A Comparative Study of Unsupervised Machine Learning Algorithms for Recognition of Change in Vegetation, Water and Built up Land Areas

Kumar Santosh[1], Singh Pitam[2], Das Bharti[3], Gaur Monika[4] and Singh Priyamvada[5*]

1. GIS Cell, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, INDIA
2. Department of Mathematics, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, INDIA
3. Department of Defence and Strategic Studies, University of Allahabad, Prayagraj, INDIA
4. Department of Applied Science, SCRIET, Ch. Charan Singh University, Meerut, INDIA
5. Department of Earth and Planetary Sciences, University of Allahabad, Prayagraj, INDIA

*priyameps@allduniv.ac.in

Abstract

Change detection is a trendy and versatile research topic in the area of application of remote sensing that includes disaster assessment, forest monitoring, urban sprawl and many more. In a growing era, it has become necessary to introduce more and more machines to the industry for better efficiency and accuracy. Each and every task of image processing is tedious and needs a lot of concentration for better accuracy. In remote sensing while finding the Land cover change detection, generally manual process is applied which consumes too much time and unworthy effort. Therefore, here arises the need of automatic change detection which originates different algorithms for image analysis. In this study we have tried to analyze the effectiveness of two type of unsupervised machine learning algorithm and to detect the change in major classified classes in satellite imagery.

The image has been classified in three major classes of vegetation, Built up and water bodies. A comparative study of unsupervised machine learning algorithms has been carried out. Landsat satellite imageries of Allahabad have been used to fulfill the required objectives. Indices majorly NDWI, MSAVI2 and NDBI have been calculated for dominant classes. Two different algorithm K-means and FCM based on machine learning concept of partition and fuzzy have been used. Different types of Land cover (vegetation, Built up and water) have been identified while implementing in MATLAB. The percentage change has been observed and compared with finding decreased percentage trend of vegetation and water while increased percentage of Built up class.

Keywords: Change detection, Machine learning, Supervised and unsupervised classification, Environmental security, Human security.

Introduction

Change detection is the process that allows us to identify differences in the state of an object or phenomenon between different times. Remote sensing techniques of change detection primarily focus on the extraction of information about the changed or transformed entity. Although they are distinct features of the Earth's surface, land usage and land cover are closely related. Originally used to describe the status of vegetation such as forests or grass cover, the term "land cover" has since expanded to encompass other elements such as soil, biodiversity, man-made structures and surface and ground water. An essential method for land surveying, disaster assessment and environmental monitoring is the study of land usage.

According to Kiswanto et al[7], research on land change is essential for managing natural resources and updating land cover maps. The study encourages the use of sophisticated change detection techniques based on information models from remote sensing. For better results, use a support vector machine classifier in conjunction with a zy clustering technique. Research on the effects of the change relations between human and natural activity is also made possible by this type of inquiry. Mukhopadhyay and Maulik[15] proposed a fresh approach that combine's SVM classifier with multi-objective fuzzy clustering scheme for the improvement of solutions.

Machine Learning plays a significant role in modern change detection techniques of computer age. It has proven helpful for various numbers of applications in many parts of the earth system as surface, ocean and atmosphere. Problem of image segmentation, image registration and image classification, scene understanding and object recognition use ML techniques to infer the visual data for extracting information. The variety of objectives and special distinctiveness of data gave rise to the use of a wide range of machine learning algorithms. Many algorithms have been developed to deal with such problems like as post classification, image differencing and principle component analysis but each algorithms has own advantages and limitations. With the evolution of time, new machine learning algorithms were developed. Now the most widely used algorithm is K-means clustering and nearest neighbor[16].

K-means is a simplest unsupervised learning algorithm that is well known to solve the clustering problem. The provided dataset is categorized into a predetermined number of

**** Author for Correspondence***

clusters. Based on the distance between each cluster center and the data point, the fuzzy C-mean algorithm assigns a membership value to each data point that corresponds to that cluster center. The membership value would lean toward the specific cluster center if the data point is close to it[6].

Methods of change detection can be classified as supervised or unsupervised according to nature of data. In method of supervised classification, we need a training sample set with the knowledge of ground truth for classification of dataset. But in this method, application requires ground data which is many times very difficult to access the ground. For such application, unsupervised classification may come to be very useful. This technique uses automatic analysis of change in data obtained by multi-temporal images. In this study, we have adopted the unsupervised method because it gives the desired change result by direct comparison of two different time multi-temporal images[11].

Method of unsupervised classification also referred as clustering is an efficient method of clustering and partitioning remotely sensed image data in multispectral feature space for extraction of land cover information. This requires a minimal amount of input from user as compared to supervised classification because it does not need training the data. This process results in a classified map consisting of k number of classes[16].

Study Area: Study area is Allahabad district which lies between 24^0 45˙ N to 25^0 45˙ N and 81^0 30’ E to 82^0 30’ E of Uttar Pradesh, India. It covers an area around 5300 km^2 (approx 530000 ha) (Fig. 1). Allahabad district is largest district of Uttar Pradesh in terms of population (as per census 2011). It has a diverse variety of Land cover but study focuses majorly on three principle classes such as vegetation, Built up and water.

Data: Landsat data have been used to solve the purpose that has been acquired from USGS (United State Geological Survey) earth explorer website of year 1995, 2005 and 2015. Data was acquired for a gap of ten year to find the change. The more details about data are given in table 1.

Methodology adopted: In this study, two methods have been used to generate the desired results, figure 2 is showing flow chart of methods.

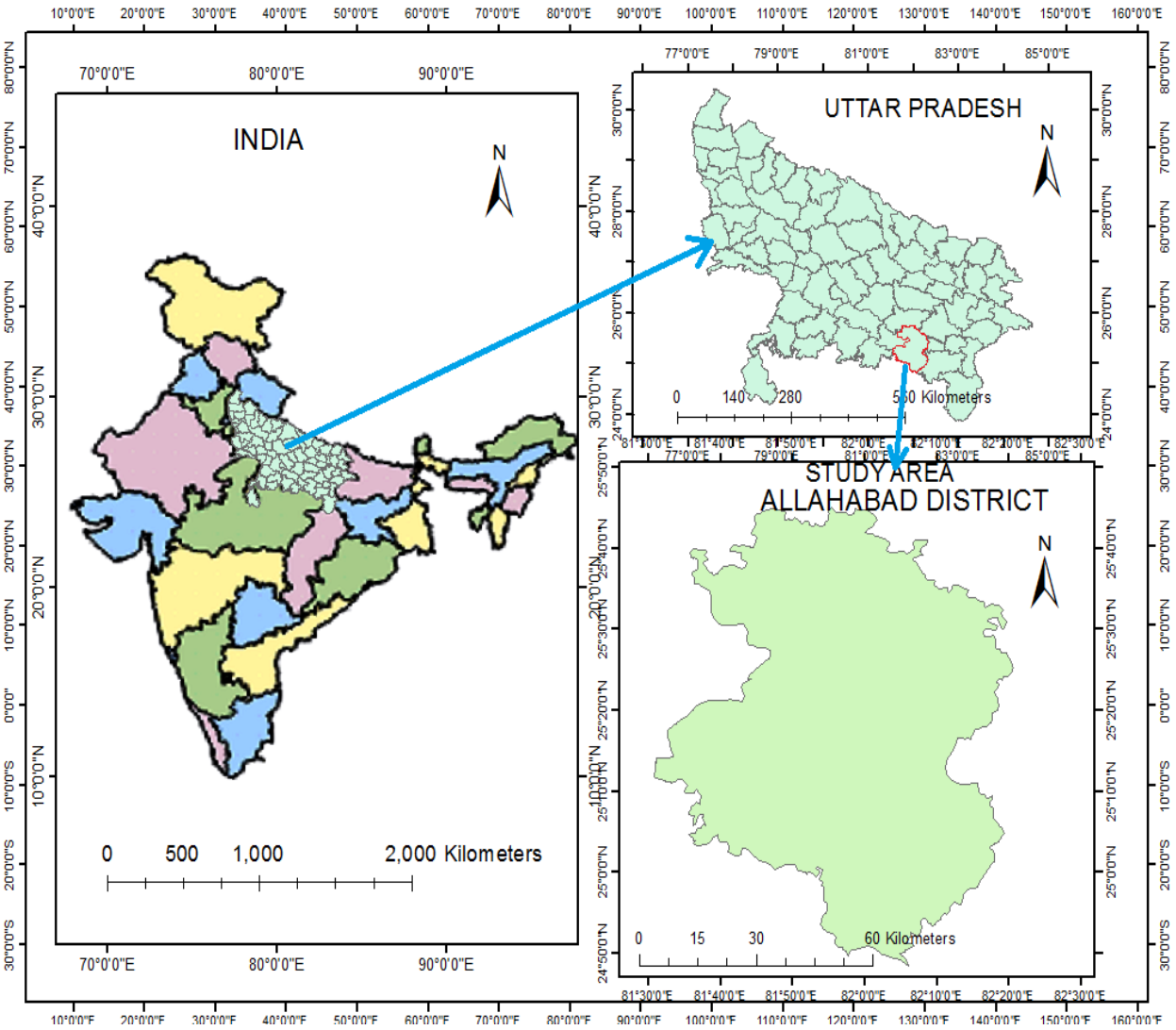

Figure 1: Location map of the study area, Prayagraj, formerly known as Allahabad, India

Supervised Classification Method: The method most frequently employed for the quantitative analysis of satellite image data is supervised classification. The idea behind supervised classification is that a user can select representative sample pixels from an image and instruct the image processing program to utilize these training sites as a guide for classifying all other pixels in the image. In general, supervised classification takes longer than uncontrolled classification[11,16].

Machine Learning Technique: A data analytics method called machine learning trains computers to perform tasks more quickly and intelligently than humans. For change detection, two algorithms: fuzzy C-means and K-means, also known as fuzzy partitioning, have been employed. Two distinct ideas within the machine learning category are represented by these two algorithms. One is based on probability and the other is based on fuzzy concepts[1,2]. The methodology describes the certain numbers of steps applied in the process to generate the result. A brief idea about the methodology has been shown in the form of flowchart figure 2. The machine learning algorithms use an automatic approach to recognize patterns in data.

Table 1
Description of dataset

S. N.	Data Type	Produced Date	Scale	Source
1	Landsat TM	April 5, 1995	30 meter	USGS Earth Explorer
2	Landsat TM	April 16, 2005	30 meter	USGS Earth Explorer
3	Landsat TM	May 30, 2015	30 meter	USGS Earth Explorer
4	Toposheet Number 63G/16, 63K/2, 63K/7	2010	1:50,000	Survey of India
5	Toposheet Number 63G/10, 63G/11, 63G/12, 63G/14, 63H/13, 63K/3, 63K/4, 63L/1, 63K/6, 63K/8	2011	1:50,000	Survey of India
6	Toposheet Number 63G/15	2012	1:50,000	Survey of India

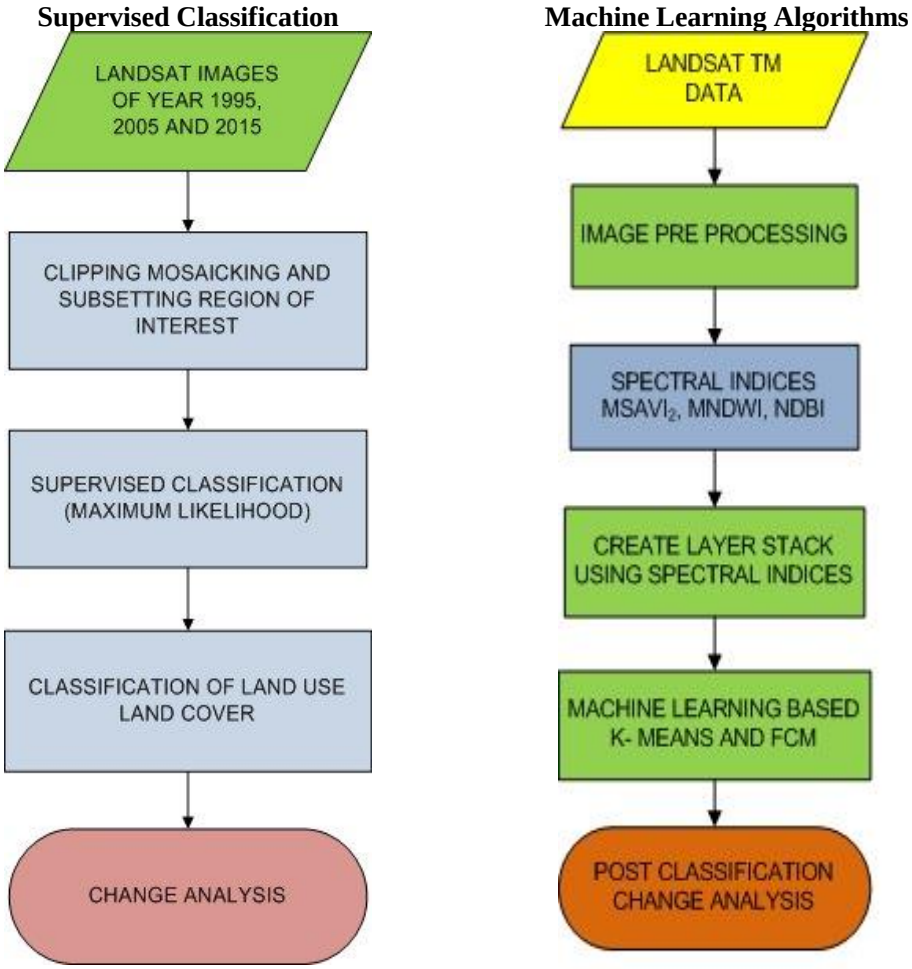

Figure 2: Flow chart for change analysis by supervised classification and machine learning algorithms

Data Pre-processing: The data processing includes correction of different types errors (cosmetic, geometric and radiometric) present in the images. The data was radiometric calibrated to reduce the effect of noise due to atmosphere, sensor calibration and Sun angle error for different dates. After processing it was geo-referenced to "GCS-Everest-India-Nepal" coordinate system using topographical sheet and Google earth image for identifying control point.

Generation of Spectral indices images: Grouping of Land use in the study area has been done in three general categories that are Built-up land, vegetation and open water. These three major classes were represented by Normalized Difference Built-up Index (NDBI), Modified Soil Adjusted Vegetation Index (MSAVI2) and Modified Normalized Difference Water Index (MNDWI) respectively. The extraction is mainly based on new images derived from thee thematic indices NDBI, MSAVI2 and MNDWI.

Normalized Difference Built up Index (NDBI): Normalized Difference Built up Index is used to extract the Built up feature. It ranges from -1 to 1 and is given as:

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

SWIR and NIR represent the reflectance in short wave infrared and near infrared band respectively.

Modified Soil Adjusted Vegetation Index (MSAVI2): MSAVI2 has been used in a number of land studies where it has often been correlated to field data on vegetation cover and as an input layer for mapping land covers or vegetation classes. The output of MSAVI2 is a new image layer representing vegetation greenness with values ranging from -1 to +1. Its formula is given as:

$$MSAVI2 = \frac{2*\mathrm{NIR}+1-\sqrt{(2*\mathrm{NIR}+1)^2-8(\mathrm{NIR}-\mathrm{RED})}}{2}$$

Modified Normalized Difference Water Index (MNDWI): Modified normalized difference water index enhances open water feature while efficiently suppressing other feature noise in the images. It can be calculated as:

$$MNDWI = \frac{GREEN - SWIR}{GREEN + SWIR}$$

Machine learning Algorithms: A data analytics method called machine learning trains computers to perform tasks more quickly and intelligently than humans. To detect changes, two algorithms; fuzzy C means and K-means have been employed. Two distinct ideas within the machine learning category are represented by these two algorithms[5].

K-means: Hugo Steinhaus came up with the concept of "K means" in 1957, but James Mac Queen[13] utilized it for the first time in 1967. Because of its universality, it is also known as Lloyd's algorithm or the K-means algorithm. Finding the groups in unleveled data, that is, data without a defined category or group of k classes is the aim of this approach. Each data point is iteratively assigned to one of the k groups according to the given features[6].

The K-means clustering method is a vector quantization technique that originated in signal processing and is widely used in data mining for cluster analysis. K-means clustering aims to separate n observations into k clusters, each of which is a prototype cluster that belongs to the cluster with the closest mean. The clustering concerns are resolved using Lloyd's approach, also known as k means. Mathematically this algorithm targets to minimize the objective function also called squared error function and this function is defined as:

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\left\|x_i - v_j\right\|\right)^2$$

where '$\|x_i - v_j\|$' is the Euclidean distance between x_i and v_j, 'c_i' is the number of data points in i^{th} cluster and 'c' is the number of cluster centers.

Cluster centers are recalculated by the formula:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i$$

where 'c_i' represents the number of data points in i^{th} cluster.

Fuzzy C-means (FCM): The fuzzy concept was first proposed by Zadeh[21]. The fuzzy set targets to deal with unclear boundaries, representing vague concept and working with linguistic variable. In this way fuzzy sets emerged as an alternative way to deal with uncertainties. A clustering technique called fuzzy C-means enables a single piece of data to be a member of two or more clusters. J.C. Dunn created fuzzy C-means clustering in 1973 and J.C. Bezdek[2] refined it in 1981. Often employed in pattern recognition, this technique is the most popular and is often referred to as soft clustering. The procedure is an iterative clustering technique that minimizes the weighted within group sum of squared error objective function to generate the ideal number of partitions. Mathematically it can be expressed as:

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(\mu_{ij}\right)^m \left\|x_i - v_j\right\|^2$$

where '$\|x_i - v_j\|$' is the Euclidean distance between i^{th} data and j^{th} cluster center.

The working of this algorithm is based on assigning to fuzzy membership value for every data point regarding every

cluster center based on Euclidian distance between data point and cluster. As much as data is nearer to cluster center, more is the membership function for that cluster center. The sum of membership value should be equals to 1. It takes number of steps and after every step, cluster center and membership are accordingly updated to mathematical equations written as:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\left(\frac{2}{m}-1\right)}}$$

$$v_j = \left(\sum_{i=1}^{n} (\mu_{ij})^m x_i\right) \Big/ \left(\sum_{i=1}^{n} (\mu_{ij})^m\right), \forall j = 1, 2, 3, \ldots, c$$

where n represents the data points, v_j represents the j^{th} cluster center, m is the fuzziness index m € [1, ∞], c represents the number of cluster center, μ_{ij} represents the membership of i^{th} data to j^{th} cluster center and D_{ij} represents the Euclidean distance between i^{th} data and j^{th} cluster center.

Post Classification change analysis: Following clustering, the total numbers of pixels are grouped or combined in each year's class. For example, the percentage of land that a class covers each year is calculated.

Results and Discussion

Outcome of change analysis through supervised classification method: Major classes of land cover and land use have been put on the map and the area regarding each class has been calculated as given in table 2. The maximum change occurs in agriculture and forest with area 1033 km^2 in duration 2005 to 2015 which is near about 20 percent of total area and minimum change occurs in water body with area 314 km^2 which is 6.10 percent of total area. From the table, we can observe that there is continuous decreasing trend of agriculture and forest land. Water body has increasing trend from 1995 to 2005 and decreasing trend from 1995 to 2015. Urban and built up land along with barren and bare land are increasing in every decade. Percentage distribution of each class is plotted in figure 3.

Table 2
Area of different land use land cover classes and transformed area by supervised classification

LULC classes	Area (km^2)			Difference (km^2)	
	1995	2005	2015	2005-1995	2015-2005
Agriculture and forest	2540	1840	807	700	1033
Barren and bare land	913	1299	2227	386	928
Water body	369	533	219	164	314
Urban or Built-up land	1323	1472	1892	149	420

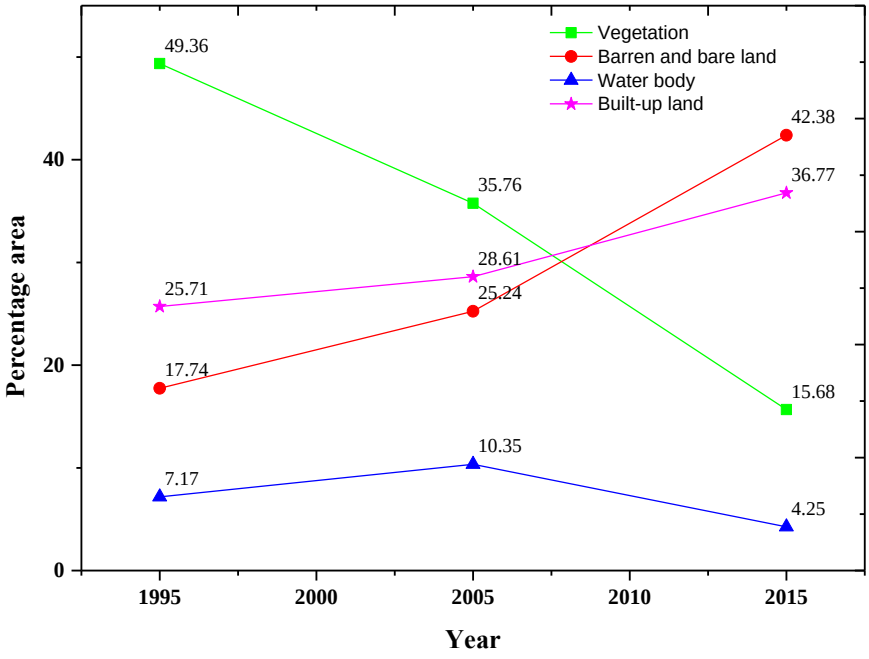

Figure 3: Percentage distribution of different classes by Supervised Classification

Outcome of change analysis through machine learning method: Land use in the study area has been divided in three general categories that are Built-up land, vegetation and open water. These three major classes were represented by NDBI, MSAVI2 and MNDWI respectively. These three indices were generated for year 1995, 2005 and 2015 as shown in figure 4. Percentage distribution of three classes of land use, extracted by Fuzzy C-means and K-means algorithms, is shown in table 3. Significant changes were found throughout the land use classifications in the research area based on the data processing results. According to the report, the rate of built-up areas grew generally between 1995 and 2015.

Notably, the built up regions class showed the biggest gain whereas the vegetation class regularly experienced losses in both time frames. The increase in built up area was observed 28.0 % to 56.5 % during the years 1995 to 2015 by C-means algorithm. When data was processed by Fuzzy C-means algorithm, the increase of built up area observed 30.2 % to 53.5%. The vegetation cover decreased from 51.4 % to 35.3 % in the time period of 1995 to 2015. These results clearly show that the area covered with vegetation was converted into built up area. Decline rate of vegetation also highlights that there is no resilience of vegetation in the present modified environment. The area of surface water body of the study area is decreasing with very slow rate. There is not much variation of percentage area of water body in both of the time intervals.

Comparison of results: Table 4 shows percentage of land use area estimated by different methods. The declining trend of vegetation is observed in both time periods by all methods as shown in figure 5.

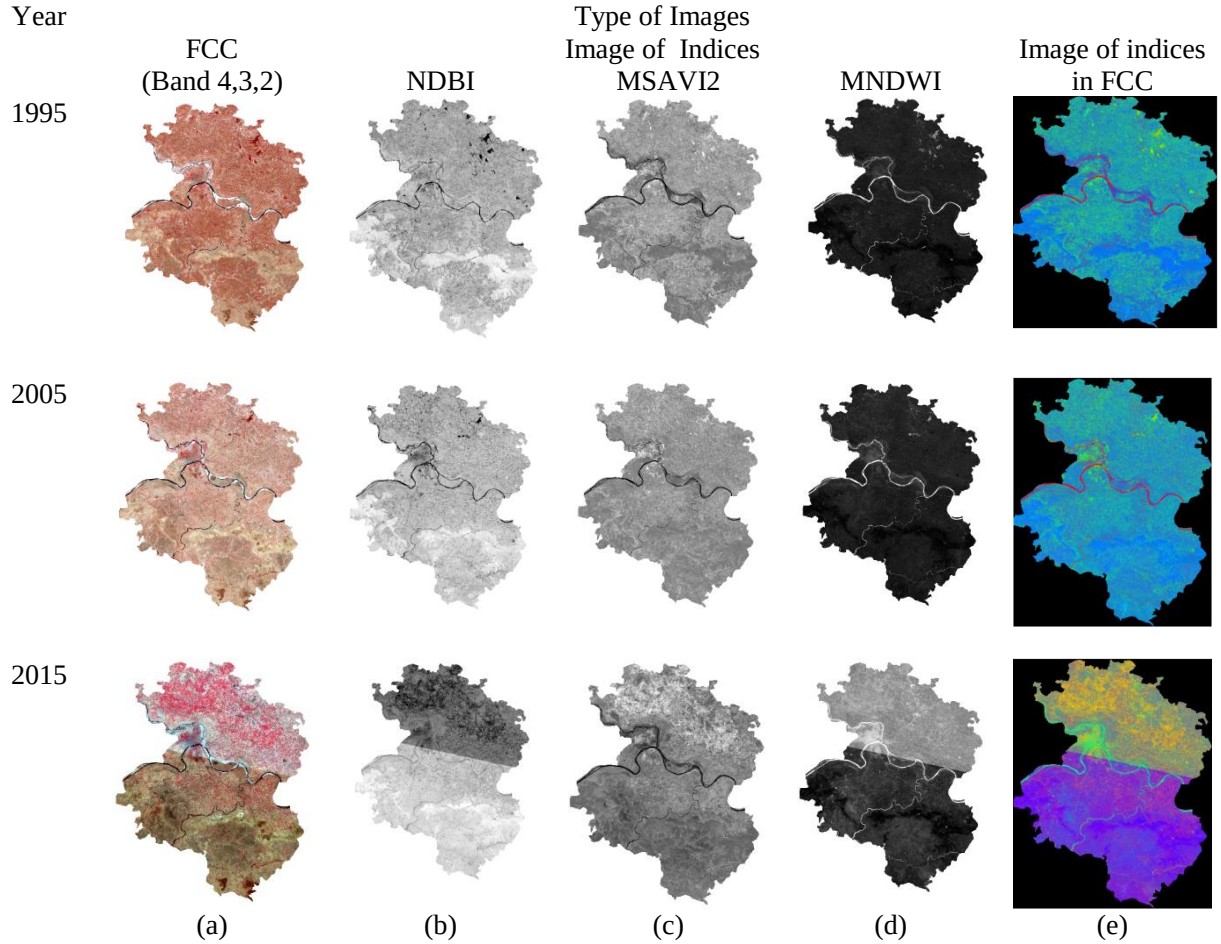

Figure 4: Indices image generated for the year 1995, 2005 and 2015.

Table 3
Percentage Distribution of land use area of three classes (1995-2015)

Year	Vegetation		Built up		Water	
	K-means	***FCM***	***K-means***	***FCM***	***K-means***	***FCM***
1995	51.4	53.1	28.9	30.2	19.7	16.7
2005	43.7	41.2	39.8	41.2	16.5	14.2
2015	31.4	35.3	56.5	53.5	12.1	11.2

Table 4
Percentage of land use obtained from different methods

Year	Vegetation			Built up			Water		
	Supervise	K-means	FCM	Supervise	K-means	FCM	Supervise	K-means	FCM
1995	49.36	51.4	53.1	43.45	28.9	30.2	7.17	19.7	16.7
2005	35.76	43.7	41.2	53.85	39.8	44.6	10.35	16.5	14.2
2015	15.68	31.4	35.3	79.15	56.5	53.5	4.25	12.1	11.2

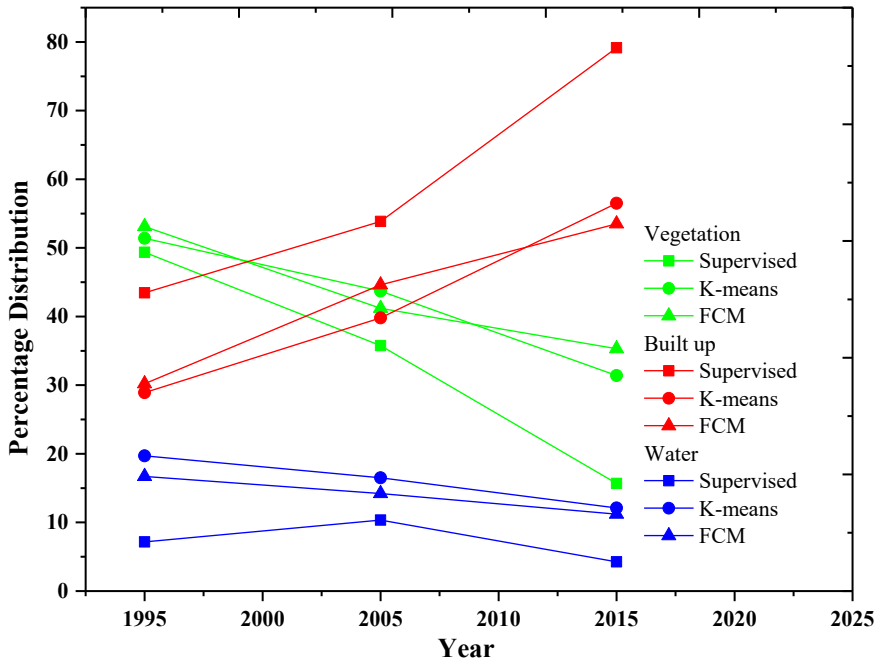

Figure 5: Comparison of results

The results show that vegetation is continuously replaced by the built up land areas. In figure 5, the built up area is 79.15 % by supervised classification which include 36.77 % built up land and 42.38 % barren and bare land as shown in figure 3. Presence of surface water is also declining by a slow rate in the study area which shows a good supply of water in the area.

Conclusion

Use of land and land cover has a significant impact and is crucial to understanding how human activity interacts with the environment, it is vital to track and identify changes in order to preserve a sustainable ecosystem. The technique of determining variations in an object's or phenomenon's status by monitoring it at various points in time is known as "change detection."

A key element of contemporary approaches to managing natural resources and tracking environmental change is the study of land use and land cover change. As a result, research using remote sensing methods offers a precise assessment of the distribution and condition of resources and land use. Mapping trends and changes in land use and land cover over time have been made possible by the repeating coverage of satellite remote sensing data. Environmental, natural and socioeconomic elements, as well as how humans have consumed them across time and space, determine a region's land use and land cover pattern. In order to meet the growing demands for basic human needs and welfare, land use information is therefore crucial for the selection, planning and implementation of a specific site. Land use monitoring aims at people's food security, environmental security and water resource development. It is also significant for improvement of transportation, business development and protection of environmental and cultural heritage[17-20]. Satellite communication or analysis of satellite images would be a powerful tool for management of any environmental or anthropogenic disaster[3,4,7,10].

Proper monitoring of land change may lead to sustainable land use organization and protection of environment and natural resources for the interest of future generation. It may also provide new solution for strengthening environment and improvement of livelihood of all people from urban and rural areas of the country[8,9]. Monitoring the dynamics of land arising from shifting demands from an expanding population is another benefit of this type of land use and land change research. The present study concludes that modern remote sensing techniques provide fast processing of surface data and repetitive coverage of data at small time interval is a

concise method of summarizing the modifications seen in every land use category.

References

1. Arya S., Mount D.M., Netanyahu N.S., Silverman R. and Wu A.Y., An Optimal Algorithm for Approximate Nearest Neighbor Searching, *J. ACM*, **45**, 891-923 **(1998)**

2. Bezdek J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, USA, 256 **(1981)**

3. Chitale V.S. and Behera M.D., Analysing land and vegetation cover dynamics during last three decades in Katerniaghat wildlife sanchuary, India, *J. Earth Syst. Sci*., **123(7)**, 1467-1479 **(2014)**

4. Das B. and Singh U.P., Russia-Ukraine war strategic conundrum, Pentagon Press, New Delhi, 265 **(2025)**

5. Ehsani A.H., Efficiency of Landsat etm+ thermal band for land cover classification of the biosphere reserve 'Eastern Carpathians' (central Europe) using Smap and ML algorithms, *Int. J. Environ. Res*., **4(4)**, 741-750 **(2010)**

6. Kanungo T., Mount D.M., Netanyahu N.S., Piatko Silverman R. and Wu A.Y., An efficient k-means clustering algorithm: analysis and implementation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **24(7)**, 881–892 **(2002)**

7. Kiswanto Santoshi T. and Mardiany Sumaryona, Completing yearly land cover maps for accurately describing annual changes of tropical landscapes, *Global Ecology and Conservation*, **13**, 1-12 **(2018)**

8. Kumar M., Singh P. and Singh P., Fuzzy AHP based GIS and Remote Sensing Techniques for the Groundwater Potential Zonation for Bundelkhand Craton Region, India, *Geocarto International*, **37(22)**, 6671-6694 **(2021)**

9. Kumar M., Singh P. and Singh P., Integrating GIS and remote sensing for delineation of groundwater potential zones in Bundelkhand Region, India, *The Egyptian Journal of Remote Sensing and Space Science*, **25(2),** 387-404 **(2022)**

10. Latoo W.N., Land acquisition and national security, *International Journal of Law, Management and Humanities*, **4(2)**, 199 **(2021)**

11. Lillesand T.M. and Kiefer R.W., Remote sensing and image interpretation, Wiley, 724 **(2000)**

12. Ma G., Ding J., Han L., Zhang Z. and Ran S., Digital mapping of Soil Stalinization based on Sentinal-1 and Sentinal-2 data combined with machine learning algoriths, *Regional Sustainability*, **2,** 177-188 **(2021)**

13. MacQueen J., Some methods for classification and analysis of multivariate observations, fifth Berkeley Symposium, 1-17 **(1967)**

14. Marcal A.R.S. and Castro L., Hierarchical clustering of multispectral images using combined spectral and spatial criteria, *IEEE Geosci. Remote Sens. Lett.*, **2**, 1-14 **(2005)**

15. Mukhopadhyay A. and Maulik U., Unsupervised pixel classification in satellite imagery using multiobjective fuzzy clustering combined with SVM classifier, *IEEE Trans. Geosci. Remote Sens*., **47**, 1132–1138 **(2009)**

16. Singh A., Digital change detection techniques using remotely sensed data: Review article, *Int. J. Remote Sens.*,**10(6)**, 989– 1003 **(1989)**

17. Singh P. and Hasnat M., A review on groundwater investigations using remote sensing in India, *Disaster Advances*, **11(3)**, 34-11 **(2018)**

18. Singh P., Hasnat M. and Pitam P., Fuzzy Criteria based Recognition of Groundwater Prospective Zones using GIS and Remote Sensing, *Disaster Advances*, **11(8)**, 1-10 **(2018)**

19. Singh P., Hasnata M., Rao M.N. and Singh P., Fuzzy analytical hierarchy process based GIS modelling for groundwater prospective zones in Prayagraj, India, *Groundwater for Sustainable Development*, **12**, 1-23 **(2021)**

20. Vidya KM., Manoharan AN., Suchitra B. and Shyni M., Combination of remote sensing, GIS, AHP techniques and geophysical data to delineate groundwater potential zones in the Shiriya river basin, South India, *Geosystems and Geoenvironments*, **3**, 100294 **(2024)**

21. Zadeh L.A., Fuzzy sets, *Information and Control*, **8(3)**, 338-353 **(1965)**.

(Received 24th December 2024, accepted 01st February 2025)