SINET: Ethiop. J. Sci., 38(1):29– 42, 2015 ISSN: 0379–2897 (Print)

2520–7997 (Online)

WHY DOES ACCURACY ASSESSMENT AND VALIDATION OF MULTI-RESOLUTION-BASED SATELLITE IMAGE CLASSIFICATION MATTER? A METHODOLOGICAL DISCOURSE

Berhan Gessesse [1], Woldeamlak Bewket [2] and Achim Bräuning [3]

[1] Entoto Observatory and Research Centre (EORC), PO Box 33679, Addis Ababa, Ethiopia.
E-mail: berhanavu@gmail.com
[2] Department of Geography and Environmental Studies, Addis Ababa University, Ethiopia
[3] Institute of Geography, Friedirich-Alexander University of Erlangen-Nürnberg, Germany

ABSTRACT: This study presents a methodological discourse about how to validate the reliability of thematic maps derived from multi-resolution satellite-based image classification. Besides, the paper examines unbiased estimates of accuracy assessment using known sampling units. Landsat and SPOT images were used for LULC thematic layer extraction. These thematic layers together with reference data extracted from panchromatic aerial photo interpretation and ground survey were used as input datasets for accuracy assessment and validation analysis. For each LULC unit, a minimum of 50 reference samples were derived using a stratified random sampling scheme. Consequently, error matrices were generated to validate the quality of the 1973, 1995 and 2007 LULC maps. To improve sampling biases introduced due to the stratified random sampling reference data collection scheme, accuracy assessment indices including the producer's, user's and overall accuracy as well as Kappa coefficient of agreement were adjusted to the known areal proportion of map categories. The computed overall accuracy, corrected for bias using known marginal proportions of the 1973, 1995 and 2007 thematic layers were 88.12%, 89.95% and 92.27%, respectively. Also, 81.20%, 82.17% and 83.11% of Kappa coefficient of agreement were achieved from the 1972, 1995 and 2007 classifications, respectively. The findings show that high resolution aerial photos are good sources of reference datasets in the absence of historical ground truth data for accuracy assessment analysis and the LULC classifications fulfilled the minimum of LULC classification standards of overall accuracy and Kappa coefficient of agreement. Consequently, all the LULC classifications could be used as an input for policy options for integrated land resource management practices in the watershed studied.

Key words/phrases: Accuracy assessment, Ethiopia, land use/land cover maps, probability estimate sampling bias

INTRODUCTION

Why mapping environmental catastrophes and geo-environmental resources as well as assessing the validity of these maps? Environmental disasters and natural resource removals are very critical globally. As a result, natural resources are scarce; and land resource exploitation is often initiated with de-vegetation and accompanied by land degradation and thus contributes to conflicts over resource uses (Hunter, 2000; Steffen *et al.*, 2005). Exploitative resource use and consequent degradation is a threat to rural livelihood and sustainable development especially in developing countries. To counteract inappropriate resource exploitation, mapping of the complex geo-system elements is an essential input for policy decisions (Metzger *et al.*, 2006; Meng, 2010). Accordingly, Congalton and Green (2009:1) argued that "decisions about environmental issues require maps, and effective decisions require accurate maps or at least maps of known accuracy". In this regard, optical, hyperspecteral and microwave-based remotely sensed data provide opportunities to produce maps, which portray the landscape characteristics with reasonable quality. Since the first aerial photograph was captured from a balloon in 1858 (Congalton and Green, 2009) and the launch of the first Landsat satellite in 1972 (De Jong *et al.*, 2005), efforts have been made to extract landscape level information from remotely sensed data. Providing input datasets for environmental resource mapping in general and Land-use/Land-cover (LULC) thematic layer generation in particular is the most widespread application of Earth observation satellite science and technology (Foody, 2002). LULC information allow contrasting the land diversity across geographic dimensions, quantifying LULC dynamic processes and input datasets for climate change predictions, soil erosion analyses, hydrological models and policy options (Jensen, 2005; Thapa and Murayama, 2008).

Although the recent advancement of geovisualization techniques is tremendous, automated digital image classification is still unable to produce maps with optimal accuracy level. In this regard, Meng (2010) argued that map making by using remotely sensed data poses a challenge for cartographers to publicize their knowledge to the largest possible number of users. Digital satellite image interpretation mostly contains classification errors resulting from the lack of a complete one-to-one agreement between remotely sensed data and features on the earth's surface (Foody, 2002; Congalton and Green, 2009). Errors in thematic maps derived from remotely sensed datasets could be termed as the deviation of classified data value from the considered or measured true value. These errors and limitations are the result of the very lengthy and complicated remotely sensed data acquisition and interpretation processes (Congalton and Green, 2009; Gao, 2009). However, maps derived from remotely sensed data should be validated based on widely adopted thematic map quality control system standards for their end users (Litwin and Rossa, 2011; ISO, 2002). In the light of this, prior to using for various applications, remote sensing-derived products (*e.g.*, LULC maps, environmental resource thematic layers, natural-disaster hotspot inventory maps, *etc.*) should be subjected to classification accuracy assessment (CAA) analysis (Stehman and Czaplewski, 1998). Lillesand *et al*. (2008:585) noted that "… a classification is not complete until its accuracy is assessed". Accuracy assessment begins with the definition of the entire geographical dimension of the study area from which LULC thematic layers and reference sample units in the form of polygons (patches) or points (pixels) are extracted.

Data from field survey, airborne videos and high resolution aerial photographs/satellite images are the main sources of ground truth reference datasets for CAA. In addition, a wide range of sampling designs have been proposed in the literature for the extraction of reference samples including random, systematic, clustered and stratified samples or a combination of them (Stehman and Czaplewski, 1998; Congalton and Green, 2009). The error matrix also called a confusion matrix or a contingency table is the most appropriate tool for validating the accuracy of remotely sensed data classification and reporting classification errors. Thus, thematic map accuracy assessment is used to express the degree of correctness of pixels in remote sensing based thematic maps as compared to the ground truth reference information (Jensen, 2005).

Accordingly, univariate statistical measures such as producer's, user's and overall accuracies as well as omission and commission errors are frequently used to report classification errors. In addition, a discrete multivariate statistical model of 'Kappa Coefficient Index' has become a widely used standard tool for CAA procedures (Congalton *et al.*, 1983; Congalton, 1991; Jensen, 2005; Lillesand *et al*., 2008; Gao, 2009; Congalton and Green, 2009). However, researchers such as Stehman and Czaplewski (1998); Pontius (2000) and Foody (2002) criticized the Kappa coefficient index application to report CAA because this index may overestimate the chance agreement that may result in an underestimation of classification accuracy. Pontius and Millones (2011) also recommended that quantity disagreement and allocation disagreement indices could be used to report CAA instead of using the Kappa index.

On the other hand, after acknowledging the limitations of Kappa agreement, Congalton and Green (2009:115) noted that "given the very powerful properties of the Kappa coefficient, including the ability to test for significant differences between two independent coefficients, it must still be considered a vital accuracy assessment measure". The value of the Kappa coefficient usually ranges from 0 to 1 (Lillesand *et al*., 2008). Although a Kappa coefficient of 0.85 is considered as a benchmark for acceptable CAA without detailed explanation (Anderson *et al.,* 1976; Van Genderen *et al*., 1978; Brown, 2005; Jensen, 2005). In contrast, Landis and Koch (1977) and Congalton and Green (2009) argued that a computed Kappa value greater than 0.80 represents strong (almost perfect) agreement; a value between 0.40 and 0.80 represents moderate to substantial agreement and a value below 0.40 represents fair to poor agreement.

LULC classification validation has established renewed interest parallel to the advent of remote sensing technology, however, “failure to know these techniques and considerations can severely limit one's ability to effectively use remotely sensed data” (Congalton, 1991). These days, many scientific communities published quite a number of articles using remotely sensed data to address various environmental and social related issues. Unfortunately, evaluating and validating the accuracy of thematic maps, derived from remotely sensed datasets, have been incomplete with the majority of studies and the findings of these studies not verified using CAA. Examples on LULC classification studies include Bruzzone and Serpico (1997) in Italy, Sierra (2000) in Napo region of the western Amazon, Reid *et al.* (2000) in Ethiopia, Kibrom Tekle and Hedlund (2000) in Ethiopia, Lopez *et al.* (2001) in Mexico, Gete Zeleke and Hurni (2001) in Ethiopia, Woldeamlak Bewket (2002) in Ethiopia, Talukdar *et al.* (2004) in India, Comber *et al.* (2004) in Scotland, Bezuayehu Tefera and Geert Sterk (2008) in Ethiopia, Thapa and Murayama (2008) in Nepal, Wu *et al.* (2008) in China and Attua and Fisher (2011) in Ghana. The findings and conclusions in the above studies were not supported by a full-fledged statistical map classification accuracy and validation assessments.

The main purpose of this study was to discourse the scientific application of accuracy assessment methods used to analyze CAA to enhance the knowledge of assessing the quality of thematic maps. The three dates (1973, 1995 and 2007) of the Modjo watershed LULC classification maps as map accuracy assessment sample datasets and independently derived ground truth facts as reference accuracy assessment sample datasets were considered as input for this discourse. Therefore, this article contributes to addressing the CAA methodological gap and aims at presenting a methodological argument for the validation of LULC classification by using retrospective aerial photographs as reference datasets, where historical ground truth reference data relevant to CAA are almost lacking, a situation typical in many developing countries. The specific objectives of the study were: i) to examine the potential of reference datasets obtained from historical high resolution aerial photographs for evaluating LULC classification accuracy assessment; ii) to evaluate the usefulness of the 1973, 1995 and 2007 satellite images to generate high-quality LULC thematic layer information and validating the quality of these maps using accuracy assessment; and iii) to compare and inspect the accuracies of these classified thematic maps derived from multi-date satellite images having different spatial resolution as per the thematic layer mapping quality control standards.

DATA AND METHODS

Application to the Modjo River watershed

The Modjo River watershed (1,478 km^2), considered as a case study site to derive thematic layers for this discourse, is found in the upper *Awash* River Basin of Ethiopia, stretching over 8° 35' 00'' to 9°05' 11'' N and 38°54' 35'' to 39°15' 30'' E (Fig. 1). The watershed is characterized by undulating topography with hills, mountains, plains and river valleys. Elevation ranges from 1,740 to 3,060 masl and the slopes are generally steeper in the north-western part of the watershed. The central and downstream parts of the watershed are characterized by relatively flat landforms and gentle slopes.

The Modjo River is a major perennial river in the study area and drains into the Koka Reservoir. Hora-Kilole, Hora-Hado, Bishoftu Guda, Hora Arsedi and Bishoftu crater lakes found in the watershed. Chefe Donsa, Godino, Debre Zeit (Bishoftu), Ejeri and *Modjo* are the major urban settlements in the watershed. Trees, shrubs and grasses are the major plant morphological types in the study watershed under consideration. A mixed crop-livestock system is the typical economic activity for the rural population of the watershed.

Data inputs and image classification

Classification input datasets

Input datasets such as the Landsat and SPOT imagery, aerial photographs and field survey were used for LULC classification and CAA (Table 1). All these datasets were imported into geospatial software for further processing. The spatial reference local coordinate systems including geodetic datum of Adindan and UTM Zone 37 North projection were employed to register all the spatial datasets used in this study. Image preprocessing, classification and CAA were undertaken using ERDASIMAGINE 9.3 software. A minimum of 10 ground control points were digitized from digital aerial photographs and topographic maps, which were orthorectified and georeferenced by Ethiopian Mapping Agency (EMA). These ground control points were used for orthorectifying the satellite imageries. A total 'Root Mean Square (RMS)' error of 0.49, 0.43 and 0.32 pixel values for Landsat MSS, Landsat TM and SPOT, respectively, were reported. False Colour Composite (FCC) in Red-Green-Blue (RGB) order, linear contrast stretching and histogram equalization enhancement techniques were used to enhance visualization and image interpretation of the various features. Ground visit was carried out in October 2011 for training and reference sample data collection for image classification and CAA.

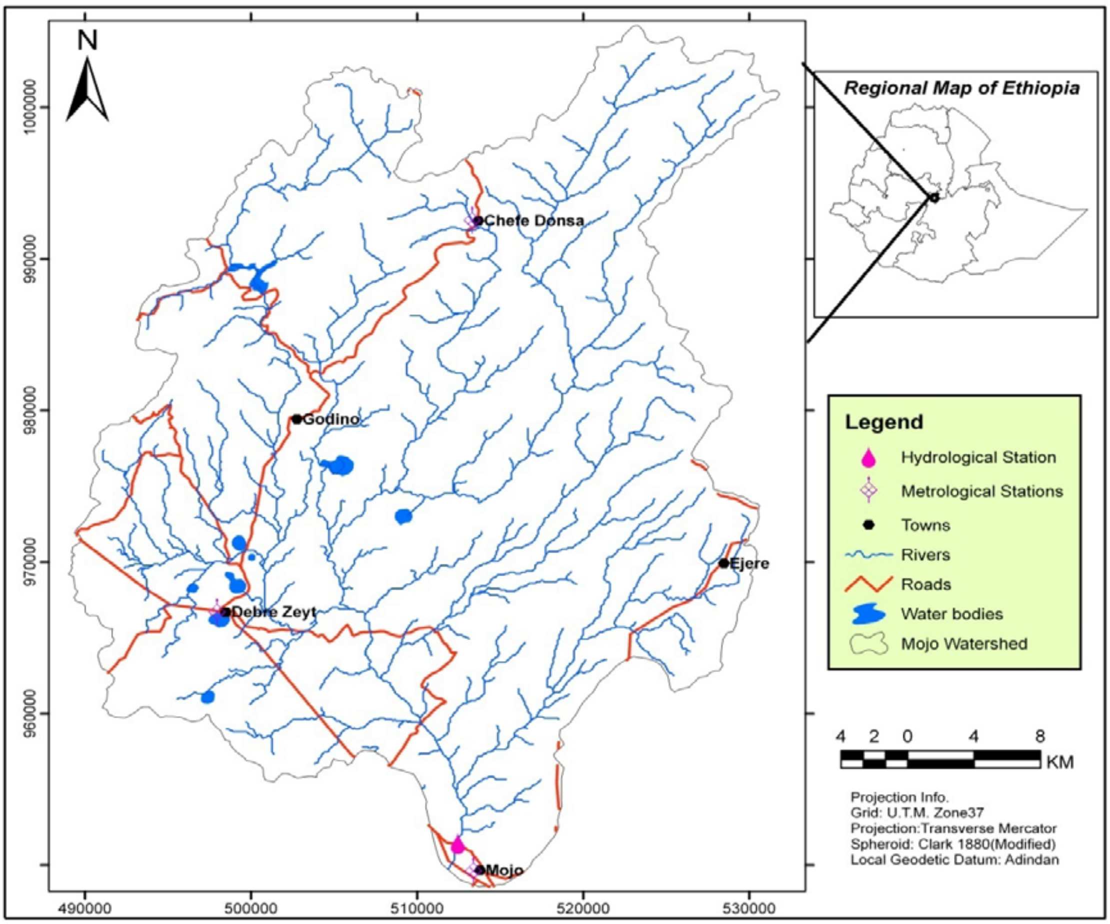

Figure 1. Location of Modjo watershed in the central highlands of Ethiopia.

Table 1. Input data used for LULC classification and accuracy assessment validation.

Dataset	Acquisition Date	Resolution(m)	Purpose	Source
Landsat_1 MSS (P180/1/R54)	01.30.1973	57 X 57(m)	Classification	GLCF*
Landsat _5 TM (P168/R54)	03.19.1995	28.5 X 28.5(m)	Classification	GLCF*
SPOT_5	03.08. 2007	10 X 10(m)	Classification	Commercial data provider
5 Scene Aerial Photos	Junauary,1972	Scale- 1:50,000	Reference	EMA**
11 Scene Aerial Photos	12.12 1994	Scale- 1:10,000	Reference	EMA**
GPS Field survey	October 2010	-	Reference	Own Survey
Topographic Maps; Sheet No. 0939_C4,0938_C3, 0838_B2, 0839_A1, 0838_B4,0839_A3, 0839_A4, 0938_C3,	1973,1975, 1976	Scale-1:50,0000	Field Survey	EMA**

Global Land Cover Facility,* * Ethiopian Mapping Agency*

After preparing satisfactory training site signatures based on aerial photo interpretations and field survey, the images of three years (1973, 1995 and 2007) were classified into nine generalized LULC types (Figs 2 A, B & C). The images were classified using supervised image classification in line with the Maximum Likelihood Classification Algorithm (Jensen, 2005; Lillesand *et al.*, 2008; Gao, 2009). Nine dominant LULC classes in the Modjo watershed, namely bare, cultivated, forest, grassland, marshland, plantations, shrub, urban and water bodies were identified and considered in this study. However, the classified images contained 'salt-and-pepper' (pixel speckling) noise caused by mixed pixels in some classes within another dominant class. As suggested by Lillesand *et al.* (2008), a 5 × 5 window majority neighbourhood filter was run to minimize such pixels noise and eliminate small patches of isolated pixels. Then, LULC maps and statistics were extracted to explain the detailed characteristics of Modjo watershed historical landscape fragmentation.

Reference sample size determination

Reference sample data extraction depends upon two types of sampling units, *i.e.,* pixels (points) and polygons (patches). In this study, the LULC maps were derived by pixel-based image classification so that pixel level sampling units were adopted. Aerial photographs were used as sample reference data source for the 1973 and 1995 LULC classification accuracy assessment. On the other hand, field survey samples were used as reference data for the 2007 LULC CAA. The digital aerial photos were manually interpreted and reference samples for each LULC category were precisely identified, digitized and recorded along with their attributes. The field survey samples were identified and encoded with the assistance of a positioning device, a Trimble Handheld GPS data logger. After LULC classification and derivation of representative reference data, overlay analyses were carried out using ERDAS IMAGINE 9.3 software to evaluate the accuracy of the classification.

Reference sample size should be kept as small as possible, but still be capable of meeting requirements to assess the accuracy of the classified maps (Gao, 2009). Although various mathematical theories were developed to determine adequate sample sizes for CAA (Jensen, 2005; Gao, 2009), no standardized consent has been reached yet regarding the adequate sample size determination. However, a general "rule of thumb" approach is recommended by researchers. For example, Gao (2009) proposed that the minimum sample size for each LULC class, necessary for 85% and 90% accuracy interpretation is to be set to 20 and 30, respectively. On the other hand, Jensen (2005), Lillesand *et al.* (2008) and Congalton and Green (2009) argue that any sample size of less than 50 will be unsatisfactory for error estimations and in most cases a minimum of 50 samples for each map class should be collected for maps of <4,000 km^2 in size and fewer than 12 classes. Thus, by considering the size of Modjo watershed (1,478 km^2) and the identified nine LULC classes, an economical and satisfactory sample size *i.e.,* a minimum of 50 samples per LULC category was used (Tables 3, 5 and 7). Among other sampling techniques, a stratified random sampling scheme was used to generate reference sample datasets. The main advantage of this type of sampling scheme for this study is that no matter how small a LULC class is in size or limited in its spatial extent, smaller areas can be adequately represented (Gao, 2009). However, bias can be introduced into the error matrix due to this sampling technique (Card, 1982).

Bias correction and indices of accuracy assessment

Although the principle of error matrix analysis and indices of CAA have been dealt with in the literature (Card, 1982, Lillesand *et al.,* 2008, Congalton and Green, 2009; Gao, 2009), the mathematical notations of the error matrix coupled with estimated unbiased sampling procedures are briefly described below. Let, the sample of n points are derived and the results are tabulated in a two-way square contingency table (m x m) where n_{ij} denotes the number of points in the sample whose reference data category is i (i = 1,2,…, m) and whose LULC map category is j (j= 1,2,…, m) (Equation 1). The diagonal elements in the error matrix represent correctly classified pixels whereas the off-diagonal cell values show misclassified pixels in the form of either 'errors of omission' (non-diagonal column elements) or 'errors of commission' (non-diagonal row elements).

Reference data(*i*)

	Class	1	2	...	m	Map total ($n_{j.}$)
LULC Map(j)	1	n_{11}	n_{12}	...	n_{1m}	$n_{1\cdot}$
	2	n_{21}	n_{22}	...	n_{2m}	$n_{2\cdot}$
	.	.	.		.	.
	.	.	.	...	.	.
	.	.	.		.	.
	m	n_{m1}	n_{m2}	...	n_{mm}	$n_{m\cdot}$
	Reference total ($n_{.i}$)	$n_{\cdot 1}$	$m_{\cdot 2}$	...	$n_{\cdot m}$	$\boldsymbol{n}$

....(1)

where,

m stands for the number of LULC classes,

n_{ij} is individual cell value in the error matrix whose true category is i and whose map category is j,

n_{ii} are values in the major diagonal line,

$n_{.i}$ is the column marginal sum in the true reference dataset,

$n_{j.}$ is the row marginal sum in the classified images and n is the total number of samples examined.

However, direct interpretation of the error matrix result from classified map and reference samples generated using the stratified random sampling method cannot be possible due to the problem of sampling bias (Card, 1982). Owing to this, the probability estimate method developed by Card (1982) was used and tested to compute unbiased summary statistics. The sampling bias was first corrected using the known (true) map marginal proportions (π_j). Rearranging the error matrix and the sampling design of stratified random sampling methods and computing the individual cell probabilities are the initial phases in CAA.

$$P_{ij} = \frac{\pi_j n_{ij}}{n_{j\cdot}} \quad \text{...(2)}$$

where,

P_{ij} stands for individual cell probabilities,

π_j is the known map marginal proportions for each LULC map category j,

n_{ij} and $n_{j.}$ as previously defined.

After the individual cell probabilities correction, the true marginal proportions (marginal column value) for given true class *i* and the remotely sensed image classification marginal proportions (marginal row value) for *j* is computed using the following equations:

$$\hat{P}_i = \sum_{i=1}^{m} \frac{\pi_j n_{ij}}{n_{.i}} \text{...(3)}$$

$$\hat{P}_j = \sum_{j=1}^{m} \frac{\pi_j n_{ij}}{n_{j.}} \text{...(4)}$$

where,

$\hat{P}_i$ is true marginal proportion?

$\hat{P}_j$ is the remotely sensed data classification marginal proportion,

π_j, n_{ij}, $n_{.i}$ and $n_{j.}$ as previously defined.

Producer's accuracy (probability correct, given true class *i*) indicates how well the samples from the reference data can be mapped on the LULC thematic layers derived from the remotely sensed image. Computing the producer's accuracy is the next step in CAA and can be calculated using the following equation:

$$\hat{\theta}_{ii} = \frac{\hat{P}_{ii}}{\hat{P}_i} \quad \text{..(5)}$$

where,

$\hat{\theta}_{ii}$ stands for probability correct of the true class 'i',

$\hat{P}_{ii}$ is the diagonal cell value of each category in the cell probability matrix, and

$\hat{P}_i$ as previously defined.

User's accuracy (estimates of probability correct, given map class *j*) indicates that the probability of a sample from the classification represents an actual class in the ground reference data. This computation is made exactly by taking the diagonal cell values of the error matrix and dividing by the row (*j*) marginal values.

$$\hat{\lambda}_{jj} = \frac{n_{jj}}{n_{j.}} \text{ or } \hat{\lambda}_{jj} = \frac{\hat{P}_{ii}}{\hat{P}_j} \text{............................... (6)}$$

where,

$\hat{\lambda}_{jj}$ is the probability correct given map class j,

n_{jj}, $n_{j\cdot}$, $\hat{P}_{ii}$ and $\hat{P}_j$ as previously defined.

Overall accuracy (overall probability correct) is determined by dividing the total correct pixels (sum of the main diagonal elements) by the total number of accuracy assessment pixels in the error matrix (*n*). Overall probabilities corrected for stratified random sampling were computed using the formula developed by Card (1982).

$$\hat{P}_c = \sum_{j=1}^{m} \frac{\pi_i n_{jj}}{n_{j.}} \text{... (7)}$$

where,

$\hat{P}_c$ is the overall probability correct,

m, π_i, n_{jj}, and n_j as previously defined.

The Kappa coefficient of agreement ($\hat{K}$) is the other commonly used discrete multivariate statistical measure that can be used to test LULC classification accuracy based on remote sensing derived datasets and the reference data. The estimator Kappa for stratified sampling is computed as:

$$\hat{K} = \frac{n\sum_{j=1}^{m} \hat{P}_{ii} - \sum_{i=1}^{m} \hat{P}_{j.}\hat{P}_{.i}}{n^2 - \sum_{i=1}^{m} \hat{P}_{j.}\hat{P}_{.i}} \text{........................... (8)}$$

where,

$\hat{K}$ is the Kappa coefficient;

m, n, $\hat{P}_{ii}$, $\hat{P}_{j.}$ and $\hat{P}_{.i}$ as previously defined.

RESULTS AND DISCUSSION

LULC mapping and classification summary

The following analysis is mainly focused on LULC classification derived from multi-level spatial resolution remotely sensed images and validation of these thematic maps. The Modjo watershed historical LULC maps coupled with its areal extent and statistics are presented in Figures 2A, B & C and in Table 2.

In general, extents of bare, cultivated and urban lands have expanded at the expense of other land cover types since 1973. On the other hand, the spatial extent of forest, grassland, shrub and marshy areas have declined primarily due to the conversion of these cover types into cultivated lands. The watershed was dominated by cultivated, forest, grass, shrub and plantations lands in 1973. In 1995, the share of bare, cultivated and built-up areas increased whereas the various vegetation covers such as forest, grass and shrub land categories declined drastically.

Subsequently, the size of cultivated land has expanded at the expense of other LULC classes, and 74.9% of the watershed was under cultivated land category in 2007. These transformation processes have largely been due to human activities. Typically, the increasing demands for cultivated land, fuel wood, construction materials and grazing land aggravated the trends of plant cover fragmentation and degradation. The long history of human settlement, coupled with high population growth and density has been putting escalating stress on vegetation covers and agricultural land productivity.

Accuracy assessment analysis

For this analysis, 562 and 563 reference samples, derived from aerial photo interpretation, were used to validate 1973 and 1995 LULC classifications, respectively (Tables 3 and 5), whereas the 2007 classification was tested using 565 field survey ground truth data (Table 7). In line with the three LULC maps, six error matrices were generated by cross tabulating each LULC classification with the respective reference sample. However, the computed values in Tables 4, 6 and 8 differ from the original error matrices in Tables 3, 5 and 7 because these later values have been corrected for bias by incorporating the true marginal proportions using Equation 2. These tables are populated by the product of individual cell probabilities (P_{ij}) and the known map marginal proportions (π_j) divided by the raw marginal of the original error matrices. Subsequently, various CAA indices were computed from these adjusted error matrices. However, it should be noted that estimates of probability correct, given the map category (user's accuracies) are straightforward using Equation 6 and not affected by this bias adjustment calculation.

Tables 3 and 4 summarize error matrices for the 1973 LULC classification. However, Table 4 is populated by the individual cell probabilities of year 1973 classification after original error matrix (Table 3) bias is corrected. The overall accuracy (overall probability correct) and the $K^\wedge$ agreement of this classification were 88.12% and 81.20%, respectively.

The individual map categories were categorized under high user's accuracies (between 80.33% and 98.08%) whereas producer's accuracies of the individual LULC classes range from 37.54% to 100%. Cultivated land, grassland, shrub land and water bodies are the most accurate classes with combined user's and producer's accuracies of above 80% at individual

Table 2. LULC **types and area measurements of the Modjo watershed for 1973, 1995 and 2007.**

Types	1973		1995		2007	
	Area(km²)	%	Area(km²)	%	Area(km²)	%
Bare	41.48	2.81	46.32	3.13	53.34	3.61
Cultivated	812.75	55	973.24	65.86	1107.15	74.92
Forest	16.87	1.14	7.50	0.51	4.34	0.29
Grassland	319.10	21.59	182.12	12.32	80.50	5.45
Marsh	6.30	0.43	5.22	0.35	4.50	0.30
Plantation	7.86	0.53	23.21	1.57	18.07	1.22
Shrub land	212.74	14.4	161.25	10.91	125.64	8.50
Urban	53.91	3.65	67.42	4.56	74.36	5.03
Open water	6.75	0.46	11.48	0.78	9.86	0.67
Total	1477.76	100	1477.76	100	1477.76	100

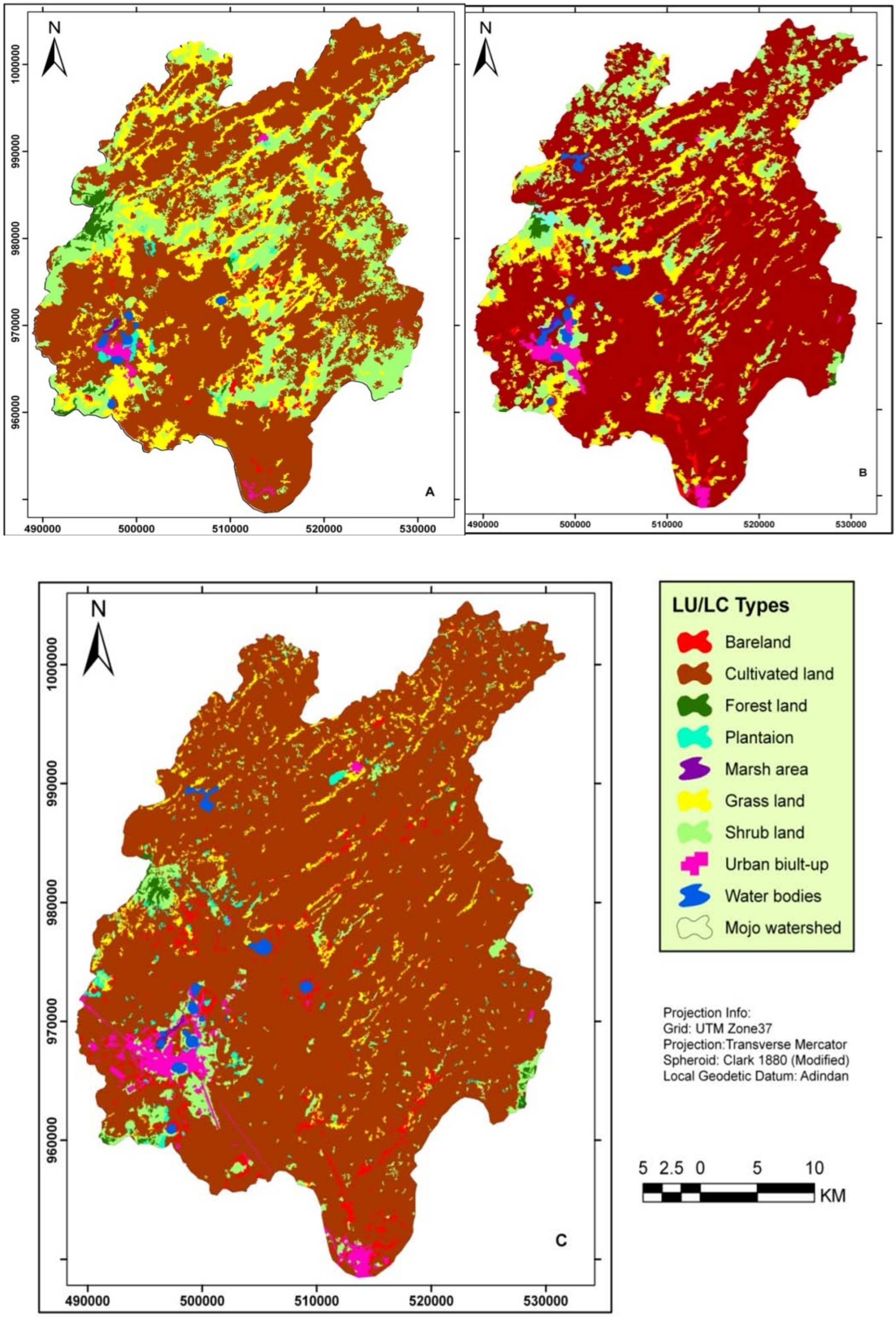

Figure 2. LULC classification maps for the Modjo watershed during the three reference periods based on A) 1973 (MSS Landsat, B) 1995 (TM Landsat), and C) 2007 (Spot) imageries.

class level (Table 4). High classification confusion is observed mainly from bare, cultivated, grassland, shrub and urban lands. For example, classification ambiguities were identified between cultivated land and bare land, cultivated land and grassland, cultivated land and marshland and cultivated land and urban lands. In this regard, 17.65% of the reference samples that should have been classified as cultivated land were excluded from this land use category due to error of omission, whereas 11.58% of sample classified pixels were inappropriately included into the cultivated land category due to error of commission from bare land, grassland, shrub land and urban lands (Table 3).

Although information regarding the type of sampling scheme and bias correction procedure is not clearly stated in many reviewed articles, this finding was comparable to other studies carried out in different regions using remotely sensed data sources and classification approach. For instance, Reis (2008) analyzed the year 1972 MSS Landsat image, which covers about 2,179 km² in the Rize region of northeast Turkey and reported slightly similar overall accuracy and $\widehat{K}$ agreement values of 84.4% and 82.3%, respectively. After analyzing the 1972 MSS Landsat image in Rhode Island (2,808 km²), Novak and Wang (2004) also reported 95% overall classification accuracy for the three generalized LULC classes *i.e.*, urban, forest and agriculture. However, this overall accuracy value confirmed that there is a considerable deviation (higher than about 7%) from the present findings. This is probably due to the fact that the fewer the LULC classes (three classes only), the higher the accuracy value achieved would be.

Table 3. Error matrix of the 1973 LULC classification of the Modjo watershed (in pixels).

Classified samples	Reference Data										
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$n_{j.}$	UAcc. (%)
BL	47	3	-	3	-	-	1	-	-	54	87.04
CL	3	84	-	4	-	-	1	2	-	95	89.36
FL	-	-	47	-	-	-	4	-	-	51	92.16
GL	1	6	-	65	1	-	-	-	-	73	89.04
MA	-	5	-	5	49	-	2	-	-	61	80.33
PL	-	-	-	2	-	47	2	3	-	54	87.04
SL	1	-	4	3	-	3	57	1	-	68	82.61
UL	-	4	-	-	-	2	2	46	-	54	85.19
WB	-	-	-	-	1	-	-	-	51	52	98.08
$n_{.i}$	52	102	51	82	51	52	69	52	51	n=562	OAcc= 87.72%
PAcc (%)	90.38	82.35	92.16	79.27	96.08	90.38	82.61	88.46	100		$\widehat{K}$= 86.09%

Note: BL= Bare land; CL= Cultivated land; FL= Forest; GL= Grassland; MA= Marshy land; PL= Plantation; SL= Shrub land; UL= Urban land; WB= Open water bodies; UAcc.= User's Accuracies; PAcc. Producer's Accuracies, OAcc =Overall Classification Accuracy; $\widehat{K}$ = Kappa Statistics; $n_{.i}$ is the column marginal sum in the reference datasets; $n_{j.}$ is the row marginal sum in the classified image; n is the total number of samples examined and blank cells indicate 0 occurrences.

Table 4. Error matrix of the 1973 LULC classification of the Modjo watershed using individual cell probabilities.

Classified samples	Reference Data											
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$\hat{P}_j$	UAcc (%)	π_j
BL	0.0244	0.0016	-	0.0016	-	-	0.0005	-	-	0.028	87.04	0.028
CL	0.0174	0.4863	-	0.0232	-	-	0.0058	0.0117	-	0.550	89.36	0.550
FL	-	-	0.0105	-	-	-	0.0009	-	-	0.011	92.16	0.011
GL	0.0030	0.0177	-	0.1923	0.0030	-		-	-	0.216	89.04	0.216
MA	-	0.0003	-	0.0003	0.0034	-	0.0001	-	-	0.004	80.33	0.004
PL	-	-	-	0.0002	-	0.0046	0.0002	0.0003	-	0.005	87.04	0.005
SL	0.0021	-	0.0085	0.0064	-	0.0064	0.1207	0.0021	-	0.144	82.61	0.144
UL	-	0.0027	-	-	-	0.0014	0.0014	0.0311	-	0.036	85.19	0.036
WB	-	-	-	-	0.0001	-	-	-	0.004	0.005	98.08	0.005
$\hat{P}_i$	0.0469	0.5087	0.0190	0.2239	0.0065	0.0123	0.1296	0.0487	0.0045	1	OAcc = 88.12%	
PAcc (%)	52.12	95.6	55.4	85.88	52.93	37.54	93.14	68.81	100		$\widehat{K}$ =81.20%	

Note: All abbreviations as defined in Table 3.

Error matrices for year 1995 classification are presented in Tables 5 and 6. Table 6 is populated by individual cell probabilities entries. Based on this error matrix, individual level user's (probability correct, given LULC classified categories) and producer's (probability correct, given true classes) accuracies are ranging from 79.8% to 100% and 46.97% to 100%, respectively. Cultivated, grassland, shrub, urban and water bodies' classes had the highest producer's and user's accuracies (>80%). Regarding user's accuracy, seven LULC categories had the highest values (>90%) whereas grassland and shrub lands have 82.43% and 78.38% user's accuracy values, respectively (Table 6). In this LULC classification, vivid misclassifications (confusions) were identified with LULC classes of bare, cultivated, grassland, and shrub land. Some of this confusion occurs because of the mixed-pixel problem, in which a given entity (a pixel) may have partial membership in more than one LULC category. For instance, some portion of shrub lands were classified as cultivated, grassland and plantation LULC categories. On the other hand, 21.62% of the LULC sample category was incorrectly included into the shrub land category from other classes including bare, cultivated, forest, grassland, plantation and urban land. On the other hand, due to error of omission, 10% of the reference samples that should have been classified as shrub land are excluded from this land use category.

Table 5. Error matrix of the 1995 LULC classification of the Modjo watershed (in pixels).

Classified data	Reference data										
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$n_{j.}$	UAcc (%)
BL	46	3	-	2	-	-	-	-	-	51	90.20
CL	3	98	-	2	-	-	2	1	-	106	92.45
FL	-	-	49	-	-	2	-	-	-	51	96.08
GL	2	3	1	61	2	2	3	-	-	74	82.43
MA	-	2	-	3	49	-	-	-	-	54	90.74
PL	-	-	1	-	-	46	4	-	-	51	90.20
SL	1	5	1	4	-	1	58	4	-	74	78.38
UL	-	-	-	-	-	-	-	48	-	48	100.00
WB	-	-	-	-	2	-	1	-	51	54	94.44
$n_{.i}$	52	111	52	72	53	51	68	53	51	n= 563	OAcc=89.88%
PACC (%)	88.46	88.29	94.23	84.72	92.45	90.2	79.71	92.31	100		$\hat{K}$ = 88.47%

Note: All abbreviations as defined in Table 3.

Table 6. Error matrix of the 1995 LULC classification of the Modjo watershed using individual cell probabilities.

Classified data	Reference data											
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$\hat{P}_j$	UAcc (%)	π_i
BL	0.0283	0.0018		0.0012						0.031	90.20	0.031
CL	0.0186	0.6089		0.0124			0.0124	0.0062		0.659	92.45	0.659
FL			0.0049			0.0002				0.005	96.08	0.005
GL	0.0033	0.0050	0.0017	0.1016	0.0033	0.0033	0.0050			0.123	82.43	0.123
MA		0.0001		0.0002	0.0032					0.004	90.74	0.004
PL		0.0003	0.0003			0.0142	0.0012			0.016	88.46	0.016
SL	0.0015	0.0075	0.0015	0.0060		0.0015	0.0867	0.0045		0.109	79.45	0.109
UL								0.0456		0.046	100.00	0.046
WB					0.0003		0.0001		0.0073	0.008	94.44	0.008
$\hat{P}_i$	0.0517	0.6236	0.0083	0.1214	0.0068	0.0192	0.1055	0.0563	0.0073	1	OAcc=89.95%	
PACC	54.65	97.63	58.44	83.67	46.97	73.82	82.18	81.01	100		$\hat{K}$ = 82.17%	

Note: All abbreviations as defined in Table 3.

The overall accuracy for the 1995 TM Landsat image classification was 89.95% with a $\hat{K}$ agreement of 82.17%, which is by far better than the 1973 LULC classification (Table 5). The computed accuracy assessment statistical measures were more or less similar with results of Kuemmerle *et al.* (2009), who analyzed a 1995 TM Landsat image and achieved 92.5% overall accuracy and 89% of $\hat{K}$ agreement values. Yuan *et al.* (2005) used TM Landsat images of 1986, 1991 and 1998 and an ETM+ image of 2002 to generate seven LULC layers and achieved a better classification result of overall accuracy ranging from 92.6% to 95.5% and $\hat{K}$ values between 90.6% and 94.5%.

Finally, the 2007 LULC classification validation error matrices are presented in Tables 7 and 8. Table 8 shows the adjusted individual cell probability values which are the product of Equation (2). The computed user's accuracy values of year 2007 classification, corrected for bias using the known map marginal proportions, range from 84.13% to 100%. Except shrub land class (84.13%), all LULC categories have high user's accuracies of more than 90%. The user's accuracy of shrub land is the lowest (84.13%). On the other hand, the probability correct value of each true class (the producer's accuracies) is ranging from 28.54% to 99%. Cultivated, plantation, shrub, urban and water bodies' classes have high producer's accuracy >80%. The lowest user accuracy occurs for bare, forest and marsh land categories as a result of spectral confusion with other LULC classes, such as mostly cultivated and grass land classes, because of errors of omission (Table 8).

Table 7. Error matrix of the 2007 LULC classification of the Modjo watershed (in pixels).

Classified Data	Reference Data										
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$n_{j.}$	UAcc (%)
BL	47	2	-	1	-	-	1	1	-	52	90.38
CL	4	119	-	2	1	-	2	-	-	128	92.97
FL	-	-	49	-	-	1	3	-	-	53	92.45
GL	-	3	-	56	-	-	1	-	-	60	93.33
MA	-	1	-	2	51	-	-	-	-	54	94.44
PL	-	-	1	1	-	48	-	1	-	51	94.12
SL	2	1	2	1	1	1	53	1	1	63	84.13
UL	-	1	-	-	-	1	1	49	-	52	94.23
WB	-	-	-	-	-	-	-	-	52	52	100.00
$n_{.i}$	53	127	52	63	53	51	61	52	53	565	OAcc=92.74%
PAcc	88.68	93.70	94.23	88.89	96.23	94.12	86.89	94.23	98.11		$\hat{K}$=91.48%

Note: All abbreviations as defined in Table 3.

Table 8. Error matrix of the 2007 LULC classification of the Modjo watershed using individual cell probabilities.

Classified Data	Reference Data											
	BL	CL	FL	GL	MA	PL	SL	UL	WB	$\hat{P}_j$	UAcc (%)	π_i
BL	0.0326	0.0014		0.0007			0.0007	0.0007		0.036	90.38	0.036
CL	0.0234	0.6965		0.0117	0.0059		0.0117			0.749	92.97	0.749
FL			0.0027			0.0001	0.0002			0.003	92.45	0.003
GL		0.0027		0.0508			0.0009			0.054	93.33	0.054
MA		0.0001		0.0001	0.0029					0.003	94.44	0.003
PL			0.0002	0.0002		0.0115		0.0002		0.012	94.12	0.012
SL	0.0027	0.0013	0.0027	0.0013	0.0013	0.0013	0.0715	0.0013	0.0013	0.085	84.13	0.085
UL		0.0010				0.0010	0.0010	0.0474		0.050	94.12	0.050
WB									0.0067	0.007	100	0.007
$\hat{P}_i$	0.059	0.703	0.006	0.065	0.010	0.014	0.086	0.050	0.008	1	OAcc=92.27%	
PAcc (%)	55.54	99.07	48.02	78.29	28.54	82.80	83.18	95.40	83.18		$\hat{K}$ = 83.11%	

Note: All abbreviations as defined in Table 3.

Based on this image classification result, 92% of overall accuracy and 83.11% of $\hat{K}$ agreement were achieved for the 2007 SPOT image classification (Table 8). This overall accuracy value agrees well with results of other studies. For instance, Yang *et al.* (2009) carried out LULC mapping using SPOT -5 images to identify crop types and crop areas in two different sites of south Texas (USA) and achieved overall accuracy values of 87% and 91% for the two study sites.

In general, the calculated producer's accuracies as well as $\hat{K}$ agreement values for the bare, forest, marshy and plantation layers in the original error matrix (Tables 3, 5 and 7) are higher than the computed values of bias correction using known map marginal proportion (Tables 4, 6 and 8). The reason for this difference is that the adjusted producer's and overall accuracies as well as $\hat{K}$ agreement values have been corrected using known map marginal proportions for the areal bias introduced by the stratified random sampling scheme. This illustrates that the known map marginal proportion (π_j) is very important to remove sample bias introduced in line with random sampling design. However, there is a slight difference in overall accuracy values among the original and adjusted matrices using known map marginal proportion. Besides, the 2007 LULC classification was achieved at the highest level of accuracy, as compared with the 1973 MSS and 1995 TM Landsat image classification (Tables 4, 6 and 8).

CONCLUSIONS

Earth observation remote sensing satellites provide essential inputs for LULC classification at watershed scales, which are not commonly inventoried through ground surveying methods. However, in order to use the LULC thematic layers for a particular application, users should know how accurate these maps are. The reason is that the importance of any thematic layer derived from remotely sensed data depends on its fitness for various applications and the quality of the derived map should be proved in terms of the accuracy of classification, which is determined during the CAA stage and is the focus of this study. Although the applicability of historical thematic map accuracy assessment is still a challenge, due to lack of retrospective ground truth or field survey referenced data (Foody, 2002; Congalton and Green, 2009), the present study has addressed a vital issue for earth observation science practitioners and remotely-sensed data users regarding how to adequately perform CAA, using historical aerial photographs as reference data sources in the absence of ground truth datasets.

Reliable LULC quality maps were extracted using a deterministic (pixel based) supervised image classification approach. Although some classification errors, mainly between bare land, cultivated, shrub and grasslands were observed from the three years classified maps, the computed overall accuracies and Kappa agreements fulfilled the minimum thematic mapping CAA quality management standard (Anderson *et al.*, 1976, Landis and Koch, 1977; Arora *et al.*, 2005; Brown, 2005; Jensen, 2005; Congalton and Green, 2009). Relatively speaking, high producer's, user's and overall accuracies as well as $\hat{K}$ agreement values were achieved from the 2007 SPOT image classification, as compared with the 1973 and 1995 LULC classification. The reason is the higher spatial resolution (10 m) of the 2007 SPOT image compared with the 1995 (30 m) and 1973 (57 m) TM and MSS Landsat images, respectively. In addition, this study confirmed that economical, accurate and dependable LULC thematic layers can be derived from multi-temporal satellite images having different spatial resolution with acceptable quality. Therefore, these thematic layers can be used as input to: i) describe and quantify the structures, compositions and distributions of landscape elements; ii) use various environmental modelling applications; or iii) serve as bases for policy options for devising integrated landscape and land resource management. However, to minimize the problem of classification confusion resulting from mixed pixels and to obtain optimal level classification, research using a fuzzy set theory (sub-pixel) classification approach is still desirable.

ACKNOWLEDGMENTS

This research was partially supported by the Graduate School of Addis Ababa University and the German Academic Exchange Service (DAAD). Special thanks go to the Global Land Cover Facility (GLCF) for offering free access to the Landsat satellite imagery. Finally, we are grateful to the reviewers and editors of *SINET* for their insightful and helpful comments.

REFERENCES

1. Anderson, J.R., Hardy, E.E., Roach, J.T. and Witmer, R.E. (1976). A land use and land cover classification system for use with remote sensor data. United States of Geological Survey Professional PaperUS Government Printing Office, Washington, 964 pp.
2. Arora, M.K., Shalan, M.A. and Elgy, J. (2005). Crisp and fuzzy classification accuracy measures for remotely sensed derived land cover maps. In: *Geodynamics,* pp 11–28, (Atkinson, P.M., Foody, G.M., Darby, S.E. and Wu, F., eds). CRC Press, Boca Raton.
3. Attua, E.M. and Fisher, J.B. (2011). Historical and future land-cover change in a municipality of Ghana. *Earth Interactions* **15:**1–26.
4. Bezuayehu Tefera and Sterk, G. (2008). Hydropower-induced land-use change in Fincha'a watershed, western Ethiopia: Analysis and impacts. *Mountain Research and Development* **28(1):**72–80.
5. Brown, K. (2005). Per-pixel uncertainty for change detection using airborne sensor data. **In:** *Geodynamics,* pp 103–120, (Atkinson, P.M., Foody, G.M., Darby, S.E. and Wu, F., eds). CRC Press, Boca Raton.
6. Bruzzone, L. and Serpico, S.B. (1997). Detection of changes in remotely sensed images by the selective use of multi-spectral information. *International Journal of Remote Sensing* **18:**3883–3888.
7. Card, D.H. (1982). Using known map categorical marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing* **48:**431–439.
8. Comber, A.J., Law A.R. and Lishman, J.R. (2004). Application of knowledge for automated land cover change monitoring. *International Journal of Remote Sensing* **25:**3177–3192.
9. Congalton, R.G. and Green, K. (2009). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices.* 2nd ed. CRC Press, Taylor & Francis Group, Boca Raton.
10. Congalton, R.G., Oderwald, R.G. and Mead, R.A. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing* **49:**1671–1678.
11. Congalton, R. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* **37:**35–46.
12. De Jong, S.M., Van der Meer, F.D. and Clevers, J. G.P. (2005). Basics of remote sensing. **In:** *Remote Sensing Image Analysis: Including the Spatial Domain*, pp. 1–16, (de Jong, S.M. and van der Meer, F.D., eds). Springer Science and Business Media, Inc., Dordrecht.
13. Foody, G.M. (2002). Status of land covers classification accuracy assessment. *Remote Sensing of Environment* **80:**185–201.
14. Gao, J. (2009). *Digital Analysis of Remotely Sensed Imagery.* McGraw-Hill Companies, Inc., New York.
15. Gete Zeleke and Hurni, H. (2001). Implications of land use and land cover dynamics for mountain resource degradation in the Northwestern Ethiopian highlands. *Mountain Research and Development* **21(2):**184–191.
16. Hunter, L.M. (2000). The Environmental implications of population dynamics. RAND publications, Santa Monica. http://www.rand.org/-pubs/monograph_reports/MR1191.html. Accessed on 26 March 2015.
17. International Organizations for Standardizations (ISO/TC-211) (2002). Geographic information/Geomatics, 19107: Geographic information-Spatial schema, text for FDIS, N. 1324. http://www.isotc211.org/pow.htm. Accessed on 7 December 2013.
18. Jensen, J.R. (2005). *Introductory Digital Image Processing: A Remote Sensing Perspective*. 3rd ed. Prentice Hall, Upper Saddle River, NJ.
19. Kibrom Tekle and Hedlund, L. (2000). Land-cover changes between 1958 and 1986 in Kalu District, Southern Wello, Ethiopia. *Mountain Research and Development* **20**(1):42–51.
20. Kuemmerle, T., Muller, D., Grifths, P. and Rusu, M. (2009). Land use change in southern Romania after the collapse of socialism. *Regional Environmental Change* **9:**1–12.
21. Landis, J.R., and Koch, G.G. (1977). The measurements of observer agreement of categorical data. *Biometrics* **33:**159–174.
22. Lillesand, T.M., Kiefer, R.W. and Chipman, J.W. (2008). *Remote Sensing and Image Interpretation.* 6th ed. Wiley, New York.
23. Litwin, L. and Rossa, M. (2011). Geoinformation metadata INSPIRE and SDI. Understanding. Editing. Publishing. Lecture Notes in Geoinformation and Cartography, Springer-Verlag Berlin Heidelberg.
24. Lopez, E., Bocco, G., Mendoza, M. and Duhau, E. (2001). Predicting land-cover and land-use change in the urban fringe: A case in Morelia city, Mexico. *Landscape and Urban Planning* **55:**271–285.
25. Meng, L. (2010). Cartography. **In**: Handbook of geo-mathematics, Chapter 16 (Freeden, W., Nashed, M.Z. and Sonar, T., eds). Springer-Verlag, Berlin.
26. Metzger, M.J., Rounsevell, M.D., Acosta-Michlik, L., Leemans, R. and Schroter, D. (2006). The vulnerability of ecosystem services to land use change. *Agriculture, Ecosystems and Environment* **114:**69–85.

27. Novak, A.B. and Wang Y.Q. (2004). Effects of suburban sprawl on Rhode Island's forests: A Landsat view from 1972 to 1999. *Northeastern Naturalist* **11:**67–74.

28. Pontius Jr., R.G. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing* **66:**1011–1016.

29. Pontius Jr., R.G. and Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing* **32:**4407–4429.

30. Reid, R.S., Kruska, R.L., Muthui, N., Andualem Taye, Wotton, S., Wilson, C.J. and Woudyalew Mulatu (2000). Land-use and land-cover dynamics in response to changes in climatic, biological and socio-political forces: The case of southwestern Ethiopia. *Landscape Ecology* **15:**339–355.

31. Reis, S. (2008). Analyzing land use/land cover changes using remote sensing and GIS in Rize, northeast Turkey. *Sensors* **8:**6188–6202.

32. Sierra, R. (2000) Dynamics and patterns of deforestation in the western Amazon: the Napo deforestation front, 1986–1996. *Applied Geography* **20:**1–16.

33. Steffen, W., Sanderson, A., Tyson, P.D., Jäger, J., Matson, P.A., Moore III, B., Oldfield, F., Richardson, K., Schellnhuber, H.J., Turner, B.L. and Wasson, R.J. (2005). *Global Change and the Earth System: A Planet Under Pressure.* Springer, Berlin. ISBN 3-540-40800-2.

34. Stehman, S.V. and Czaplewski, R.L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of the Environment* **64:**331–344.

35. Talukdar, G., Ghosh S. and Roy, P.S. (2004). Landscape dynamics in north east region of India (Meghalaya State) using spatial decision tree model. *Geocarto International* **19:**11–18.

36. Thapa, R.B. and Murayama, Y. (2008). Spatial structure of land use dynamics in Kathmandu valley. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **37:**11–16.

37. Van Genderen, J.L., Lock, B.F. and Vass, P.A. (1978). Remote sensing: statistical testing of thematic map accuracy, *Remote Sensing of the Environment* **7:**3–14.

38. Woldeamlak Bewket (2002). Land cover dynamics since the 1950s in Chemoga Watershed, Blue Nile basin, Ethiopia. *Mountain Research and Development* **22:**263–269.

39. Wu, X., Shen, Z., Liu, R. and Ding, X. (2008). Land use/cover dynamics in response to changes in environmental and socio-political forces in the upper reaches of the Yangtze River, China. *Sensor* **8:**8104–8122.

40. Yang, C., Everitt, J.H. and Murden, D. (2009). Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture* **75:**347–354.

41. Yuan, F., Sawaya, K.E., Loeffelholz, B.C. and Bauer, M.E. (2005). Land cover classification and change analysis of the Twin Cities (Minnesota) metropolitan area by multi-temporal Landsat remote sensing. *Journal of Remote Sensing of Environment* **98:**317–328.