for the same sample, producing regularly updated area estimates. In contrast, monitoring using temporary samples requires a new set of samples and reference labels to produce area estimates for each unique monitoring period. The sampling design considerations of permanent and temporary (or a combination) are explained in Section 2.4. Here, we focus on the process of retrieving reliable reference labels for permanent and temporary samples and on how to update the reference labels for permanent samples for repeated monitoring.

An advantage of permanent samples is greater consistency in the interpretations, with previous interpreters' classification logic embedded in the label assignments. However, the high costs and time-consuming nature of visual interpretations make it a challenge to keep reference data labels up-to-date for regular reporting. Therefore, efficient methods for updating reference labels are essential for operational monitoring.

For a temporary sample, the reference label is collected for each reporting period, independently of past samples from previous reporting periods. However, changes in the number and locations of the sample locations over the reporting periods add a source of variability that may interfere with assessing land use change or the implementation of the definitions of REDD+ activities defined by the country. For example, forest management within plantations and shifting agriculture in natural forest can lead to loss in tree cover, but not in forest land use, as a previously cut forest will regenerate. If the sample units vary over time, the observed changes could be misinterpreted as land use change or deforestation, forest degradation or enhancement of carbon stocks, depending on the timing of the survey (Figure 23).

Figure 23 Characterization of a unit of land subject to conversion from natural forest to a cyclical system.



Notes:
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".
The same unit of land is classified as deforestation (A, B) or enhancement (C) depending on the part of the system cycle covered by the monitoring period.
Source: Author's own elaboration.

The use of temporary sample units could also lead to multiple detections of land use transitions if the timing of assessments happens to coincide with consecutive cycles of a forest plantation or agricultural practices (Figure 23).

When using such an approach, the history of each sample unit in the years prior to the monitoring period needs to be assessed to correctly determine whether a land use change has occurred or not. It also allows effective application of specific change definitions 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-forested lands) that occurred once in an area. This implies that deforested areas that regenerate to again meet the forest definition are not considered a second time in the emission calculation from deforestation. However, if sample units vary among successive surveys, there is the risk of a "double detection", as previously described. To resolve this, maps can be used to mask out areas previously reported as deforested to ensure that no losses have occurred previously.

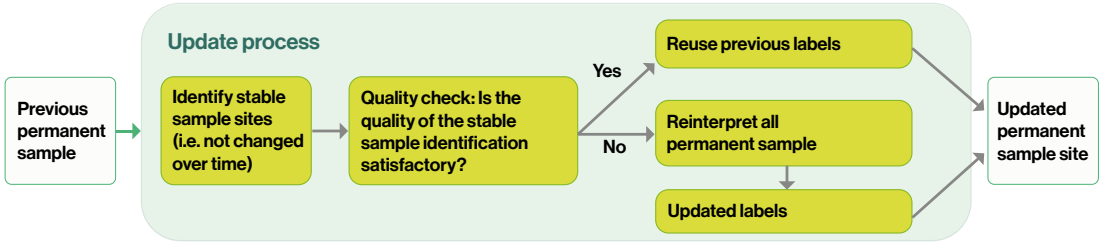
As shown in Section 2.4, it is possible to combine a permanent sample with a temporary sample for each monitoring period. In this case, both updating reference labels for the permanent samples and obtaining historical context (such as historical land-use transitions) for temporary sampling units are relevant.

Overview of existing good practices and potentials

For monitoring updates based on SBAE, new reference data interpretation and collection are unavoidable. Specifically, reference labels for permanent samples must be updated or reinterpreted for the new monitoring period, whereas temporary samples require entirely new reference data collection. However, for permanent samples, it is plausible to consider the reuse of reference labels from a previous monitoring period, as part of the permanent sample. In particular, there can be potentially efficient approaches to keep the no-change or stable permanent sample labels up to date.

For operational validation of annual land cover maps, Tsendbazar *et al*. (2021) implemented partial rechecking of the original permanent reference sample to update the reference labels for new monitoring. The approach aimed to maximize resource efficiency by focusing efforts on sample sites in the original sample with a high likelihood of land cover change while relying on the labels of stable sites, which were hence deemed to be unchanged. To identify unstable sample sites (high possibility of land cover change), they applied the BFAST change detection algorithm (Verbesselt *et al*., 2010) on the full permanent sample. Sample sites with detected breaks were flagged as unstable and then reinterpreted to update the reference dataset. Re-checking a 10 percent random subset of the remaining sample sites was also employed to assess the extent to which the change detection algorithm omitted occurrences of land cover change.

Figure 24 A potential approach to update permanent sample labels for operational monitoring of forest change using sample-based area estimation



Source: Author's own elaboration.

Considering the efforts and time required for manual image interpretation, reusing previous labels of permanent sample units can be an efficient approach for operational updates of forest change monitoring, provided that the permanent sampling unit has not undergone change (Figure 24). The following are potential approaches to identify stable sample units for reuse. A key cautionary note here is to assess the quality of the approach used to identify stable sample units. If the quality of the approach is not sufficient, reinterpretation of the entire permanent sample is necessary.

- Using a combination of change detection algorithms such as BFAST and CCDC. (Verbesselt *et al.*, 2010; Zhu and Woodcock, 2014). Here, stable sample sites without breaks or changes detected by all the algorithms can be flagged as "stable" and reused for monitoring updates. Any other unstable sites require reinterpretation. The confluence of evidence on stable (unchanged) sample units using multiple change detection algorithms will ideally lead to higher confidence in using the previous reference labels, particularly considering that change detection algorithms tend to overestimate actual changes (Xu *et al.*, 2022). A careful quality assessment (QA/QC) of this approach still needs to be conducted. If QA/QC discovers true change within the algorithmically classified "stable" sample units, it is likely that reinterpretation of all permanent sample sites is required. In dry forested areas in particular, many time-series algorithms tend to struggle (Bos *et al*., 2019).

- Using a combination of multiple forest change or land cover products. Identification of stable sample sites according to multiple annual maps for forest change such as the GFC and Tropical Moist Forests (TMF) could also be considered (Hansen *et al*., 2013; Vancutsem *et al*., 2024; Brown *et al*., 2022; Karra, 2021). This can be supplemented by annual land cover maps such as ESRI-LULC and Dynamic World Cover annual aggregates (Brown *et al.*, 2022; Karra, 2021). The confluence of multiple maps indicates a high likelihood that unrevisited reference data have indeed not changed. Hence, their labels are reusable for the update period, provided a thorough quality assessment is conducted. A disadvantage of this approach is the delay in the availability of the annual maps, owing to the time required for map generation.

- Using a custom-made algorithm. Besides existing change detection algorithms and maps, a customized methodology can be developed to identify sample sites that are stable over time. To reuse historical reference labels, Liu *et al*. (2022) assessed the spectral and phenological similarity of the historical sample with the target year. Approaches originally developed to filter out poorly labelled training data could also be investigated to identify sample sites that are not stable and need reinterpretation. Previous research (Stanimirova *et al*., 2023) can be relevant here, as it updated and filtered existing data to create global land cover training data from 1984–2020 using a combination of machine-learning-based cross-validation procedures. These approaches can be adjusted to identify stable sample sites for reuse after quality assessment.

- Leveraging new technologies such as interpretation guided by artificial intelligence (AI). As AI technologies improve, and tools such as temporal vision transformers and other systems can use human-like logic to incorporate context and label high-resolution imagery, opportunities exist to lower costs and improve accuracy. For example, a multi-temporal geospatial foundation model like Prithvi-EO 2.0 (Szwarcman *et al*., 2026) can be used to pre-screen plots based on change category, in a manner similar to the BFAST/CCDC approach described previously. In addition, using vision language models (VLMs) such as SkyScript (Wang *et al*., 2023) to guide human interpreters or to serve as quality assurance can enable improvements in the speed and quality of reference label collection.

These approaches have the potential to identify stable permanent sample sites and reuse the reference labels. This can be an efficient strategy for updating reference labels over time and balancing resource constraints with the need for high-quality data. Note that the approach of reusing reference labels of stable sample sites is a shortcut, as opposed to reinterpreting all sample sites, and it comes with a risk of omitting changes that are identified as stable by change algorithms and maps. Although using the confluence of multiple sources of evidence (maps and algorithms) can increase confidence in stable sample sites, a thorough quality assessment of this approach is required, as the accuracy of change algorithms and maps varies across regions. For example, a certain subset (for example, 10 percent used by Tsendbazar *et al*. [2021]) of the identified stable sample can be reinterpreted to check for errors. If the approach accurately identifies stable sample units without omitting changes, the previous reference labels can be reused for monitoring updates.

Future considerations

Although there are several sources in the literature suggesting methods for detecting unstable sample points or poorly labelled reference samples (Liu *et al*., 2022; Tsendbazar *et al*., 2021), there are currently limited examples of their use for repeated forest monitoring. In addition, regardless of their potential in effectively keeping permanent samples up-to-date, their quality and impact on SBAE require further research.

References

Bos, A.B., De Sy, V., Duchelle, A.E., Herold, M., Martius, C. and Tsendbazar, N.-E. 2019. Global data and tools for local forest cover loss and REDD+ performance assessment: accuracy, uncertainty, complementarity and impact. *International Journal of Applied Earth Observation and Geoinformation*, 80: 295–311. https://doi.org/10.1016/j.jag.2019.04.004

Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R. and Tait, A.M. 2022. Dynamic World, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1): 251. https://doi.org/10.1038/s41597-022-01307-4

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O. and Townshend, J.R.G. 2013. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160): 850–853. https://doi.org/10.1126/science.1244693

Karra, K. 2021. Global land use/land cover with Sentinel-2 and deep learning. *IEEE International Geoscience and Remote Sensing Symposium IGARRS:* 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499

Liu, Z., Zhang, L., Yu, Y., Xi, X., Ren, T., Zhao, Y., Zhu, D. and Zhu, A. 2022. Cross-year reuse of historical samples for crop mapping based on environmental similarity. *Frontiers in Plant Science*, 12: 761148. https://doi.org/10.3389/fpls.2021.761148

Stanimirova, R., Tarrio, K., Turlej, K., McAvoy, K., Stonebrook, S., Hu, K.-T., Arévalo, P., Bullock, E.L., Zhang, Y., Woodcock, C.E., Olofsson, P., Zhu, Z., Barber, C.P., Souza, C.M., Chen, S., Wang, J.A., Mensah, F., Calderón-Loor, M., Hadjikakou, M. and Friedl, M.A. 2023. A global land cover training dataset from 1984 to 2020. *Scientific Data*, 10(1): 879. https://doi.org/10.1038/s41597-023-02798-5

Szwarcman, D., Roy, S., Fraccaro, P., Gislason, T.E., Blumenstiel, B., Ghosal, R., De Oliveira, P.H. *et al.* 2026. Prithvi-EO-2.0: a versatile multitemporal foundation model for Earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*: 64. https://doi.org/10.1109/TGRS.2025.3642610

Tarko, A., Tsendbazar, N.-E., de Bruin, S. and Bregt, A.K. 2021. Producing consistent visually interpreted land cover reference data: learning from feedback. *International Journal of Digital Earth*, 14(1): 52–70. https://doi.org/10.1080/17538947.2020.1729878

Tsendbazar, N., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R. and Duerauer, M. 2021. Towards operational validation of annual global land cover maps. *Remote Sensing of Environment*, 266: 112686. https://doi.org/10.1016/j.rse.2021.112686

Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragão, L.E.O.C. and Nasi, R. 2021. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. *Science Advances*, 7(10): eabe1603. https://doi.org/10.1126/sciadv.abe1603

Verbesselt, J., Hyndman, R., Newnham, G. and Culvenor, D. 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1): 106–115. https://doi.org/10.1016/j.rse.2009.08.014

Wang, Z., Prabha, R., Huang, T., Wu, J. and Rajagopal, R. 2024. SkyScript: a large and semantically diverse vision-language dataset for remote sensing. In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, 645:* 5805–5813. https://doi.org/10.1609/aaai.v38i6.28393

Xu, L., Herold, M., Tsendbazar, N.-E., Masiliūnas, D., Li, L., Lesiv, M., Fritz, S. and Verbesselt, J. 2022. Time series analysis for global land cover change monitoring: a comparison across sensors. *Remote Sensing of Environment*, 271: 112905. https://doi.org/10.1016/j.rse.2022.112905

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


© FAO/Guillermo Carlos


© FAO/Brent Stirton

3.6 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 12 summarizes key considerations and recommendations highlighted in the section.

Table 12 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 LULC class recorded for each sample unit *(Section 3.1)*	Land use and LULC classes recorded at point-level within sample unit *(Section 3.3)*	LULC class proportions recorded for each sample unit *(Section 3.3)*	LULC 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 LULC class recorded for each sample unit *(Section 3.1)*	LULC classes recorded at point-level within sample unit *(Section 3.3)*	LULC class proportions recorded for each sample unit *(Section 3.3)*	LULC 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 LULC classes of interest. Also, small changes in the plot's LULC 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 LULC 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


© FAO/Kallio Enni



Section 4

Quality assurance/ quality control



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 SRS 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



Annexes



Annex I
Sampling strategies

by

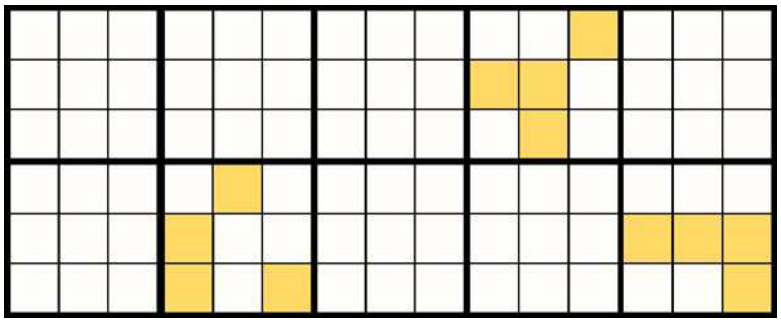
Paul Patterson and Andrew Lister

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 SRS and SYS 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 Section 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 35)

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

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

(Equation 36)

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

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

(Equation 37)

Since $0 \le \hat{P}_i \le 1$, the product $\hat{P}_i(1-\hat{P}_i)$ is less than or equal to 0.25 and hence $\frac{1}{n}\sum_{i=1}^{n}\hat{P}_i(1-\hat{P}_i)$ 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:

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

(Equation 38)

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 39)

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 40)

$V(\hat{P}_h)$ 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 41)

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 42)

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.

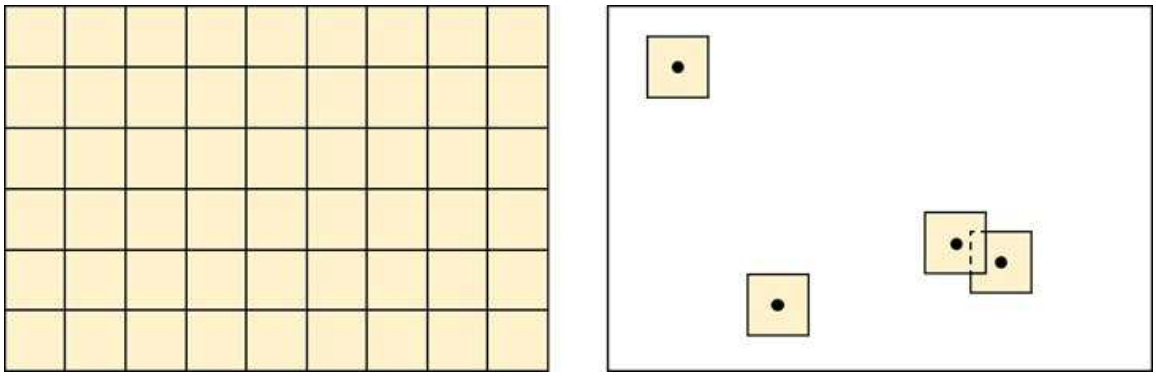
In the two-step strategy, what we have referred to as the "sample units" in the body of the document are defined as support regions for the points. The points are the true sample units in this strategy and the proportion of the attribute of interest within the support region is assigned to the centre 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, $s_1, \ldots, s_i, \ldots, s_n$ 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 $s_{i1}, \ldots, s_{ij}, \ldots, s_{im}$. 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 43)

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:

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

(Equation 44)

$\|R\|$ 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 45)

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 46)

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 47)

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 48)

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, $v\left(\hat{P}_R\right)$, 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):

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

(Equation 49)

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.

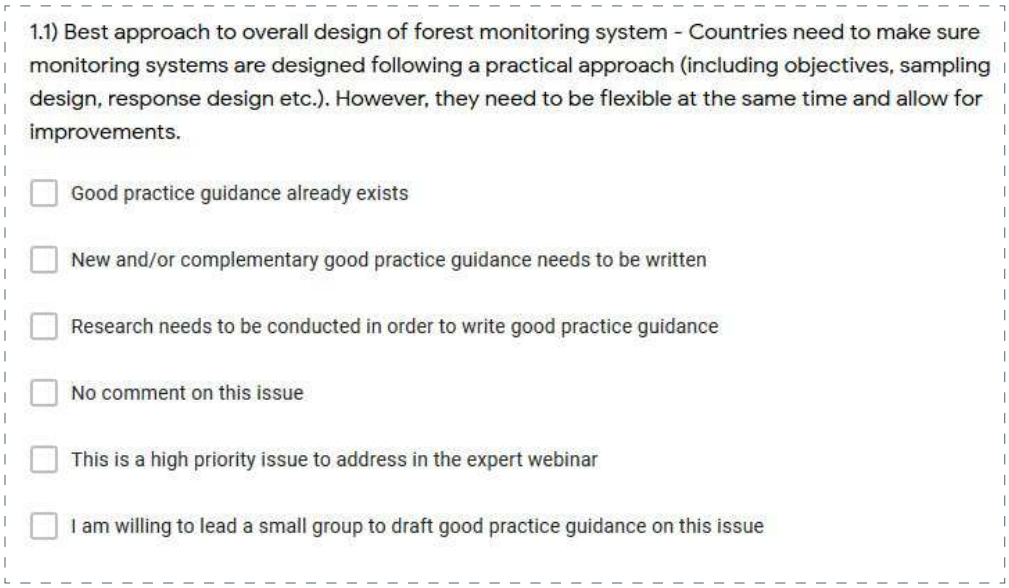
Annex II
Baseline surveys and results

by

Inge Jonckheere

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



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 STR versus SYS? 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 SYS 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 LULC – Some countries use only a land use classification system, others use both LULC, and others use a single mixed LULC system. What are the implications and the pros/cons? When would a country want to interpret both LULC? For example, some countries have expressed interest in characterizing tree cover within pasture and agricultural land, which requires both LULC.
- 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?


© FAO/Ramiro Blázquez González


© FAO/Ramiro Blázquez González


© FAO/Ramiro Blázquez González

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

Estimating the area of land use dynamics, especially those of forests and forest change, is fundamental for both shaping and evaluating climate and conservation policy. It reveals the impact of climate change and human activity, as well as that of mitigation and adaptation actions taken in response. It further underpins the achievement of targets and the market mechanisms that reward management and conservation efforts, as well as associated emission reductions.

Although area estimation is often associated with maps and remote sensing, it is rigorous statistics, grounded in sampling theory, that turn these into defensible products, with validation procedures that are directly tied to the area estimation process itself. As with the first edition, this updated guidance addresses technical and methodological issues that sample-based area estimation (SBAE) raises; it is grounded in country experience, expert consultations and recent scientific advances. In order to reflect on the growing landscape of reporting frameworks, accounting standards and market-based mechanisms, greater weight is given to multipurpose design decisions and how to improve consistency over multiple reporting cycles. It further elucidates optimization strategies that can reduce costs and duration of an assessment, or lead to more precise estimates. Rather than prescribing a single method, this publication focuses on practical recommendations and theoretical options, giving practitioners a firmer basis for area estimates that are accurate, credible and comparable.

For more information, please contact:

Forest Monitoring and Data Platforms team

forest-monitoring@fao.org

https://www.fao.org/forest-monitoring/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 978-92-5-141096-7

9 789251 410967

CE2179EN/1/10.26