

Annexure -II**ACCURACY OF ASSESSMENT****II.01 Introduction**

No assessment is complete until its accuracy has been assessed. Assessments are carried out with the help of technology and field survey to assess the actual population parameter values as closely as possible. The assessment, which provides the true status of population is said to be accurate. However, since the value of population parameter is not known it is very difficult to ascertain whether the assessment is accurate or not. It is but obvious, that due to technological limitations, sampling, fieldwork and human errors, inaccuracies would creep into the assessed results. Therefore, it is essential for any assessment to quantify the inaccuracy contained in assessed values.

SFR-2001 contains information generated from two assessments; (i) forest cover and (ii) tree cover. The forest cover assessment has been carried out by interpreting remote sensing derived data followed by their interpretation and classification of forest cover. The tree cover assessment has been carried out through field surveys following a suitable sampling design. The nature of assessment is entirely different in both the cases, which necessitates two different approaches for quantifying their accuracy.

II.02 Accuracy of Forest Cover Assessment

Accuracy of forest cover assessment is hampered primarily due to errors contained in the satellite data and wrong interpretation and/or classification of the imageries. The remote sensing systems have their own limitations, whereby radiometric and geometric errors creep in and deteriorate the quality of remotely sensed data. The radiometric errors may be caused by the malfunctioning of the sensor itself or by the intervening atmosphere between the terrain and remote sensing system i.e. the radiant flux reflected by the terrain

may not resemble the energy recorded by the sensor. The geometric errors may be caused by the variations in altitude, velocity of sensor platform, panoramic distortions, earth curvature, atmospheric refraction, relief displacement, etc. There are procedures to minimize these errors but these cannot be totally ruled out and in turn have an impact on accuracy of assessment. Remote sensing system also has limitations on account of spatial, spectral, temporal and radiometric resolution. Errors in interpretation and classification may be caused due to cloud or shadow effects, or seasonal variation in the canopy of deciduous trees or bushy and agricultural vegetation getting mixed with forest crop, etc. All these errors while classifying the remote sensing data influence the accuracy of the assessment.

Often, an error matrix (also termed as confusion matrix) is prepared for assessing the accuracy of classification of remotely sensed data. It compares occurrences of agreements and disagreements between remote sensing derived classification with the reference data (ground truth) on a class-by-class basis at randomly selected locations. The data on the class a particular location belongs to on the ground and what it has been classified into are recorded and put in the form of a matrix (called error matrix). It is an array of numbers arranged in rows (generally, classification) and columns (generally, ground truth). It is a square matrix as both the numbers of rows and columns in the matrix are equal, representing the classes (dense forest, open forest, etc.) whose classification accuracy is to be assessed. The randomly selected locations or sampling units, which are presented in the matrix, can be pixels or a group of pixels or a polygon. In this study, groups of pixels are the sampling units. An entry made along the major diagonal of the error matrix implies agreements i.e., that the classification at a

sampling unit matches with the corresponding ground truth and hence suggests that the classification is proper. The non-diagonal elements indicate disagreements or wrong classification.

The percentage of correctly classified sampling units (i.e. sum of all diagonal elements) out of the total considered sampling units in the error matrix provides measure of “over all accuracy” of assessment. Similarly, accuracies of each class can be measured by calculating the percentage of properly classified sampling units (diagonal element) out of the total sampling units considered for that class in that row or column.

II.03 Methodology

The sampling design is very crucial for assessing the accuracy of classification. It should ensure representation of the entire spatial population. Likewise, choosing the appropriate sample size is also very important. It is suggested in the literature that if the area for which the assessment is being carried out is large or the classification has a large number of vegetation or land use classes then the minimum number of samples should be more than 50 sampling units per class. However, it may be adjusted according to relative importance of the category and variability within each category. For the present study a minimum of 2000 sampling units was considered adequate for preparation of an error matrix.

Ideally, the sampling units should be randomly selected from the whole assessment area, i.e. the country, but there are certain limitations in this approach. Firstly, it is difficult to collect data physically for a large sample of randomly distributed locations scattered all over the country, which would require massive manpower, time and cost. Secondly, there is a time lag of about 1 to 2 years between the date of satellite data used and the ground truth period. Thirdly, it is very difficult to precisely register the location on the ground with that on image data. Fourthly, since the forest cover is not distributed uniformly all over the country, the sampling design should be such that

more sampling units are located in the regions or strata having larger proportion of forest cover.

In order to overcome these difficulties, an unconventional approach of error matrix preparation was followed where scenes from high-resolution satellite data were used to provide ground truth. The forest cover assessment presented in this report is based on data from sensor LISS-III of satellite IRS 1C/1D (resolution of 23.5 m x 23.5 m). The same satellite has another sensor, namely IRS PAN, that provides data at a higher resolution of 5.8 m x 5.8 m. The PAN scenes provide much clearer picture of the classes of forest cover on the ground. One can easily differentiate between dense forest, open forest, scrub and non-forest areas, the classes used in the present assessment. Hence, it was decided that instead of going to the field for ground truth, PAN data would be used as a proxy for field data. This view is also supported in literature as well.

A detailed sampling design was developed on the basis of LISS and PAN data by stratifying the LISS scenes using forest cover information of past assessment (1999). LISS scenes having 25 percent or more forest cover formed one stratum and the scenes having less than 25 percent forest cover formed the other stratum. In order to provide spatial representation of the whole area, systematic sampling with random start in both the strata has been followed. Since the stratum that has LISS scenes with 25 percent or more forest cover is relatively more important from forest cover classification point of view, 35 scenes from this stratum and only 5 scenes from the other stratum were selected. The selection was made systematically with random start. In each selected LISS scene, the PAN scenes which had more than 50 percent forest cover, were considered and then one PAN scene was randomly selected. If there was no PAN scene with more than 50 percent forest cover then the PAN scene that had the maximum forest cover was selected for this exercise. (A LISS scene covers about 20,000 km² of ground area whereas a PAN scene covers about 5,000 km², thus four PAN scenes correspond to each LISS scene.

The selected PAN scenes were procured from NRSA, Hyderabad and rectification was carried out after verification. It was decided that the classification would be verified considering cluster of 16 pixels of LISS image, which is approximately 1.0 ha at the surface of earth. Grids of size 4'x 4' (latitude x longitude) were prepared for each PAN scene. Cross section of each grid represented the central point of a sampling unit. Observations of ground truth from PAN scene were made at all these points to be compared with the classification made at corresponding locations in the classified LISS image. The ground truth data were recorded along with their geographic coordinates. The corresponding classified pseudo image, based on LISS image, was checked at these geographic coordinates and also recorded. These two sets of data were compared to prepare the error matrix.

In order to check bias in preparation of error matrix, this whole task was assigned to the Forest Inventory Unit of FSI. The staff responsible for forest cover classification work was not involved in this exercise.

II.04 The Error Matrix

Using the method described above, a total of 3,608 sampling units were selected for the whole country. The error matrix given in Table II.01 was prepared on the basis of observations made at these points.

For the sake of illustration, the diagonal element, say number 884 for dense forest at row 1 and column 1 implies that at 884 sampling units, dense forest on the ground was correctly classified as dense forest. Whereas, the off-diagonal number 5 in row 2 (Open Forest) and column 4 (Non Forest) implies that at 5 sampling units, non forest on the ground was wrongly classified as open forest.

A simplified error matrix can be created by considering only two classes of land use i.e. "forest" and "non forest". This is done by combining dense and open forest classes into "forest" and scrub and non forest classes into "non forest". The simplified error matrix is given in Table II.02.

Table II.01 Error Matrix

Classification Classes	Ground Truth (based on PAN scenes)				Row Total
	Dense Forest	Open Forest	Scrub	Non- Forest	
Dense Forest	884	56	1	3	944
Open Forest	47	455	0	5	507
Scrub	0	1	49	1	51
Non-Forest	5	22	7	2,072	2,106
Column Total	936	534	57	2,081	3,608

Table II.02 Simplified Error Matrix

Classification Classes	Ground Truth (based on PAN scenes)		Row Total
	Forest	Non-Forest	
Forest	1,442	9	1,451
Non-Forest	28	2,129	2,157
Column Total	1,470	2,138	3,608

II.05 Accuracy Levels (Findings)

The error matrix (Table II.01) reveals that out of the total sampling units i.e., 3,608 where observations were made, classification was found to be correct at $884 + 455 + 49 + 2,072 = 3,460$ sampling units, i.e., the sum of elements along the main diagonal of the matrix. The “over all accuracy” of classification works out to be 95.9 percent = $(3,460/3,608) \times 100$. This is quite high implying that classification procedures followed at FSI are satisfactory.

Considering the simplified error matrix (Table II.02), classification was found to be correct at $1,442 + 2,129 = 3,571$ sampling units out of 3,608 sampling units, yielding an “over all accuracy” of 99.0 percent $(3,571/3,608) \times 100$. It shows even higher capability in FSI to distinguish areas between forest and non forest by interpreting satellite data.

Class wise accuracy can be estimated in two ways: ratio of correctly classified sampling units in that class, i.e. the diagonal element, with respect to total number of sampling units in the corresponding column or row. An off-diagonal number along a column indicates that of the total locations on the ground belonging to that class, this number of locations were excluded or omitted during classification giving rise to “omission error”. The ratio of correctly classified

sampling units in a class with respect to column total provides a measure of omission error and is termed as “producer’s accuracy”. It indicates how well or correctly the producer or the analyst has interpreted that class. The off-diagonal number along a row indicates that of the total points on the map shown to represent a particular class, these do not belong to that class but were included or committed into the class due to wrong classification. This provides a measure of “commission error”. The ratio of correctly classified sampling units in a class with respect to the row total provides a measure of commission error and is termed as “user’s accuracy”. For the user of the map it is the probability that a location classified on the map actually represents that class on the ground.

The Producer’s and User’s Accuracies for different classes as computed from the error matrix (Table II.01) are given in Table II.03.

From Table II.03 it is found that the producer’s accuracy for Dense Forest, Open Forest, Scrub and Non Forest classes are 94.4, 85.2, 86.0 and 99.6 percent, respectively. Similarly, user’s accuracy for these classes are 93.6, 89.7, 96.1 and 98.4 percent, respectively. These levels of accuracy are satisfactory and acceptable.

Table II.03 Producer’s and User’s Accuracy based on Table II.01

	Class	Dense Forest	Open Forest	Scrub	Non Forest
Producer’s Accuracy	Operation	884 / 936	455 / 534	49 / 57	2,072 / 2,081
	Accuracy (%)	94.4	85.2	86.0	99.6
User’s Accuracy	Operation	884 / 944	455 / 507	49 / 51	2,072 / 2,106
	Accuracy (%)	93.6	89.7	96.1	98.4

Table II.04 Producer’s and User’s Accuracy based on Table II.02

	Class	Forest	Non-Forest
Producer’s Accuracy	Operation	1,442 / 1,470	2,129 / 2,138
	Accuracy (%)	98.1	99.6
User’s Accuracy	Operation	1,442 / 1,451	2,129 / 2,157
	Accuracy (%)	99.4	98.7

Table II.04 based on Simplified Error Matrix (Table II.02) shows even higher accuracy levels. The producer's accuracy for forest and non-forest classes are found to be 98.1 and 99.6 percent, respectively while user's accuracy for these classes are 99.4 and 98.7 percent, respectively.

The KAPPA analysis, which is a multivariate technique, provides a statistic known as K_{HAT} . This statistic usually ranges between 0 and 1. It is used to indicate whether the correct values of the error matrix are due to true agreement or due to chance agreement. K_{HAT} calculated from the error matrix given at table II.01 is equal to 0.93, which indicates that an observed classification is 93 percent better than one resulting from chance.

II.05 Accuracy of Tree Cover Assessment

The tree cover assessment is based on stratified random sampling. The country has been stratified into 14 physiographic zones as has been described in Chapter 4 and in each zone simple random sampling has been followed. The sampling units used for rural areas are the sampled villages and for urban areas, UFS blocks. Zone wise estimates have been developed following ratio method of estimation. The precision of an estimate is adjudged by standard error of estimate, which is calculated as positive square root of variance of the estimate.

The estimate of trees per ha (R_i) for the i th physiographic zone is the sample ratio:

$$\hat{R}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{j=1}^{n_i} x_{ij}}$$

Its variance is estimated by:

$$\hat{V}(\hat{R}_i) = \frac{1}{n_i(n_i - 1)\bar{x}_i^2} \left[\sum_{j=1}^{n_i} y_{ij}^2 - 2\hat{R}_i \sum_{j=1}^{n_i} y_{ij}x_{ij} + \hat{R}_i^2 \sum_{j=1}^{n_i} x_{ij}^2 \right]$$

and its standard error (in percent) is estimated by:

$$SE(\hat{R}_i)\% = \frac{\sqrt{\hat{V}(\hat{R}_i)}}{\hat{R}_i} \times 100$$

The estimate of total number of trees (T_i) for the i th physiographic zone is calculated as:

$$\hat{T}_i = A_i \times \hat{R}_i$$

and its variance is estimated by:

$$\hat{V}(\hat{T}_i) = A_i^2 \times \hat{V}(\hat{R}_i)$$

where,

y_{ij} = Number of stems in j^{th} sampling unit in i^{th} physiographic zone

x_{ij} = CNFA of j^{th} sampling unit in i^{th} physiographic zone

\bar{x}_i = Average CNFA of sampling units of i^{th} physiographic zone

n_i = Number of sampling units in i^{th} physiographic zone

A_i = CNFA of i^{th} physiographic zone

Likewise, the estimate of total number of trees (T) for all the 14 physiographic zones at the country level is calculated as:

$$\hat{T} = \sum_{i=1}^{14} \hat{T}_i$$

and its variance is estimated by;

$$\hat{V}(\hat{T}) = \sum_{i=1}^{14} A_i^2 \times \hat{V}(\hat{R}_i)$$

The estimate of trees per ha (R) at the country level is given by:

$$\hat{R} = \frac{\hat{T}}{A}$$

Its variance is estimated by:

$$\hat{V}(\hat{R}) = \frac{1}{A^2} \times \hat{V}(\hat{T})$$

and standard error percent is estimated by:

$$SE(\hat{R})\% = \frac{\sqrt{\hat{V}(\hat{R})}}{\hat{R}} \times 100$$

where, A = CNFA of the country.

II.06 Findings

The following table gives the precision of estimates (SE percent) at the level of physiographic zones.

Table II.05 Physiographic zone wise precision of estimates

Physiographic zone	Tree Cover (km ²)	SE percent
Western Himalayas	3,069	11.12
Eastern Himalayas	392	22.47
North East	642	11.86
Northern Plains	10,098	5.87
Eastern Plains	8,323	7.50
Western Plains	3,875	21.13
Central Highlands	7,077	14.32
North Deccan	6,905	6.53
East Deccan	9,760	9.37
South Deccan	11,468	15.03
Western Ghats	3,957	19.79
Eastern Ghats	1,788	21.44
West Coast	3,699	20.31
East Coast	10,419	20.02
Total	81,472	1.83

Overall precision of estimate at the country level = 1.83 percent.

It may be observed that for the physiographic zones, Eastern Himalayas, Western Plains, Western Ghats, Eastern Ghats, West Coast and East Coast the standard error is relatively higher i.e. around 20% in comparison with other zones. This is because higher variability in tree cover in these zones and the number of samples taken in these zones were just adequate to capture 80 percent of the variability. This also indicates that a larger sample size is required for these zones if higher level of accuracy is required. However, tree cover at the national level has been estimated quite precisely with a standard error as low as 1.83 percent.