Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 6 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data Preprocessing in Multi-Temporal Remote Sensing Data for Deforestation Analysis

By Dr. Manjula. K.R, Dr. Jyothi. Singaraju
& Prof. Anand Kumar Varma. Sybiyal
SASTRA University, Tamilnadu

Abstract - In recent years, the contemporary data mining community has developed a plethora of algorithms and methods used for different tasks in knowledge discovery within large databases. Furthermore, algorithms become more complex and hybrid as algorithms combining several approaches are suggested, the task of implementing such algorithms from scratch becomes increasingly time consuming. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data pre-processing steps are often necessary to prepare data for further usage. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results.

Keywords : *data preprocessing, data mining, remote sensing images, deforestation analysis.*

GJCST-C Classification : J.1

Strictly as per the compliance and regulations of:

Data Preprocessing in Multi-Temporal Remote Sensing Data for Deforestation Analysis

Dr. Manjula. K.R [α], Dr. Jyothi. Singaraju [σ] & Prof. Anand Kumar Varma. Sybiyal [ρ]

Abstract - In recent years, the contemporary data mining community has developed a plethora of algorithms and methods used for different tasks in knowledge discovery within large databases. Furthermore, algorithms become more complex and hybrid as algorithms combining several approaches are suggested, the task of implementing such algorithms from scratch becomes increasingly time consuming. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data pre-processing steps are often necessary to prepare data for further usage. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results.

This paper introduces and defines the study area and throws light on the data preprocessing on both collateral and image data. Under data preprocessing, the non spatial data are preprocessed with normalization, generalization and other techniques. For the satellite image, the preprocessing is done both at the image dissemination and during feature extraction process. These data are preprocessed to fill data gaps and correct data anomalies. This paper provides a brief description of local maximum likelihood method, pepper salt method, boundary clean method and edge matching methods which are used while classifying the image.

Keywords : data preprocessing, data mining, remote sensing images, deforestation analysis.

I. Introduction

The technical progress in computerized data acquisition and storage results in the growth of vast databases. With continues increase and accumulation, the huge amount of the computerized data have far exceeded human ability to completely interpret and use. Users need adequate search tools in order to quickly access and filter relevant information. The development of novel technique and tools in assist for humans aiding in the transformation of data into useful knowledge, has been the heart of the comparatively new and interdisciplinary research areas called “Knowledge Discovery in Databases (KDD)”. With rapid growth in development of research in data mining order to quickly access and filter relevant information. The development of novel technique and tools in assist for humans aiding in the transformation of data into useful knowledge, has been the heart of the comparatively new and interdisciplinary research areas called “Knowledge Discovery in Databases (KDD)”. With rapid growth in development of research in data mining and data warehouse, many systems were emerged in those fields.

It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results. What variables should be selected? What measurement framework, such as Euclidean space or non-metric network space, should be used? What spatial relations or contextual information should be considered? Can the chosen data adequately represent the complexity and nature of the problem?

a) Study Area

The setting of this study spans an area of 5000 Square Kilometers and it includes the mandals of Chittoor such as Thirupathi, Kalahasthi, Yerpedu, Renigunta and major portion of Kadapa Mandals such as Nandalur, Chitvel, Rajampet, Pullampet, Obulavari Palli, Kodur, and Nellore District mandals such as Venkatagiri, Rapur, Kaluya and Takkili. The study area boundary in lat-long is E 79 39" to E 78 45" and N 13 35" to N 14 33". The study area district outline is specified in the Figure 1:

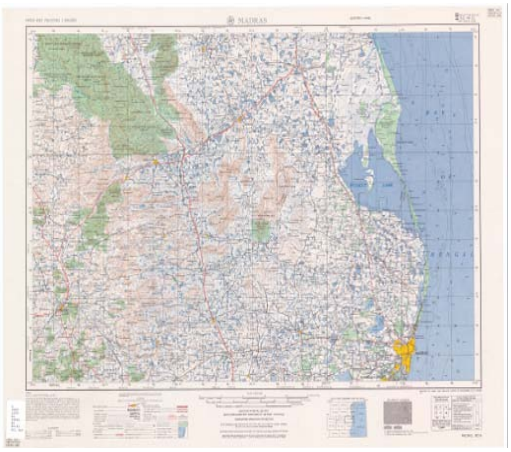
Figure 1.1 : Map Showing the District Outline Containing the Study Area

Beside the undamaged natural environment in some parts, a big part of the area has been changed by

Author α : SAP, School of Computing, Dept. of CSE, SASTRA University, Tirumalaisamudram, Tamil Nadu.
E-mail : manju_sakvarma@yahoo.co.in
Author σ : Professor & BOS Chairperson, Dept. of CS, SPMVV, Tirupati, Chittoor District, Andhra Pradesh.
E-mail : jyothi.spmvv@gmail.com
Author ρ : Professor, Dept. of Civil Engineering, SIETK, Puttur.
E-mail : manju_sakv@yahoo.co.in

agriculture and grazing activities. The following figures 1.2(a) and (b), 1.3.(a), (b) and (c) and 1.4 represents top sheets, satellite images path row of the study area and scene and satellite image along with scanned mandal boundary map of Cuddapha district.

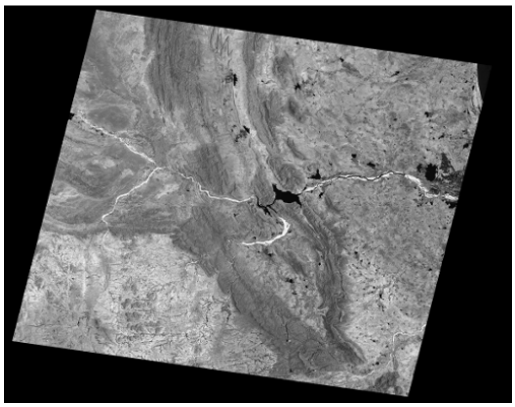
(a)



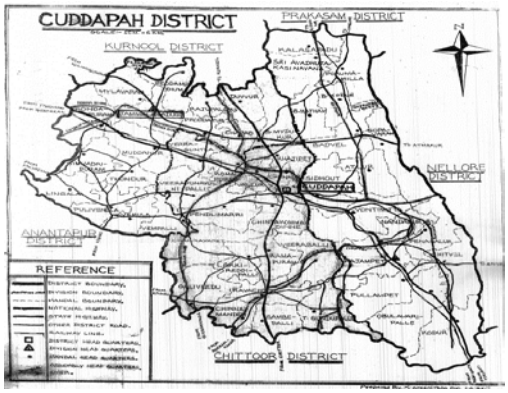
(b)

Figure 1.2 : Topography and Forest Area of Study area is shown in the Map

(a)



(b)



(c)

Figure 1.3 : Reference Map of the study: (a) Path Row boundary, (b) Scene of the study area, (c) Scanned paper map of Cuddapah District

i. *Data Collection*

Spatial information on Land use / Land cover is a necessary prerequisite in planning, utilizing and management of the natural resources. The study area is based on the secondary data, the satellite imagery which is downloaded from the Global Land cover Mapping web site and National Remote Sensing Agency and Survey of India. Three kinds of data are used for this study:

- Satellite data,
- Topographic and thematic maps and
- Descriptive data.

The necessary *satellite data* is selected and acquired after visits to the three test areas. The main criteria for selection of satellite data are:

- Date of acquisition of images according to the local calendar;
- Weather conditions (cloud cover) during the acquisition of images;
- Spectral and spatial resolution of images.
- After examine the above list, the following satellite images available for study area, one Landsat Thematic Mapper, Enhanced Thematic Mapper and IRS P6 LISS III images are ordered.

Table 1.1 : Various Inputs used in the Study

S.No	Type of Map	Resolution/ Scale	Date/Year of Acquisition	Source
1	Topsheet	1:50,000	57 O/5 – 1973-79	SOI
		1:50,000	57 J/11 – 1973	SOI
		1:50,000	57 O/6 – 1973	SOI
		1:50,000	57 N/9 – 1973	SOI
2	Landsat – TM	28.5 mt	1991	GLCF
3	Landsat – ETM+LISS 3	Medium 250,000	05th April, 2001	NRSC
4	IRS P6. LISS 3 101 – 63	Medium 250,000	6th Feb, 2010	NRSC
5	Mandal Maps	A3 Size	-	Mandal HQ

ii. *Review of Literature*

Amos Storkey [2] proposed various data preprocessing methods applied on any data before applying data mining techniques to ensure the quality of decision making. Aleksandar Lazarevic et al [3] proposed the software system for spatial data analysis and modelling (SDAM) which provide flexible machine learning tools for supporting an interactive knowledge discovery process in large centralized or distributed spatial databases. Caroline M. Bruce and David W. Hilbert [4] suggested a Pre-processing methodology for application to Landsat TM/ETM+. This report details the various pre-processing techniques either to derive multitemporal and multispatial image classifications or to use in biophysical/geochemical modelling. P.S. Roy et al [10] proposed a multilevel land use land cover classification system, wherein LULC information can be accessed Nationwide, State wide and at the intrastate, regional or municipal level. Stefan Erasmi et al [11] evaluated available satellite data sets and established a transparent work flow for the monitoring of past and future land cover dynamics at a regional scale based on medium resolution satellite data while *mapping* deforestation and land cover conversion at the rainforest margin in central Sulawesi, Indonesia.

II. Data Preparation

One of the methods for change detection using satellite images is to compare the results of classified images. The advantage of the classified-map comparison method is that not only the location but also the nature and type of the changes are determined in the study area. In this method, first, the images of different times are classified according to the purpose of change detection. Afterward, by overlaying these classified images with a proper overlay condition, the location and amount of these changes that are interested is determined. As the goal is to determine the deforestation, the only two classes that are considered are the forest and non-forest.

III. Data Preprocessing

Under data preprocessing, the non spatial data are preprocessed with normalization, generalization and other techniques. For the satellite image, the preprocessing is done both at the image dissemination and during feature extraction process. These data are preprocessed to fill data gaps and correct data anomalies. This paper provides a brief description of various preprocessing methods that is applied on the collected images in order to achieve the data quality of the study while classifying the image.

a) Part I - Collateral Data Preprocessing

Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data pre-processing steps are often necessary to prepare data for further usage. The following figure explores the pre-processing steps generally used for all type of data [3].

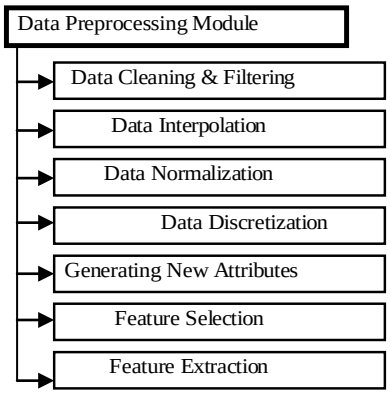


Figure 3.1 : Data Preprocessing Functions

i. *Data Cleaning and Filtering*

Due to the high possibility of measurement noise present in collected data sets, there is a need for data cleaning. Data cleaning consists of removing duplicate data points, and removing value outliers, as well spatial outliers. Data can also be filtered or smoothed by applying a median filter with a window size specified by the user.

ii. *Data Interpolation*

In many real life spatial domain applications, the resolution will vary among data layers and the data will not be collected at a common set of spatial locations. Therefore, it is necessary to apply an interpolation procedure to the data to change data resolution and to compute values for a common set of locations. Deterministic interpolation techniques such as inverse distance and triangulation can be used but they

do not take into account a model of the spatial process or variograms.

iii. *Data Normalization*

The system supports two normalization methods: the transformation of data to a normal distribution and the scaling of data to a specified range. In this work, normalization is applied for the image data while georefrencing the three time period based images.

iv. *Data Discretization*

This step is necessary in some modeling techniques like association rules, decision tree learning and all classification problems and includes different attribute and target splitting criteria. In this work discretization is applied for collateral data that includes population data with diversity in data. So this data is discretized into three ranges of groups as 'High', 'Low' and 'Medium'.

Table 3.1 : Discretization of population in the ranges

Population Size & Growth(Difference)	Density = Population / Area	Range Label
500 <	<100	Low
>=500<1000	>=100 and < 200	Medium
>=1000	> 200	High

v. *Generating New Attributes*

Users can generate new attributes by applying supported operators to a set of existing attributes. The density range, population range etc., are created as new attributes for the study.

vi. *Feature Selection*

In domains with a large number of attributes this step is often beneficial for reducing attribute space by removing irrelevant attributes. Several selection techniques (Forward Selection, Backward Elimination, Branch and Bound) and various criteria (inter-class and probabilistic selection criteria) are supported in order to identify a relevant attribute subset.

In this thesis, while preparing a single table input for association rule mining, some of the attributes in individual tables are removed as irrelevant. For deriving rules the attributes such as Gridcode, Area etc are removed as it does not give any meaningful information while deriving rules.

vii. *Feature Extraction*

In contrast to feature selection where a decision is target-based, variance-based dimensionality reduction through feature extraction is also supported. The transformed data can be plotted in d-dimensional space and resulting plots can be rotated, panned and zoomed to better view possible data groupings.

viii. *Data Partitioning*

Partitioning allows users to split the data set into more homogenous data segments, thus providing better modeling results.

b) Part – II - Image Data Preprocessing

Availability, Accessibility, and Affordability of Remote Sensing Data, a range of airborne and space-borne sensors has acquired remote sensing data, with the number of sensors and their diversity of capability increasing over time. Ideally, the following image characteristics are required for studying defore -station [1][4].

- Cloud free and clear atmosphere during the time of data acquisition;
- Availability of imagery for the optimum date or dates;
- Spatial resolution fine enough for accurate mapping and course enough so image size is manageable;
- Band selection (band width, placement, and number of bands) optimized to identify features of interest;
- Study area covered on a single image;
- Same sensor and sun position when images were acquired similar atmospheric conditions.
- Pragmatically, it is rather difficult to acquire the data with the above characteristics. Instead, the following problems are common in data acquisition process:
- Unavailability of data for specific time period;
- Persistent cloud coverage throughout the year and for many years;
- Cost is too high specially for commercial satellite data;
- Availability of data in usable format (digital or hard copy);
- Cost of processing, in producing value added product,
- Lack of expertise, equipment/software for analysis;
- Significant improvements have been made in terms of spectral, spatial, temporal and radiometric resolutions. More specifically, improvements have been observed in
- Visibility and clarity that includes more detailed image of a smaller piece of land;
- Clear definition involving more precisely the specific colours or light responses reflecting off of the field; and
- Frequent data acquisition on a regular interval of every other day or every 5-7 days.
- The background environment reflected through the remote sensing image obtained in different instant is different because of the influence of various factors in the acquisition process. These factors can be divided into two categories: remote sensing system factors and environmental factors.
- The remote sensing system factors are: the impact of temporal, spatial, spectral and radiation resolution.
- The environmental factors are: The impact of atmospheric conditions, soil moisture and phonological characteristics.

The impact at different times and the influence of these factors on the images must be fully taken into account in the change detection. The influence may be eliminated as much as possible by the geometric registration and radiometric correction on the remote sensing images.

Preprocessing and Analysis of the Satellite Images

Prior to data analysis, initial processing on the raw data is usually carried out to correct for any distortion due to the characteristics of the imaging system and imaging conditions. Depending on the user's requirement, some standard correction procedures may be carried out by the ground station operators before the data is delivered to the end-user. Figure 3.20 derives the processing procedure applied to image data [1] [4] [10].

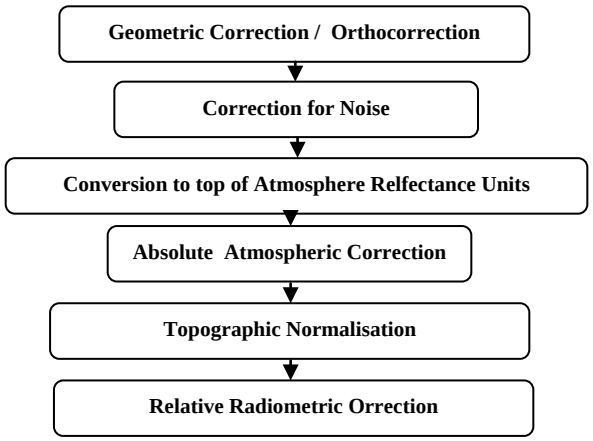


Figure 3.2 : Processing Procedures Applied to Imagery Data

Usually three types of errors occur when a satellite image is generated by the satellite sensor. The first is the sensor error. The second is the error created by the atmospheric parameters, which affect the amount of radiation received by the sensor. The third one is the geometric errors related to the curvature of the Earth surface, the Earth rotation, elevation differences, location and situation of the satellite etc. Therefore, these errors should be considered and managed before using the data:

1. *Sensor Errors:* The two images used were already corrected by their providers. Therefore, there was no need for any processing in this regard.
2. *Radiometric Correction:* The Earth atmosphere scatters the shorter wavelengths in a selective manner and this reduces the contrast of the image. The numerical value of each pixel in the image is not a realistic representation of the amount of radiation from the ground surface. These values are changed either by atmospheric absorption or by scattering throughout the atmosphere.
3. *Atmospheric Effects:* Scattering and absorption of EM radiation by the atmosphere have significant effects that impact sensor design as well as the processing and interpretation of images. When the concentration of scattering agents is high, scattering produces the visual effect we call haze. Haze increases the overall brightness of a scene and reduces the contrast between different ground materials. In general, atmospheric errors are discussed in three parts: the Haze, Sun angle and Skylight errors.

Atmospheric corrections are required in the following situations:

- When user want to compare the images related to different times.
- When using methods such as image subtraction and image division for change detection, the effect of atmosphere on the two images related to different times are quite different.
- When the ratio of two bands of an image is needed to be calculated, because the atmosphere has different effects on different wavelengths.
- When user want to study spectral characteristics of different phenomena.

If user wanted to use the division or subtraction of images for determining the changes in forest land use, then user would have to correct for the haze, sun angle and skylight errors [4]. In this approach the results of the land use classification maps extracted from the three images are compared. The classification of land use can be done better and more accurate with the raw (unprocessed) images. Therefore, there was no need for the above corrections for images used in this study.

- Geometric Corrections: The process and analysis of multi-temporal data can be done only when they are geo-referenced similarly, or in another words, when they are geo-referenced to each other [11]. The images of this study had to be geo-referenced to each other with an accuracy of one pixel. Otherwise, the error coming from different coordinates for similar objects in the two images can be wrongly accepted as a land use change. To prevent such a problem, in comparison of multi-temporal images, geo-reference one of the images using the available topographic maps and then geo-referencing the other images according to the first one, i.e. using image-to-image registration is done.

In photo/image registration (geo-referencing), the most important task is the proper selection of control points, especially when there is a long time period between the map and the image. In this work, the first order polynomial equations for geo-referencing of the images is used, which remove the errors related to the rotation and scaling of the image. The image may also be transformed to conform to a specific map projection system. Furthermore, if accurate geographical location

of an area on the image needs to be known, ground control points (GCP's) are used to register the image to a precise map (geo-referencing).

In this study, the ETM+ image of the year 2001 was first geo-referenced using the information in its header approximately. Then, it was geo-referenced accurately using the available 1:25000 digital maps and the digitized features of the 1:50000 maps of the area. Afterward, the TM image of 1991 was geo-referenced using the already registered TM image. For geo-referencing the 2001 image 18 control points were used initially. Every control point with an RMSE or residual error bigger than a pixel size was removed from the calculation and the process of registration was repeated with the rest of the control points. Finally, 10 points with the average error of 01.00 meters remained and were used for registration. For image-to-image registration of the 1991 image 20 control points were initially used. Finally, 6 points were removed and the image was geo-referenced using the remained 14 points with the RMSE of 0.92 meters.

All images and aerial photographs were rectified to UTM zone 39 N with at least 25 well distributed ground control points. At first geometric correction was carried out using topographic maps with the scale of 1:25000 to geo code aerial photos. Also for geometric correction of the 2001 IRS-1C land sat image, topographic maps with the scale of 1:25000 were used and then this rectified image was employed to register the 2011 LISS-III image. Geometric correction of Land sat TM image of 1990 was carried out by the use of IRS-P6 LISS-III image. Finally, a first-order polynomial model was applied and all data were resampled to a 30 m pixel size using the nearest neighbour method. After geometric correction of aerial photos, all photos for each year were mosaic ked to prepare one image for land cover mapping.

- *Image Enhancement:* In order to aid visual interpretation, visual appearance of the objects in the image can be improved by image enhancement techniques such as grey level stretching to improve the contrast and spatial filtering for enhancing the edges [4]. The goal of image enhancement is to improve the visual interpretability of an image by increasing the distinction between features. In this study, two false colour composites (FCC) are produced for selecting training samples. Also image fusion was done to increase spatial resolution of the LISS-III image. LISS-III image was fused with IRS-1C PAN image to generate an image with high spatial resolution [1]. Land sat TM enhanced false colour composites RGB (red, green, blue) 4,5,3; 5,3,2; 4,5,7 and 4,3,2 are used for the interpretation and delimitation of the land cover classes [11]. Interpretation and vectorization on the screen, available in Arc Info format was the preferred methodology because polygons created have vector format and can be directly transformed to a land cover map.
- *Neighborhood Filling:* This method has been used to clean and fill the missing cell in the image while doing image classification.
- *Edge Matching:* This features is carried out to maintain the continuity of classes between adjoining mandals/districts/states. Generation of seamless geo data set at district/state level, creation of metadata, class wise area statistics are prepared.
- *Aerial Photos Interpretation:* Land cover pattern is interpreted visually on black and white aerial photographs and simultaneously digitized with the Arcmap software. Identifying features in aerial photos is performed based on tone, texture, pattern, size and shape.
- *Post-Classification Change Detection:* Post-classification comparison change detection algorithm is used to determine changes in urban areas in 3 decades from 1991 to 2011. Finally, due to anthropogenic activity, changes such as the reduced vigour of forest vegetation, urbanization, mining etc are noticed in the area.

IV. Conclusion

Data mining is data-driven but also, more importantly, human-centered, with the user controlling the selection and integration of data, cleaning and transformation of the data, choice of analysis methods, and the interpretation of results. The abundance of spatial data provides exciting opportunities for new research directions but also demands caution in using these data. The data are often from different sources and collected for different purposes under various conditions, such as measurement uncertainty, biased sampling, varying area unit, and confidentiality constraint. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results.Pre-processing improves performance, but massive data volumes associated with encoding spatial relationships for all combinations of geographic objects prohibits the storage of all spatial relationships.

References Références Referencias

1. Arms ton J.D, Danaher, T.J, Goulevitch, B.M and Byrne, M.I, "Geometric Correction of Land sat MSS,TM and ETM+Imagery for Mapping of Woody Vegetation Cover and Change Detection in Queensland", Climate Impacts and Natural Resource Systems, ISBN0-9581366-0, www.nrm.-qld.gov.au/slats, 2002.
2. Amos Storey, "*Data Mining and Exploration: Preprocessing*", School of Informatics, http://www. -

inf.ed.ac.uk/teaching/ courses/dme/, January 23, 2006.
3. Aleksandra Lazarevic, Tim Fiez and Zoran Obradovic, *"A Software System for Spatial Data Analysis and Modeling",* INEEL University Research Consortium project No: C94-175936, www.ist. -temple.edu/~zoran/papers/lazarevic00.pdf.
4. Caroline M. Bruce and David W. Hilbert, "*Pre-processing Methodology for Application to Land sat TM/ETM+ Imagery of the Wet Tropics",* Cooperative Research Centre for Tropical Rainforest Ecology and Management. Rainforest CRC, Cairns. (44 pp.), ISBN: 0864437609, www.rainforest-crc.jcu.edu.au, March 2006.
5. Principles of Remote Sensing- Centre for Remote Imaging, Sensing and Processing, CRISP, www.crisp.nus.edu.sg/~research/tutorial/rsmain.ht ml.
6. Hutchinson C, *"Techniques for Combining Land sat and Ancillary Data for Digital Classification Improvement",* Photogrammetric Engineering and Remote Sensing, Vol.48, No.1, 123-130, 1982.
7. Luis Otavo Alvares, Gabriel Oliveira, Vania Bogorny, "*A Framework for Trajectory Data Preprocessing for Data Mining*", http://www.inf.ufsc.br/~ vania/artigos/ seke2009_6.pdf.
8. Lilles and TM and Keifer W, *"Remote Sensing and Image Interpretation",* New York: John Wiley, 1994.
9. Loveland T.R, Sohl T.L, Stedman S.V, Gallant A.L, Saylor K.L and Nap ton D.E, "A strategy for estimating the rates of recent United States land-cover changes. *Photogrammetric Engineering and Remote Sensing", 68*, 1091–1100, 2002.
10. Roy P.S, Dwivedi R.S and Vijay an P, "*Remote Sensing Applications-Land Use Land Cover Analysis",* National Remote Sensing Centre, 2011.
11. Stefan Erasmi, Andre Twele, Muhammad Ardiansyah, Adam Malik and Martin Kappas, *"Mapping Deforestation and Land Cover Conversion at The Rainforest Margin in Central Sulawesi, Indonesia",* EAR SeL proceedings 3, 2004.