IIASA

International Institute for Applied Systems Analysis • A-2361 Laxenburg • Austria
Tel: +43 2236 807 • Fax: +43 2236 71313 • E-mail: info@iiasa.ac.at • Web: www.iiasa.ac.at

INTERIM REPORT IR-98-081 / November 1998

A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory

Gebhard Banko (banko@edv1.boku.ac.at)

Approved by
Sten Nilsson (nilsson@iiasa.ac.at)
Leader, *Forest Resources Project*

Foreword

IIASA, the Russian Academy of Sciences and several Russian governmental agencies, signed agreements in 1992 and 1994 to carry out a large-scale study on the Russian forest sector. The overall objective of the study is to focus on policy options that would encourage sustainable development of the sector.

The first phase of the study concentrated on the generation of extensive and consistent databases for the total forest sector.

In its second phase, the study encompassed assessment studies of the greenhouse gas balances, forest resources and forest utilization, biodiversity and landscapes, non-wood products and functions, environmental status, transportation infrastructure, forest industry adn markets, and socio-economics.

The remote sensing activities within this project aims at the following three main objectives:

- to produce an up-to-date forest information database of the Russian forest sector;
- to develop and test methods to produce an up-to-date land use and land cover database for Russia; and
- to develop and test operative forest information and decision support system, with monitoring and revision capabilities, in a GIS environment.

This work, carried out by Gebhard Banko during his participation in the YSSP (Young Scientists Summer Program), deals with the accuracy assessement of classifications of remotely sensed data and provides a review of current European forest inventory systems. Based on this review recommendations are given for the design of the accuracy assessment for the forest variables derived within the SIBERIA-Project. Examples from current national forest inventory systems demonstrate how such data — derived from remotely sensed images — can be integrated in a forest information system.

Acknowledgements

It was a pleasure for me to have had the opportunity to attend the Young Scientists Summer Program and to get involved in the forest activities at IIASA. I would like to thank Alf Öskog for his contribution to and supervison of my work. In addition to all the other colleagues of the Sustainable Boreal Forest Resource project, it was namely Professor Sten Nilsson and Cynthia Festin who encouraged my work in their typical manner. Special thanks are also given to Michael Gluck, who helped me a lot with his constructive comments to my work.

Special thank is also to be given to my colleagues of the Institute of Surveying, Remote Sensing and Land Information at the University of Agricultural Sciences Vienna. They took over most of my work and thus enabled me to work in a completely new environment.

Besides the scientific work, a major part of my stay at IIASA consisted of the discussion and communication with my YSSP-friends. Thanks to all the YSSPers for this unique time.

About the author

Gebhard Banko attended the Young Scientists Summer Program at IIASA in 1998. This work has been accomplished during this three months. He studied Forestry at the University of Agricultural Sciences in Vienna, and currently works as a graduate assistant at the Institute of Surveying, Remote Sensing and Land Information at this university. In his thesis he discussed the possibilities for using Landsat-TM data for forest applications in Austria.

Contents

A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory

Gebhard Banko (banko@edv1.boku.ac.at)

1 Introduction

1.1 Scope and Objectives

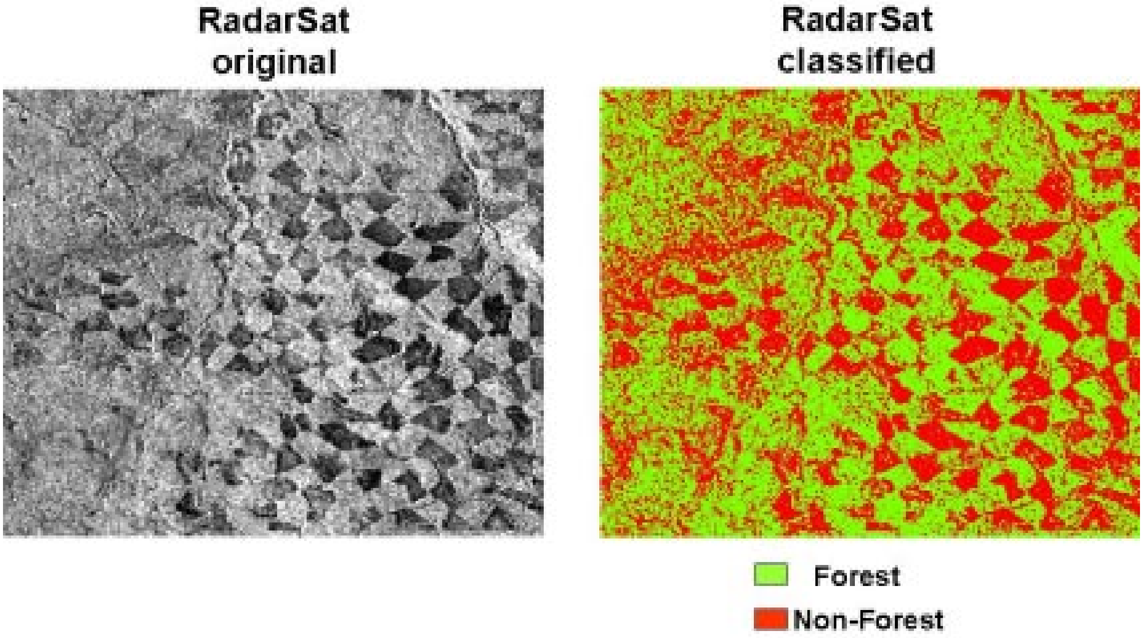
The scope of this study is to review the methods of assessing the accuracy of the classifications of remotely sensed data and to review current forest inventory systems in Europe. Special emphasis is given to forest inventory systems which include remotely sensed data in the assessment of forest variables. The aim of these reviews is to provide recommendations for the current SIBERIA-project in which the Forest Resources Project at IIASA is involved. The overall objective is to contribute to the development of a forest inventory system for Russia.

1.2 Accuracy Assessment

The accuracy of spatial data has been defined by the United States Geological Survey (USGS) as: "The closeness of results of observations, computations, or estimates to the true values or the values accepted as being true" (USGS, 1990). Nevertheless it must be stated that "truth" has a certain subjective dimension (Janssen and van der Wel, 1994).

Accuracy assessment is an important step in the process of analyzing remote sensing data. Remote sensing products can serve as the basis for political as well as economical decisions. Potential users have to know about the reliability of the data when confronted with maps derived from remote sensing data. A RADARSAT-scene of Whitecourt, Alberta, from the 6 April 1996, was used to derive a map showing a forest and non-forest classification (*Figure 1*). The difference between the backscatter from forest areas and clearcuts is due to the change in surface roughness. Because Radarsat operates at C-Band (5,6 cm wavelength), only directional reflectance occurs over a clearcut area compared to mostly diffuse reflectance over forested areas. Thus, Radar-Images are a useful tool to detect clearcut areas in

Figure 1: RADARSAT-Image of Whitecourt, Alberta, and a unsupervised classification showing clearcuttings in boreal forest (RADARSAT image copyright Canadian Space Agency, 1996).



boreal forests. However, their utility depends on how well a user of the map can rely on these classification results.

Users with a variety of applications should be able to evaluate whether the accuracy of the map suits their objectives or not (Aronoff, 1982). Therefore error matrices, also known as confusion matrices (see 4.2, have become a widely accepted method to report the error of raster data. Different methods have been developed to evaluate these error matrices (see 4). Non-statistical methods are included, those based on coefficients of agreement and those based on the binomial distribution. Although these methods provide a powerful tool to evaluate error matrices, they all have certain assumptions concerning the collection of data filling the error matrices. It is further assumed that misclassification of a given area can be unambiguously determined (Ginevan, 1979).

The overriding assumption in the entire accuracy assessment procedure is that the error matrix must be representative of the entire area mapped from the remotely sensed data (Congalton, 1988a). The question is whether the proper sampling approach has been used on which future analyses are based (see 3. If this assumption is violated then all the results of the accuracy assessment are void. Therefore, not only the error matrix itself has to be evaluated but also the whole procedure of data collection for the accuracy assessment.

Congalton (1991) suggests that the following factors should be considered:

- error sources;

- sampling scheme;
    - sample scheme,
    - number of samples,
    - sample unit (ground data collection and sample size).

Each of these factors provide essential information for the quality of accuracy assessment. There are many opinions regarding accuracy assessment approaches. The methodology should be chosen in accordance with the specific objectives and requirements of an investigation.

In many investigations not all ground reference areas collected for classification must be used for training areas. Consequently, it is common practice to use these areas for accuracy assessment. These test areas, though, are designed to work as reference areas for the image classification and as such present large homogeneous areas. However, reference data derived from these homogeneous areas might not provide a valid indication of classification accuracy of land cover variability at the individual pixel level. The accuracy obtained from such reference data can represent at least a first approximation to classification performance throughout the scene (Lillesand and Kiefer, 1994).

1.3 Forest Inventory Systems

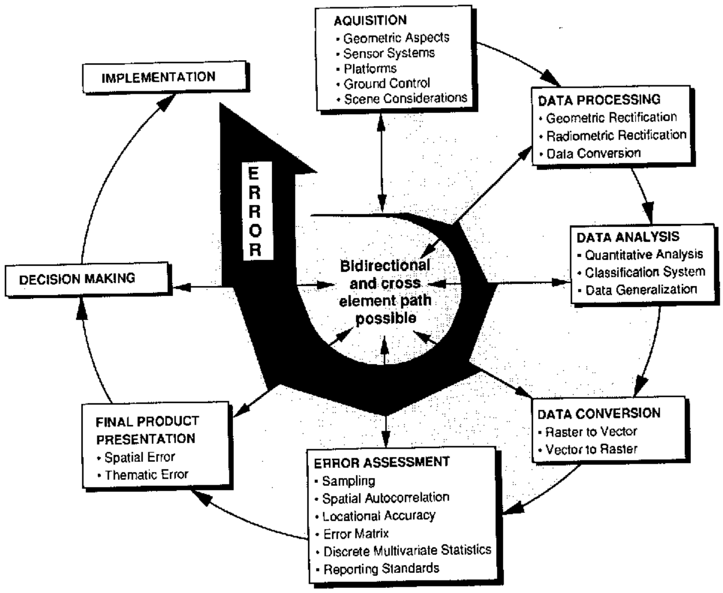
Accuracy assessment receives special attention due to the fact that remote sensing data are a prefered information source for national forest inventories (e.g., Finland). For such inventory systems the accuracy of every data acquisition level must be known. Although most of the national European forest inventories are not yet based on remote sensing data, great efforts are currently underway to integrate this data into the process of collecting information on forests. The current European approaches for national forest inventories are discussed in 5.

2 Error Sources

Although this review focuses on the thematic aspect of accuracy, one should bear in mind that the thematic error is only one error source out of a variety of spatial error data sources. Lunetta *et al.* (1991) list the errors associated with GIS data acquisition, processing, analysis, conversion, and final product presentation (*Figure 2*). They stated: “In theory, the amount of error entering the system at each step can be estimated. In practice, however, error is typically only assessed at the conclusion of data analysis, if it is assessed at all”.

Constraints in both time and cost lead to a two-step accuracy assessment in most projects. The first step is usually to evaluate the geometric accuracy and to assess the thematic accuracy afterwards. This approach considers the two main transformations of remote sensing images (Janssen and van der Wel, 1994):

Figure 2: Error sources and accumulation of error in a typical remote sensing information process flow (Lunetta *et al.*, 1991).



1. **Registration** of the image coordinate system into a certain map projection, enabling other geodata to be used; and

2. **Classification** of the continuum of spectral data into nominal user-desired classes (the most subjective transformation).

Geometric accuracy assessment benefits from the experience available in photogrammetry. These methods to assess geometric accuracy are already well defined and can be operationally used. Non-parametric methods are generally accepted as the most realistic option as orbital geometry models used to describe the errors (parametric method) are still incomplete and cause geometric distortions (Janssen and van der Wel, 1994).

Ground Control Points (GCPs) are used to calculate a transformation from an image coordinate system into the specific ground coordinate system. The accuracy assessment is done by calculating the root-mean-square (RMS) error. It is combined from the error in x-direction (RMS$_x$) and the error in y-direction (RMS$_y$):

$$RMS_{xy} = \sqrt{RMS_x^2 + RMS_y^2} \tag{1}$$

where the RMS error in one direction can be calculated as:

$$RMS_x = \left[\frac{1}{n} \sum_{i=1}^{n} (\delta_{xi})^2 \right]^{\frac{1}{2}} \tag{2}$$

where δ_{xi} = the residual of the i^th GCP and n = the number of GCPs. Statistically, it is more sound to calculate a standard deviation. The sum of the residuals is divided by the redundancy (r) which depends on the degree of freedom determined by the applied polynomial

$$s_x = \left[\frac{1}{n} \sum_{i=1} n(\delta_{xi})^2 \right]^{\frac{1}{2}} \tag{3}$$

for large numbers of GCPs the RMS error and standard deviation converge.

There are many methods to incorporate the positional uncertainty in field surveys and subsequent cross-tabulation of reference and classified data. A distinction between positional and thematic accuracy can be achieved by involving the contextual information in the identification of the land-cover class of a pixel or to use only homogeneous areas for sample points (Warren *et al.*, 1990).

The US National Map Accuracy specifications (NMAS) for cartographic products are used as a reference for the allowable error in many applications (Welch *et al.*, 1985). An accurate map in the terms of the US NMAS must fulfill the following conditions: "For maps on publication scales larger than 1:20.000 not more than 10 percent of the points tested shall be in error more than 0,8 mm (1/30 inch), measured on the publication scale; for maps on publication scales of 1:20.000 or smaller, 0,5 mm (1/50 inch)" (Hord and Bronner, 1976).

3 Sampling Scheme

3.1 Sample Design

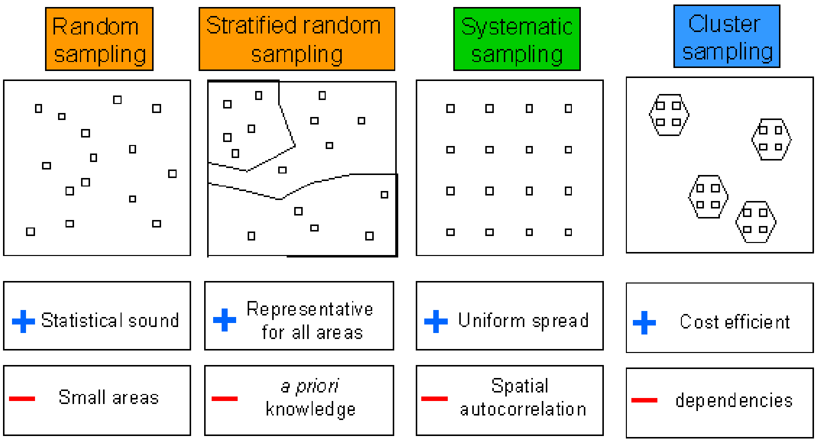
Assessing the accuracy of maps derived from remote sensing data is both time- and money-consuming. Due to the fact that it is not possible to check whole mapped areas, sampling becomes the means by which the accuracy of land-cover maps can be derived (Congalton, 1988a). As stated by Ginevan (1979) any sampling scheme should satisfy three criteria:

1. It should have a low probability of accepting a map of low accuracy.
2. It should have a high probability of accepting a map of high accuracy.
3. It should require a minimum number, N, of ground truth samples.

Therefore researchers have published formulas to calculate the numbers of sample plots which are dependent on the objectives of the project (van Genderen and Lock, 1977; Rosenfield, Fitzpatrick-Lins and Ling 1982; Rosenfield, 1982; Congalton, 1991). These formulas are discussed in 3.3. The sampling schemes that have been used are:

- Simple Random Sampling (SRS).
- Stratified Random Sampling (STRAT).

Figure 3: Sampling designs used for accuracy assessment and rough evaluation of the different approaches.



- Systematic Sampling (SYS).
- Stratified Systematic Unaligned Sampling (SSUS).
- Cluster Sampling (CLUSTER).

Figure 3 illustrates the various sampling approaches and lists their major advantages and drawbacks. The choice of sampling technique will depend upon several factors, including the size of the study area, the type and distribution of features being mapped, and the costs of acquiring verification data.

3.1.1 Random sampling

The simple random sampling (SRS) yields too many samples in larger areas and too few samples in smaller areas (Congalton, 1988a). As the SRS is area-weighted, it is generally accepted that some kind of stratified sampling should be used, thereby ensuring that each class is adequately tested (Aronoff, 1985). The definition of strata requires knowledge of the population that will be assessed; classification of the remotely sensed data must therefore be performed before field verification (Fenstermaker, 1991). This can lead, in some projects, to serious problems because of the temporal change of land-cover between the time of image acquisition and field verification (Congalton, 1991). In these cases, only a spatial random distribution can be used for sampling as it cannot be based on the distribution of the individual classes (Janssen and van der Wel, 1994).

3.2 Systematic Sampling

In systematic sampling approaches, the sampling unit is selected at an equal interval over space. The advantage of systematic sampling is the convenience of obtaining

Table 1: Sample schemes discussed by different authors.

SRS[1]	STRAS[2]	SYS[3]	SSUS[4]	CLUSTER[5]
Congalton, 1988a	Card, 1982	Congalton, 1988a	Berry and Baker, 1968	Congalton 1998a
Aronoff, 1985	Hay, 1979	Warren *et al.*, 1990	Aronoff, 1985	Stehman, 1997
	Ginevan, 1979			Janssen *et al.*, 1994
	Congalton 1988a			
	Fitzpatrick-Lins, 1981			
	Hord and Bronner, 1976			
	Van Genderen and Lock, 1977			

1 = simple random sampling, 2 = stratified random sampling, 3 = systematic sampling, 4 = stratified systematic unaligned sampling, 5 = cluster sampling.

the sample and the uniform spread of the sampled observations over the entire population (Cochran, 1977). An obvious problem in systematic sampled population is the bias that exists if the population shows some kind of spatial autocorrelation. If the presence, absence, or degree of certain characteristics affects those in neighboring units, then the phenomenon is said to exhibit spatial autocorrelation (Cliff and Ord, 1973). Work by Congalton (1988b) on Landsat MSS data from three areas of varying spatial diversity (agricultural land, range land, and a forest site) showed a positive influence, as much as 30 pixels. If spatial autocorrelation analysis indicates periodicity within the data, then the use of systematic sampling schemes may result in poor estimates of classification accuracy (Fenstermaker, 1991).

3.2.1 Stratified systematic unaligned sampling

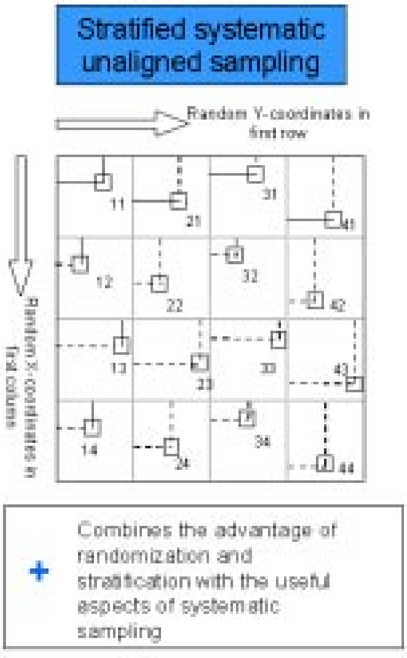
A systematic sampling ensures that sample points of one class are sampled from the entire area (see *Figure 4*). This assumes that the class areas are randomly distributed over the area, but commonly most classes show some form of clumped distribution (e.g., urban areas), or regular distribution (e.g., regular road network) (Aronoff, 1985). If the distribution of the polygons tends toward a direction parallel to the transects of the systematic sampling, a significant bias can be introduced. An unaligned systematic sample can be used to eliminate this bias. As described by Berry and Baker (1986), a stratified systematic unaligned sampling combines the advantage of randomization and stratification with the useful aspects of systematic sampling, while avoiding the possibilities of bias due to the presence of periodicities.

Table 1 lists the use of different sample schemes by author.

3.2.2 Cluster sampling

Cluster sampling is a technique of sampling in which units are not single pixels but groups of pixels. The idea is that it is much easier and cheaper to visit a few large areas than many small areas. Congalton (1988b) suggest that the rate of homogeneity, a coefficient of intraclass correlation, determines whether the chosen clusters are useful for accuracy assessment. The more heterogeneous the pixels within one cluster, the higher the intraclass coefficient; which is favorable when

Figure 4: Stratified systematic unaligned sampling approach.



Stratified systematic unaligned sampling

Random Y-coordinates in first row
Random X-coordinates in first column
11
21
31
41
12
22
32
42
13
23
33
43
14
24
34
44
+

Combines the advantage of randomization and stratification with the useful aspects of systematic sampling

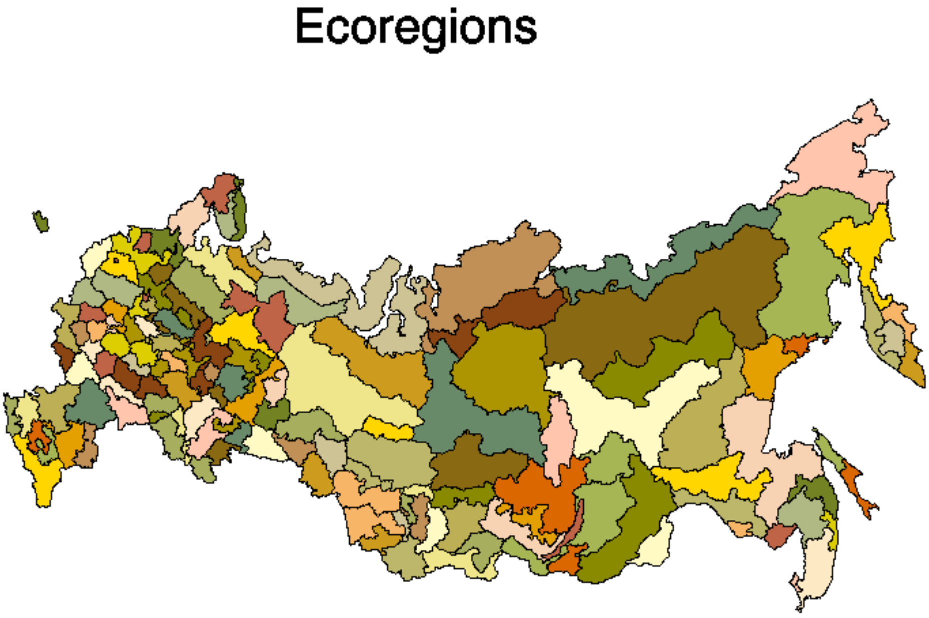
using cluster sampling. The size of the clusters should be smaller than 10 pixels and should never exceed 25 pixels.

3.2.3 Recommendations for the sampling design

Congalton (1991) suggested a combination of stratified and random sampling. The stratified sampling can be done in conjunction with training data collection in an early phase of the project. After the first classification results, stratified random sampling completes the data collection necessary for accuracy assessment. Fenstermaker (1991) proposes a multistage sample approach for large area sampling. Ecoregion types partition the area in the first phase of sampling. The stratifying of a large population into homogeneous primary, secondary, etc., strata enables the description of the entire population with a smaller number of samples. For the sampling of the whole Russian area one should use an *apriori* stratification. This stratification can be made on existing auxiliary data, e.g., based on ecoregions. Ecoregions represent aggregated geographical information based on several factors like climate, human impact, hydrology, etc. *Figure 5* demonstrates one possible stratification of the Russian territory.

Within those strata covering representative areas all classes are selected and stratified random sampled, ensuring at least 50 points per class. At each verification site, an area at least 3 × 3 pixels in size is examined. For the purpose of using the Kappa coefficient (see 4.4) in the analysis of the confusion matrix, a random sampling approach should be chosen.

Figure 5: Ecoregions of Russia.



3.3 Number of Samples

Once the sample scheme has been fixed, the exact number of samples to be taken should be decided. The number of samples is a compromise between the effort to minimize the costs of field sampling and the requirement of a minimum sample size to be representative and statistically sound. In general, the larger the sample size, the greater the confidence one can have in assessments based on that sample (Dicks and Lo, 1990). Depending on the goal of the accuracy assessment the number of sample plots can be calculated with different methods.

If only a right and wrong assessment is needed then binomial distribution may be used to calculate the sample size (van Genderen and Lock, 1977; Hay, 1979; Rosenfield, Fitzpatrick-Lins and Ling, 1982). However, if the project objective is to test not only right versus wrong but also to look at the multiple classes of wrong then a multinomial distribution should be used to calculate the sample size (Rosenfield, 1982).

3.3.1 Binomial distribution

The binomial probability density function can be used to calculate the number of pixels for wrong and right assessment.

$$f(Y,N,Q) = \frac{N!}{(N-Y)!Y!}Q^{N-Y}(1-Q)^{Y} \tag{4}$$

This function describes the probability of getting Y misclassifications in a sample of N drawn from a population with a parametric accuracy proportion Q (Ginevan, 1979). However, these techniques are not designed to choose a sample size for filling in an error matrix (Congalton, 1991). Formulas for the extension of the binomial distribution — the multinomial distribution — can be found in Rosenfield (1982).

In an investigation to compare sampling procedures Fitzpatrick-Lins (1981) used the following equation (5) to calculate the sample number (cumulative binomial probability distribution):

$$N = \frac{Z^2pq}{E^2} \qquad (Z=2) \tag{5}$$

where p is the expected percent accuracy, $q = 100 - p$ and E is the allowable error. Congalton (1991) remarks on this work that it is not possible to fill an error matrix of 30 categories with 319 samples resulting from equation 5. Only 35 of the 900 cells had a greater value than zero. These computations allow only the calculation of overall accuracy.

3.3.2 Rule of thumb

As a rule of thumb Congalton (1991) recommends at least 50 samples per class. If the area exceeds 500 km^2 or the number of categories is more then 12, than at least 75–100 samples should be taken per class. These recommendations coincide with those recommended by Hay (1979) and Fenstermaker (1991). The number of samples for each category might be adjusted based on the relative importance of that category for a particular application. Furthermore, sampling might be allocated with respect to the variability within each category (Congalton, 1991).

3.3.3 Recommendations for the number of samples

As previously demonstrated, various approaches exist to calculate the number of samples for accuracy assessment. Although the *rule of thumb* of having 75–100 samples per class is just an empirical approach, it should be favored. This method provides sufficient measures in each stratum for later calculations. Especially the error-matrix (4.2) and the Kappa-calculation (4.4) require a sufficient number of samples.

3.4 Sample Unit

Sampling units applied to the accuracy testing of maps include points, transects, and areas. Transect units have been used by Skidmore and Turner (1992). The most common units are area units. Aronoff (1985) states that a point unit is, in practice, also an area unit because a point can not be accurately verified. Sampling units should be at least the size of one pixel but, in order to take geometric distortions also into consideration, these sampling units should be more than one pixel. Fenstermaker (1991) recommended, for a multistage approach, a 3 × 3 pixel environment as the sampling unit.

A sampling unit should be at least as large as the minimum mapping unit (Aronoff, 1985). The sampling unit occupies an area and therefore more than one map class can be found within this area. Aronoff (1985) suggests using plurality rules and other more complex methods to describe the class at the test site. Other authors recommend methods to guarantee the homogeneity of the area where the sample unit is located (Warren *et al.*, 1990). One way to overcome this problem is to sample only those pixels whose identity is not influenced by potential registration errors (e.g., points at least several pixels away from a field boundary) (Lillesand and Kiefer, 1994).

4 Error Reporting

4.1 Overview of Error Reporting Methods

The most common way to express classification accuracy is the preparation of a so-called *error matrix* also known as *confusion matrix* or *contingency matrix.* Such matrices show the cross tabulation of the classified land cover and the actual land cover revealed by sample site results. Different measures and statistics can be derived

Table 2: The recommended layout of an error matrix as presented by Congalton (1991).

		Reference Data					
		F	I	U	W	Σ	users acc.
Classified Data	Forest (F)	68	7	3	0	78	87.2%
	Industrial (I)	12	112	15	10	149	75.2%
	Urban (U)	3	9	89	0	101	88.1%
	Water (W)	0	2	5	56	63	88.9%
	Σ	83	130	112	66	391	
	prod. acc.	81.9%	86.2%	79.5%	84.8%		
					overall accuracy:	84%	

from the values in an error matrix. The basic form of an error matrix and non-statistical measures are described in 4.2. Procedures based on multivariate analysis are described in 4.4 and methods based on the binomial distribution are described in 4.5.

4.2 Confusion Matrix

A confusion matrix lists the values for known cover types of the reference data in the *columns* and for the classified data in the *rows*. The main diagonal of the matrix lists the correctly classified pixels. Some confusion exists concerning the layout of the matrix. Most researchers use a layout as demonstrated in *Table 2*.

One basic accuracy measure is the **overall accuracy**, which is calculated by dividing the correctly classified pixels (sum of the values in the main diagonal) by the total number of pixels checked. Besides the overall accuracy, classification accuracy of individual classes can be calculated in a similar manner. Two approaches are possible:

- user's accuracy, and
- producer's accuracy.

The **producer's accuracy** is derived by dividing the number of correct pixels in one class divided by the total number of pixels as derived from reference data (column total in *Table 2*. The producer's accuracy measures how well a certain area has been classified. It includes the error of omission which refers to the proportion of observed features on the ground that are not classified in the map. The more errors of omission exist, the lower the producer's accuracy.

$$producer's\ accuracy(\%) = 100\% - error\ of\ omission(\%) \tag{6}$$

If the correct classified pixels in a class are divided by the total number of pixels that were classified in that class, this measure is called **user's accuracy**. The user's accuracy is therefore a measure of the reliability of the map. It informs the user how well the map represents what is really on the ground. One class in the map can have two types of classes on the ground. The 'right' class, which refers to the same land-cover-class in the map and on the ground, and 'wrong' classes, which show a different land-cover on the ground than predicted on the map. The latter classes are referred to as errors of commission. The more errors of commission exist, the lower the user's accuracy.

$$user's\ accuracy(\%) = 100\% - error\ of\ commission(\%) \tag{7}$$

The difference between these two measures is quite substantial and will be discussed with an example (*Figure 6*). In this example a forested area was classified into four different classes: coniferous, mixed, and deciduous forest, and non-forest. Let us assume that two persons would like to get information from the classified map. One is a forest-owner who wants to know if his forest is actually mapped; the other is a biologist, who would like to investigate a coniferous o forest and must, therefore, plan a field-trip using the map.

Given the example's probabilities, the coniferous forest has 81 percent probability (producer's probability) of being classified as an coniferous forest on the map. This means that almost one fifth of all coniferous forests have not been mapped as coniferous forests and almost 20 percent of all the forest owner's property will not be mapped. On the other hand, the biologist will be more successful. Due to a user's accuracy of 98 percent, he will encounter a coniferous forest in almost every case when selecting a point on the map. He will be disappointed in only 2 percent of all of his visits in the field because he will not find a coniferous stand on the specific place marked on the map.

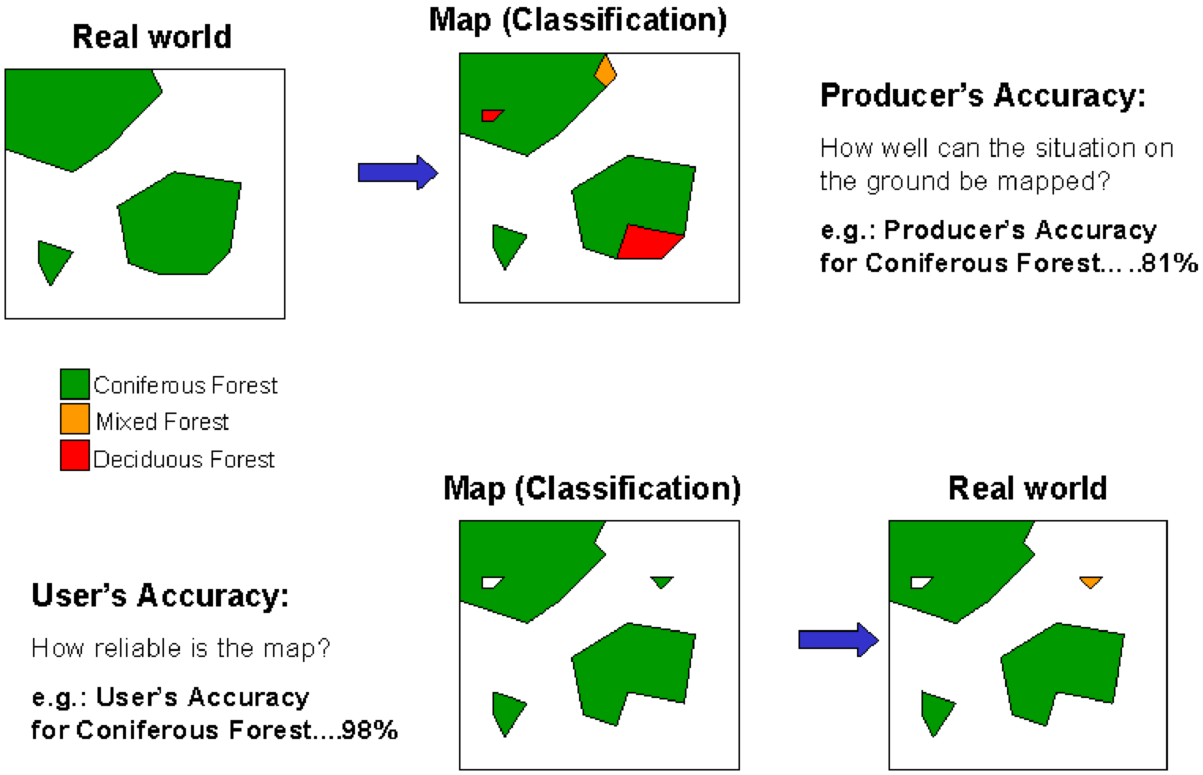
4.3 Matrix Normalization

With the descriptive techniques described in 4.2 it is rather impossible to compare matrices generated from different numbers of samples or to compare cellvalues of matrices derived from different interpretation or classification approaches. Congalton (1991) developed a method to *"normalize"* the values. This technique forces the sum of the rows and the sum of the columns to one. The values in the rows and in the columns are iteratively balanced. This method should only be used if there are not too many zero cell values, because the algorithm will change these values. The **normalized accuracy** can then be computed analogue to the overall accuracy.

4.4 Kappa Coefficient

The Kappa coefficient is a measure of overall agreement of a matrix. In contrast to the overall accuracy — the ratio of the sum of diagonal values to total number of cell counts in the matrix — the Kappa coefficient takes also non-diagonal elements into account (Rosenfield and Fitzpatrick, 1986).

Figure 6: Examples for user's and producer's accuracy.



Developed by Cohen (1960) the Kappa coefficient measures the proportion of agreement after chance agreements have been removed from considerations. Kappa increases to one as chance agreement decreases and becomes negative as less than chance agreement occurs. A Kappa of zero occurs when the agreement between classified data and verification data equals chance agreement (Fenstermaker, 1991). The Kappa coefficient was introduced to the remote sensing community in the early 1980s (Congalton and Mead, 1983; Congalton *et al.*, 1983) and has become a widely used measure for classification accuracy. It was recommended as a standard by Rosenfield and Fitzpatrick-Lins (1986).

Due to a typographical error in Congalton *et al.*, (1983) there has been some confusion about the correct computation of the Kappa coefficient. Hudson and Ramm (1987) have clarified this confusion. Most articles cite Bishop *et al.* (1975) as a source of formulation:

$$\hat{K} = \frac{N \sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r} X_{i+}X_{+i}}{N^2 - \sum_{i=1}^{r} X_{i+}X_{+i}} \tag{8}$$

where

r = number of rows and columns in error matrix,
N = total number of observations,
X_{ii} = observation in row i and column i,
X_{i+} = marginal total of row i, and
X_{+i} = marginal total of column i.

To interpret the formula of the kappa coefficient the following formulation is more useful:

$$\hat{K} = \frac{p_0 - p_e}{1 - p_e} \tag{9}$$

where

p_0 = accuracy of observed agreement, $\frac{\sum X_{ii}}{N}$,
p_c = estimate of chance agreement, $\frac{\sum X_{i+}X_{+i}}{N^2}$,

It should be stated that the Kappa coefficient assumes a multinomial sampling model (e.g., a simple random sampling). The influence of sampling schemes other than simple random sampling have not been investigated and, as Congalton (1991) noted, "an interesting project would be to test the effect on the kappa analysis of using a sampling scheme other than simple random sampling". Nevertheless, researchers have also used the Kappa coefficient for other sampling models (e.g., Dicks and Lo, 1990). Stehman (1997) provides formulas for the calculation of the Kappa coefficient under cluster sampling. There are only a few authors expressing critical remarks on the calculation of the Kappa coefficient with equation 8. Foody (1992) states that the degree of chance agreement may be overestimated. He proposes, as an alternative, a Kappa-like approach discussed by Brennan and Prediger (1981). The probability of chance agreement is considered to be $1/n$ and the alternative to the Kappa coefficient is calculated as defined in equation 10.

$$k_n = \frac{p_0 - 1/n}{1 - 1/n} \tag{10}$$

The approximate large sample variance of Kappa, $\hat{\sigma}^2$, can then be used to construct a hypothesis test for significant difference between error matrices (Cohen, 1960). The equation for computing the variance of Kappa can be formulated as follows (Hudson and Ramm, 1987):

$$\hat{\sigma}^2[\hat{K}] = \frac{1}{N}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^3(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}\right] \tag{11}$$

where

$\theta_1 = \sum\limits_{i=1}^{r} \frac{X_{ii}}{N}$,
$\theta_2 = \sum\limits_{i=1}^{r} \frac{X_{i+}X_{+i}}{N^2}$,
$\theta_3 = \sum\limits_{i=1}^{r} \frac{X_{ii}(X_{i+}+X_{+i})}{N^2}$, and
$\theta_4 = \sum\limits_{i=1,j=1}^{r} \frac{X_{ij}(X_{i+}+X_{+i})^2}{N^3}$.

Cohen (1960) describes the test of significance between two independent Kappa's by:

$$Z = \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma_1} + \hat{\sigma_2}}} \tag{12}$$

This test is possible because the large sample asymptotic distribution of Kappa is normal. It can be used to test whether the classification accuracy differs significantly from chance agreement or if there is significant difference between various classification approaches. To demonstrate the use of the Kappa coefficient data for error matrices of Congalton (1991) are used, which are listed in *Table 3.*

Table 4 provides a comparison of the overall accuracy, the normalized accuracy (as discussed in 4.3) and the Kappa coefficient. All three measures agree about the relative ranking of the two classification approaches. Nevertheless, their absolute values are quite different, because each measure incorporates different levels of information (Congalton, 1991). The Kappa coefficient provides the lowest accuracy

Table 3: Error Matrices (pixel counts) for two classification approaches (Congalton, 1991).

Supervised Approach					
		Reference Data			
		F	I	U	W
	Forest (F)	68	7	3	0
Classified	Industrial (I)	12	112	15	10
Data	Urban (U)	3	9	89	0
	Water (W)	0	2	5	56

Unsupervised Approach					
		Reference Data			
		F	I	U	W
	Forest (F)	60	11	3	4
Classified	Industrial (I)	15	102	14	8
Data	Urban (U)	6	13	90	2
	Water (W)	2	4	5	52

Table 4: Comparison of three accuracy measurements for the supervised and unsupervised classification approach presented in *Table 3.*

Classification Algorithm	Overall Accuracy	Kappa Accuracy	Normalized Accuracy
Supervised classification	84%	77%	83%
Unsupervised classification	78%	70%	78%

Table 5: Results of the Kappa analysis test of significance for individual error matrices and for comparison between the matrices of *Table 3*.

Classification	Kappa Coefficient	Z Statistic	Result[a]
Supervised appr.	0.7687	29.41	S[b]
Unsupervised appr.	0.6956	24.04	S[b]
Superv. vs. unsuperv.		1.8753	NS[b]

[a] At the 95 percent confidence level.

[b] S = significant, NS = not significant.

A significant result of the test means that the results of the error matrix are significantly better than a random result (i.e., the null hypothesis: Kappa = 0)

in this example; but the ranking of the different error measures strongly depends on the form of the error matrix. Many off-diagonal cells with values of zero can force the normalized accuracy to differ significantly from the others. The Kappa coefficient for the supervised classification approach results in a value of 77 percent. This implies that the accuracy of the classification is 77 percent better than the accuracy that would result from a random assignment.

The Kappa coefficient belongs to the family of bivariate agreement coefficients, in the form:

$$Agreement = 1 - \frac{observed\ disagreement}{expected\ disagreement}$$

These agreements, like the Kappa coefficient, are zero for chance agreement, one for perfect agreement, and negative for less than chance agreement. The Kappa coefficient can be used to test accuracy in two ways: between the classified matrix and a random classification, and the difference between two classification approaches (*Table 5*).

Rosenfield and Fitzpatrick-Lins (1986) presented further coefficients to describe category accuracy, which belong to the family of bivariate coefficients. Their review included the following coefficients:

- ground truth index by Turk (1979).
- mean accuracy index by Hellden (1980).
- mapping accuracy index by Short (1982).

As they conclude in their article: "A family of such coefficients exist which correct for chance agreement, but the Kappa coefficient is one of few which also are defensible as intraclass correlation coefficients".

4.5 Procedures Based on Binomial Distribution

The user of remote sensing may be interested on whether the map failed or passed a certain level of accuracy and/or in the confidence intervals of the overall accuracy.

Based on probability density functions, Ginevan (1979) started his work to develop a sound statistical methodology for map accuracy validation. Aronoff (1982, 1985) endorsed this work and developed the **minimum accuracy value**.

The sampling problem defined for the use of the probability density functions is the determination of the optimal number, N, of ground truth samples and an allowable number, X, of misclassifications of these samples (see equation 4). If X of fewer ground truth samples are misclassified the map is accepted as accurate on this confidence level (Ginevan, 1979).

4.5.1 Confidence level

Confidence limits for the **overall accuracy** from the sample size N, and a significance level α can either be read from the standard binomial nomograms as given in statistical textbooks, or calculated using the exact binomial distribution (Janssen and van der Wel, 1994). The confidence limits of the overall accuracy show that this is the center of an interval in which the overall accuracy can be found with $1 - \alpha$ confidence.

4.5.2 Hypothesis testing

For many applications the classification should at least have a minimum overall accuracy. For these cases a statistical method — the hypothesis testing — is appropriate. For hypothesis testing the following parameters are defined:

- H_0...null hypothesis.
- H_1...alternative hypothesis.
- α...significance level.

There are always two types of error. The *type I error* is defined by Aronoff (1982) as the **consumer risk** and the *type II error* is defined as the **producer risk**. It should be noted that the terms consumer's risk and producer's risk have a completely different meaning to producer's and consumer's accuracy.

Consumer Risk (α): is the probability that a map of unacceptable accuracy will pass the accuracy test (wrongly accepting H_0); it has the largest consequence for the user of the map.

Producer Risk (β): is the probability that a map of some acceptable accuracy Q_H will be rejected (wrongly rejecting H_0); it has the largest consequence for the producer of the map.

The consumer's risk can be calculated as follows:

$$CRISK = \sum_{Y=0}^{X} \frac{N!}{Y!(N-Y)!} Q_L^{(N-Y)} (1 - Q_L)^Y \qquad (13)$$

where $CRISK$= consumer risk,
Q_L = the minimum accuracy required,
X = number of allowable misclassifications,
N = total number of points sampled, and
Y = number of misclassifications.

Q_L can be calculated iteratively. The producer's risk can be calculated as follows:

$$PRISK = \sum_{Y=X+1}^{N} Q_H^{(N-Y)} (1 - Q_H)^Y \tag{14}$$

where $PRISK$= producer risk and
Q_H = a selected high accuracy level.

Values of consumer and producer risk have been tabulated by Ginevan (1979) and Aronoff (1985) for specific sample designs (total number of samples and number of misclassifications).

Ginevan used these tables to calculate the sample size to reduce field survey. As Aronoff (1982) stated: "The selection of values for consumer's and producer's risk depend on the value of information and cost of errors in a specific application". Both consumer's and producer's risk should be minimized, which is difficult because of their interdependency: a smaller producer's risk can be obtained by increasing consumer's risk or increasing the number of samples (Janssen and van der Wel, 1994).

The producer's risk is closely related to the number of sample points. Assume a situation demonstrated in *Table 6*. The producer's risk is tabulated for 90 percent, 95 percent and 99 percent map accuracies. The least expensive accuracy test to perform would use the smallest sample size (in this case: 19 points). The producer's risk for a class that has been mapped with an accuracy of 95 percent would be 0.6226. This means that if the map would be repeatedly tested with 19 points the result would fail the test approximately 62 percent of the time. Selecting a higher sample size (e.g., 93 sample points) reduces the producer's risk (as low as 43 percent). High producer risk has additional costs:

- re-checking a sufficiently accurate map; and
- cost of the delay in providing information to the user.

These costs have to be balanced with the costs of increased sampling with lower producer's risk.

Aronoff (1985) introduced the **minimum accuracy** value, which is defined as the lowest expected accuracy of a map given an observed accuracy test result and the

Table 6: Critical value table for a required accuracy of 85 percent with a consumer risk of 0.05 (Aronoff, 1985). The first six columns are for accuracy test design, and the remaining columns are for interpretation purposes of the test result.

			Producer Risk values			Minimum Accuracy Values in					
			(Accuracy levels in%)			Percent for Deviations from X					
X	N	CRISK	90.0%	95.0%	99.0%	-6	. . .	2	4	. . .	26
0	19	0.0456	0.8649	0.6226	0.1738	—	...	70.4	58.0	...	—
1	30	0.0480	0.8163	0.4465	0.0361	—	...	76.1	68.1	...	2.7
2	40	0.0486	0.7772	0.3233	0.0075	—	...	78.5	72.5	...	18.3
3	40	0.0460	0.7497	0.2396	0.0016	—	...	80.1	75.3	...	30.1
:	:	:	:	:	:	:	:	:	:	:	:
7	85	0.0478	0.6247	0.0624	0.0000	93.3	...	82.2	79.4	...	57.7
8	93	0.0496	0.5919	0.0432	0.0000	93.3	...	82.4	79.9	...	54.4
9	102	0.0471	0.5746	0.0318	0.0000	93.3	...	82.7	80.5	...	57.1

X= maximum allowable misclassifications; N=required sample size; CRISK=consumer's risk

user selected consumer risk. It is a useful index to compare the results of accuracy tests using different sample sizes and for use in a loss function for comparing the relative expected maximum cost of alternative classification results.

The tables presented by Aronoff (1985) list the sample size (N), and exact consumer risk ($CRISK$) for critical values (X) — the maximum allowable number of misclassifications. The values of each table are calculated at a specific level of accuracy (producer's risk) (*Table 6*).

The use of the binomial distribution and the hypothesis test methods enables users, when the accuracy test results are represented in an aggregate form (error matrix) to evaluate if the map is suitable for a specific applications.

Aronoff (1985) illustrates the table with the following example: suppose 10 out of 93 points were incorrectly classified. The critical value ($X = 8$) for 93 sample points (N) for a required accuracy of 85 percent is exceeded, the map fails the test. A user may be interested not only on whether the map failed the test, but how it failed or how well it passed the test. The deviation listed in the interpretative columns is the observed misclassifications minus the critical value ($10 - 8 = 2$). The minimum accuracy value is 82.4 percent. This means that there is only a small chance (consumer's risk) of 5 percent that a map with an accuracy level as low as 82.4 percent would give a test result as ten misclassifications out of 93 sampled points.

4.6 Recommendations for the SIBERIA-Error Reporting

Performing Accuracy Assessement should not be the end of a project, but rather the start for a discussion between the producer of spatial information and the user of this information. Confusion matrices are relatively easy to understand and can function as a basis for further accuracy discussion. The calculation of producer's and user's accuracy guarantee an individual accuracy assessment based on the needs of the customers. In addition, the Kappa coefficient is a statistical measurement to test whether the current classification approach is better than a classification solely based on chance agreement or not.

5 Forest Inventory Systems

The methods used for assessing the thematic accuracy of classifications of remote sensing data are closely related to the methods of forest inventory. Forest inventory is defined as both the method of estimating the biomass and the estimates (the inventory results) themselves (Cunia, 1981). Comparable to accuracy assessment, the sampling design, the number of samples and stratification considerations are important aspects of the inventory methodology. Cunia (1981) classifies the inventory systems as *operational, management, or national (or regional).*

Cunia assumes that an operational inventory is intended to supply information regarding the current values of forest biomass, whereas management or national inventories are primarily designed to ensure a continuous flow of information about the general conditions of the forest. According to these objectives the sampling design for the two types of inventory can be quite different. Cunia points out that a sampling design should present both a space and time dimension. The spatial dimension regulates the distribution of the samples in the forest area. The time dimension considers the selection of sampling units over successive points in time. Temporary sample plots are commonly used in operational inventories, because they are most efficient to gain information on current values of biomass. Permanent plot techniques are more expensive than temporary ones, but allow for estimates for rates of change to be derived more precisely from this sampling technique.

5.1 Inventory Sampling Design

5.1.1 Stratification

Stratification is a method to divide the population into smaller units called strata. The homogeneity within the strata is greater than the homogeneity between the different strata. For long term monitoring the strata, one may use prestratification which would have to be defined on permanent geographic units (Cunia, 1981). Due to the change of strata boundaries in time most inventory methods are stratified through poststratification.

One exception is the use of a design named double sampling for stratification. It has been used in the inventory system of the North-eastern United States (Cunia, 1981) and for a resource inventory in Central America (Dorigan, 1981).

5.1.2 Double sampling

In literature, double sampling is also referred to as two-phase sampling. This inventory system belongs to the group of multiphase inventory systems. The basic idea is, that in many cases the variable of interest is rather expensive to assess. But if these variables of interest exhibit a strong relationship to a so-called auxiliary variable, then it is economically much more effective to sample the auxiliary variable in a first level and to specify the variable of interest in a second level.

Remote sensing data is the ideal solution for data assessment of the first phase, as it considerably reduces the assessment cost for auxiliary variables compared to the assessment cost of the variable of interest using non-remote sensing means (European Commission, 1997). Double sampling is used in the national forest inventories of Belgium, France, Greece, the Netherlands, Portugal (two-stage sampling), Switzerland, the United Kingdom (two-stage sampling), and for the Northern part of Finland.

Double Sampling can be divided into:

- Double sampling for stratification; and
- Double sampling for regression estimators.

Double sampling for stratification. The basic idea of the double sampling or two-phase sampling is to reduce the number of points for ground measurements. Therefore in the first phase a large number of points are located on remote sensing data and classified according to a previous defined forest strata. The number of points are used to calculate the area coverage of the particular forest strata. A subsample of the points of the first phase are selected and the corresponding area on the ground is measured for tree and stand characteristics. The measurement on these ground plots is used to estimate the means and variances within each stratum. The number of points can be calculated according to Neyman's optimum allocation formula (Cunia, 1981).

Double sampling for regression estimators. This double sampling method is useful if an attribute, which is costly to assess, is closely correlated to a variable which can be economically assessed.

5.2 Permanent vs. Temporary Plots

There are currently three different methods concerning the time dimension of a forest inventory:

1. permanent plots — Continuous Forest Inventory (CFI);
2. permanent and temporary plots — Sampling with Partial Replacement (SPR); and
3. temporary plots.

5.2.1 Permanent plots

The continuous forest inventory (CFI) consists of regularly spaced permanent sample plots. The design is efficient for estimating both current values and rates of

Table 7: Plot design in current European forest inventories (European Commission, 1997).

	Plots	
Permanent	Mixed	Temporary
Austria	Finland	France
Belgium	Norway	Greece
Germany	Sweden	Iceland
Italy	Switzerland	Ireland
Netherlands		Portugal
Spain		United Kingdom

change. It is also expensive, however, since the permanent plots must be well maintained in the field, and the same sample plots are located and measured on every occasion. Another disadvantage is that the system can not easily be adapted to future estimates of different precision (Cunia, 1981). Besides the European countries listed in *Table 7* the CFI was widely applied for more than 40 years in the United Sates, Canada, and Mexico.

5.2.2 Permanent and temporary plots

The efficiency of the CFI can be improved by remeasuring only some of the permanent plots and installing a set of new plots. This system makes use of the basic statistical concept *sampling with partial replacement (SPR)*. It is used in Scandinavia and Switzerland.

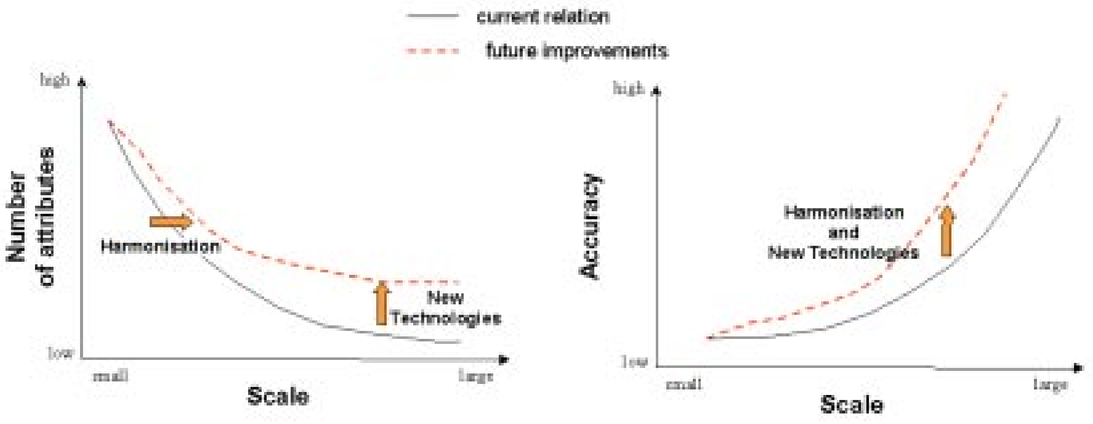
Table 7 lists the currently used plot designs in European forest inventories (European Commission, 1997).

Besides the use of two-phase sampling techniques, many researchers demonstrated the usefulness of the sampling with partial replacement method also for large-scale forest inventories. Most of them used aerial photos in the scale between 1:6000 and 1:25000 (Akça, 1989 and 1995, Kätsch, 1990, Wolff, 1992). Akça (1995) showed that two-phase sampling required only 74 samples to achieve the same accuracy in timber volume estimation, as the 130 samples needed for terrestrial sampling. The conversion from aerial sampling plots to terrestrial sampling plots is calculated using a cost ratio of 1:15 (aerial plot vs. terrestrial plot).

6 The Use of Remote Sensing Data in Forest Inventories

In most of the national forest inventories of the European countries remote sensing data is already used or will be used in the next inventory period. Until now satellite data has been used in the national forest inventory of Northern Finland and in the regional inventory of Liguria (Italy). Because of the importance of the Finnish

Figure 7: Effect of harmonization and new technologies on the number of attributes and the accuracy on different scales (adapted from Köhl and Päivinen, 1996).



Multi-Source Inventory for applying remote sensing data, this inventory system will be described in 6.3. In general, additional data received from remote sensing sources can be used in several ways in the inventories (Köhl and Päivinen, 1996):

- for stratification for field inventory designs;
- for parameter estimating; and
- for map production.

The objective for using remote sensing data in forest inventory is to improve the accuracy and efficency of estimates of forest attributes especially for broad scales. Köhl and Päivinen (1996) state that the number of attributes for which results can be estimated depend on the scale of the assessment. Thus, global inventories report only a small number of attributes. There are two possibilities to improve this situation:

- Harmonization, and
- New technologies (remote sensing).

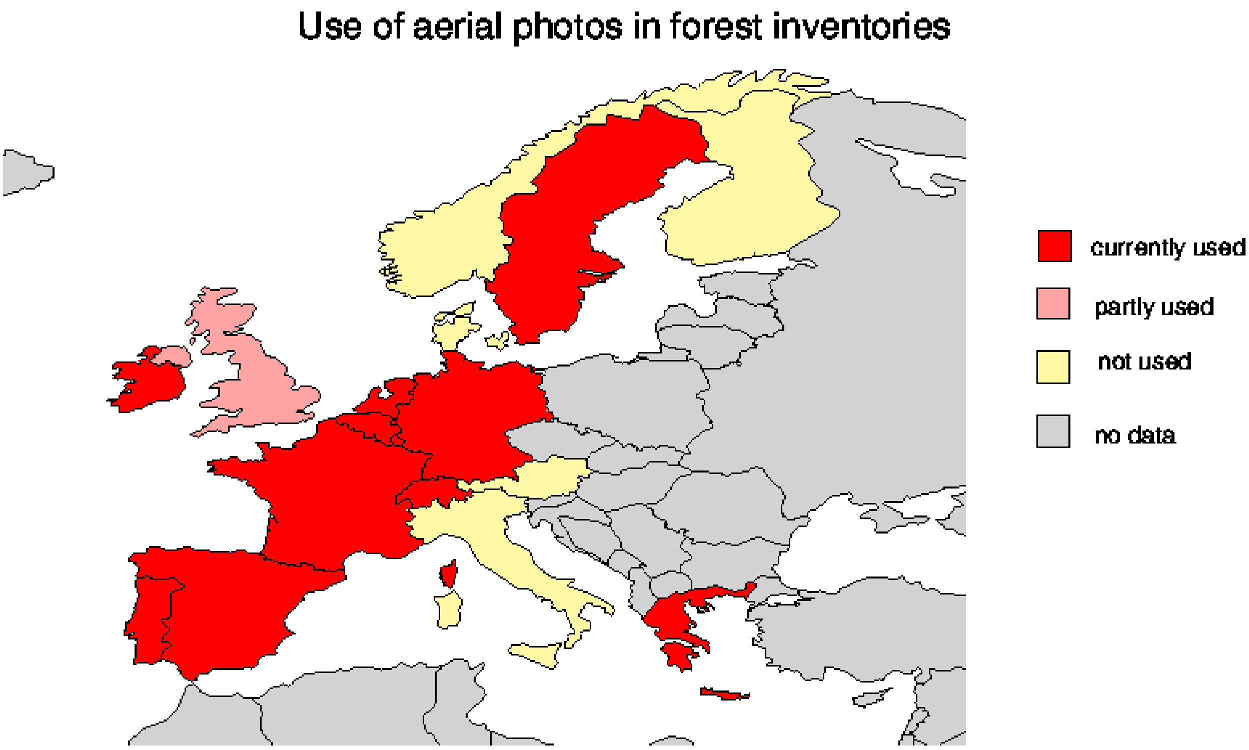
The relation between the scale and both the number of attributes and the accuracy can be seen in *Figure 7*.

In their review Köhl and Päivinen (1996) summarize the feasibility of remote sensing techniques for the assessment of different forest attributes. The results are presented for different forest attributes, sensor resolutions, and different area units.

6.1 Use of Aerial Photos

The main use of aerial photos has been to classify forest and non-forest areas, or in the first phase of a double phase sampling approach (*Figure 8*).

Figure 8: Use of aerial photos in national forest inventories (European Commission, 1997).



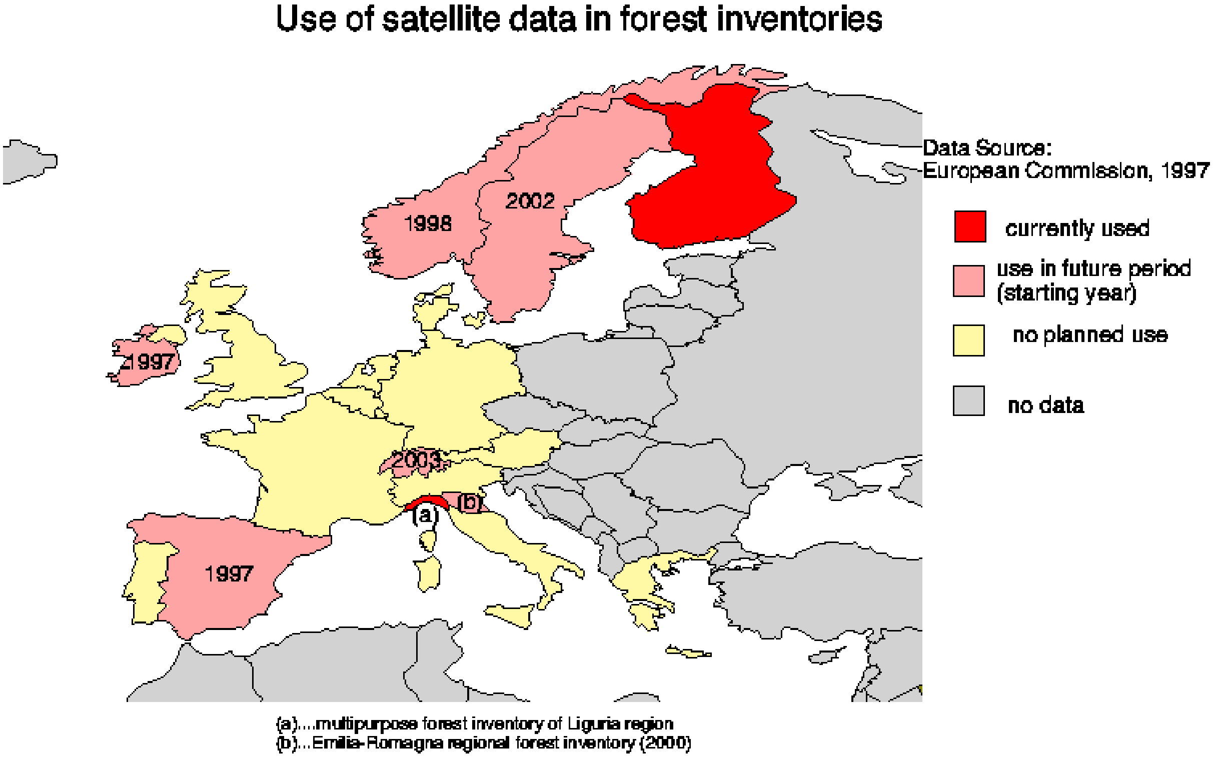
6.2 Use of Satellite Data

Satellite images are becoming an increasingly important data source in forest inventories. The use of satellite images will be considered in the inventory of Private Woodlands in Ireland, in Italy's National Forest Inventory and regional inventory of Emilia Romagna, and national forest inventories of Norway, Spain, Sweden, and Switzerland (see *Figure 9*).

6.2.1 Italy

Landsat-TM data is already used in the multipurpose forest inventory of the Liguria Region. A proposal for creating an inventory structure for the national forest inventory enabling the acquisition and integration of information from different sources at various levels has been made. For the regional forest inventory of Emilia-Romagna

Figure 9: Use of satellite data in national forest inventories (European Commission, 1997).


Use of satellite data in forest inventories
Data Source:
European Commission, 1997
currently used
use in future period (starting year)
no planned use
no data
1998
2002
1997
2003
(a)
(b)
1997
(a)....multipurpose forest inventory of Liguria region
(b)...Emilia-Romagna regional forest inventory (2000)

the integration with satellite and other remote sensing data is highly desirable and should be deeply investigated (Tosi and Marchetti, 1997).

6.2.2 Norway

Experiments will be carried out in the near future to evaluate the possibility of combining satellite images and other georeferenced information with sample plot data. GPS will be used for an exact location of the sample plots (Tomter, 1997). However in a case study in Norway, Gjertsen (1996) came to the conclusion that the use of satellite data (Landsat-TM and SPOT) in forest inventory do not yet meet national standards. The potential of volume stratification through satellite data within a two-phase sampling approach was tested.

6.2.3 Spain

The possibility of using of satellite images for the actualization of forest areas will be studied in the next inventory period. Due to recent changes in the Forest Administrations Structure, however, the funding of the third inventory period is not yet guaranteed. The sampling design is stratified systematically, in which the strata are derived from existing maps by grouping polygons of the forest map. Principal criteria for the stratification are species and forest types (Martínez-Millán, 1997).

6.2.4 Sweden

The eighth Swedish National Forest Inventory is planned to be conducted between 2003 and 2012. New information demands on biodiversity and other environmental information require new methods for field sampling and combination with remote sensing (Söderberg, 1997). Currently the sampling design is a systematic cluster sampling with partial replacement of plots. The attributes derived from aerial photos are used for the simulation of real field plot data. Aerial photos are primarily used for the assessment of plots in high mountains.

6.2.5 Switzerland

Satellite imagery might be used in addition to aerial photography. Aerial photography is combined with field-assessed attributes by a double sampling for stratification approach. The current interpretation of aerial photographs by analytical instruments is likely to be replaced by digital photogrammetry (Köhl, 1997).

6.2.6 Other countries

There are also investigations to establish remote sensing technologies as a basis for the national inventory and for the system of monitoring the forest status in Croatia (Kusan, 1995). And as Hocevar *et al.*, (1995) stressed the monitoring system which includes the analysis of satellite images, enforced in the Slovenian forestry, seems to be an important contribution to the Slovenian environmental science and policy. For Russia satellite images have been used since the 1960s as a forest information source. Besides the *lesoustroistvo* — the forest inventory and planning (FIP) — they are the most important data source for the Russian State Forest Account (SFA) (Shvidenko and Nilsson, 1997).

6.3 Forest Inventory in Finland

In addition to the traditional *field inventory*, a new inventory system was developed in 1989 (eighth inventory) and applied in the Northern part of Finland. The so-called *Multi-Source National Forest Inventory* is a remote sensing and digital map information aided extension of the field inventory, allowing an accurate estimation of the results for small areas (Tomppo *et al.*, 1997). So far, one fifth of the former temporary plots have been substituted by permanent plots. Image analysis methods are applied in a way that estimates of forest variables of the inventory are calculated for each pixel (e.g., growing stock volume by tree species, increments by tree species, site fertility class and mean age).

The basic classification method is a **K-nearest neighbor classification** (Tomppo, 1996). Field sample plots are used to form so-called *plot pixels* which are used as ground truth information for the classification. The classification contains both a feature and a geographic dimension, and can be interpreted in such a way that each pixel represents a proportion of each of the K-neighboring pixels. The k-nearest pixels are derived by calculating the Euclidian distance $d_{p(i),p}$ in the spectral space of the satellite image channels from the pixel p (to be classified) to each field plot pixel $p(i)$ within a radius of 50 to 100km. The difference in elevation is restricted to less than 50–100 m. The limitations in the geographic space avoid using sample plots from different vegetation zones (Tomppo, 1996). The k-nearest field plot pixels are denoted by $p(1), ..., p(k)$. The field plot pixels are weighted inversely proportionally to the squared distance in feature space.

$$w_{i,p} = \frac{\sum\limits_{j=1}^{k} d^2_{p(i),p}}{d^2_{p(i),p}} \tag{15}$$

So the weight for each field plot in a certain stratum (municipality), for matters of area calculations of the variables, is calculated by summation over the weights of each pixel in the stratum:

$$c_i = \sum_p w_{i,p} \tag{16}$$

The forest variable M can be estimated for the pixel p as:

$$\hat{m}_p = \sum_{j=1}^{n} w_{j,p} \times m_{j,p} \tag{17}$$

where $m_{j,p}, j = 1, ..., k$ is the value of the variable M in the sample plot j corresponding the pixel p_j which is the j^th closest pixel in the spectral space to the pixel p (Tomppo, 1991).

6.4 Examples for Combining Traditional Field Inventories with Remote Sensing Methods

Additional data sources are important for forest inventory for various reasons. Since field inventories constitute the major part of forest mapping costs (Hagner, 1990), one interest is to reduce this cost through the combination of traditional field inventories and remote sensing methods.

6.4.1 Compartment based methods

Although Gjertsen (1996) published non-promising results for the extraction of forest variables on a per pixel basis, Hagner (1990) developed a method to improve traditional forest variables estimates on a per stand basis.

Hagner (1990) used merged SPOT-data for the purpose of segmentation. A t-ratio method was applied for the merging of regions after the initial segmentation. This method tests the hypothesis that the spectral intensities of the two regions are in fact the observations on the same population (equation 18).

$$t - ratio = \frac{\bar{X}_1 - \bar{X}_2}{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^{\frac{1}{2}}} \tag{18}$$

For multiple bands Hagner (1990) uses a value calculated as the square-root of summed squared t-ratios:

$$T^2 = \sum j = 1bt_j^b \tag{19}$$

where t is the $t - ratio$ in band j and
b is the number of bands.

The results of the segmentation are comparable to visual interpretation. Regression analysis was carried out using NFI-plots (National Forest Inventory of Sweden). The models used are described in equation 20.

$$log(y) = b_o + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n \tag{20}$$

where y represents the forest variables mentioned in the following, and x_i are the band intensities of the original bands or derivates. On the basis of the stand, the following estimates were determined:

- volume/ha;
- mean diameter;
- mean age; and
- tree species mixture.

The average value of the pixelwise calculations within each stand were corrected for the logarithmic bias of the regression with a correction factor obtained from the regression analysis and the mean of real observed values were divided by the mean of real predicted values.

The standwise calculations were combined with traditional field inventory data. The combination of the two data sources could improve the result under the assumption of uncorrelated errors. The estimates of both the field data and the remote sensing data are inversely weighted according to their variance. With this method other auxiliary data such as previous field inventory data and photo interpretation can also be included in the combined estimates. The accuracy of estimates obtained from satellite and NFI-data were found to be comparable to those obtained by traditional field inventory, except for tree species composition.

6.4.2 Sample based methods

To overcome several problems linked with the mostly subjective delineation of forest compartments, Poso and Waite (1995) demonstrated a sample-based forest inventory and monitoring system as a substitution or supplement to compartment-based analysis. In this approach a certain number of sample units (points) are systematically selected over the whole area based on required accuracy. The necessary field data for the units are acquired by stratifying the sample plots using auxiliary data sources such as remote sensing data. In Finnish applications the common point distance is 25 m.

6.5 Conclusions for the Russian Forest Inventory

The use of remote sensing data must be strongly recommended for the future National Russian Forest Inventory, although their use in National Forest Inventories is still rather the exception than the rule. However, compared to European forest inventory systems, the Russian system is also the exception, because remote sensing data has been integrated since 1948 (Shvidenko and Nilsson, 1997). Since the 1960s a three-stage stratified sampling procedure has been used to record an average annual area of 10 to 25 million ha. The design of a new inventory system must make use of knowledge and experience that has been gathered in Russia in the fields of remote sensing.

Remote sensing data becomes essential when mapping forest variables of the whole area is of interest. For deriving estimates of forested area, volume/ha, tree species composition, etc., remote sensing data can be integrated in multi-phase forest inventory designs. Multi-phase inventory designs (see 5.1 are most flexible for the integration of different data sources. This aspect is especially important for the integration of data from different sensors, because the variety of remote sensing data will increase in the near future.

A new forest inventory system in Russia should reveal information about the current status of the forest and of the dynamic processes. As expressed in 5.2.1 permanent plots are the ideal design for monitoring changes. The sample design of these permanent plots has to be adapted to the varying information needs. The more or less unmanaged forests in the north will show a rather wide sampling distance compared to forests whose principal function is production of industrial wood. The permanent plots should be designed as cluster samples. In addition to permanent plots, the use of temporary plots enables forest management to be flexible on future information demands (see 5.2.2).

References

[1] Akça, A. (1989) Permanente Luftbildstichprobe. Allg. Forst- und Jagdzeitschrift, 160(4):65–69 (in German).

[2] Akça, A. (1995) Two-phases sampling method using regression estimators and small scale aerial IR-photographs in volume and increment estimation. In: Remote Sensing and Computer Technology for Natural Resource Assessment, (IUFRO XX World Congress), Saramäki-Koch-Gyde Lund (eds.), University of Joensuu, Faculty of Forestry, Research Notes 48, p. 255–264.

[3] Aronoff, S. (1982) The map accuracy report: A user's view. Photogrammetric Engineering and Remote Sensing, 48(8):1309–1312.

[4] Aronoff, S. (1985) The minimum accuracy value as an index of classification accuracy. Photogrammetric Engineering and Remote Sensing, 51(1):99–111.

[5] Berry, B. and Baker, A. (1986) Geographic sampling. Marble (ed.), Spatial Analysis — A Reader in Statistical Geography. Prentice-Hall, Englewood Cliffs, New Jersey.

[6] Bishop, Y., Fienberg, S. and Holland, P. (1975) Discrete Multivariate Analysis — Theory and Practice, MIT Press, Cambridge, MA.

[7] Brennan, R. and Prediger, D. (1981) Coefficient Kappa: Some Uses, Misuses and Alternatives. Educational and Psychological Measurement, 41:687–699.

[8] Campbell, J. (1997) Introduction to Remote Sensing. Guilford Press, New York.

[9] Card, D. (1982) Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing, 48(3):431–439.

[10] Cliff, A. and Ord, J. (1973) Spatial Autocorrelation, Pion Limited, London, England, p. 178.

[11] Cochran, W. (1977) Sampling Techniques. John Wiley & Sons, New York.

[12] Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20, No. 1, 37–40.

[13] Congalton, R. and MEAD, R. (1983) A quantitative method to test for consistency and correctness of photointerpretation. Photogrammetric Engineering and Remote Sensing, 49(1):69–74.

[14] Congalton, R., Oderwald, R. and Mead, R. (1983) Assessing Landsat classification accuracy using discrete multivariate statistical techniques. Photogrammetric Engineering and Remote Sensing, 49(12):1671–1678.

[15] Congalton, R. (1988a) A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing, 54(5):593–600.

[16] Congalton, R. (1988b) Using spatial autocorrelation analysis to explore the error in maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing, 54(5):587–592.

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

[18] Cunia, T. (1981) A review of the methods of forest inventory. In: Woodpower — New Perspectives on Forest Usage, Talbot and Swanson (eds.), Pergamon Press, Internationla Science and Technology Institute, pp. 9–20.

[19] Dicks, S. and Lo, T. (1990) Evaluation of thematic map accuracy in a land-use and land-cover mapping program. Photogrammetric Engineering and Remote Sensing, 56(9):1247–1252.

[20] Dorigan, C. (1981) The role of remote sensing and information systems in forest resource management. In: Woodpower — New Perspectives on Forest Usage, Talbot and Swanson (eds.), Pergamon Press, International Science and Technology Institute, pp. 31–48.

[21] European Commision (1997) Study on European forestry information and communication system. Reports on forestry inventory and survey systems, Luxembourg.

[22] Fenstermaker, L. (1991) A proposed approach for national to global scale error assessments. In: Proceedings GIS/LIS '91, ASPRS, ACSM, AAG, AM/FM International and URISA, Vol. 1, pp. 293–300.

[23] Fitzpatick-Lins, K. (1981) Comparison of sampling procedures and data analysis for a land-use and land-cover map. Photogrammetric Engineering and Remote Sensing, 47(3):343–351.

[24] Foody, G. (1992) On the compensation of chance agreement in image classification accuracy assessment. Photogrammetric Engineering and Remote Sensing, 58(10):1459–1460.

[25] Ginevan, M. (1979) Testing land-use map accuracy: another look. Photogrammetric Engineering and Remote Sensing, 45(10):1371–1377.

[26] Gjertsen, A. (1996) Two-phase sampling for forest inventory based on satellite imagery. In: Progress in environmental remote sensing reseearch and applications, Parlow (ed.), Balkema, Rotterdam, pp. 63–71.

[27] Hagner, O. (1990) Use of digital satellite data for stand delineation and estimation of stand variables by regression analysis and field inventories. Mid-Term Symposium, ISPRS Commission VII, 17–21 September, Victoria, Canada.

[28] Hay, A. (1979) Sampling designs to test land-use map accuracy. Photogrammetric Engineering and Remote Sensing, 45(4):529–533.

[29] Hellden, U. (1980) A test of Landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya. Sweden, Lund University, Natural Geography Institute Report No. 47.

[30] Hocevar, M., Kovac, M. and Hladnik, D. (1995) Ecological monitoring of preserved forested landscapes in Slovenia by means of remote sensing and GIS. In: Remote Sensing and Computer technology for natural resource assessment (IUFRO XX World Congress), Saramäki-Koch-Gyde Lund (eds.), University of Joensuu, Faculty of Forestry, Research Notes 48, p. 95–118.

[31] Hord, M. and Bronner, W. (1976) Land-use map accuracy criteria. Photogrammetric Engineering and Remote Sensing, Vol. 42, No. 5, pp. 671–677.

[32] Hudson, W. and Ramm, C. (1987) Correct formula of the Kappa coefficient of agreement. Photogrammetric Engineering and Remote Sensing, 53(4):421–422.

[33] Janssen, L. and van der Wel, F. (1994) Accuracy assessment of satellite derived land-cover data: A review. Photogrammetric Engineering and Remote Sensing, 60(4):419–426.

[34] Kätsch, C. (1990) Zweiphasige Stichprobeninventur für Zwecke der Betriebsinventur auf der Basis einfacher Luftbildasuwertung. Dissertation, Göttingen (in German).

[35] Köhl, M. (1997) Country report from Switzerland. In: European Commision (1997): Study on European forestry information and communication system. Reports on forestry inventory and survey systems, Luxembourg, pp. 1019–1094.

[36] Köhl, M. and Päivinen, R. (1996) Definition of a system of nomenclature for mapping European forests and for compiling a pan-European forest information system. EUR 16416, Office for official publications of the European Communities, Luxembourg.

[37] Kusan, V. (1995) Remote sensing and GIS in Croatian forestry. In: Remote Sensing and Computer technology for natural resource assessment (IUFRO XX World Congress), Saramäki-Koch-Gyde Lund (eds.), University of Joensuu, Faculty of Forestry, Research Notes 48, p. 67–79.

[38] Lillesand, T. and Kiefer, R. (1994) Remote sensing and image interpretation. John Wiley & Sons, Inc., New York.

[39] Lunetta, R., Congalton, R., Fenstermaker, L., Jensen, J., McGwire, K. and Tinney, L. (1991) Remote sensing and geographic information system data integration: Error sources and research issues. Photogrammetric Engineering and Remote Sensing, 57(6):677–687.

[40] Martínez-Millán, J. (1997) Country report from Spain. In: European Commision (1997): Study on European forestry information and communication system- Reports on forestry inventory and survey systems, Luxembourg, pp. 905–954.

[41] Poso, S. and Wait, M.-L. (1995) Sample based forest inventory and monitoring system using remote sensing. In: Remote Sensing and Computer technology for natural resource assessment (IUFRO XX World Congress), Saramäki-Koch-Gyde Lund (eds.), University of Joensuu, Faculty of Forestry, Research Notes 48, p. 21–28.

[42] Rosenfield, G. (1982) Analysis of variance of thematic mapping experiment data. Photogrammetric Engineering and Remote Sensing, 47(12):1685–1692.

[43] Rosenfield, G., Fitzpatrick-Lins, K. and Ling, H. (1982) Sampling for thematic accuracy testing. Photogrammetric Engineering and Remote Sensing, 48(1):131–137.

[44] Rosenfield, G. and Fitzpatrick-Lins, K. (1986) A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing, 52(2):223–227.

[45] Short, N. (1982) The Landsat tutorial workbook-Basics of satellite remote sensing. Greenbelt, Md., Goddard Space Flight Center, NASA Reference Publication 1078.

[46] Shvidenko, A. and Nilsson, S. (1997) Are the Russian forests disappearing? Unasylva 188, Vol. 48, pp. 57–64.

[47] Skidmore, A. and Turner, B. (1992) Map accuracy assessment using line intersect sampling. Photogrammetric Engineering and Remote Sensing, 58(10):1453–1457.

[48] Söderberg, U. (1997) Country report from Sweden; In: European Commision (1997): Study on European forestry information and communication system-Reports on forestry inventory and survey systems, Luxembourg, pp. 995–1017.

[49] Stehman, S. (1997) Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sensing of Environment, 60:258–269.

[50] Story, M. and Congalton, R. (1986) Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing, 52(3):397–399.

[51] Tomppo, W. (1991) Satellite Image-Based National Forest Inventory of Finland. International Archives of Photogrammetry and Remote Sensing, Vol. 28, Part 7-1, pp. 419–424.

[52] Tomppo, E. (1996) Multi-Source National Forest Inventory of Finland. In: New thrusts in Forest Inventory, EFI Proceedings 7, pp. 27–41.

[53] Tomppo, E., Varjo, J., Korhonen, K., Ahola, A., Ihalainen, A., Heikkinen, J., Hirvelä, H., Mikkola, E., Salminen, S. and Tuomainen, T. (1997) Country report from Finland. In: European Commision (1997): Study on European forestry information and communication system. Reports on forestry inventory and survey systems, Luxembourg, pp. 145–226.

[54] Tomter, S. (1997) Country report from Norway;; In: European Commision (1997): Study on European forestry information and communication system. Reports on forestry inventory and survey systems, Luxembourg, pp. 799–860.

[55] Tosi, V. and Marchetti, M. (1997) Country report from Italy. In: European Commision (1997): Study on European forestry information and communication system. Reports on forestry inventory and survey systems, Luxembourg, pp. 423–645.

[56] Turk, G. (1979) GT Index: A measure of the success of prediction. Remote Sensing of Environment, Vol. 8:65–75.

[57] USGS (1990) The spatial data transfer standard. United States Geological Survey, Draft, January 1990.

[58] van Genderen, J. and Lock, B. (1977) Testing land use map accuracy. Photogrammetric Engineering and Remote Sensing, 43(9):1135–1137.

[59] Warren, S., Johnson, M., Goran, W. and Diersing, V. (1990) An automated, objective procedure for selecting representative field sample sites. Photogrammetric Engineering and Remote Sensing, Vol. 56, No. 3, pp. 333–335.

[60] Welch, R., Jordan, T. and Ehlers, M. (1985) Comparative evaluations of the geodetic accuracy and cartographic potential of Landsat-4 and Landsat-5 thematic Mapper image data. Photogrammetric Engineering and Remote Sensing, 51(11):1799–1812.

[61] Wolff, B. (1992) Betriebs- und bestandesweise Holzvirratsinventur auf der Basis von permanenten terrestrischen und Luftbildstichproben. Dissertation, Göttingen (in German).