LULC IMAGE CLASSIFICATIONS USING K-MEANS CLUSTERING AND KNN ALGORITHM

Y. Baby Kalpana[1*]And S.M. Nandhagopal[2]

[*1]Associate Professor, Department of CSE, P.A. College of Engineering and Technology, Coimbatore, Tamilnadu, India. E-mail: drybk.pacet@gmail.com

[2]Assistant Professor-Grade III, Department of CSE, Chitkara University, Chandigarh, Punjab, India.

E-mail: nandhagopalsm@gmail.com

ABSTRACT. The human consumption of ecological aspects includes soil, water resources and greenery. Normally, on the earth, the lands may be used and covered by the human being and there are numerous changes in spatial distribution during a period of time. The physical characteristics of a particular land area may be detected and monitored using special cameras or sensors and such technique is called remote sensing. The fields like geography, ecology, land surveys, oceanography and all other earth science disciplines may use remote sensing to obtain the required information's accordingly. It is also used for military, intelligence, commercial and economic human applications. The Geographic Information System (GIS) is used for encapsulating, accumulating, verifying, and displaying data related to positions on Earth's surface. GIS helps the researchers to find special kinds of images on a single satellite pictures like lanes, constructions, and plants. Supervised Classification technique is used for the quantitative analysis with the theory of spectral domain segmentation into regions which is associated with ground cover classes of a defined application. A pixel based classification is employed to specify the number of spectral regions based on the number of information's of the sensed data. This technique is called unsupervised classification in which the information's are created with the help of pixel values for each spectral bands. In this work, K means clustering algorithm and KNN algorithms are used to tabulate the Land used for Coimbatore district at a specific range of 11.01° N and 76.9°E. The two algorithms are employed for the two different classification techniques which are compared with the Tamilnadu gazette data.

Keywords: Image Differencing, Classification Approaches, Multi-Scale Amalgamation, Land Use, Remote Sensing Images.

1. INTRODUCTION

Image classification analyzes the properties of various images and organizes data into defined categories. There are two basic phases of classification process: training and testing data. The basic image features are isolated and created a unique description about it is called training class of data. This will be used to partition the image feature to classify further [4]; [5]. The training classes description is very important component of the image classification. In supervised classification the prior knowledge is required to extract the image features. But in unsupervised classification, a clustering technique is required to train the data. Normally it is noted that image classification is a complex process where the accuracy is related to the dataset characteristics, complexity and the robustness of the algorithm used [4]; [5]; [7]. Spectral transforms and spatial transformations are

Received JUL 03, 2021
www.dynamicpublishers.com;
www.dynamicpublishers.org;

ISSN1056-2176(Print); ISSN 2693-5295 (online)
$15.00 ©Dynamic Publishers, Inc.
https://doi.org/10.46719/dsa202130.10.07

useful in this thematic classification which is collectively called as "feature spaces" .Image classification can be used to identify the land use extensively for urban planning.

Figure 1:Overview of Remote Sensing Process for Land Use

In supervised classification, the information classes like forest, urban etc., are identified and called as training data (Figure 1). The statistical characterization of the information class is created and is named as "Signature analysis". Each characterized sample may involve as the mean, variances and covariance over all bands [6]; [7].

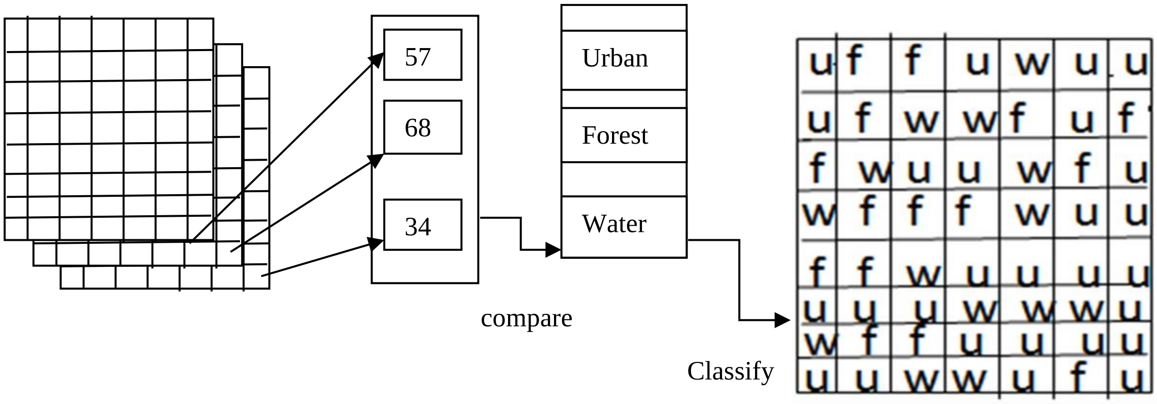


Figure 2: Steps of Supervised Classification

A quantitative analysis of tenuously sensed data is done with the help of supervised classification technique (Figure 2). The spectral domain may be segmented into regions that can be associated with Land cover classes (urban, forest and water) to a particular supervised or unsupervised classification.

2. METHODOLOGY

Normally, the image processing techniques include edge detection techniques, image acquisition, image enhancement, segmentation, classification, data modeling. In this work, the image acquisition using LANDSAT 7 and clustering methods are described [3].

2.1. Image Acquisition

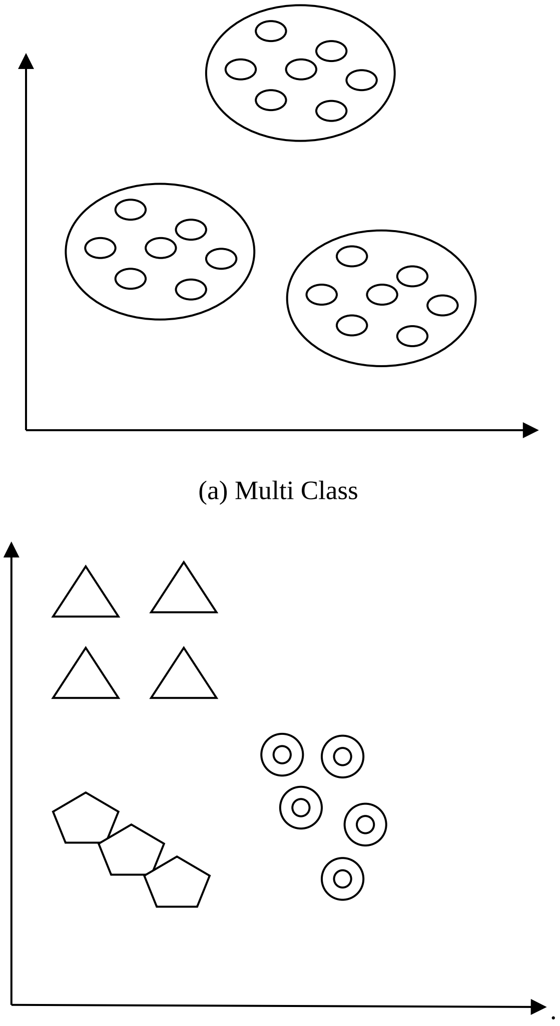
Special request must be made for acquiring data according to LANDSAT 7 or 8. The request will be considered and evaluated for further approval (Ref: Landsat.usgs.gov/landsat-data-acquisition...). A large quantitative landscape of the sensed data and photo interpretation techniques are used for the analysis and classification processes. The image acquired is completely unprocessed image.

Figure 3: Satellite Image of Coimbatore

Figure 3 shows the satellite image of coimbatore. There are few types of sensors called single image sensor, line sensors and array sensors previously. But at present 6 organizations like India, Russia, United States, China, Japan and ESA have the capability to launch many satellites in a very single mission. The available remote sensing resources help to have the comprehensive survey on acquiring spatial data. For this work, the LANDSAT 8 satellite images are used. This new version was launched in the year 2013. It is the eighth satellite launch program by NASA which has the orbit height 705km of having 5 years of mission duration and elapsed 8 years[1]; [3].

2.2. Clustering Technique

Data Clustering technique is the common data combining method used to get the structure of the data acquired through remotely sensed mage. The assignment of identifying subgroups are formed that the same cluster contains similar data while different data points are grouped into variable cluster group as in figure 4. In other words, the uniform subgroups within the large sensed data are found and grouped together figure 4(b). The decision of measuring similar data point is to use an application-specific cluster data for the final classification.



(a) Multi Class

(b) Similar Class of Data

Figure 4: Clustering Technique Involved

Clustering analysis is based on the features of samples like forest, urban, bare soil. Multi Class data Clustering is considered for unsupervised learning method if the ground truth is not available [2]; [3]. An investigation is done by grouping the data points into different groups (Figure 4 (a)) to evaluate and compare the output of the unsupervised algorithm.

3.UNSUPERVISED CLASSIFICATION

The classifiers involve algorithms to inspect the unknown pixels(called False positive) in gray scale image and combine them into a number of land use classes based on the normal clustering which are nearby the image values [10]. The process may be as follows:

1. Choose the number of classes (land use)
2. Cluster pixels into spectral classes (water, forest, urban, wastelands)

3. Label the informational cluster classes and
4. Estimate results.

Unsupervised classification is done with the tasks of detecting and combining of spectral image pixels. It considers only spectral distance (the local projection neighborhood (LPN) which measures a region between a data pair and involves minimum user interaction. This approach involves interpretation after classification. The user often specifies the number of classes to find the user to define their physical meaning. The Land cover information's are not described in advance for the unsupervised classification methods. But some of the statistical clustering algorithms like Hierarchical clustering, Centroids-based Clustering, Fuzzy Clustering and Supervised Clustering may be considered as the information described to arrange pixels into different spectral classes [10]. The number of classes may or may not be defined earlier [1]. The responsibility of the clustering algorithms is to resolve the association between the defined spectral pixel classes. The efficient land use and land cover categories are found by organizations named Survey of India, Remote sensing Geological Survey agencies.

3.1 K-Means Clustering

K-Means clustering is one of the simplest technique to solve the clustering problem for unlabelled or uncategorized image data [1]. The algorithm is used to find number of groups of data within 'k' number of clusters (assume k clusters).

1. First K centers are defined and these centers should be placed in such a way that different location produces different results (Figure 4.1). So, it is better to locate and arrange the same pixels together.

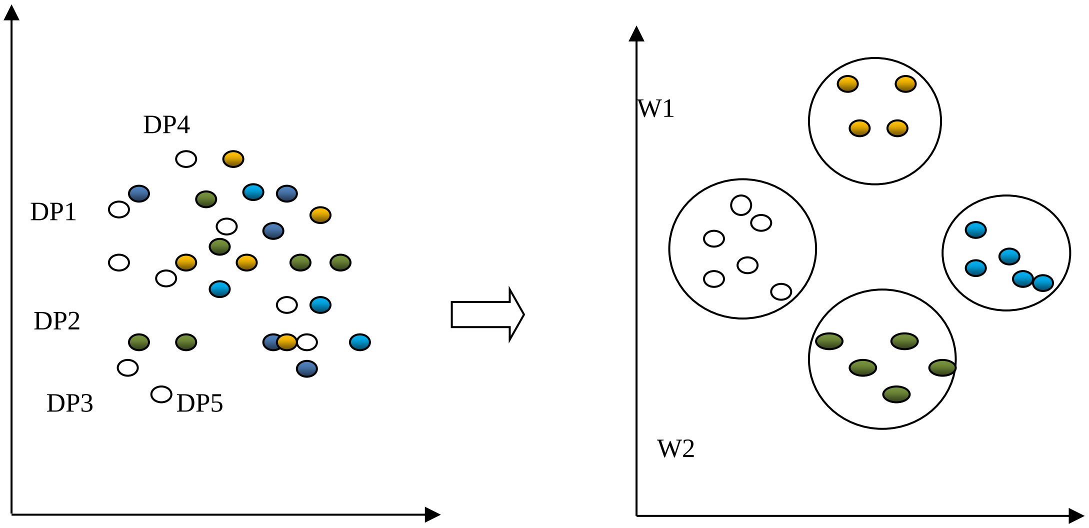


Figure 4.1: 'K' Different Clusters

2. The next step is to associate the given same data set to the nearest cluster center. When no pixel point is found pending for further clustering, re-calculate k new centroids of the clusters resulting from the previous step.

3. Here K defines the number of pre-defined clusters that need to be created in the process.

If K=3, there are three clusters and for K=5, there are five clusters. It is known as an iterative algorithm which divides the unlabeled dataset into K different clusters where each dataset belongs to one group with similar properties like green, blue and white as shown in figure 4.1.

There are 2 ways to define K-means clustering algorithm

1. An iterative process is used to determine the best value for K centroids.
2. Create a closest k-center point for nearer data points.

Like the way the nearest points are calculated and centered using the common cluster data points[10]. The K-means clustering algorithm is as follows:

1. Initialize the number of clusters K.
2. Allocate the data point and calculate the cluster center.
3. Calculate the distance between the data points and centroids.
4. Find the sum and average of the distances.
5. Continue from step 2 to 4 until all the data points are reassigned.

. The distance between the data points is calculated as ,

$$distance = \sum_{i=1}^{n} \sum_{j=1}^{m} Wij \mid x^{i} - \mu_{j} \mid \quad (1)$$

Where transformed space wij=1 for data point x^i , (where i=1,2,3,...n), if x^i belongs to cluster *K*; otherwise, w_{ij}=0. Also, μ_j is the mean of the centroid of xi's cluster. There are two parts in minimization problem.

(i). Minimize distance with respect to wij and treat μj fixed and minimize distance with respect to μj and treat wij fixed.

(ii) Now differentiate distance with respect to wij. Then differentiate distance with respect to μj and re-compute the centroids .

$$\partial distance/\partial \text{wij} = \sum_{i=1}^{n} \sum_{j=1}^{m} Wij \mid x^{i} - \mu_{j} \mid^{2} \quad (2)$$

The data point x^i is assigned to the closest cluster by its sum of squared distance from cluster's centroids.

Sample Calculation of Centroids and nearest Cluster

From figure 5(b), the data points are taken as DP1,DP2,DP3,DP4 and DP5. The x-axis value W1 (Figure 4.1) are 2,1,3,2,4. The y-axis values, W2 (Figure 4.2) , are 0,3,5,2,6. Dp2 and Dp4 are considered as the centroids and the distance is calculated as ,

$$\sqrt{(x1-y1)^2+(x2-y2)^2} \tag{3}$$

The distance is calculated between the initial centroid points with other data point.

(i) calculation of distance between DP1 and DP2

$$\sqrt{(2-1)^2+(0-3)^2} = \sqrt{10} = 3.1$$

(ii) Calculation of distance between DP1 and DP4

$$\sqrt{(2-2)^2+(0-2)^2} = \sqrt{4} = 2$$

After all the calculation of distances, the mean is calculated for the grouped values of centroids as,

(i) For cluster points DP1 and DP4,

mean = 2.1 and distance between DP4 and other data points = 1.03

(ii) For clusters DP2, DP3, DP4

mean=2.83 and distance = 5.1

So the centroids for cluster 1 and 2 are (2,1) and (2.8,5.1) respectively. Similarly the Euclidean distances are calculated from the centroids.

3.2 Implementation

K-means clustering assigns the data points to the nearest cluster and computes the centroid of each cluster as described in the above calculation. The following code describes the implementation of K-means with 2 sample data.

```
# calculate clustering from dataset
cluster import KMeans
 import metrics
import numpy as npi
import matplotlib.pyplot as plott

DP1 = np.arr1([2,3,1,1, 6, 5,5, 6,  8])
DP2 = np.arr2([5, 4, 6, 6,5, 6,2, 1, 2])

# create new plot and data

plott.plot()

X = npi.array(list(zip(x1, x2))).reshape(len(x1), 2)
colors = ['r', 'g', 'b']
```

```
markers = ['d', 'h', 'n']

# KMeans algorithm

K = 3
k_means_model_1 = KMeans(n_clusters=K).fit(X)

print(k_means_model_1.cluster_centers1_)

centers1 = np.arr1(kmeans_model.cluster_centers_)

plott.plot()

plott.title('Calculate_centroids__k')

for k, l in enumerate(kmeans_model.labels_):

   plott.plot(x1[i], x2[i], color=colors[l], marker=marks[l],ls='None')
   plott.xlimi([0, 10])
   plott.ylimi([0, 10])

plt.scatter(centers[:,0], centers[:,1], marker="O", color='r')
plt.show()
```

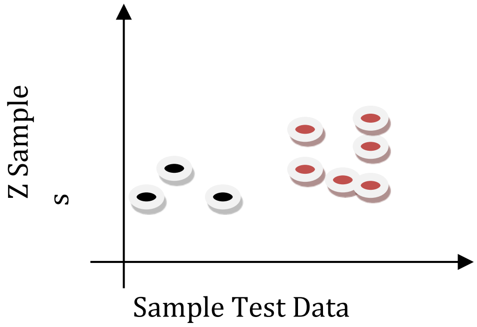


Figure 4.2: Sample Testing Data

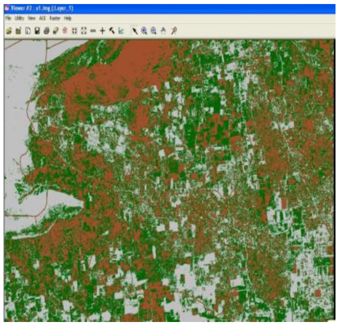


Figure 4.3: Resultant Image of K Means Clustering for Unsupervised Classification

Advantages

- Clustering algorithms using K-Means may regulate the data

- The K-Means fix in a local optimum and use different centroids to find the distance easily.

4. SUPERVISED CLASSIFICATION

Supervised classification is based on the sample pixels given by the user in an image representation of specific classes and then train the data. these are otherwise called as input classes which are selected on the users knowledge. The user is authenticated to set the similar pixels and group them. Once if the user achieve the statistical characterization for every information class, the image can be classified then [9]; [10].

The performance is controlled by the number of features and samples and the different classes assigned by the user. There are 6 steps to do the supervised learning :

1. Data collection
2. Select a success measure of cluster data.
3. Set an evaluation procedure.
4. Prepare the data to be combined as dataset
5. Develop model for training set
6. **6.** Check for better result

4.1 K-Nearest Neighbor (KNN) Algorithm

K-nearest neighbor algorithms or KNN classifies the unknown data by finding K-closest data from the image. It learns from the user training data set (Figure 5.1). This algorithm consists K nearest pixels to be classified with the help of Euclidean distance.

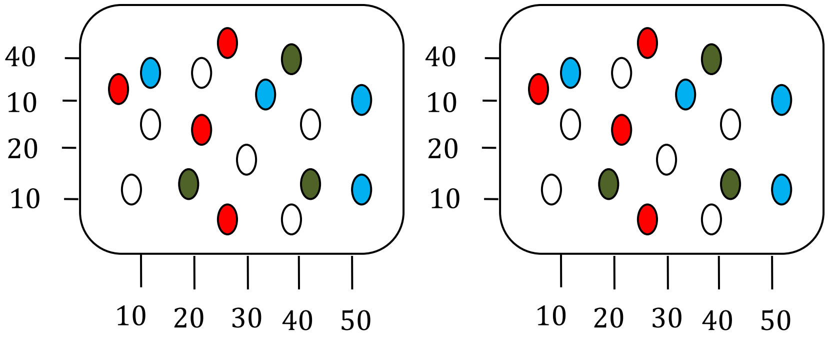


Figure 5.1: Locating the Nearest Data Point

The algorithm

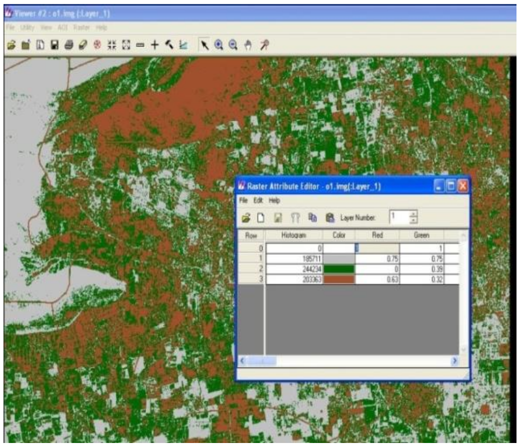
1. Select the data set to implement the KNN algorithm [8].

2. Locate the K nearest neighbor of the data to classify. Let K is any integer. (Eg.(K=3)

3. Calculate the distance between each training data using Euclidean distance formula.
4. Arrange the data in ascending order , based on their distance value
5. Choose the top K rows in the arranged data array
6. Now, assign a class to test data point in the selected row.
7. The right K value is chosen in order to have a good balance between the neighboring pixels to establish an appropriate distance metric. Euclidian distance calculation is employed to find the distance among the nearest neighbor pixels (Figure 6). There are three categories called Bare soil, Forest and Urban are indicated by Red, Green and Gray colors respectively in the figure and each data point within each category are grouped close together in an n dimensional space [7]. The distance metric is defined in order to apply the KNN classification algorithm [2]. This distance is defined as follows:

$$dis\left(x^{i}, x^{j}\right) = \sqrt[2]{\sum_{k} |x^{i} - x^{j}|^{2}}$$

Where K-number of sample data, x^i, x^j are the pixel position to define the distance and p is the class label.


Viewer #2 : o1.img (:Layer_1)
File Utility View AOI Raster Help
Raster Attribute Editor - o1.img(:Layer_1)
File Edit Help
Layer Number: 1

Row	Histogram	Color	Red	Green
0	0			1
1	185711	⬜	0.75	0.75
2	244234	🟩	0	0.39
3	203363	🟫	0.63	0.32

Figure 6: Sample Image of Supervised Classification

The above figure (Figure 6) demonstrates the nearest neighbor pixel calculation. The particular land cover areas are defined as training sites [8]. The spectral signature of the pixels is determined within each training area and the statistics including mean, variance of each of the classes are calculated. The classifies each class that cover full range of variables within the class accurately. Various training data set are selected for each class. It took more time and effort in collecting and selecting training sites to establish a better classification result.

5.CONCLUSION

Table 5.1. shows the report generated on land use. The following are the sample calculated data table and classified images using the above said two classification techniques with the help of the mentioned two algorithms. Python is a structured language which is taken for implementation for this work. Figure 7.1 shows the clustered image & figure 7.2 shows the supervised classification.

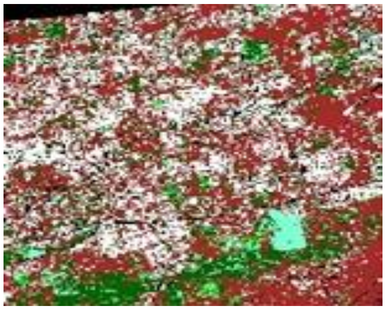


Figure 7.1: Clustered Image

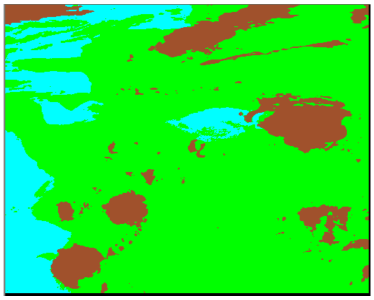


Figure 7.2: Supervised Classification

5.1 Calculated Data

Table 5.1: Report generated on Land Use

Land Use(Hectares)	Previous report(2016)	As per TN Gazette report (2019)	Calculated (2020)
River/ground water	14256	116846	117846
agriculture	31043	177313	178913
Urban	9623	10553	11553
Greenland	1601	2452	2652
Waste land	173	237	226
Forest	241	246	250
Total	56937	307647	311440

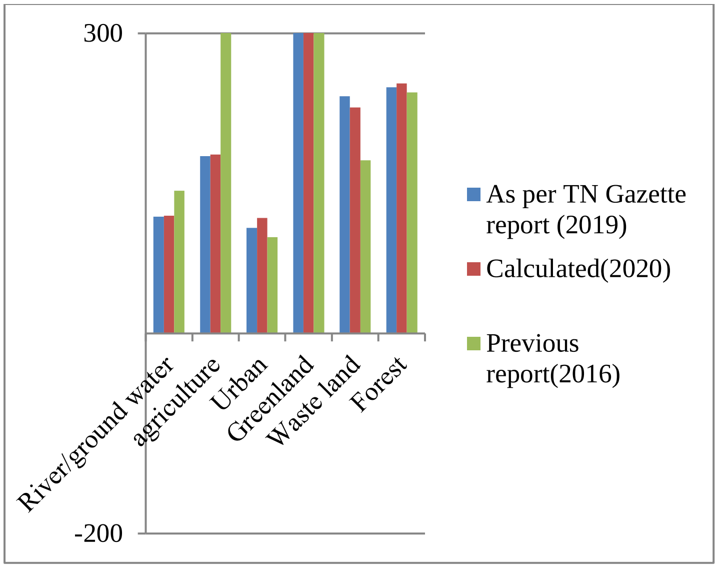


Chart 5.1 Comparison of the Calculated Area for Three Different Years

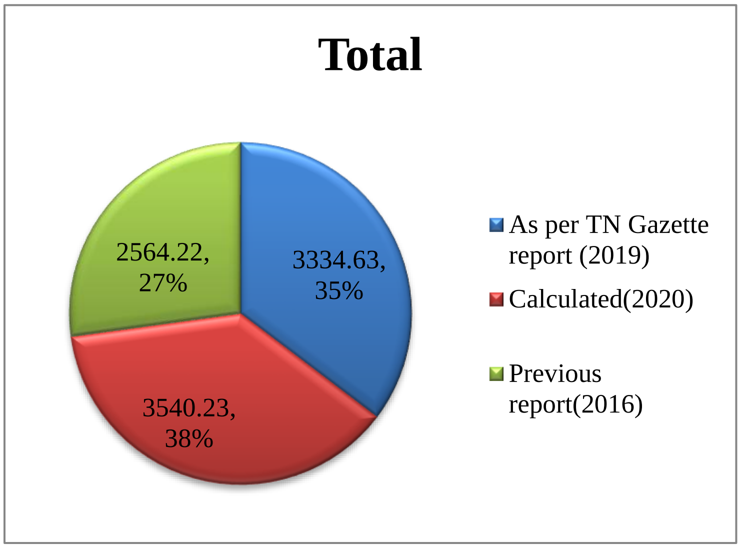


Chart 5.2: Comparison of the Total Area Covered Three Different Years

The remote sensing data for the land used have been analyzed with the help of TN gazette report to fix the Land Use classification of entire Coimbatore District, according to above mentioned year. The supervised and unsupervised classification methods are used for calculations. The K means and KNN algorithms are employed for such calculation of Land use. The above table and chart compared the data which shows drastic changes in urban, water and agriculture area of the district. The Tamilnadu and Coimbatore District administrations have taken necessary actions to retain the green lands which are highly appreciable.

REFERENCES

[1]. Liu, J., & Gao, M. (2008). An Unsupervised Classification Scheme Using PDDP method for network Intrusion Detection. IEEE *Second International Symposium on Intelligent Information Technology Application*, pp. 658-662).

[2]. Cipar, J., Lockwood, R., Cooley, T., & Grigsby, P. (2007). Testing an automated unsupervised classification algorithm with diverse land covers. *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2589-2592.

[3]. Jia-Cun, L., Shao-Meng, Q., & Xue, C. (2003). Object-oriented method of land cover change detection approach using high spatial resolution remote sensing data. *IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, pp. 3005-3007.

[4]. Chang, K. T. (2008). Introduction to geographic information systems, Boston: McGraw-Hill, pp. 247-248.

[5]. Srivastava, S. K., Singh, H. K., & Sinha, A. K. Agriculture land estimation from Satellite Images Using Hybrid Intelligence: A case study on Meerut City. Serials publications, 4 (1), 92-99.

[6]. Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons, pp. 473-488, 532-558.

[7]. Kurosu, T., Yokoyama, S., Fujita, M., & Chiba, K. (2001). Land use classification with textural analysis and the aggregation technique using multi-temporal JERS-1 L-band SAR images. *International Journal of Remote Sensing*, *22*(4), 595-613.

[8]. Thomas M.Liiesand, R.W.Kiefer, Remote Sensing and Image Interpretation, 4th edition, John Wiley & sons, Inc., pp. 473-488, 532-558.

[9]. Yingian Cai, Xiaohua Tong, Rong Shu, Multi-Scale Segmentation of remote sensing image based on Watershed Transformation, IEEE Transactions on Urban Remote Sensing, Sep’ 2009, pp 978-982.

[10]. Yokoyama, Kurosu T., SHI., and Fujita, M., 2001: Land Use Classification with Textural Analysis and the Aggregation Technique Using Multi-temporal JERS-1 L -band SAR Images. International Journal of Remote Sensing, 22, No. 4, pp. 595 -613.