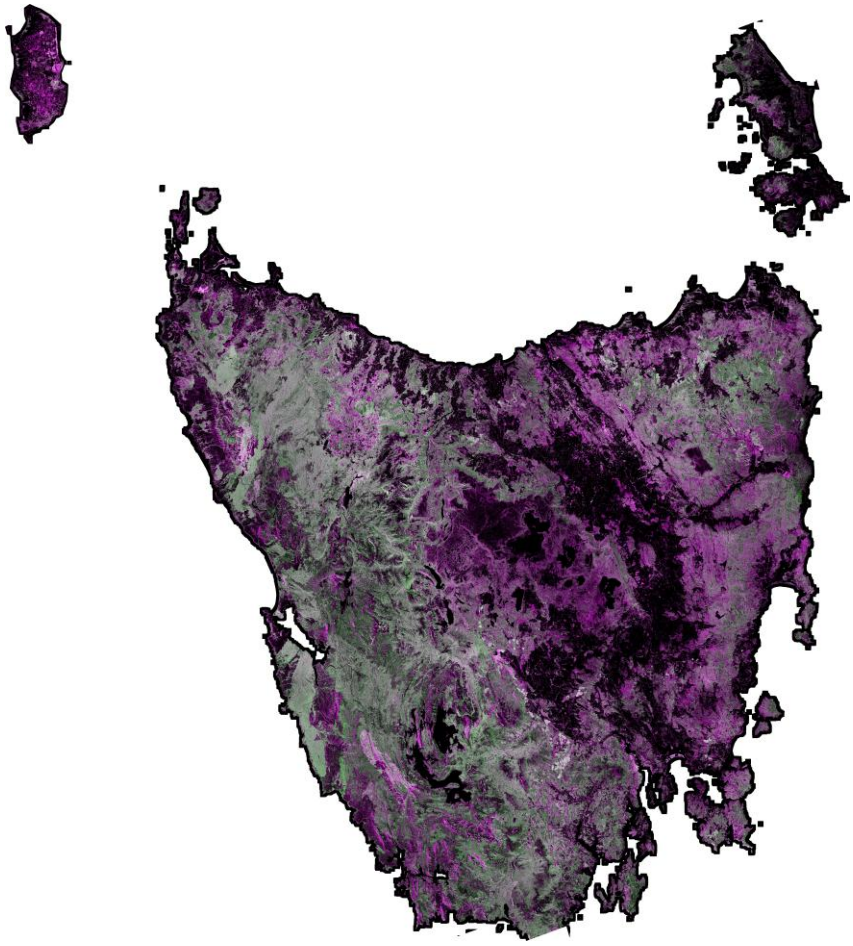
GEO FOREST CARBON TRACKING

Tasmania National Demonstrator

Verification, continuous improvement and reporting methods for national demonstrators



Volume IV

Sept 2012

List of contents

Executive Summary *2*

1 Introduction *3*

Context **3**

Scope **5**

Contents **5**

2 Verification and Reporting Needs *6*

End-users **6**

Verification Approaches **6**

Scaling Issues **10**

Reporting **11**

3 Verification and Continuous Improvement Data Requirements and Sources *13*

Ground-based information **17**

Non-digital Aerial Photography **19**

Digital Imagery **22**

4 Conceptual Framework for Verification and Reporting *25*

Reporting Emphases for Verification vs. Continuous Improvement **25**

Difficulties Associated with Conventional Approaches to Classification Verification **26**

Verification Protocol Principals **28**

4..1 Sampling 28

4..2 Evaluation of Each Sample 34

Spatial Resolution 34

Horizon Level 36

Single Date or Change Maps 38

Evaluation of Verification Sample Units and Analysis 41

Point Samples for Single Date Maps 41

Point Samples for Change/Trend Maps 43

Areal Samples for Single Date Maps 53

Areal Samples for Change/Trend Maps 57

5 Summary *62*

6 Bibliography *65*

Authors

- Kim Lowell, Anthea Mitchell, Tony Milne, Ian Tapley, Cooperative Research Centre for Spatial Information, Melbourne Australia
- Alex Held, Peter Caccetta, Eric Lehmann, Zheng-Shu Zhou, Commonwealth Science and Industrial Research Organisation, Canberra, Australia.

For further information, please contact: Kim Lowell (klowell@crcsi.com.au)

Executive Summary

The GEO Task on Forest Carbon Tracking seeks to demonstrate the feasibility of forest monitoring information generated from coordinated Earth observation as input to future national forest and carbon monitoring systems. This Task has been assigned the highest priority in 2009. The Group on Earth Observations (GEO), in particular to demonstrate to the UNFCCC COP-15 in Copenhagen in December, the value of linking coordinated acquisition of satellite data with standardised processing methods, forest inventory and ecosystem models.

A major part of this task is to provide early proof that satellite data collected by space agencies, using several optical and SAR satellite sensors, can be processed in an interoperable, consistent and repeatable manner, and to provide a set of annual, border-to-border forest change coverages across multiple countries world-wide.

Critical to the process is a capacity to verify the change maps produced. Consistent with the way that the GEO task on Forest Carbon Tracking is being undertaken, a broad set of standardised guidelines for undertaking such verification is required. Implicit in such standards must be a reporting mechanism that allows rapid assessment by the international community of the quality of remote sensing inputs to national carbon accounting systems. An equally important part of the reporting is feedback to the image data analysts in order to enable continuous improvement of the processing and mapping methodology.

This document is an initial step in the dialogue required among international policy makers, agencies that undertake operational image processing for national carbon accounting, and science experts. The ultimate goal is to achieve common principles for verification and continuous improvement of processed digital imagery products and a standardised reporting framework in support of future global measurement approaches of forest carbon in the post-Kyoto framework.

1 Introduction

Context

Urgent coordinated international action and monitoring has been called for by the Intergovernmental Panel on Climate Change (IPCC), which has documented the need for mitigation of global warming driven by anthropogenic greenhouse gas emissions. The IPCC has shown that global carbon emissions could be reduced by as much as 20% by reducing deforestation and forest degradation alone. To achieve this, global agreement on robust and comparable national monitoring, reporting and verification (MRV) systems will be necessary, so that certainty can be provided (i.e., in terms of robustness and consistency) in the various national forest carbon emissions estimates. The Group on Earth Observation (GEO) has risen to the challenge in late 2008 and formally established a new Forest Carbon Tracking Task, to demonstrate that satellite observations of forests can provide such robust and reliable contributions to national forest monitoring programs and reporting schemes.

Therefore the Task has been assigned the highest priority in 2009 by GEO and the Committee on Earth Observation Satellites (CEOS). If successful, the initiative could see global forest monitoring emerge as the first international application of Earth observation outside meteorology to achieve operational status. This would effectively establish an integrated GEO forest carbon monitoring "system of systems" capability, that combines national *in-situ* forest inventory information and wall-to-wall remote sensing and ecosystem carbon modelling that would be subject to an internationally robust verification methodology that also provides for ongoing system improvement.

In brief, Forest Carbon Tracking is seen by many as the biggest opportunity for the promotion of satellite Earth observations for public good applications for the foreseeable future – and with the potential of providing a major success story for GEO in its attempts to implement the Global Earth Observation System of Systems (GEOSS), including through recognition within United Nations (UN) frameworks.

The following key steps are sought as guiding objectives for this task:

1. Achieve "best efforts" commitment from space agencies with data of value for forest carbon tracking purposes to data supply on a continuous basis, in support of an optical and SAR satellite data acquisition strategy agreed to, and established by space agencies.
2. Definition and establishment of "National Demonstrator" sites from the three major global tropical forest regions: South East Asia, Africa and South America - in consultation with national governments, Food and Agriculture (FAO), Non-governmental Organisations (NGOs) and expert teams.
3. GEO Documents identifying the derived satellite-data products to be delivered, and methods for interoperable processing from the various satellite sensor sources. The goal is to be able to provide guidance on globally consistent generation of mid-resolution, wall-to-wall, time-series information products ("maps") on forest-change and derived forest degradation information. This includes common steps such as consistent ortho-

rectification against a common digital elevation model (DEM) – notably including synergistic Optical and Synthetic Aperture Radar (SAR) sensors.

4. Technical frameworks and guidance documents for standard ground measurements at national country levels, to link forest inventory, ecosystem modeling and remote sensing data, as a basis for generation of consistent forest carbon stocks at project and national scales, as required in support of national policy needs.
5. Agreed validation procedures and accuracy assessment for remote sensing of forest area and carbon stock estimates including a standardised reporting framework that provides guidance for continual improvement of the image processing, modelling, and carbon accounting methodologies.
6. Compelling visualisations of progress and demonstrations as inputs to the GEO-VI and COP-15 events, making clear the policy implications of the new technical capabilities.

Achievement of these objectives requires a strong focus on production of global satellite image coverage with a sufficient update frequency (preferably annual) to support the verification of post-Kyoto climate agreements, as well as transparent carbon markets. This will require CEOS and private space industry commitment for suitable and timely data provision. Achieving the required coverage, and facilitating participation of as many contributing space agencies and private satellite operators and missions as possible will require the technical process to include a wide range of both SAR and optical data. Historic, Landsat-class optical data are essential for establishing baselines of deforestation rates in the early 1990s, and can now be augmented with the increasing operational availability of wide-coverage, multi-date SAR data that are a key option for remote sensing observation in tropical areas where cloud cover is prevalent.

It is therefore of paramount importance that CEOS can demonstrate it has the capability to coordinate coverage of satellite data in support of climate policy frameworks, and for this GEO Task on Forest Carbon Tracking to show that interoperability among different kinds of sensors is very practical for achieving consistent and repeatable results – all to provide confidence to the climate negotiators that the space agencies are up to the challenge ahead, and that negotiated agreements can include monitoring, reporting, and verification systems which will be supported by CEOS agency programmes.

The GEO Forest Carbon Task will test the ability and willingness of the space agency coordination efforts by requesting that in 2009 coordinated coverage be achieved for a number of national demonstrator activities – meaning wall-to-wall coverage is required for a country (or large region at least) from each of South America, Africa, Asia and Oceania. Further, intensive data acquisitions will be requested for certain test sites within those countries in support of technical comparison and validation work.

Scope

This document provides a set of principles and conceptual approaches for verification of map products produced by image processing methodologies that will have been accepted by a country for carbon accounting. These principles are established in accordance with fundamentals of a reporting system that satisfies the dual needs of international evaluation of the reliability of country-based carbon accounting and providing for improvement of map products across spatial, temporal, and taxonomic (for landcover) scales. These verification principles are not concerned with the imagery and or processing techniques used to produce the map products as these will have been independently selected by a country. These principles also recognise that suitable verification data will vary in quality and spatial and temporal abundance not only from country to country, but also within a country.

The methods outlined in this document will initially be applied in support of verification and continuous improvement related to the GEO Forest Carbon Task (FCT) 2009 National Demonstrators. These National Demonstrators are intended to show the process for generating quantitative time-series forest carbon products including comparable and interchangeable products from different sensors as well as products derived from combinations of sensors. This document addresses the need for estimates of errors, and attribution of sources of errors, for each product.

Contents

This document lays out a framework for consistently robust and rigorous verification of image-based map products and associated reporting for both international carbon accounting compliance and efficient continuous improvement of image-processing methodologies. **Section 2** covers verification and reporting needs, **Section 3** addresses verification and continuous improvement data requirements and sources, and **Section 4** describes the conceptual framework for undertaking verification and reporting dual purpose results – i.e., international carbon accounting and system improvement.

2 Verification and Reporting Needs

End-users

The ultimate use of results of a verification program is not limited to a single group. The most obvious end-user is the collective of the international community – i.e., countries who have agreed to participate in efforts to reduce greenhouse gas emissions. The United Nations Framework Convention on Climate Change (UNFCCC) requires countries to report on their greenhouse gas emissions and removals. Though countries have flexibility in the methods used to produce information to comply with this requirement, their accounting must be internationally comparable, nationally complete, accurate, transparent, and temporally consistent. Verification is of prime importance for individual countries to demonstrate to the international community that the methods adopted for carbon accounting produce reliable information.

Additional end-users are designers and developers of national carbon accounting systems. For these individuals, verification is a critical part of a continuous feedback loop (Figure 1). In the carbon accounting system designed, validity of outputs is quantified through verification, weaknesses in the carbon accounting system are identified, and this information is used to re-design and improve the system; the process then continues.

There are important implications for the design of a verification program in order to enable use of its outputs for continuous improvement. Whereas international reporting necessitates only an estimate of carbon across an entire country, the need to use the verification program for continuous improvement of image-based system inputs requires a verification program that can target weaknesses in the image-based map products. This means that an ideal verification program will indicate how well a country's image-based map products perform relative to different forest types, landcovers, climatic zones, topography, soil types, and any other factors that may impact the reliability of carbon estimates.

This dichotomy of users is indicative of different types of needs for each group. This is addressed at the end of this section.

Verification Approaches

Conceptually there are two approaches to verification of carbon accounting systems.

The first is through direct measurement such as forest inventory coupled with additional ground-based sampling (for soil organic matter, for example). Direct measurement in essence places a confidence limit on national estimates of carbon stocks through quantification of statistical sampling error. Though this is the most intuitive way to undertake verification, there are a number of associated issues and difficulties.

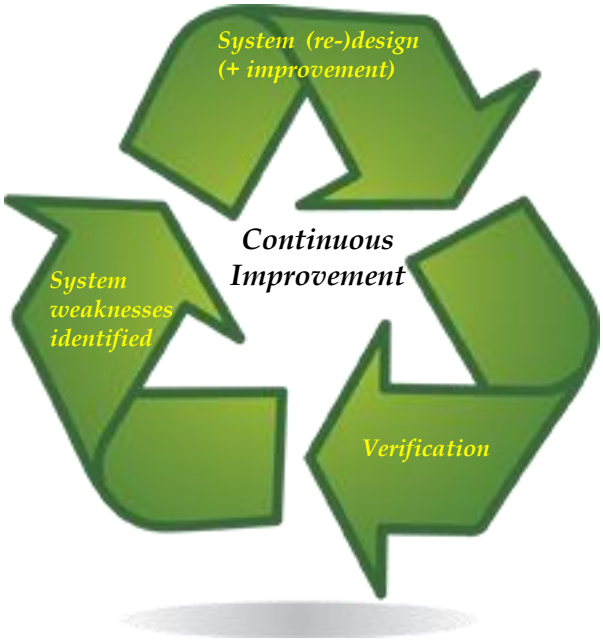


Figure 1. Continuous improvement loop for carbon accounting verification programs.

Though it may be economically feasible for smaller countries to achieve sufficient sampling density to satisfy international verification requirements under a direct approach, it is doubtful that larger countries would be able to do so. Of equal importance is that such an approach is ill-suited to using a verification program for continuous improvement. Given that it is difficult to achieve a sufficient sampling density for national reporting under a direct verification approach, the likelihood is remote of also achieving a sufficient sampling density to allow system evaluation for specific geographic regions, topographic characteristics, or vegetative landcover and soil types.

Another important issue is that "direct measurement" of carbon stocks at individual locations is itself subject to sampling error and variability associated with allometric relationships. For example, the carbon stocks at a forest location can be estimated via biomass studies that sample various parts of a number of trees – i.e., stems, leaves, branches – and then statistically expand the information obtained to produce information considered to be representative of the entire forest. An alternative and less costly option is to measure tree diameter and height and relate this to biomass and carbon using mathematical relationships developed for this purpose. Regardless of the approach, however, generalised numerical relationships are an integral part. Hence "direct measurement" does not actually produce true measurements of carbon stocks as is desired.

The second verification methodology is indirect. In this approach, the inputs to the carbon accounting system are evaluated for reliability and the models, algorithms, and processes used to convert the input data into estimates of carbon stores are examined for individual validity and systemic coherence. Though such an approach may partially rely on direct estimates of outputs (as does the direct measurement approach), but the required density of direct observations is far lower than in a verification program based entirely on direct measurement. Moreover, ground-based measurements may not be required at all. If they are, the type of information taken at individual locations may not be as costly to obtain as that associated with the direct measurement verification methodology.

Verification of system input data through an indirect approach can be done a number of ways that are dependent on the nature of the input data. Generally these will be landscape-wide input data that address the characteristics of the vegetative landcover. Forests are of particular interest because of the relatively high amounts of carbon stored. At the coarsest level, systems for estimating carbon stocks are based on maps containing two classes – forest and non-forest. Such maps can be verified a number of ways – e.g., using aerial photographs, processing independent digital imagery, independent processing of already-used imagery. As systems and technology evolve, however, it is anticipated that input maps will become more refined and will include a measure of forest density and other relevant characteristics of forest structure. The verification of such maps will have to rely on more advanced technologies such as laser/lidar potentially coupled with a greater number of ground-based observations.

Some of the inputs into a system for estimating carbon stocks will be difficult or impossible to verify. For example, the estimation of soil carbon in such systems may depend upon a soil map as input. Soil maps are an artificial construct of the real world that are useful for a variety of purposes. Nonetheless, they are created using a particular taxonomy, spatial scale, and minimum mapping unit (MMU) – i.e., the smallest individual soil polygon that will be mapped. It is quite common in soils mapping to have a taxonomic class such as "Soil C is composed of at least 70% Soil B with not more than 20% inclusions of Soil C." The result is that no two cartographers will produce the same soil map, nor is it possible to compare a soil map against "truth."

An indirect verification approach also requires an evaluation of the system of models, algorithms, and statistical relationships used to convert data inputs into statements of carbon stocks. Though the sophistication of such systems will vary, the principles for verification of models are the same across systems and involve two steps.

First, the allometric relationships of individual components must be evaluated for logical consistency and predictive strength across the geographic and ecophysiographic ranges to which they will be applied. Logical consistency means, for example, having relationships that indicate there is no soil carbon on bare rock, or that forest carbon does ot perpetually increase with increased rainfall in an exponential manner. Consideration of geographic-ecophysiographic ranges means that relationships are equally applicable in tropical and temperate regions, and low lying and alpine areas; this is of limited concern in countries with little climatic and ecophysiographic variability.

Second, the interactions among individual components must have integrity. This means the output of each component is mutually conditioned by outputs of other components. At its simplest this means that an area estimated to have high amounts of woody carbon from forests must necessarily have low amounts of herbaceous carbon. It is notable, however, systems for estimating carbon stocks are likely to be of sufficient complexity that evaluation of interactions among system components is a difficult and time-consuming task.

These last two paragraphs indicate that the indirect verification process associated with evaluation of the system of models, algorithms, and statistical relationships used to estimate carbon stocks will require independent review of the system by domain-specific experts.

Scaling Issues

Scaling issues are most relevant to indirect verification approaches that are expected to be most common in operational verification; these are the focus of this protocol. Scaling issues are of considerable importance given that different countries will employ different approaches for estimating carbon stocks that rely on a myriad of different data. There are three types of scaling that must be considered: spatial, temporal, and taxonomic.

Spatial scale is the one that is most familiar. In essence, if carbon stocks are estimated using spatial units that are 100 ha in size – i.e., 1 km^2 – verification spatial units must be targeted at the same size. This means that a single 0.25 ha forest inventory plot, for example, cannot be used for verification of a 100-ha spatial unit. Similarly, it would be inappropriate to use satellite imagery composed of 5-km pixels, for example, for verification of 100-ha spatial units. Such limitations might nonetheless be overcome by clustering forest inventory plots or developing methods for upscaling forest inventory plots to an appropriate scale.

Temporal scale is also reasonably well understood. In the context of an international verification protocol temporal scale is likely to be quite problematic for large countries, and those that have heavy cloud cover because of the difficulty of getting country-wide imagery with regular periodicity. This is particularly true of systems for estimating carbon stocks that are based on passive (optical) rather than active (radar and lidar) digital imagery. Operationally, carbon stocks are likely to be estimated for a single period using digital imagery that was acquired over a range of dates. Hence the time t_n imagery will not represent a single day, week, month, or even year. Compounding this is the reality that verification data are also almost certain to not represent a single date. Hence it is critical to have agreed temporal standards for imagery used for estimating carbon stocks and undertaking verification, and imagery dates/periods for both must be documented.

Taxonomic scale and its impact on verification is generally the least understood despite the long-term existence of a number of hierarchical classification systems (e.g., Anderson *et al*. 1976). And though taxonomic scale is often not considered in verification, it will become increasingly important as carbon accounting methodologies evolve. Initially, it is anticipated that systems for estimating carbon stocks will use relatively coarse spatial resolution and a relatively long temporal interval – i.e., coarse temporal resolution. In addition, such systems are likely to initially be based on maps of forest/non-forest landcover – i.e.., a coarse taxonomic scale. As the systems evolve, however, it is likely that these classes will be broken into more refined classes. For example, non-forest might be subdivided into urban, water, agricultural with the latter being further broken into different crop or pasture types. The forest class will evolve similarly being divided into species types and combinations as well as height and density classes. This will have profound impact on the verification program as it will require better verification data, sampling schemes, and analytical methods to satisfy the dual purposes of international compliance and continuous improvement.

These three types of scale will not act in isolation and the verification program will have to evolve in a comparable "interconnected" manner. Clearly one of the appeals of using data with a high spatial resolution in carbon stock estimation systems is the possibility of also increasing the taxonomic resolution. And presumably having a high resolution –- spatially, temporally, and taxonomically -- carbon stock estimation system will provide increased system credibility internationally. However, this will only occur if the verification program applied is able to provide information that is also scale-relevant for all three type of resolutions.

Reporting

For the dual purposes of verification – i.e., international acceptance and continuous improvement – reporting needs are somewhat different. Prior work by the Global Observation of Forest and Land Cover Dynamics (GOFC-GOLD) recognised this in suggesting three priorities for map "accuracy[1]" indicators: whole map, individual class, spatial variability of map reliability (Strahler *et al*. 2006).

[1] In this document the term "verification" is favoured over "accuracy." The former suggests comparison with an accepted standard whereas the latter suggests comparison against a measurable truth. Because many land covers including "forest" are subject to definitional issues within and between countries, the position is taken herein that even the best possible map of "forest" and other land cover classes represents an accepted standard and not truth.

International acceptance of carbon estimates requires reporting outcomes of verification programs in a readily understood and standardised format. There are a number of approaches to such verification reporting. Conceptually this might be as simple as providing a confidence interval – e.g., "outputs of the carbon accounting systems of Country X are within 20% of true." While such a statement might be applicable for verification programs based on direct carbon measurements, such statements cannot be readily produced and are of limited value for indirect verification programs. Because it is anticipated that most verification programs will be indirect in nature, it will be necessary to use an alternative reporting methodology such as a grade (e.g., A/Excellent, B/Good, C/Poor, or D/Unacceptable) with the criteria for each grade based on the nature of the carbon accounting system evaluated and the specifics of the verification program employed. Such ordinal reporting systems might even be simplified to binary classes – i.e., Pass/Fail.

Continuous improvement requires a different type of reporting that is more detailed than a simplisitic country-wide verification statement. Continuous improvement requires separate statements of carbon accounting system reliability for each ecophysiographic class. For example, if there is a pronounced rainfall gradient across a country, the process of continuous improvement will be most aided by knowing the validity of carbon estimates for the high-rainfall and low-rainfall zones. If in addition to a rainfall gradient there is also a heterogeneity of vegetative types, continuous improvement will benefit most by having carbon validity estimates for "hardwood" and "softwood" within each rainfall zone. Specific ecophysiographic classes selected for reporting will be dependent on a country's individual characteristics and the nature of the system used for estimating carbon stocks. To be most useful for continuous improvement, however, the selection of classes must enable weaknesses in the carbon accounting system to be readily identified.

Though the previous paragraph applies primarily to outputs of image processing, they also have implications for continuous improvement across an entire carbon accounting system. Critical to whole-of-system continuous improvement is thorough documentation of the carbon accounting system. Poor system performance for a particular ecophysiographic class could be due to, for example, inappropriate region-specific image processing techniques or a paucity of calibration data for a particular relationship. Hence system documentation must include specifics of allometric equations, image processing techniques, density and locations of ground-based plots used for verification, and all other details that may be relevant for pin-pointing weaknesses in the verification program and the carbon accounting system. This includes literature references and citations for any relationships or process-based models incorporated into a carbon accounting system that are generic and process-driven rather than being specifically calibrated for that system.

3 Verification and Continuous Improvement Data Requirements and Sources

It is re-stated that this Verification Protocol document is focussed on the image processing component of indirect system verification – i.e., evaluation of system inputs and structure rather than comparison with "measured truth." Hence the focus of the document from this point on is the verification of digital imagery products.

Standard texts and articles on verification/validation of classified remote sensing images rely on the availability of a data set of "higher quality" than the one being evaluated. Such data can be obtained from a variety of sources – e.g., digital imagery having a higher spatial and/or spectral resolution, aerial photographs, detailed landcover maps, lidar, or forest inventory plots.

A critical point to understand, however, is that ***none of these data sources represent absolute "truth."*** This is because of issues of spatial and taxonomic scale, the nature of verification data, and how the two are analysed. Suppose that one has ground-sampled a number of 0.1 ha circular plots, and information from each plot will be used to verify if co-located image pixel is correctly processed. Figure 2a shows the location of trees on a 1-ha (100 m by 100 m) pixel that is the spatial resolution of a fictitious carbon accounting system for Country X. In that carbon accounting system, this 1-ha pixel is classified as "forest" against a definition of forest as "trees covering 20% or more of the area." Figure 2b shows three possible locations of a 0.1 ha circular plot used to verify the 1-ha pixel and the classification – forest or non-forest – of each. Clearly the position of the plot has a strong impact on how the sample plot is classified and subsequently if the image pixel is judged Correct or Not-correct; the shape of individual plots would have a comparable impact.

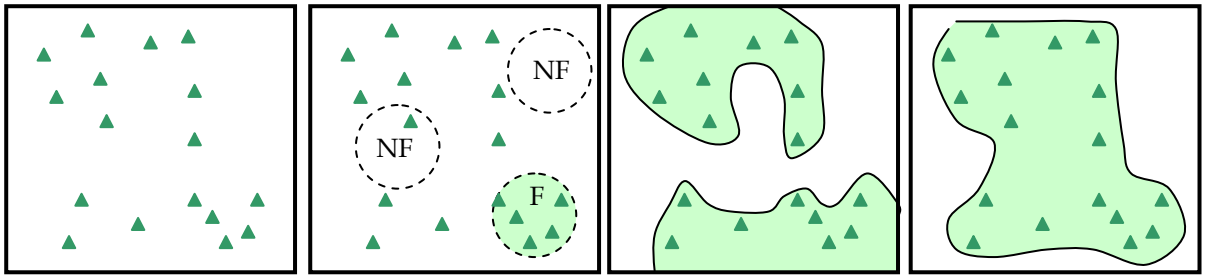


Figure 2. a(left). Fictitious spatial arrangement of trees on a 1 ha (100 m-by-100 m) block. b(middle-left). Potential locations of three 0.1 ha ground-based plots that might be selected to represent the entire 1 ha forest and their forest(F)/non-forest(NF) classification. c & d (two rightmost). Two equally correct subjective interpretations of forest/non-forest.

Alternatively, suppose that ground-based (forest inventory) plots are not available or are judged inappropriate for verification. The alternative selected is to use human interpretation of aerial photographs to identify areas of forest and non-forest or other forest characteristics. Figures 2c and 2d show two alternative interpretations of the 1-ha image pixel based. Recall that the entire pixel had been classified as forest. If one were to extract the centre of the 1-ha pixel from the verification data, 2c would indicate the pixel was incorrectly classified whereas the interpretation in 2d would indicate that it was correctly classified. Alternately, a similar outcome would result from using a rule such as "the majority class on the verification data is considered representative of an entire pixel" – i.e., 2c is less than 50% forest whereas 2d is more than 50% forest.

Note that though the example in the preceding paragraph focuses on the simplest taxonomic division – forest or non-forest – similar principles apply to more detailed taxonomies that will be employed as carbon accounting systems improve. Ultimately systems that employ attributes such as forest density or dominant tree height will have to confront similar issues that are highly dependent on the spatial arrangement of trees and locations and sizes of pixels.

What this means for image processing verification and continuous improvement is that no specific data source can be considered universally the most acceptable as a high quality reference. Instead, the emphasis must be on using validation data that are compatible with the image-based maps produced as input into the carbon accounting system. In addition, for an image product to be considered "valid," there must be an acceptance that the way validation data were used is consistent with what is required for verification and continuous improvement.

What this means generally is that there are a number of sources of data that may be suitable for verification. A critical point is that three image products will be verified for each time period – the vegetative cover at time t_1 and t_2 and the resulting change map. Hence the data employed for verification must be available or obtainable for the temporal span addressed by the carbon accounting system and must provide the capacity to evaluate landcover change. Notably whereas a given data source may be appropriate for evaluating an image-based map at one time or another, its utility in evaluating landcover change may be limited.

There a number of issues common to all data used for verification and their processing and analysis. Among the most important is the issue of geographic and taxonomic "edge" pixels. Even the best possible image registration/rectification cannot avoid problems of geographic edge pixels. Such pixels result from slight misregistrations of pixels from one time period to another (Fig. 3). The resuls is that "pixel A@t_1" will be considered geographically equivalent to "Pixel A@t_2" but also geographically equivalent to "pixel B@t_3" – i.e., an adjacent pixel. If the pixels considered are all truly the same class – i.e., forest – then this misregistration has no impact. However, if the pixels being analysed are not the same type – something that would occur at the edge of forests, for example – spurious change maps result. The impact of this on change map validity is decreased in more homogeneous and less fragmented areas. Similarly, the impact can be decreased using images of finer spatial resolution. However, its impact cannot be eliminated.

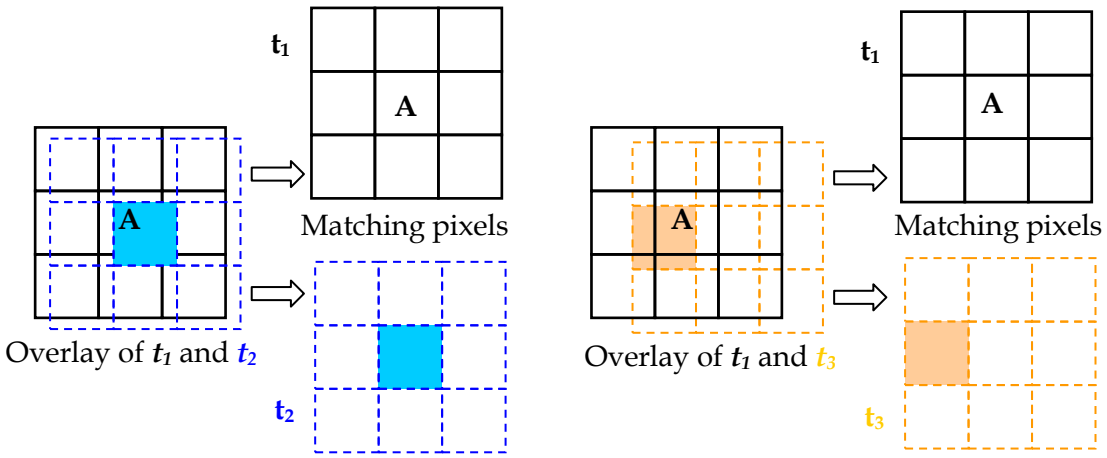


Fig. 3. Illustration of how slight image misregistration errors can cause lack of geographical consistency in pixel matching over multiple time periods.

Taxonomic edge pixels are those that straddle class boundaries. Consider an example in which non-forest is defined as being less than 20% tree cover. A pixel that has exactly 20% tree cover is by definition forest. However, the ability to consistently classify such pixels correctly is limited particularly in light of earlier comments about difficulties associated with consistently mapping forests. Difficulties associated with pixels that are taxonomically close to class edges becomes more of an issue as one moves beyond Horizon 1 data[2] – i.e., distinguishing between "high density" and "medium density" forest. This issue is also related to the concept of the minimum mapping unit (MMU) in which the size (and location) of the pixel/MMU determines if a given pixel is near a taxonomic edge. This effect will be exacerbated in the verification of Horizon 2, 3, and 4 products. For verification, the issue of edge pixels is at the heart of a verification methodology that truly detects inaccuracies rather than reflecting inherent limitations of verification data.

A second issue common to all verification data is data confidentiality. A potentially large amount of data that would be useful for verification is held in proprietary archives. Unless an appropriate data sharing agreement can be achieved, such data are of limited utility for verification. Note that such an agreement requires not only that a verification program can obtain and use these successfully, but an inability to make them available to independent "auditors" may preclude their use.

In the following sections, three general classes of potential verification information and issues associated with the use of each are discussed. In principle, there is no problem inherent in using one type of data to verify t_1 map products, a different type to verify t_2 map products, and a third to verify the change map. However, use of different data potentially complicates the verification program, and individual countries are likely to find it most convenient to use a consistent data source. An exception to this is that it may be found useful to use a mixture of information types to enhance the robustness of verification information and the statistical efficiency of the verification program.

---

[2] The data horizons of interest have been defined as:

i. Horizon 1 – Forest/non-forest map and trends
ii. Horizon 2 – Forest degradation map and trends
iii. Horizon 3 – Land use and forest type map and trends
iv. Horizon 4 – Sparse woody perennial extent map and trends

Ground-based information

Ground-based information will most often come from forest inventories developed for the management of a country's forest resources. The quality, density, and suitability for verification is likely to vary considerably and the size and shape of individual sample units are likely to have an important impact on their appropriateness. There are also likely to be issues of data confidentiality that may limit access to those data.

An overarching consideration is that obtaining such data having a sufficient sampling density to verify the change map will prove difficult. A related issue is that ground-based inventories most often address the amount of carbon/biomass present in areas known to be forest, but they often do not consider substantial amounts of trees growing on predominantly agricultural land that are in shelterbelts. That is, such plots will indicate only areas where forest is present, but do not provide verification of areas where forest is not present. This means that only certain parts of a confusion matrix can be verified (Table 1). If sample points are established in non-forest areas to address this, there is considerable difficulty in designing a sampling scheme that is equally representative for areas that are predominantly forest or predominantly non-forest areas. A final issue of considerable importance in the use of ground-based information is that in a number of countries, available ground-based information only addresses, or is readily available for, public land.

Table 1. Confusion matrix indicating the capacity of ground-based forest inventory samples to address errors of commission(C) and omission (O) for forest.

		Truth	
		Forest	**Non-forest**
Mapped	**Forest**	Verifiable	Non-verifiable(C)
	Non-forest	Verifiable(O)	Non-verifiable

Ground-based samples do have potential value in verification, of course. Perhaps their best use will result from combining them with other types of verification data in highly specialised verification programs. In addition, the value of such data is likely to increase as image-based inputs to carbon accounting systems progress from maps of forest/non-forest (Horizon 1) to maps indicating forest structure (Horizons 2, 3, and 4). Such data also have considerable value in the calibration of carbon accounting systems.

To be of use, ground-based data must have a variety of characteristics. Table 2 lists the most important.

Table 2. Characteristics of useful ground-based plots.

Characteristic	Comment
Accurately located	• Preferably recorded using Global Positioning System (GPS)
Sample units are shape-appropriate	• Thin transects are of limited value. • Forest angle-gauge samples are of limited value. • Circular plots are not ideal but probably are most common.
Sample units are size-appropriate	• Long transects are of limited value • Small plots (relative to map pixel size) are of limited value. • Large plots (relative to map pixel size) can be useful.
Appropriate information collected at each sample unit	• Ground-based information matches taxonomy of map product being verified – e.g., definition of forest/non-forest, density, etc.
Locally representative sampling scheme	• Preferable if samples are not located in consistently optimal areas – e.g., pure forest types, distant from geographic and taxonomic boundaries.
Globally representative sampling scheme	• Density and location of sample units matches needs of verification. • Sampling schemes for ground-based plots that are appropriate for forest management may not be appropriate for verification.

Sources for ground-based data will be varied and may not be uniform across countries. For example, whereas the United States has a national forest inventory system, responsibility for forest inventory in Canada resides with individual provinces. In other countries there is no standardised system of forest inventory for periodic monitoring of woody biomass, and in others standardised government-sanctioned systems are supplemented by private interests. Though gaining access to the latter may be difficult, and the likely non-uniform coverage of areas may be problematic, the potentially high density of plots may ironically make such information of greatest value for verification.

Non-digital Aerial Photography

Some countries will have available archival aerial photographs. Exceptionally, this will entail periodic coverage of an entire country held in a publically available archive. More commonly, however, when such data are available, neither geographic nor temporal coverage will be complete. Moreover, the existence of such data may not even be readily ascertained because they are not located in a central repository.

The Australian situation illustrates potentially common difficulties in obtaining aerial photographs for use in image verification. Australia has no national program for periodically obtaining standardised aerial photography. Australia is a land mass whose population is concentrated in relatively few areas, particularly in the east and near the ocean. Because of these factors, more aerial photographs are available for coastal areas than in the centre of Australia. Moreover, existing and available aerial photographs are not necessarily held in national or state-wide repositories because the photographs may have been obtained by local councils or other organisations. Hence data discovery is a potentially time-consuming task. Nonetheless, for each Australian state there does exist a public repository comprised of aerial photographs that were obtained from a variety of sources. Various agencies would have contracted to have aerial photographs taken for specific projects and then would have archived the photographs in the state repository and licensed the state to sell them. The result is an archive that varies by scale (e.g., 1:80,000 in the interior to 1:8,000 nearer the coast), tone (colour vs. black-and-white), quality, and date.

These difficulties are not highlighted to suggest that aerial photographs have little use in verification. On the contrary, despite the difficulties inherent in obtaining aerial photographs, they can be among the most useful data sources for verification. This will become increasingly true generally, and as national carbon accounting systems develop to be more sophisticated and refined, aerial photographs are likely to prove critical for evaluation of Horizons 2, 3, and 4 data. However, the issues discussed – particularly the lack of complete spatial and temporal coverage -- will mean that achieving satisfactory verification using aerial photographs will require careful consideration of appropriate sampling schemes. The primary considerations for sampling schemes are:

- The geographic coverage available may be concentrated in a relatively small area. This means that verification samples may be positively spatially autocorrelated thereby causing under- or over-estimates of the true accuracy of the carbon accounting system.

- The poor quality or coarse scale of some photography may make it impossible to use certain aerial photographs for anything more than Horizon 1 data products.

- Temporal coverage is likely to be sparse meaning that for all but the smallest countries it will be difficult to do a national verification for a single time period.

- Aerial photographs must be interpreted by human interpreters who make semi-subjective judgements about presence of forest, forest density, and land cover. These judgements are highly impacted by the minimum mapping unit employed in doing the interpretation. As an example, Figure 4 shows a synthetic black-and-white image created from a polygonal map by Edwards and Lowell (1996). This was "interpreted" by nine individuals having similar training and backgrounds. Four of the interpretations are shown in Figure 4 to indicate the magnitude of difference; areal agreement among all nine interpretations was approximately 50%. Figure 5 provides an additional example based on real aerial photographs and interpretations produced by equally trained and experienced professional photo-interpreters. Differences in these interpretations are comparable to what would be expected for Horizons 2, 3, and 4 products as forest stands are discriminated based on dominant species, forest density, and height of dominant trees.

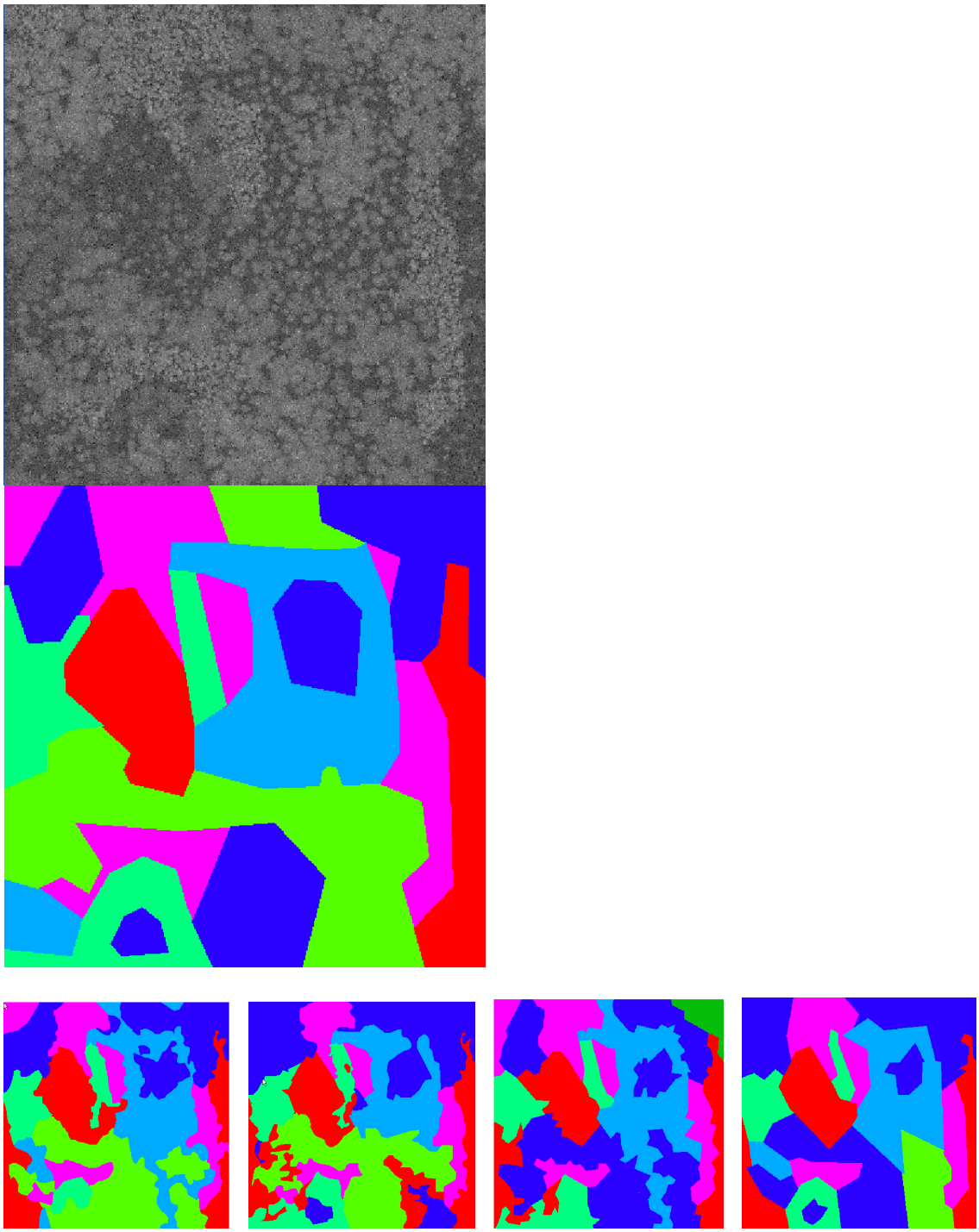


Figure 4. Synthetic image (top left) generated for polygonal map (top right) and four human interpretations of the image.

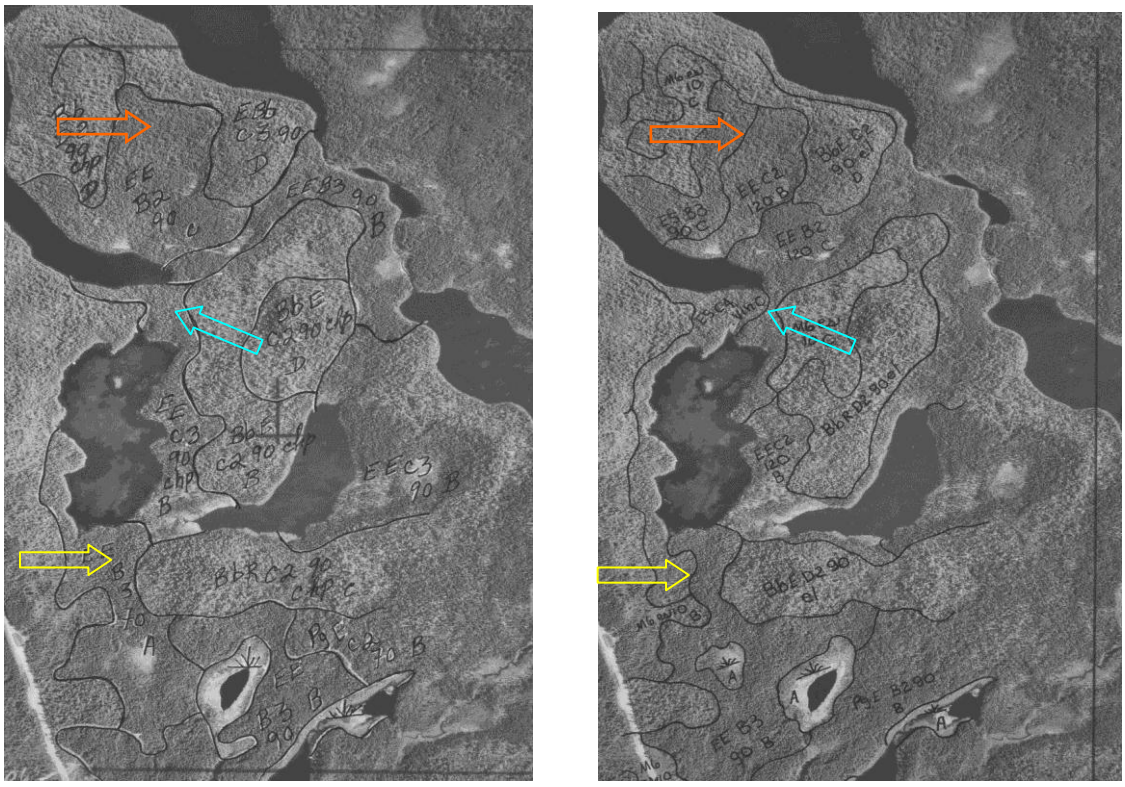


Figure 5. Examples of two different photo-interpretations with some differences highlighted.

Digital Imagery

Many types of digital imagery are available and these will continue to evolve. A variety of imagery can be useful for verification with the utility of any particular imagery being dependent on the carbon accounting system being verified. There are currently three types of imagery that may be of value: optical (passive), radar (active), and laser (active). Potentially useful and available optical and radar data are presented in the protocol "Data Requirements for National Demonstrators." Relevant tables from that document are reproduced here (Tables 3 and 4). In addition to these data, there are other existing and planned optical and radar sources. Though not mentioned in Tables 3 and 4 laser, or its more common term "lidar" ("LIght Detection And Ranging"), is another form of digital imagery that has great potential for verification, particularly of Horizon 2, 3, and 4 data products. Lidar data are currently acquired using airborne sensors and data availability is limited – particularly in widely accessible public archives -- is limited.

Table 3. Potential contributing optical sensors.

Satellite	Spectral Bands	Geometric Resolution	Swath Width	Repeat Cycle
Landsat 5,7	VNIR, SWIR, TIR	30 m / 120 m (TIR)	185 km	16 days
IRS: AWiFS	VNIR, SWIR	56 m	740 km	4 days
IRS: LISS-III	VNIR, SWIR	23 m	140 km	24 days
CBERS 2b: CCD, IRMSS, WFI	VNIR, SWIR	20 m	114 km	26 days
AVNIR-2	VNIR, SWIR	10 m	70 km	46 days
SPOT 4, 5	VNIR, SWIR	20 m / 10 m	60 km	26 days
Kompsat-2	VNIR, SWIR	1 m / 4 m	15 km	28 days

Table 4. Potential contributing radar sensors.

Satellite	Frequency / Polarisation	Geometric Resolution	Swath Width	Repeat Cycle
ALOS PALSAR	L-band (23.6 cm) / full pol	7 m – 154 m	30 – 360 km	46 days
RADARSAT-1	C-band (5.6 cm) / HH	9 m – 100 m	45 - 500 km	24 days
RADARSAT-2	C-band (5.6 cm) / full pol	3 m – 100 m	20 - 500 km	24 days
ENVISAT ASAR	C-band (5.6 cm) / dual pol	1 m – 16 m	5 - 100 km	35 days
TerraSAR-X	X-band (3.1 cm) / full pol	1 m – 16 m	5 - 100 km	11 days
COSMO-SkyMed	C-band (3.1 cm) / full pol	1 m – 100 m	10 - 100 km	16 days

A critical thing to consider when selecting digital imagery for verification is that complete geographic coverage for a single date is usually not required. Whereas complete spatial coverage is required for carbon accounting, verification is based on a program of temporal and geographic sampling. Hence data that are of limited use for establishing carbon accounts may be extremely useful for verification.

The selection of the digital imagery to be used in verification is dependent on the nature of the image data and the carbon accounting system being verified. High resolution optical imagery, and possibly even pre-processed radar, can be used as aerial photography is – i.e., via subjective human interpretation. Both optical and radar data can also be processed by automated techniques for use in verification.

The utility of lidar data depends on the available coverage and data processing as well the map product horizon being verified. Rather than providing information that primarily serves to identify forest types, lidar provides detailed information about tree height and forest density. Hence it is of most use for verifying Horizons 2, 3, and 4 map products. Lidar data are sometimes acquired in spatially disjoint strips or segments that may be of limited value depending on their size. Lidar data also require pre-processing before it is useful; many contractors can provide the necessary pre-processing.

Notably, whether digital imagery or interpreted by humans or through algorithmic techniques, the resulting maps products will have uncertainty that must be addressed in the design of the verification program. As emphasised earlier, digital imagery does not provide "absolute truth" against which carbon accounting image products can be tested for "accuracy." Instead, it must be acknowledged that though verification data are accepted as being better, a number of reasons for disagreements between them and carbon accounting image products exist including errors and limits associated with the verification data. Where relevant, this point is raised throughout this document.

4 Conceptual Framework for Verification and Reporting

Reporting Emphases for Verification vs. Continuous Improvement

As stated earlier, the verification program must be conducted in such a way that two types of reporting can be undertaken.

First, the international community must be satisfied that the true reliability of carbon accounts has been determined and documented. This entails not only adhering to internationally accepted methodological and reporting standards, but also having program transparency. Reporting for international acceptance also implies a statistically robust verification program achieved in part through appropriate sampling strategies that produce unbiased and representative global accuracy estimates.

Second, the verification program must satisfy the internal need of enabling continuous improvement of the image processing methodologies. This necessitates a verification program that identifies landscape types and/or regions where image processing results are "poor" or "good." For example, certain soil types might interfere with the information content of optical imagery on sparse forests, or forest structure and slope characteristics might interact with a radar look-angle in a deleterious manner in certain areas. The needs of the continuous improvement aspect of the program are likely to become more important as carbon accounting systems move from Horizon 1 to Horizon 4 products.

The difference in what is required for verification and continuous improvement is indicative of important differences in reporting requirements. Whereas satisfying the verification requirements of the international community necessitates a methodology with a certain level of statistical rigour, continuous improvement can be achieved through an "indicative approach." Simplistically put, the international community requires a statement such as "the carbon stocks for this country are x units $\pm$ y%." Conversely, continuous improvement can be achieved through a system that can produce statements such as "the image processing methodology appears to perform somewhat poorly in the southern regions."

The example simplistic reporting statements provided indicate another important difference in reporting requirements: spatial scale. Provided that carbon accounting is always undertaken at a national scale, verification reporting is only required over an entire country. This means that for national verification it is acceptable to have an image processing system that overestimates carbon by z% in "the south" and underestimates carbon stocks by z% in "the north." The result will be an unbiased estimate of national carbon stocks (although the "$\pm$ y%" statement will increase as z increases).

Conversely, intra-country statements of reliability are likely to be critical for continuous improvement. One of the most effective ways to target system improvement is to identify where the system performs poorly and then determine the cause – i.e., poor performance in the south, on steep terrain, in old growth forests, etc. It would also be desirable to be able to further target certain soil types and/or topographic conditions, although the verification program may not provide a sampling intensity sufficiently high to do this.

Difficulties Associated with Conventional Approaches to Classification Verification

For carbon accounting, two types of verification must be addressed. The remote sensing community will be familiar with the need to verify an image classification for a single date. And indeed methods for undertaking such verification are well established. However, a verification and continuous improvement program must also enable the explicit verification of land cover change – i.e., multi-temporal verification. This latter presents substantial logistical and statistical problems and indeed some work has been undertaken to address these (e.g., Khorram 1999, Lowell *et al*. 2005, Lowell 2001).

It may initially seem that if one can verify that image products from t_1 and t_2 are "correct" at some level, then a resulting change map is correct at the same level. It may also be assumed that change can be verified through an extension of the confusion matrix approach; the principles of doing this have been described by Strahler *et al*. (2006). Considerable difficulties arise in verifying change maps, however, because of the relative rareness of change and the nature of some change. Moreover, little work exists on validation of continuous maps such as Horizons 2, 3, and 4 products, although some work does exist on verification of continuous maps for single dates (Defries *et al*. 2000).

Classic methods of image evaluation – i.e., confusion matrix approaches – rely on point samples. However, point samples generally cause difficulties in confusion matrix approaches if a class is "rare." When a class is rare, it becomes difficult to obtain a sufficient number of point samples to obtain a reliable estimate of the accuracy of the image classification for that rare class. A seeming solution – having a high sample density for that class – runs the statistical risk of decreasing the independence of samples due to spatial autocorrelation. Moreover, sampling a rare class intensively means that accuracy results for that rare class will have a disproportionate impact on the confusion matrix and the assessment of the global accuracy of the classification.

Another difficulty with a point-based accuracy approach for change maps is the need to detect errors of omission (i.e., unmapped real change) as well as errors of commission (i.e., non-existent mapped change). Notwithstanding statistical difficulties associated with rare classes, evaluating errors of commission for change maps is relatively straightforward conceptually – one examines all areas mapped as change, and evaluates if they are truly change. However, to detect errors of omission, conceptually one must look at all areas where change was not mapped, and assess if there really was no change. Given the rareness of change, this is difficult to do, and it is difficult to obtain reliable estimates of the error(s) of omission.

Difficulties associated with these factors are further compounded by the spatial nature of some change. Though most loss of woody vegetation occurs in large blocks – e.g., forested areas that are cleared -- some deforestation nonetheless occurs in narrow strips – e.g., clearing for roads or power lines. Afforestation similarly can occur in narrow bounds at the edges of abandoned field, for example. Afforestation can also occur in narrow strips as animal farmers plant shelterbelts for livestock protection. If landcover change is linear, it is difficult to map change and to verify its accuracy due to issues associated with edge pixels (Fig. 3). Indeed, algorithms for mapping change must be appropriately formulated to distinguish differences in image registration from true change – e.g., by mapping change for a given location based on more than two time slices.

All of these difficulties will be exacerbated as carbon accounting systems advance from Horizon 1 to Horizon 4. As this occurs, it will be necessary to detect errors of commission and omission for changes in forest density. It will also be necessary to be able to verify the conversion of an open abandoned field to a woody cover type that meets the country-wide definition of "forest." Further complicating this will be the need to do this for anthropogenic as well as natural causes. For forest degradation, certain decreases in forest density will be due to forest management operations such as thinning and will occur in solid blocks. Conversely, natural degradation will occur sparsely across a forest making verification much more difficult, particularly if an approach based on a confusion matrix is employed.

Verification Protocol Principals

4..1 Sampling

Appropriate sampling is the cornerstone of producing verification information that is accepted by the international community. Given the number of factors that must be considered, sampling is by far the most complex aspect of the verification program. And it will become increasingly complex as national systems for carbon accounting move from Horizon 1 to Horizon 4.

The most important aspects of sampling are:

- Single date likelihood of error
- Multi-temporal likelihood of error
- Image registration
- Land cover type
- Regional differences

It will be incumbent on each country undertaking verification that the sampling scheme adopted respects each of these issues to produce information that provides true estimates of the validity of a carbon accounting system. Consequently, each is discussed in a bit more detail.

Single date likelihood of error is related to doing an appropriate classification for raw imagery representing a single date/time period. Horizon 1 classifications will contain two classes – forest and non-forest – whereas Horizon 4 products may be composed of hundreds of classes representing forest type and density classes as well as agriculture and other land cover classes. The classification error associated with a single date will be somewhat related to raw image/data quality and image pre-processing such as radiometric correction. In most cases, most of the classification inaccuracies will be related to similarities among classes and the correct application of suitable image processing techniques. For example, a definition of forest of "20% canopy cover" will cause difficulty in correctly classifying forest that is "19%" or "21%" – i.e., pixels that are close to a taxonomic edge. And indeed, it is important to recognise that the concept of "forest" can only be mapped relative to a specified minimum mapping unit and spatial arrangement of trees as indicated in Figure 2. Given these issues, an appropriate sample for a single date for Horizon 1 products would place an appropriate number of samples in the 2-by-2 predominantly forest/non-forest – high certainty/low certainty matrix (Figure 6). Such a sampling design assumes that the designer of the verification program has a priori knowledge about areas that may be difficult to classify.

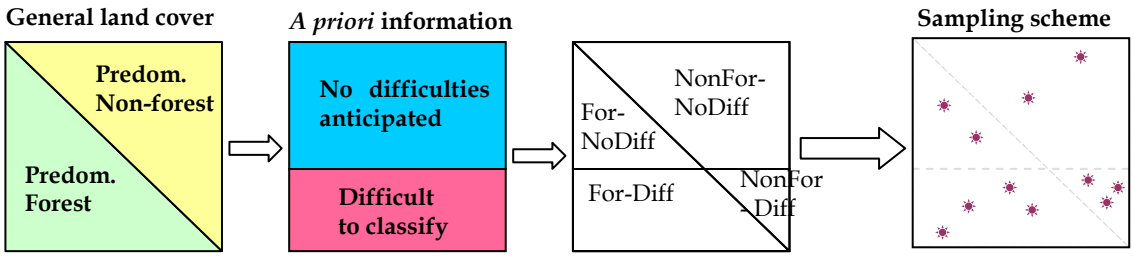


Figure 6. Example of verification sampling scheme that places most emphasis on areas expected to be difficult to classify.

Multi-temporal likelihood of error is related to a locally correct classification at times t_1 and t_2. A global approach for assessing classification correctness for a single date provides a non-localised estimate of how correct is the single date classification. However, the rareness of land cover change, the need to detect afforestation and deforestation, and the spatial nature of both types of change alter the dynamics of classification verification. Simply stated it is not necessarily true that if the classification for each date is "correct" or "valid" as determined by a global evaluation then the change map will be equally valid. The relative rareness of change requires that a change map be evaluated locally and that both errors of omission (undetected change) and commission (erroneously detected change) be addressed. Ideally this means examining every pixel of a change map and recording it in a three-by-three no change/deforestation/afforestation matrix (Table 5). However, because it is logistically impossible, the sampling scheme must be able to address it. A temptation in designing an appropriate sampling scheme may be to avoid edge areas such as forest/non-forest for Horizon 1. While it is true that some change detected at such points will be related to image co-registration, it is also true that some forest types re-establish by creeping outward from an existing seed source. And while deforestation that will be of interest for Horizon 1 products occurs in lines or large blocks, forest degradation that is of interest for Horizon 2 and higher products will not occur in geometrically regular shapes.

Some of these problems will be mitigated if more information is used to design the overall sampling scheme than only the classified t_1 and t_2 maps. This might involve examining the probabilities associated with the classes on the t_1 and t_2 maps to develop the sampling scheme. Fig. 7 provides a four-pixel example in which the sampling scheme would concentrate on those pixels with the highest likelihood of deforestation or afforestation – the southwest pixel with a change probability of 50. Such information for an entire image allows one to develop a frequency diagram of the probability of change to enable appropriate sampling of pixels with a high likelihood of deforestation or afforestation.

Using information from longer time periods than times t_1 and t_2 to develop the change map also will potentially simplify development of the verification program and sampling scheme. In Fig. 8, consideration of only t_1 and t_2 probabilities or binary classification leads to the conclusion that the two western pixels change from one period to another. However, the probability sequence $t_1 \rightarrow t_2 \rightarrow t_3$ suggests that the change is likely to be an artefact of image co-registration, or a mixed-pixel footprint, or other anomaly. For example, the northwest pixel has a low "p of forest" for t_1 and t_3 suggesting that the high "p of forest" at t_2 is an anomaly. For verification purposes, the naïve classification would suggest that the two yellow pixels are candidates for sampling whereas the multi-date classification eliminates them as potential anomalies.

Table 5. Example of a confusion matrix for Horizon 1 change products that must be adequately populated by an appropriate sampling scheme.

		Reference Information		
		No Change	**Afforestation**	**Deforestation**
Mapped Information	**No Change**	Correct	Change Omission	Change Omission
	Afforestation	Change Commission	Correct	Change Misclassification
	Deforestation	Change Commission	Change Misclassification	Correct

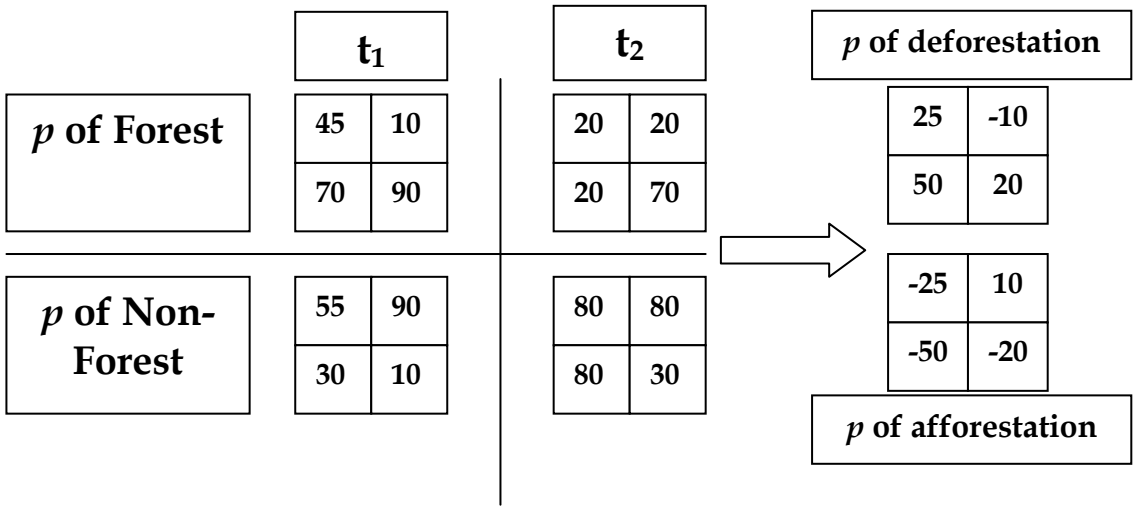


Figure 7. Pre-classification probabilities for Horizon 1 products (Forest/Non-forest maps) that lead to probability of change (right). Negative "*p* of deforestation" values indicate an increased likelihood of regeneration whereas negative "*p* of regeneration" values indicate an increased likelihood of deforestation.

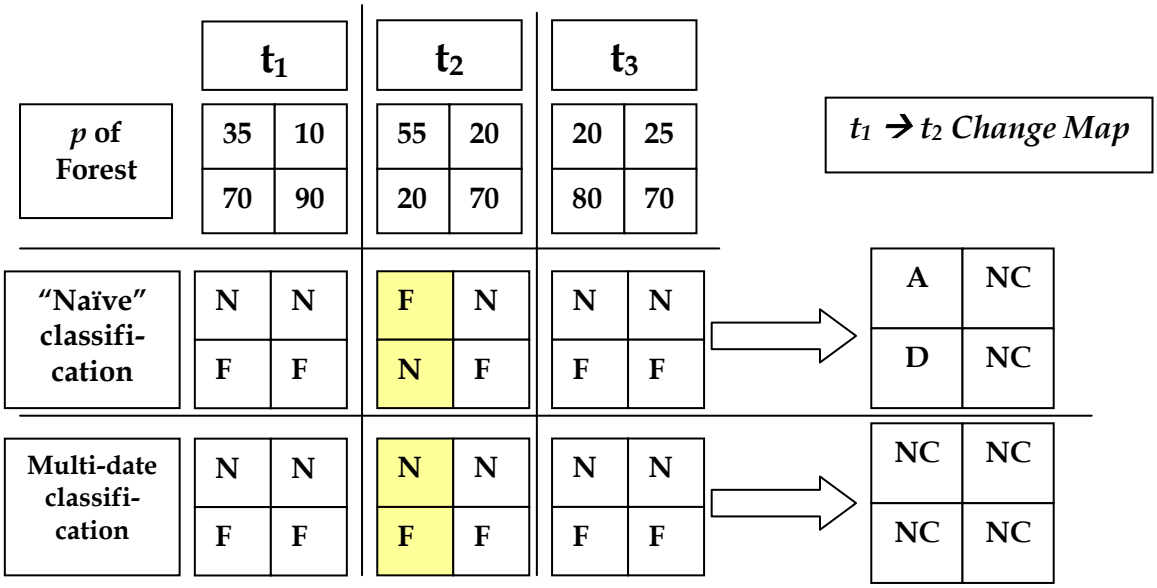


Figure 8. Example of Horizon 1 classifications and change maps based only on class probabilities at two dates (the "naïve classification") and over a longer time sequence ("multi-date classification"). The pixels in yellow are those that change depending on methodology. (NC – No Change; A – Afforestation; D – Deforestation.)

As suggested earlier, ***image registration*** can cause problems that may affect the sampling scheme by virtue of mapping spurious areas of deforestation or afforestation. Yet locally imprecise image co-registration from one time period to the next is inevitable. For Horizon 1 products, the impacts are likely to be most pronounced for deforestation at forest/non-forest boundaries. However, the validity of afforestation in such areas will depend on forest type and the local presence of cleared areas and agricultural fields that are being permitted to convert to forest naturally. And as indicated in Figure 8, the methodology used to produce change maps can have a strong effect on the number of spurious change pixels related to image registration.

Detecting spurious change pixels due to image registration for Horizon 2 to 4 products will present a greater challenge. Whereas it will be possible to detect spurious change at edges of forests, for example, detecting (and verifying) spurious change of forest density is likely to be considerably more difficult. Though this issue should be diminished by developing change maps based on multi-date comparisons, the sampling scheme developed for any verification program must account for the effects of image (mis-registration).

The verification sampling scheme must also be able to address ***land cover type***. This is not as simple as having a Horizon 1 verification sample that, for example, locates sample points in forest and non-forest in direct proportion to the total of each class in an area or entire country. Instead, the verification sample must also consider the underlying land use types. This means that even though, for example, Horizon 1 products are being considered, the major forest and agricultural types – e.g., wheat and barley, rain forest and temperate highland forest -- in the area or country being verified must be represented in the verification sample. This will considerably complicate sample design when Horizon 2, 3, and 4 products are eventually produced as this will require further stratification.

Regional differences will have a similar effect on sampling as land cover type. That is, all of the regions in a country that are considered "distinct" must be appropriately represented in the verification sample. Areas might be judged "distinct" based on topography, aridness, or political dynamics. Similarly, "distinctiveness" might be assessed based on soil characteristics that could impact the sensor employed to capture an image. Local knowledge of regional differences will be critical for addressing this.

This section has been developed based on a conceptual verification model in which:

1. Maps derived from algorithmic treatment of digital images are produced for the appropriate data horizon for two or more dates

2. Maps of change are derived from those maps

3. Samples are located on the three maps – i.e., t_1, t_2, and the change map

4. For each sample location, sample units of a size appropriate for the spatial resolution of the digital imagery and maps are compared against a data set that is accepted as being of higher quality than the maps. This may mean sampling individual pixels or areas comprised of multiple pixels.

5. Results across all samples are compiled in a way that is meaningful for internationally accepted country-wide verification and continuous improvement of the mapping process.

In doing this, it should be accepted that potentially separate sampling schemes must be developed for each of the three maps. In particular, this means that evaluation will not be limited to the two maps developed for t_1 and t_2 under an implicit assumption that if both are considered to be of acceptable accuracy that it automatically follows that the change map will be acceptable. Instead, the change map must be evaluated explicitly.

This also highlights that the sampling scheme for the three map products may be completely independent. Although sampling schemes for all three would likely reflect similar considerations of, for example, regional and landcover differences, sampling for the change map in particular must reflect the unique characteristics of change such as rareness and spatial configuration.

4..2 Evaluation of Each Sample

There are essentially three aspects to verification that must be addressed coherently:

- Sample design – Determination of the number of samples and the way they are to be allocated geographically and to different areas or classes. (This was addressed in the previous section.)

- Sample unit specifics – Establishment of the size and type of sample units and information to be collected for each. (This has been termed by others "Response Design" (Stehman and Czaplewski 1998).)

- Data analysis – How the sample design and sample unit information will combine to provide internationally accepted reports and information for internal continuous improvement.

Individual sample units must be designed and evaluated relative to overall carbon accounting system design and related specifications including spatial resolution and Horizon level. Another important consideration is if one is evaluating a single date map or assessing a change map. These will in turn impact the type of data that is accepted as reference data.

Spatial Resolution

Harmonising the spatial resolution of the carbon accounting system spatial units and the verification sample units must be a prime consideration. It makes little sense, for example, to evaluate a forest/non-forest map composed of 250 m square pixels (6.25 ha) using ground-based 0.2 ha circular plots (diameter: 50.5 m). However, it might make sense to evaluate such a map based on, for example, 1 ha strip/transect plots that are 50 m wide and 200 m long. Of course, the use of the latter may mean that one has to address using a strip plot that crosses pixel boundaries to represent one or more pixels.

This issue is not only relevant to the potential use of ground data in verification programs. Similar issues are manifest if one is using high resolution[3] aerial photography or digital imagery. That is, it makes little sense to verify a 250 m square pixel based on a single location on an aerial photograph or a single pixel from a finer resolution digital image.

It is recommended that if single entities – pixels or ground-based plots – are used for image verification, they should be no smaller than one-third, and no larger than 1.33, of the pixel size of the operational carbon accounting map. In addition, the shape of the verification sample must be concordant with the operational pixels. An acceptable alternative to these guidelines would be the use of multiple verification data types to evaluate a single operational pixel – something that may be required for the evaluation of Horizon 2, 3, and 4 products. In such an alternative, a series or cluster of ground-based plots or points/pixels would be located in the larger pixel being evaluated and accepted as being representative of the larger pixels.

An exception to this would be if the resolution of the operational pixels is such that individual ground-based samples, aerial photograph points, or digital image pixels can reasonably be considered representative of the operational pixel. For example, a circular 0.2 ha ground-based verification plot would be considered representative of an operational 30 m square pixel derived from TM imagery. A single point on an aerial photograph would be considered similarly representative provided the scale of the photography is "reasonably close" to that of the operational map.

These guidelines mean that individual ground-based samples are likely to be of limited value in verification. This is particularly true for Horizon 1 products, although clusters of such plots might prove extremely useful for evaluation of higher Horizon products. And if the ground-based samples are established in a cluster design or in relatively large transects, their value for verification is likely to increase.

Another limitation of ground-based samples for verification relates to the reality that such plots are most often established as part of operational forest inventories. Thus the sampling scheme generally locates such plots in forested areas only. Hence such plots provide a means for evaluating errors of omission – i.e., forest that was mapped as non-forest – but not errors of omission – i.e., non-forest that was mapped as forest.

---

[3] "High resolution" in this case refers to imagery used for verification that has a substantially finer resolution – i.e., smaller pixels -- than the imagery used to produce the maps that are the basis of the carbon accounting system.

Horizon Level

To date, most operational work on the use of remote sensing in carbon accounting has focussed on Horizon 1 data – i.e., forest/non-forest map. Similarly, much verification work has focussed on Horizon 1 because image evaluation techniques have evolved primarily to address categorical maps. As carbon accounting systems become more sophisticated and move towards Horizon 2 to 4 data, the complexity of sampling designs will increase, and the design of individual sample units and their analysis will similarly need to become more sophisticated.

Verification information collected to evaluate Horizon 1 products is relatively straight-forward: one requires a reliable statement of whether or not an individual pixel is forest or non-forest. This may be satisfactorily derived from ground-based observations (provided the evaluation area is comparable to the size and shape of pixels being evaluated), algorithmic analysis of digital imagery, and/or subjective analysis of high resolution imagery.

Verification of Horizon 2, 3, and 4 products requires evaluation of forest degradation (and trends) on an interval data scale which also implicitly requires verifying areas mapped as afforestation. This implies the need for a refined verification measure of forest density including regenerating areas that must address an area the size of an operational pixel.

It is recommended that for imagery composed of relatively small pixels -- i.e., 10 to 50 m or 0.01 to 0.025 ha -- individual ground-based forest sample plots be accepted for evaluating an individual pixel. This is, or course, subject to a number of considerations such as the shape of the plot – e.g., circular, square, or linear, the ability to locate the plot accurately in the field and on the operational map, and the size of the plot being in harmony with the pixel size. If operational pixels are smaller than 10 m, ground-based forest plots are likely to be useful only if contiguous operational pixels are aggregated to a size comparable to the verification pixel size. Ground-based plots might also be useful for larger plots provided the ground-based samples are established using a cluster sampling design that places multiple ground-based sample plots in a relatively small area.

An acceptable alternative to ground-based forest sample plots for verification of Horizon 2 and above products is human interpretation of high resolution imagery including aerial photographs. Country-wide operational Horizon 2 products will presumably be developed through algorithmic processing of digital imagery. Human interpretation of aerial photographs will provide an independent data source for verification. However, such information derived from aerial photographs is known to be highly variable from one interpreter to another (see Figures 4 and 5), and it is also based largely on the forest overstory – i.e., it may ignore a considerable amount of woody matter in the understory. Hence such verification information should not be viewed as "truth" but simply as an appropriate basis for comparison and allowance must be made for differences in interpretation. This is unlikely to be an issue for continuous improvement because a mismatch between the Horizon 2 operational product and aerial photography interpretation as a basis for verification will be able to indicate areas where further study is required. For international compliance, achieving acceptance that forest density information obtained from human interpretation of aerial photographs must be viewed with some scepticism is likely to be problematic due to the long-term use and acceptance of such information by a range of land management professionals.

Algorithmic processing of digital imagery might also be acceptable for verification, provided the algorithms employed can be considered independent of those used to generate the Horizon 2 products. "Independent" is clearly a subjective term, but it may suffice to have an independent operator generate Horizon 2 products using the original algorithm. This is acceptable because all image processing algorithms require human input that varies from one individual to another. For example, unsupervised classification algorithms require input concerning the number of classes sought, convergence thresholds, etc. and resulting classes are usually assigned semi-subjectively to the ultimate classes in a categorical taxonomy. In contrast to having a different operator apply the same algorithm/process to the same imagery used to generate the operational product, it is preferable to have an independent operator apply a different algorithm/process to different imagery. Regardless of the approach, however, is an underlying assumption to the use of algorithmic processing of imagery for verification: if two independent operators achieve the same result, then the final product is valid both are correct. It is recommended that this assumption be accepted as a principle of verification.

Eventually, carbon accounting systems will move past Horizon 2 products to Horizons 3 and 4 – land use and forest type map and trends, and sparse woody perennial extent map and trends, respectively. Sample unit considerations for these products are the same as for Horizon 2 products. The difference is that the sample scheme will have to be adjusted to address map taxonomies that further divide the operational map product from non-forest/{forest by density} to {non-forest by land use}/{forest by density and type} (Horizon 3) to {non-forest by land use}/{forest by density and type}/{sparse woody perennials}.

Single Date or Change Maps

The characteristics of sample units and the way that each unit is evaluated is highly dependent on whether a single date map or a change map is being verified.

Single date maps can broadly be evaluated by obtaining data that are relevant to the horizon of the map being verified on a sample unit that is spatially and temporally representative of one or more pixels on the map. Obtaining data that are temporally relevant can be an issue regardless of the type of data employed for verification. It will usually be the case that the date of ground-based plots, aerial photography, and digital imagery will not have been collected within one year of the date of the map being evaluated. Moreover, a "single date" map is likely to be a mosaic of images from different dates, particularly for larger countries. The impact of this on verification will be minimal provided that the amount of change in the area being verified is relatively low. Nonetheless, evaluation of individual sample units must be done using subjective but reasoned judgement. For example, a verification sample unit established in an area that appears to have been recently deforested, yet that is located on the single date map in an area of "high density forest" requires further examination. If further study indicates that it is clearly an error – i.e., the area in question was not cleared between the date of map and the verification data – then the sample is recorded as being erroneous. However, if it appears that the supposed error is due to deforestation that occurred between map and photograph dates, the sample should be discarded from consideration and the sample design adjusted to accommodate this.

The evaluation of individual samples for change or "trend" maps is dependent on a number of factors. An ideal situation is that spatially exhaustive verification data are available for both dates represented by the t_1 and t_2 maps. In such cases, the state/class of each verification sample unit is noted for t_1 and t_2 and then compared against the state/class of the relevant pixel(s) for the same time periods. Usually, however, there will not be such data and it will be difficult assess change at individual points. There are effectively two options to address this.

One option is to define areal sample units and to evaluate them using multi-temporal imagery as was done by Lowell (2001). The imagery used is most likely to be satellite imagery, although it is conceivable that in some situations suitable aerial photographs would be available. The imagery employed does not have to be the same imagery as was used to develop the t_1, t_2, and change maps themselves. Sample units would need to be large enough to be able to detect change. This means that it may be possible to reduce the size of such sample units as one moves from Horizon 1 to Horizon 4 products – i.e., as more subtle changes are considered – although it is likely that the sampling scheme would have to be adjusted to increase the number of samples as sample unit size decreases.

To use areal units, an independent operator processes the (independent) imagery from t_1 and t_2 to the same horizon as the change map, produces a change map, co-registers the independent imagery to the change map, "clips out" the sample areas from both, and then tabulates the areas by the relevant change class (Figure 9). For Horizon 1 products this means recording the amount or percentage of no change, afforestation, and deforestation for each areal sample unit. For Horizon 2, it means further tabulating information for non-forest and forest density for all pixels with an estimated density greater than 0%. Horizons 3 and 4 add a further stratification of land-use and forest type. Once the information from both change maps – operational and verification -- is available for each sample unit, analysis is undertaken to provide verification information for the entire carbon accounting system to the international community and for continuous improvement. Analytical approaches for this information are discussed in a subsequent section.

The other alternative is to use more familiar point samples. The number and placement of these would be determined in the sample scheme design. In the rare case that verification imagery is available for the periods t_1 and t_2, then the information collected for each point is "change" or "no change" based on the state/class at each time. However, other types of verification information – e.g., forest inventory plots, aerial photographs -- will almost always only be available for a single daate. The verification information collected for each point must accord with horizon of the product being evaluated. For Horizon 1 products this means recording forest/non-forest for each point, for Horizon 2 non-forest/forest density, etc. The way that these are analysed to undertake verification and continuous improvement is described in a subsequent section.

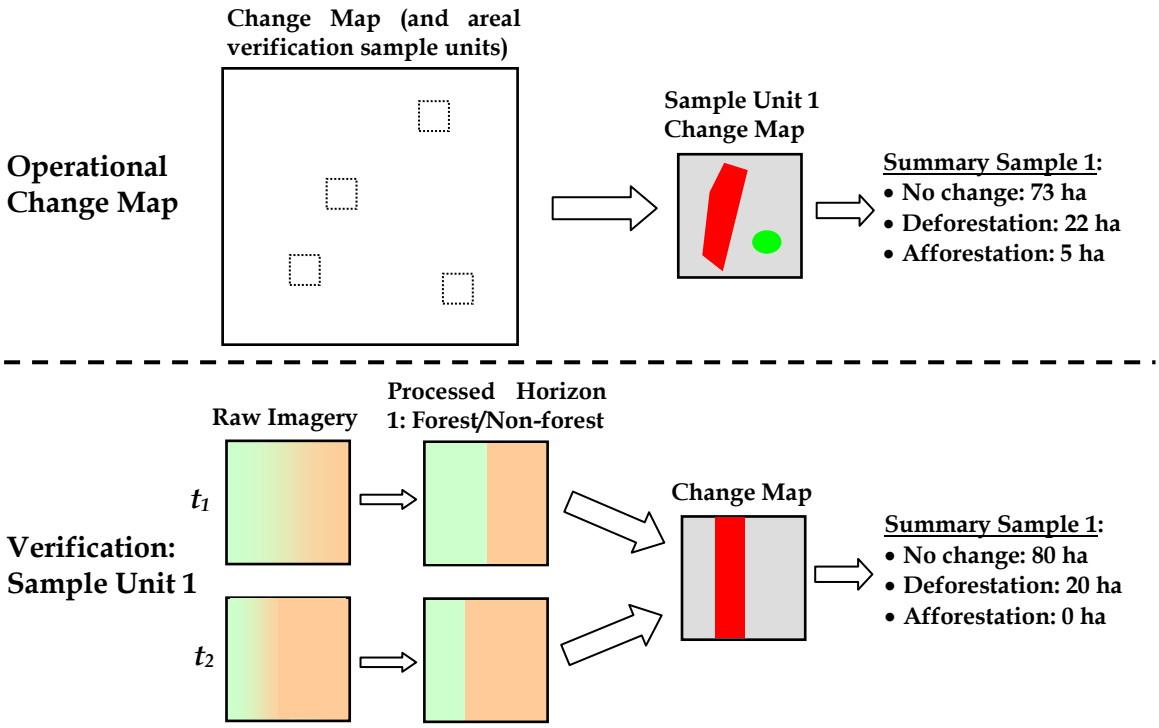


Figure 9. Example of evaluation process for areal verification sample units for Horizon 1 products.

Regardless of the sample units employed, it is re-emphasised that the collection of verification data for each sample unit will require certain subjective yet rational decisions concerning "spatial representativeness." As an example, consider a case where the validation point samples are not the same size as the pixels on the operational map. Though it is only necessary to assign a statement of "forest/non-forest" to each verification point sample for Horizon 1 products, the final assignment must be representative of the entire pixel. If the pixel is considerably larger than the verification unit size, the verification sample size must be expanded if possible – e.g., if using aerial photographs – or it must be accepted that the size mismatch will inflate the global error reported for the carbon accounting system being evaluated. If the verification unit is larger than an individual pixel, rules for considering neighbouring pixels will have to be developed, reviewed, and ratified.

Evaluation of Verification Sample Units and Analysis

The nature of the sample unit employed and the information collected for each will determine how each will be assessed for "correctness" and how all sample units will be analysed collectively to provide a statement of the validity of remote sensing products validity for international compliance and continuous improvement.

Point Samples for Single Date Maps

Point samples for Horizon 1 products at a single date are the most straightforward to address. This is because the well-documented confusion matrix approach (e.g., Table 6) to image verification is designed for such situations. Hence for both Horizon 1 single date and change/trend products, a confusion matrix approach is recommended. Such an approach nonetheless requires a decision about "temporal representativeness" of operational remote sensing products and verification data. That is, one must determine which date the verification data represent since the original imagery is likely to be a compilation of dates, and the verification data are unlikely to match any of the image dates exactly.

The analysis of point samples for Horizon 2, 3, and 4 products (and for change/trend products) are more problematic than Horizon 1 products. While verification of single date products can conceptually be addressed by a confusion matrix approach regardless of the product horizon, as one moves from Horizon 1 to 4, the number of classes increases. Consider the (arbitrary) classification of forest density into four classes for Horizon 2 products – e.g., 1 to 25% forest density, 26% to 50%, etc. For a single date this will result in five classes – non-forest plus four forest density classes – and a 5-by-5 confusion matrix. This is still quite manageable with a confusion matrix approach, although it does complicate the sampling scheme considerably. However, if 10 forest density classes are used (1 to 10%, 11 to 20%, etc.), the confusion matrix is 11-by-11. Similarly, if four non-forest land use classes, four forest types, and four forest density classes are included for Horizon 3 products, one has 20 classes – four non-forest and 4-density-by-4-forest-type classes – yielding a 20-by-20 confusion matrix. Horizon 4 adds yet more complexity. And the situation is exacerbated if one is considering change/trend maps.

Hence for verification of Horizon 2, 3, and 4 products a mixed approach is recommended when verification data are point samples. The nature of the validation data is that if a verification point falls in a forest, information about forest density (Horizon 2), forest type (Horizon 3), and sparseness of forest (Horizon 4) will be recorded for each. If the point sample falls on agricultural land, it will be recorded as non-forest (Horizon 2), and non-forest land type (Horizons 3 and 4).

For the first part of the verification, it is recommended that regardless of product horizon, an initial Non-forest/Forest confusion matrix be developed. For Horizon 2, 3, and 4, this requires lumping all forest classes together regardless of density and forest type, and grouping all non-forest types together regardless of land use.

The second part of the verification involves regression of density for forestland for Horizon 2, 3, and 4 products[4]. Available for each point will be two estimates of forest density – one from operational image analysis and one from the verification data source. These should be regressed against each other and the following items verified:

- R^2 statistically significant ($p < 0.10$)
- Slope not statistically different from 1.0 ($p < 0.10$)
- Intercept not statistically different from 0.0 ($p < 0.10$)
- Residual variance is homoscedastic

If the first condition is not met, it means that the image data have no relationship to verification ("true") forest density. The second and third conditions must be evaluated together (Table 6) to determine the accuracy – i.e., lack of bias -- of operational estimates. The fourth condition not being met is indicative that the precision of forest density estimates – estimated using the root mean square error (RMSE) of the regression – varies with forest density. This is likely to be of most concern if the forest density estimates are determined to be valid/unbiased – i.e., the first condition in Table 6. Often, the precision of regression estimates for biological phenomena decreases with increasing values of the independent variable – forest density in this case. If forest density estimates are determined to be valid/unbiased but residuals are heteroscedastic, further analysis may be required to demonstrate international compliance.

For Horizon 3, the regression method of verification must be undertaken for all forest types combined (as for Horizon 2 products) and for each forest type. In addition, a confusion matrix must be developed for the non-forest land uses; this supplements the global non-forest/forest confusion matrix analysis. In this latter, one tabulates verification point samples using the non-forest land use taxonomy employed – e.g., bare soil, crops, pasture – for operational image processing.

[4] An example of the regression analysis discussed is presented subsequently in Table 11.

Table 6. Meaning of slope and/or intercept being statistically different from 1.0 and 0.0 respectively.

Slope = 1.0?	Intercept = 0.0?	Meaning
Yes	Yes	Image estimates of forest density are valid/unbiased
Yes	No	Estimates of density are consistently biased
No	Irrelevant	Estimates of density are biased with the direction and magnitude of bias dependent on forest density

For Horizon 4, one first undertakes a global confusion matrix analysis for three classes: non-forest, forest, and sparse woody perennial. Following this, one undertakes the same analyses associated with Horizon 3 verification.

The analyses described for each horizon provide two critical types of information. First, the global confusion matrix developed initially for each horizon provides an overall statement of the validity of the forest/non-forest map for the area considered. Subsequent analyses partition potential error in such a way that it enables continuous improvement. Nonetheless, as one moves from Horizon 1 to Horizon 4 products, the interpretation of results is increasingly nuanced. For example, it is entirely possible that for Horizon 2, the global confusion matrix shows that non-forest/forest have been correctly classified. However, a subsequent regression analysis suggests that the density of the what has been classified as forest is over- or under-estimated. This is then likely to mean that carbon stocks have been similarly over- or under-estimated – despite the confusion matrix indicating that the extent of forest has been accurately mapped.

Point Samples for Change/Trend Maps

As was the case for single date Horizon 1 products using point samples, a confusion matrix approach is also appropriate for Horizon 1 change maps. However, the classic confusion matrix approach must be modified to reflect that one will only have verification point sample data for a period – "$t_{1.5}$" – between the two single-date maps – t_1 and t_2. This requires the use of a linguistic correctness scales composed of five classes – Definitely Correct, Probably Correct, Inconclusive, Probably Wrong, Definitely Wrong. Moreover, there remain the other issues mentioned that must be addressed in a reasoned manner – e.g., size differences between pixels and verification sample points, sampling schemes that need to address the rarity of change.

In general terms, a confusion matrix approach to verification of Horizon 1 change maps fundamentally requires that a point sample/pixel be determined "correct" or "not correct." However, the availability of verification data for a single date requires "reasoned evaluation" that has an implicit level of subjectivity. To evaluate correctness of point samples for change maps, a human operator must examine the state (forest or non-forest) of the pixel at t_1 and t_2 according to the operationally processed imagery, the state of the pixel according to the verification data from $t_{1.5}$ and the state of the change map – i.e.., No Change(Forest), No Change (Non-forest), Afforestation, Deforestation. Once this information is available, a determination of the correctness of each verification sample point can be determined.

This evaluation also requires a re-examination of the original imagery from t_1 and t_2 and a subjective evaluation of the correctness of each and the associated change map. Operators must look for clues that reflect the real-world conditions under which afforestation and deforestation occur. For example, whereas some afforestation occurs naturally due to woody ingrowth in low density or cleared areas, anthropogenic afforestation occurs because of the establishment of forest plantations. Hence plantation-related afforestation generally occurs in geometrically regular shapes and verification interpreters may be reasonably sure that it has occurred because of spatial context – e.g., the presence of adjacent roads. Natural afforestation will occur in areas that are "salt-and-peppered" with woody ingrowth and may be difficult to confirm with certainty, particularly if image pixels are large. Anthropogenic deforestation also generally occurs in geometrically regular shapes, whereas deforestation due to bushfires, for example, may be more difficult to confirm[5]. These factors must be considered when an interpreter is examining the forest/non-forest state of a pixel at t_1 and t_2 according to the operationally processed imagery and the state of the pixel according to the verification data at $t_{1.5}$.

Tables 7 illustrates decisions that can be made for each verification pixel if its state is non-forest. This information is employed for assessing the correctness of the change map only and not the single date maps; that will have been addressed by the single date verification. Table 8 shows comparable information for a verification pixel whose state is forest.

[5] Reductions in forest density associated with Horizon 2, 3, and 4 products are addressed subsequently.

Table 7. Example of conclusion for Horizon 1 change map verification pixel whose state is non-forest.

Class			Change Map Class	t_1 and t_2 maps judged to be correct?	Change Map Class Conclusion	Comments
Map t_1	*Verif. Data $t_{1.5}$*	*Map t_2*				
Non-forest	Non-forest	Non-forest	No Change (Non-forest)	Yes	Definitely Correct	
Non-forest	Non-forest	Non-forest	No Change (Non-forest)	No	Definitely Correct to Definitely Wrong	"No change" is correct if the t_1 and t_2 maps are both definitely wrong.
Non-forest	Non-forest	Forest	Afforestation	Yes	Definitely Correct	
Non-forest	Non-forest	Forest	Afforestation	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary
Forest	Non-forest	Non-forest	Deforestation	Yes	Definitely Correct	
Forest	Non-forest	Non-forest	Deforestation	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary
Forest	Non-forest	Forest	No change (Forest)	Yes	Definitely Correct or Inconclusive	• Change map Definitely Correct if $t_{1.5}$ data shown to be definitely wrong. • Inconclusive if $t_{1.5}$ data cannot be explained. • Sample point should probably be eliminated.
Forest	Non-forest	Forest	No Change (Forest)	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary

Table 8. Example of conclusion for Horizon 1 change map verification pixel whose state is forest.

Class			Change Map Class	t_1 and t_2 maps judged to be correct?	Change Map Class Conclusion	Comments
Map t_1	*Verif. Data $t_{1.5}$*	*Map t_2*				
Non-forest	Forest	Non-forest	No Change (Non-forest)	Yes	Definitely Correct	• Change map Definitely Correct if $t_{1.5}$ data shown to be definitely wrong. • Inconclusive if $t_{1.5}$ data cannot be explained. • Sample point should probably be eliminated.
Non-forest	Forest	Non-forest	No Change (Non-forest)	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary
Non-forest	Forest	Forest	Afforestation	Yes	Definitely Correct	
Non-forest	Forest	Forest	Afforestation	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary
Forest	Forest	Non-forest	Deforestation	Yes	Definitely Correct	
Forest	Forest	Non-forest	Deforestation	No	Inconclusive to Definitely Wrong	Difficult to assess if non-forest/forest at t_1, $t_{1.5}$, or t_2 is close to definitional forest boundary
Forest	Forest	Forest	No change (Forest)	Yes	Definitely Correct	
Forest	Forest	Forest	No Change (Forest)	No	Definitely Correct to Definitely Wrong	"No change" is correct if the t_1 and t_2 maps are both definitely wrong.

For Horizon 1, analysis of verification point samples for change maps involves tabulation of samples by the "Change Map Class Conclusion" in Tables 7 and 8 (Table 9). From this, a variety of statistics can be developed for individual classes (e.g., "% of Deforestation that is Probably or Definitely Correct") or for combinations of classes (e.g., "% of No Change that is Definitely Incorrect"). For international compliance, an agreed set of statistics and acceptable levels for each would have to be developed. For continuous improvement, such a table provides information about which image classes are most questionable over an entire country, and similar tables could be produced for smaller areas.

Table 9. Example of possible summary table for Horizon 1 products using verification information based on Tables 7 and 8.

	Definitely Correct	**Probably Correct**	**Inconclusive**	**Probably Wrong**	**Definitely Wrong**	***Total***	***% Prob. Correct +***	***% Prob. Wrong-***
No Change (Non-forest)	55	30	10	4	1	100	***85***	***5***
No Change (Forest)	35	10	2	2	1	50	***90***	***6***
Afforestation	1	3	4	2	0	10	***40***	***20***
Deforestation	7	8	1	2	2	20	***75***	***8***
Total	***98***	***51***	***17***	***10***	***4***	***180***	***83***	***8***
% of Total	***54***	***28***	***9***	***6***	***2***	***100***		

The verification of Horizon 2 (and 3 and 4) products commences with a completely different approach based on the same assumption that for change map verification, one will only have a pixel/point for a single date.

The first verification task is to "screen" or "pre-analyse" the data. Recall that for Horizon 2 products available data will be an estimate of woody/forest density for three time periods – t_1, $t_{1.5}$(verification data), and t_2. – of the form presented in the three leftmost columns of Table 10. This screening identifies all verification points that are probably anthropogenic afforestation or any type of deforestation as well as those that are clearly erroneous.

Strictly speaking no change (non-forest) points do not need to be distinguished from the no change (forest) points. This is because Horizon 2 products address density of woody vegetation; for analysis of Horizon 2 and higher products non-forest is viewed simply as a low density of woody vegetation. However, there is a risk that not separating no change (non-forest points) from subsequent regression analysis will inflate goodness-of-fit regression statistics due to an expected lower variability in the no change (non-forest) verification points than in the no change (forest) verification points. This screening should also include an assessment of the similarity of variance between no change (non-forest) and no change (forest) points using statistics such as coefficient of variation; tests for equality of variance such as the *F* test (parametric) or the Moses test (non-parametric) may also be employed. If it is judged necessary to separate the no change (non-forest) points from the no change (forest) points, the analysis subsequently described is conducted separately for both sets of points.

Assuming that no change (non-forest) points are not separated from no change (forest) points, the initial data screening separates the data into four groups – definite errors, anthropogenic afforestation, deforestation, no change. It is not necessary to separate natural from anthropogenic afforestation as the former is considered a natural increase in forest density. Similarly, natural thinning does not need to be separated from anthropogenic or catastrophic deforestation, although anthropogenic silvicultural operations such as thinning or selective forest harvesting may need to be placed in their own group and eliminated from subsequent regression analysis.

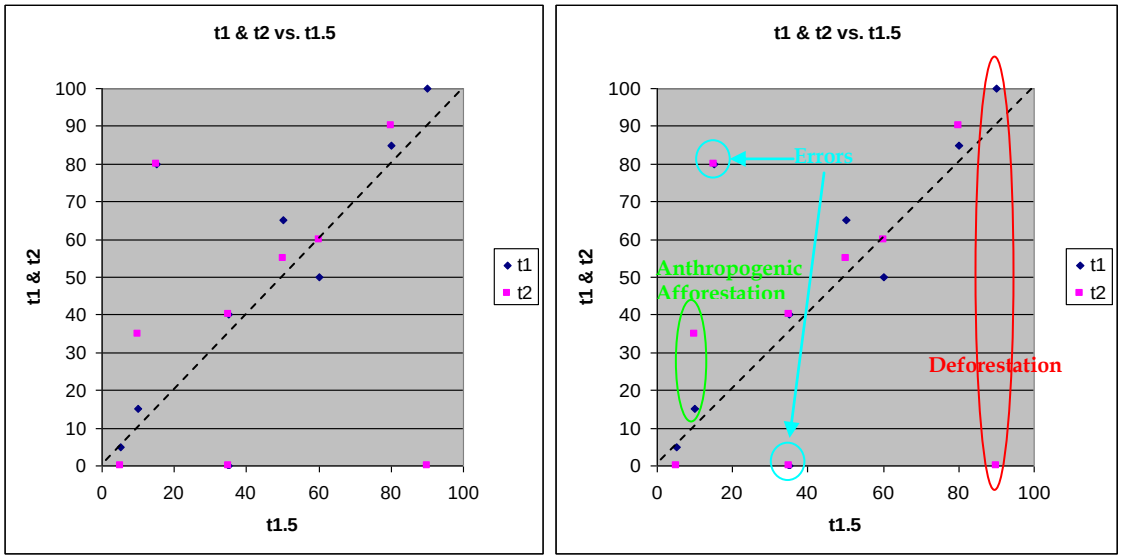
Table 10. Forest density data collected from verification sample points (hypothetical).

Point	t_1	$t_{1.5}$	t_2	**Summary**	**Analytical Comments**
1	5	5	0	No Change (NF)	Natural variability or measurement imprecision
2	100	90	0	Deforestation	Must confirm t_1, $t_{1.5}$, and t_2
3	50	60	60	Denser	Natural variability or measurement imprecision
4	65	50	55	Sparser	Natural variability or measurement imprecision
5	15	10	35	Afforestation	Must confirm t_1, $t_{1.5}$, and t_2
6	80	15	80	No change (Forest)	Error of t_1/t_2 classification if $t_{1.5}$ confirmed; possibly eliminate from regression analysis.
7	0	35	0	No change (Non-forest)	Error of t_1/t_2 classification if $t_{1.5}$ confirmed; possibly eliminate from regression analysis.
8	40	35	40	No change (Forest)	Natural variability or measurement imprecision
9	85	80	90	No change (Forest)	Natural variability or measurement imprecision

There are a number of ways this screening could be undertaken. The most efficient would be algorithmic. The verification data points of Table 10 provide an example of how indicative rules could be developed. For example, points with a high density at t_1 and low density at t_2 (e.g., Point 2) are possible deforestation. In other cases (Points 3, 4, 8, and 9) similar high density values for t_1, $t_{1.5}$, and t_2 suggest no change (Forest), and points with similar low density values (Point 1) suggest no change (non-forest).

For illustration purposes, rather than conduct this screening algorithmically, a graphical technique is employed. Figure 10 shows the density values at t_1 and t_2 plotted against the density values for $t_{1.5}$ for all points. It also shows the 1:1 line – i.e., the line on which all points would fall if densities for t_1 and t_2 were the same as those for $t_{1.5}$. A point that is "close to" the 1:1 line suggests a no change point with those at lower densities suggesting no change (non-forest) and those at higher densities suggesting no change (forest). As is often the case, it is the definition of "close to" that is problematic and that requires subjective judgement, although a numerical criteria could be established a number of ways. The critical point about Figure 10, however, is that it indicates how each point can be separated into one of the four groups.

Figure 10. Data points from Table 10 represented graphically and screening that can be conducted; raw data (left) and identified groups (right).



Possible error points are those whose t_1 and t_2 densities are both considerably different from the $t_{1.5}$ density – those that are not close to the 1:1 line. In the graphic there are two such points – one (Point 6) with a $t_{1.5}$ density of 15 and another (Point 7) with a $t_{1.5}$ density of 35. Each of these points would have to be verified to confirm the errors based on the rules presented in Tables 9 and 10. Doing so would indicate that both are problematic as the t_1 and t_2 densities for Points 6 and 7 are 80 and 0 respectively.

Deforestation points can be readily identified as those that are not close to zero along the x-axis ($t_{1.5}$ density) and for which the t_1 density is considerably higher than the t_2 density. In the graphic there is one such point – i.e., Point 2 has a $t_{1.5}$ density of 90, a t_1 density of 100, and a t_2 density of 0. Assigning Point 2 to the deforestation group would only be done after examining the raw imagery from t_1 and t_2 and the verification data for $t_{1.5}$ and making a judgement for each such point based on rules presented in Tables 9 and 10.

Possible anthropogenic afforestation points can also be readily identified by identifying points where the t_2 density is considerably higher than the t_1 density and where the t_2 density is considerably above the 1:1 line; Point 5 is an example. As always, assignment to the anthropogenic afforestation group requires Such a point exists when $t_{1.5}$ is 10 (Point 5). Each of these possible anthropogenic afforestation points would have to be verified and it might be necessary to verify a fairly large number of such points. The implicit situation illustrated by Point 5 is forest planting that occurred soon after the date associated with t_1 and $t_{1.5}$, and sufficient time elapsing before t_2 that tree canopies closed to cover 35% of the ground. Were verification data collected closer to t_2 than t_1, anthropogenic afforestation is likely to be more difficult to detect.

To assess accuracy of deforestation and anthropogenic afforestation, the points assigned to those classes would be tabulated in a table like Table 9 based on the rules provided in Tables 9 and 10. In addition, the number of erroneous pixels and their nature would have to be tabulated. An acceptable number/percent of errors could be determined for international compliance and the type of error – e.g., forest vs. non-forest, high vs. low forest density -- would be useful for continuous improvement.

The remaining points – i.e., those not subject to deforestation, anthropogenic afforestation, or error (Fig. 11). – are analysed by regression. In doing this, the t_1 data is treated as being independent of the t_2 data data (Table 11). This creates an issue of temporal correlation, that if not addressed will overestimate accuracy. Techniques for addressing this are well-established and described in statistical literature. A simplistic approach is recommended whereby a "pass/fail" statistical criteria is employed that is more stringent than the widely accepted significance level of $p = 0.95$. Yet another approach that could be recommended is to conduct two regression analyses – one for each date.

Figure 11. Points not assigned to the error, anthropogenic afforestation or deforestation data groups. (See text for filtering explanation.)

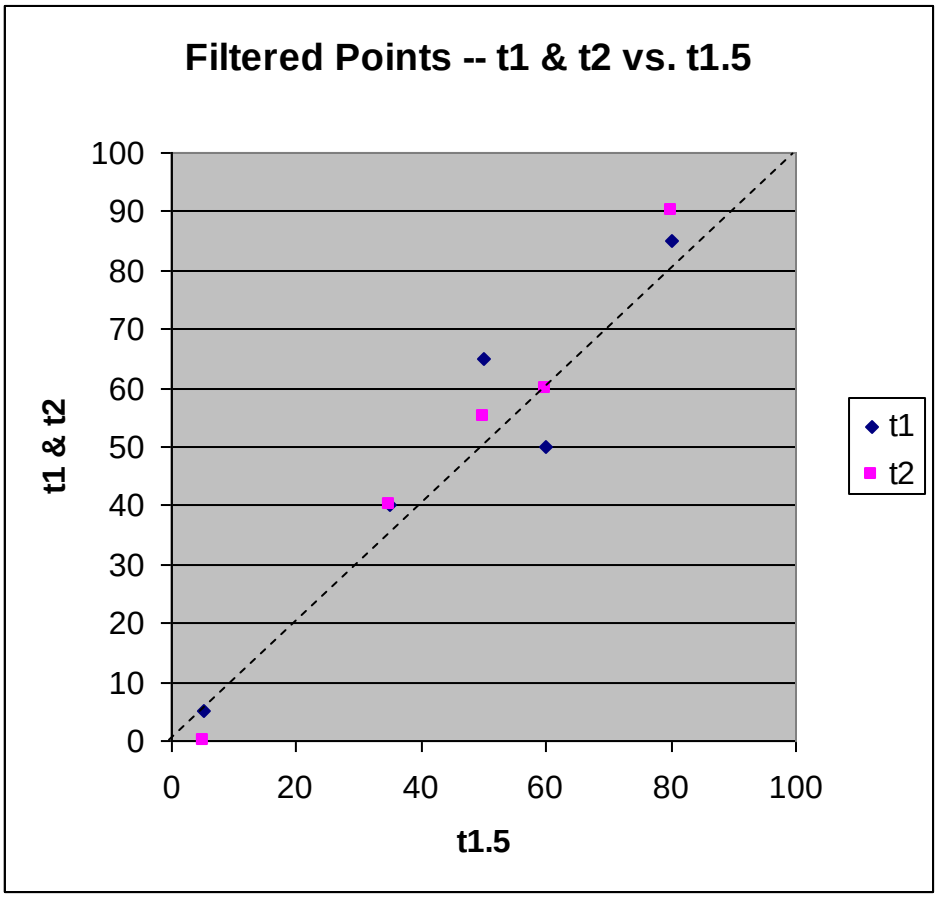


Table 11. Data set for regression analysis after point filtering.

Point	t_1,t_2	$t_{1.5}$
1(t_1)	5	5
3(t_1)	50	60
4(t_1)	65	50
8(t_1)	40	35
9(t_1)	85	80
1(t_2)	0	5
3(t_2)	60	60
4(t_2)	55	50
8(t_2)	40	35
9(t_2)	90	80

The information produced by the regression analysis (Table 14) is sufficient to assess the validity of the verification points not associated with deforestation or anthropogenic afforestation. The assessment of validity is done the same as for Horizon 2, 3, and 4 maps for individual dates. First, if the relationship between $t_{1.5}$ woody vegetation density and the t_1 and t_2 densities does not explain a statistically significant amount of variability as measured by p for R^2, the operational forest density maps are not valid. Second, if there is a statistically significant relationship, the equation of the regression line is assessed to determine if it has an intercept and slope that are not significantly different from 0.0 and 1.0, respectively. If not, the method for estimating density is not valid. Third, if there is a significant relationship and the intercept and slope are not different from 0.0 and 1.0 respectively, then a visual or statistical verification of homoscedasticity is made. If residuals are considered homoscedastic, then the regression standard error is compared to internationally mandated standards. If the standard error is excessive, the operational forest density maps are not considered valid. If the standard error is within accepted limits, the operational remote sensing data products are considered valid for Horizon 2 products.

Table 12. Results of the regression analysis conducted on the data in Table 13.

Parameter	Calculated Value	Validity Value	Standard Error	*p* for testing	Significantly Different from Validity Value?
$R^2_{adjusted}$	0.94	1.0	N/A	N/A	No (as determined by *p* for R^2 below)
***p* for R^2**	0.00	0.01	N/A	0.99	No
Intercept	-1.0	0.0	4.7	0.90	No
Slope	1.1	1.0	0.1	0.90	No
Residual Variance	Homo-scedastic	Homo-scedastic	N/A	N/A	No
Regres-sion Standard Error	7.2	Set by inter-national standards	N/A	Set by inter-national standards	To be determined

These regression results must be interpreted in parallel with the deforestation, anthropogenic afforestation, and error information compiled in parallel. It is possible that the density estimation methodology is valid, but that deforestation and anthropogenic afforestation are not estimated with sufficient accuracy to meet international standards, or that the error rate is excessive. In such cases, the continuous improvement program must provide for the development of suitable maps.

Horizon 3 and 4 change map products can be verified with point samples using the methodology described for the Horizon 2 change map products. The primary difference for Horizons 3 and 4 is that individual deforestation/anthropogenic afforestation tables, error rates, and regression analyses will have to be undertaken for each land-use class of interest.

Areal Samples for Single Date Maps

Areal verification samples for Horizon 1 products will record the amount (or percent) of forest and non-forest for each sample. Areal verification information is likely to be obtained via automated or human image interpretation rather than ground-based data. For each sample, the same area will have been extracted from the operational processed imagery and the amount/percentage of forest and non-forest tabulated for each sample.

Such data can be analysed using parametric statistics such as paired *t* tests ** and/or regression on the amount of a given type mapped on the map and the verification data for each sample unit. This is appropriate provided that the country being verified is not covered 80% to 90% by either forest or non-forest. If one cover is heavily dominant, however, there is a risk that the use of parametric statistics will be inappropriate. This is also an issue for continuous improvement that requires evaluating individual regions; smaller regions have a greater likelihood of being dominated by forest or non-forest.

The inability to use parametric analysis can be addressed a variety of ways. Lowell (2001) used a the non-parametric Kolmogorov-Smirnov two-sample test to evaluate if independent areal samples provided the same estimates of land cover change as operational image processing. Note that this approach requires areal verification samples that are independent of rather than paired with areal samples extracted from the map being verified. A similar approach could be adopted for single date Horizon 2 products (and also Horizon 2 change/trend maps). Although such an approach provides for global verification, it does not address individual regions and therefore has limited utility for continuous improvement. This was addressed by Barson *et al.* (2004) who used approximate confidence intervals and z-scores to enable interval/ratio assessment of the validity of individual sample units. Undertaking such analysis and summarizing them by region provides for continuous improvement and could be employed for that purpose.

Information collected for areal validation samples and associated areas on operationally processed imagery for Horizon 2, 3, and 4 products is considerably more complex. This in turn makes analysis for verification and continuous improvement more complex. Whereas for point samples a given parameter – land cover, woody vegetation density – is characterized by a single location, with areal samples a parameter is characterized by multiple pixel. An additional complexity is that Horizon 2, 3, and 4 products will contain continuous interval/ratio data – i.e., woody vegetation density – instead of the nominal forest/non-forest measures of Horizon 1 single date products.

To validate areal samples for single date Horizon 2 products, it is assumed that the verification image data have the same form as the Horizon 2 map product being validated. This means that one has two maps – the original and the validation data -- composed of co-registered pixels of the same size and that employ the same cartographic taxonomy. If validation data are derived from images with a different spatial resolution, they must be resampled to the same resolution of the operational map product being validated. Similarly, if validation data are broad woody vegetation density classes obtained from human interpretation of aerial photographs, these must be converted to the same data format and taxonomy as the operational map product being validated.

This will result in a number of paired pixels for each areal sample unit. These can be analysed to determine validity of individual areal sample units using the non-parametric Wilcoxon matched-pairs signed-ranks test ("the Wilcoxon test"; Daniel 1978). The Wilcoxon test evaluates if the locations – i.e., the central values -- of two distributions is different by testing for differences in the median. Because the Wilcoxon test is not widely familiar, its use is described briefly here.

The data in Table 13 represent all five pixels of a single fictitious areal sample unit. Column (2) shows the woody vegetation mapping density on the map being validated and Column (3) shows the forest density from the validation data. Column (4) shows the difference and the sign of the difference. Once these have been calculated for all pixels in the areal sample unit, the absolute values of the differences are calculated (Column (6)) and ranked (Column (7)) with the sign of the difference being noted (Column (8)). The ranks for each sign are then summed to obtain two values T_- and T_+ (Row (9)) with the smaller of the two being the test statistic T; in this example the test statistic is T_- with a value of two (2; Row (10)). The probability *p* of observing a particular value of T given a sample size of *n* (five in this case) is then obtained from a table for the Wilcoxon test in the case of small sample sizes, or by using a numerical approximation for large sample sizes (Row (1)1). The conclusion for this sample unit is that it is valid if we use a 95% confidence level. This means that the validation sample unit and the operational map product have the same forest density. To be considered "non-valid" with 95% confidence the *p* value (Row (1)1) would have to be less than 0.05.

Table 13. Example of evaluation of individual sample units based on the Wilcoxon matched-pairs signed-ranks test.

Pixel (1)	***Mapped Density (2)***	***Validation Data (3)***	***Diff. (4)***		***Pxl (5)***	***Ranked Abs. Value (6)***	***Rank (7)***	***Sign (8)***
1	5	4	+1		1	1	1	+
2	50	60	-10		5	2	2	-
3	65	50	+15		4	5	3	+
4	40	35	+5		2	10	4	+
5	85	87	+2		3	15	5	+
					Summary		T_-	T_+
					Sum of ranks (9):		2	13
					Test statistic T (10)	2		
					***p* for n=5/T=2 (11)**	0.125		

In undertaking the same analysis for each areal sample unit, a *p* value is obtained for each. International compliance can be demonstrated by having no more than (1-*a*)% sample units considered non-valid where *a* is the desired level of confidence. That is, if one has 200 areal validation samples and wants validity with a confidence level of *a* = 0.95, then having 10 units or less -- i.e., (1-0.95 =) 5% of 200 -- whose *p* value is less than 0.05 would indicate that the mapping method produces the same values as the validation data.

The *p* values for each areal unit also provide for continuous improvement. Low *p* values indicate areas where the original data and validation data are most different. Such information can be summarized by regions or other classes, or by undertaking spatial interpolation[6] on the *p* values to produce a surface showing areas of high and low validity.

Though it is non-parametric, the Wilcoxon statistic is subject to certain assumptions and potential constraints. The only assumption that is potentially problematic for verification is that the differences (Column 4 in Table 13) are assumed to be independent. This will not be true in the procedure described since there is likely to be spatial autocorrelation among the pixels in the same areal sample unit. This will have the impact of making the original mapping for each areal unit appear "more valid" than it truly is. It is recommended that this be addressed by using a more stringent statistical level than would otherwise be accepted.

Another issue associated with the use of the Wilcoxon statistic for verification involves "ties." There are two that must be addressed.

[6] The ability to produce a meaningful interpolation is dependent on the *p* value having a continuous spatial structure. If strong positive spatial autocorrelation is not present among *p* values, any resulting surface will be meaningless.

The first occurs when the original woody vegetation density is the same as the validation density for a pixel – i.e., the difference in woody vegetation density is zero (0). For the Wilcoxon test such pixels must be eliminated. In practice, this is most likely to occur for pixels in areas where there is no woody vegetation – i.e., density is zero (0) – rather than in heavily forested areas, although it is also potentially an issue in areas where forest density is high and consistently near 100%. Because it is supposed that these will be areas of high validity because they will include areas of bare soil and water that are relatively easy to classify, eliminating them will have the effect of making the original mapping appear less valid than it truly is. Because this puts the focus on mapping forest density accurately – one of the most important inputs to carbon accounting systems – this is not viewed as a problem. However, there must be an awareness of this issue with respect to the dominant land cover – forest or non-forest -- in the country being validated and for individual areal samples.

The second type of tie occurs when the absolute difference values (Column 6 in Table 13) are the same for two or more pixels. In this case, one simply averages the ranks for the tied pixels and assigns the average rank to all of them.

The validation of single date Horizon 3 and 4 products using areal units is similar to what has been described for Horizon 2 products. The primary difference is that the same analysis must be repeated for each land use class. No methodological modification needs to be made for the definition of Horizon 4 products including sparse woody vegetation; these are addressed as a matter of course using the technique described that is based on evaluating forest density in a way that includes sparse woody vegetation.

Areal Samples for Change/Trend Maps

Validity of areal samples for Horizon 1 change maps can be assessed using the techniques described for single date Horizon 1 maps at the beginning of the previous section. The only difference is that because of the relative rarity of change, frequency distributions of areal change samples will almost never be normally distributed. Hence it will almost always be inappropriate to analyse areal change map samples using parametric statistics. This is in contrast to analysing the validity of single date Horizon 1 products using areal samples in which the amount of forest and non-forest may be normally distributed. As a consequence, Horizon 1 change products must be analysed by using non-parametric techniques for three classes – no change, deforestation, afforestation. It may also be desirable to divide the no change class into two groups – no change (forest) and no change (non-forest). The Kolmogorov-Smirnov test employed by Lowell (2001) for this purpose is likely to be of limited value here because of the large number of sample units for which the mapped amount of deforestation and afforestation will be equal to the amount estimated using the validation data – i.e., both zero; the K-S test has difficulties in such situations. Hence it will be preferable to use approximated confidence intervals as done by Barson *et al.* (2004).

Validation for Horizon 2 change maps using areal samples will follow the same general method as the evaluation of Horizon 2 change maps using point samples. This procedure starts with evaluation of individual areal units. First, the pixels comprising an areal sample unit are divided into three groups – no (anthropogenic) change, deforestation, anthropogenic afforestation – using algorithmic or graphical filtering techniques illustrated in Figure 10. The validity of the deforestation and anthropogenic afforestation groups is then determined using techniques that lead to the development of information presented in Table 9 – i.e., a modified confusion matrix with classes such as "probably correct" and "inconclusive." The validity of the density estimated for the no change class is then determined using regression techniques described in sections related to for analysing data in Table 12. These procedures determine the validity of individual areal samples.

Summarising information for Horizon 2 change maps across all areal samples is more difficult than for Horizon 2 single date maps. For change maps, one must consider individual sample validity for three classes. It is conceivable -- and even likely -- that a particular areal unit will be valid for anthropogenic afforestation and deforestation but not for woody vegetation density in the no change class. Moreover, a binary statement of validity – i.e., valid or not valid – for afforestation or deforestation pixels is not readily apparent from confusion matrix-like information presented in Table 9. Mreover, there are multiple reasons that woody vegetation density estimates may not be considered valid:

- Bias -- slope of the validation line equal to 1.0 but an intercept that is not equal to 0.0.

- Accuracy varying with forest density -- slope and intercept of validation line not equal to 1.0 and 0.0, respectively.

- No correlation between mapped and validation woody vegetation density.

- Excessive error as determined by international compliance standards.

Each of these suggest different things relative to international compliance and continuous improvement. Interpretation of validity statistics may therefore be contentious.

Despite the difficulty, ultimately some statement of validity for individual areal samples is required. For anthropogenic afforestation and deforestation, this requires an accepted standard such as "no more than x% of pixels are definitely wrong" or "at least y% of pixels are probably or definitely correct." The forest density of the no change class could be made more straightforward by only requiring a statistical level such as 95% (a=0.05). This would then be used to assess difference of the slope and intercept of the validation line from 1.0 and 0.0, respectively, and to evaluate if the relationship of mapped density to validation data density is statistically significant. Determination of individual sample validity could be made more informative through increasing complexity – e.g., "the slope and intercept of the validation line must be 1.0 and 0.0, respectively, the relationship must be statistically significant, and the standard error of regression must be less than x."

For international compliance, once a definition of validity for individual areal sample units is determined, summarising information across all areal samples can be addressed a number of ways. Which is preferable or the most appropriate can only be determined after more experience is gained with the development and validation of Horizon 2 products. Conceptually, however, the are two ways to summarise areal samples for international compliance of change maps.

One way is to assess validity for international compliance separately for each of the three classes anthropogenic afforestation, deforestation, and no change. For this, standards would have to be set for each group. For anthropogenic afforestation and deforestation, this might entail a statement such as "x% of the areal samples on which anthropogenic change was observed in the mapped or validation data must be valid and the data errors (points circled in blue in Fig. 10) must beless than y%." Such a standard would eliminate the inflation of validity statistics due to many/most areal samples having no recorded anthropogenic change, but it might not adequately address errors of omission – i.e., unmapped anthropogenic change. For the no human change class, validity might be determined using an acceptable statistical level – e.g., "z% of areal samples must be considered 'valid'" with the definition of "valid" requiring further clarification.

The second possible way of summarising validity information across all areal samples is by not considering the three groups separately. In this scenario, each areal sample would be determined to be valid or not valid, with no consideration of which group(s) caused a sample to be evaluated as "not valid." In addition to defining what constitutes a valid areal sample, the only additional requirement would be the determination of an internationally accepted standard for the percent of areal samples that must be considered valid for the mapping methodology to be determined to be internationally compliant.

The second aspect of summarising validation data across all areal samples for Horizon 2 is continuous improvement. For this, the information for each areal sample must be considered in a reasoned manner. It is recommended that a standardised methodology or protocol for doing this be developed as more experience is gained in Horizon 2 change maps. Such a protocol might include the following:

- Assess the units judged as not being valid for anthropogenic change for spatial pattern.

- Assess the units judged as not being valid for anthropogenic change relative to and land use or other class of interest.

- For the no change group, examine the spatial pattern of bias and lack of significant relationship for mapped vs. validation woody vegetation density.

The validation of Horizon 3 and 4 change maps using areal units is similar to what has been described for Horizon 2 products. The primary difference is that similar analysis must be repeated for each land-use class of interest. No particular allowance needs to be made for the definition of Horizon 4 products including sparse woody vegetation; these are addressed as a matter of course using the validation technique described that is based on evaluating forest density in a way that includes sparse woody vegetation.

5 Summary

Among the most important inputs to internationally accepted carbon accounting systems will be maps developed from the processing of remotely sensed imagery. Verification of processed imagery must be an integral part of any carbon accounting system that relies on such imagery. The importance of such products and their validity will be accentuated as such systems become more sophisticated and move from Horizon 1 to Horizon 4 products. This in turn will increase the complexity of verification procedures.

Verification programs for processed imagery will also provide critical information for the continuous improvement of image processing procedures. Because internationally compliant carbon accounting will be an ongoing long-term activity, it is in the interest of all countries to continually improve their image processing procedures in order to easily demonstrate international compliance of their carbon accounting system. In a well designed validation program, intelligent and reasoned use of information developed primarily for international compliance will also provide the information necessary for individual countries to identify and rectify weaknesses in their image processing procedures.

This document describes multiple approaches to validation that can be implemented operationally. These are presented in the context of a recognition that there are numerous ways to evaluate if inputs to carbon accounting systems are internationally compliant. The options presented also reflect the reality that different countries will have different types of data available for validation. Hence, the various options presented are tailored to adapting validation procedures to country-wide data realities rather than mandating a single system that forces the establishment of an underlying validation-specific data collection effort.

It is acknowledged that some of the suggestions made are speculative. That is, statistical procedures, sampling considerations, and other aspects of the recommended validation methodologies will have to tested and modified for operational validation as international experience increases. This is particularly true for single date maps for Horizon 2, 3, and 4 products as well as for change map products for Horizons 1 to 4. Because such map products are not widely available, only a limited amount of effort has been expended in the scientific community to develop and test appropriate validation methods. Extending fundamentally sound validation techniques for these products to an operational context will require additional work.

Implicit in any validation program is independent evaluation. The most critical aspect of independence is personnel. Individuals and even organizations involved in validation personnel must be completely independent of those who did the operational image processing. The validation and operational processes must also be kept independent – i.e., operational personnel must not be informed about sampling schemes, validation data, etc. Maintaining independence of the validation program is critical for demonstrating that a country's carbon accounts are internationally compliant.

Some of the techniques suggested require the establishment of internationally acceptable thresholds for determination of compliant or non-compliant data. The establishment of these thresholds is a policy-based rather than scientific question. After more experience with country-wide carbon accounting systems is gained, the international community will have to establish these.

Among the issues that will also have to be addressed by the international community are temporal and spatial compliance. For international carbon accounting, each country will produce Horizon 1, 2, 3, or 4 maps for the entire country at time periods t_1, t_2, t_3, …. Subsequent change maps will be produced for periods of interest which might be limited to $t_1 \rightarrow t_2$ and $t_2 \rightarrow t_3$, but could also include, for example $t_1 \rightarrow t_3$. In this document, it has been implicitly assumed that validation will be undertaken for an entire country each time a new single date map and a change map is produced, and that the only change map to be evaluated will be the one associated with the time period $t_n \rightarrow t_{n+1}$. In fact, it may be agreed that validation is not required for each time period – i.e., a single successful validation may provide "international certification" for an entire reporting cycle even if maps are produced for more than one date within that reporting cycle. Similarly, it may eventually be mandated that only certain geographic regions within a country must be validated. How the validation program is structured at the international level is a policy-based issue rather than a scientific one.

Finally, it is critical to recognize that the procedures described address the validation of a single part of carbon accounting systems: land cover data derived from remotely sensed imagery. Other critical inputs into carbon accounting system will have to be similarly validated. Furthermore, input data that are considered valid addresses only one aspect of system validity since valid input data can lead to spurious carbon estimates because of the model-based nature of carbon accounting systems. The ultimate system validation is a carbon accounting system that produces accurate carbon estimates. Yet because system accuracy is difficult to evaluate through direct measurement, and because carbon accounting systems that convert input data into estimates of carbon are composed of algorithmic and statistical models, system-wide validation must be addressed by domain-specific modeling specialists. Such specialists must verify that the system being validated includes all relevant major components of the carbon cycle, each has been properly conceived and appropriately calibrated, and that the separate components are linked in a way that correctly addresses critical interactions among them.

6 Bibliography

Anderson, J., Hardy, E., Roach, J., Witmer, R., 1976. *A Land Use and Land Cover Classification System for use with Remote Sensor Data*. Geological Survey Professional Paper 964. United States Government Printing Office, Washington, D.C. 41 pp.

Barson, M., Bordas, V., Lowell, K., Malanfant, K., 2004. An independent reliability assessment for the Australia agricultural land-cover change project 1990/91-1995. Chapter 8: *Remote Sensing and GIS Accuracy Assessment* (Eds. R. Lunetta and J. Lyon). CRC Press, pp. 105-224.

Daniel, W., 1978. *Applied Nonparametric Statistics*. Houghton Mifflin Company, pp. 135-139.

Defries, R., Hansen, M., Townshend, J., 2000. Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data. *International Journal of Remote Sensing* **21**:1389-1414.

Edwards, G., Lowell, K., 1996. Modelling uncertainty in photointerpreted boundaries. *Photogrammetric Engineering and Remote Sensing* **62**(4):377-391.

Khorram, S., 1999. *Accuracy Assessment of Remote Sensing-Derived Change Detection*. American Society for Photogrammetry and Remote Sensing Monograph Series. 65 pp.

Lowell, K., 2001. An area-based accuracy assessment methodology for digital change maps. *International Journal of Remote Sensing* **22**(17):3571-3596.

Lowell, K., Richards, G., Woodgate, P., Jones, S., Buxton L., 2005. Fuzzy reliability assessment of multi-period landcover change maps. *Photogrammetric Engineering and Remote Sensing* **71**:939-945.

Stehman, S., Czaplewski, R., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of the Environment* **64**:331-334.

Strahler, A., Boschetti, L., Foody, G., Friedl, M., Hansen, M., Herold, M., Mayayx, P., Morisette, J., Stehman, S., Woodcock, C., 2006. *Global Land Cover Validation: Recommendations for Evalution and Accuracy Assessment of Global Land Cover Maps*. GOFC-GOLD Report No. 25. GOFC-GOLD is a Panel of the Global Terrestrial Observing System (GTOS). www.fao.org/gtos/gofc-gold.