Component-I(A) - Personal Details

Role	Name	Affiliation
Principal Investigator	**Prof.MasoodAhsanSiddiqui**	Department of Geography, JamiaMilliaIslamia, New Delhi
Paper Coordinator, if any	**Dr. Mahaveer Punia**	BITS, Jaipur
Content Writer/Author (CW)	**Dr. Mahaveer Punia**	BITS, Jaipur
Content Reviewer (CR)	**Prof.MasoodAhsanSiddiqui**	Department of Geography, JamiaMilliaIslamia, New Delhi
Language Editor (LE)		

Component-I (B) - Description of Module

Items	Description of Module
Subject Name	Geography
Paper Name	**Remote Sensing, GIS, GPS**
Module Name/Title	**Unsupervised Classification**
Module Id	RS/GIS 08
Pre-requisites	
Objectives	
Keywords	

Unsupervised Classification

Introduction:

Image classification refers to process of extraction of information from satellite image. The purpose of Image classification is to categorize all pixels in a digital image into one of several land use classes or themes. In order to make use of the multitude of digital data available from satellite imagery, it must be processed in a manner that is suitable for the end user. For many projects this processing includes categorizing the land into its various use functions. Depending on the interaction between computer and interpreter during classification process, there are two types of classification. These two main categories used to achieve classified output are called Supervised and Unsupervised Classification techniques

Unsupervised classification (commonly referred to as *clustering*) is an effective method of partitioning remote sensor image data in multispectral feature space and extracting land-cover information.

• Compared to supervised classification, unsupervised classification normally requires only a minimal amount of initial input from the analyst.

• This is because clustering does not normally require training data.

• The process where numerical operations are performed that search for natural groupings of the spectral properties of pixels, as examined in multispectral feature space.

• The clustering process results in a classification map consisting of m spectral classes. The analyst then attempts a posteriori to assign or transform the spectral classes into thematic information classes of interest (e.g., forest, agriculture).

In unsupervised classification, the software does most of the processing on its own generally resulting in more categories than the user is interested in. This is the point where the user has to make decisions on which categories can be grouped together into a single land use category. In

either case additional image processing may be used to determine which method is better for a given situation. It must be kept in mind that maps are simple attempts to represent what actually exists in the world and are never completely accurate.

The image used for demonstration is Landsat 5 Thematic Mapper (TM) scene. Ground resolution for this image is 30 meters. Landsat TM records data in seven different bandwidths. These bandwidths covers parts of the visible, infrared, and thermal infrared regions of the electromagnetic spectrum (Table No. 1).

Table No.1

Landsat 5 TM Band Descriptions(Jensen 2000)		
Band	**Wavelength(μm)**	**Spectral region**
1	**0.45-0.52**	**visible blue**
2	**0.52-0.60**	**visible green**
3	**0.63-0.69**	**visible red**
4	**0.76-0.90**	**reflective infrared**
5	**1.55-1.75**	**mid-infrared**
6	**10.40-12.50**	**thermal infrared**
7	**2.08-2.35**	**mid-infrared**

Fig. 1 Perry Lake natural color

Source: http://academic.emporia.edu/aberjame/student/banman5/perry3.html

This composition is a window from LANDSAT image. This scene is centered on Perry Reservoir located in Jefferson county, northeastern Kansas and is shown in natural color (Fig no. 1).

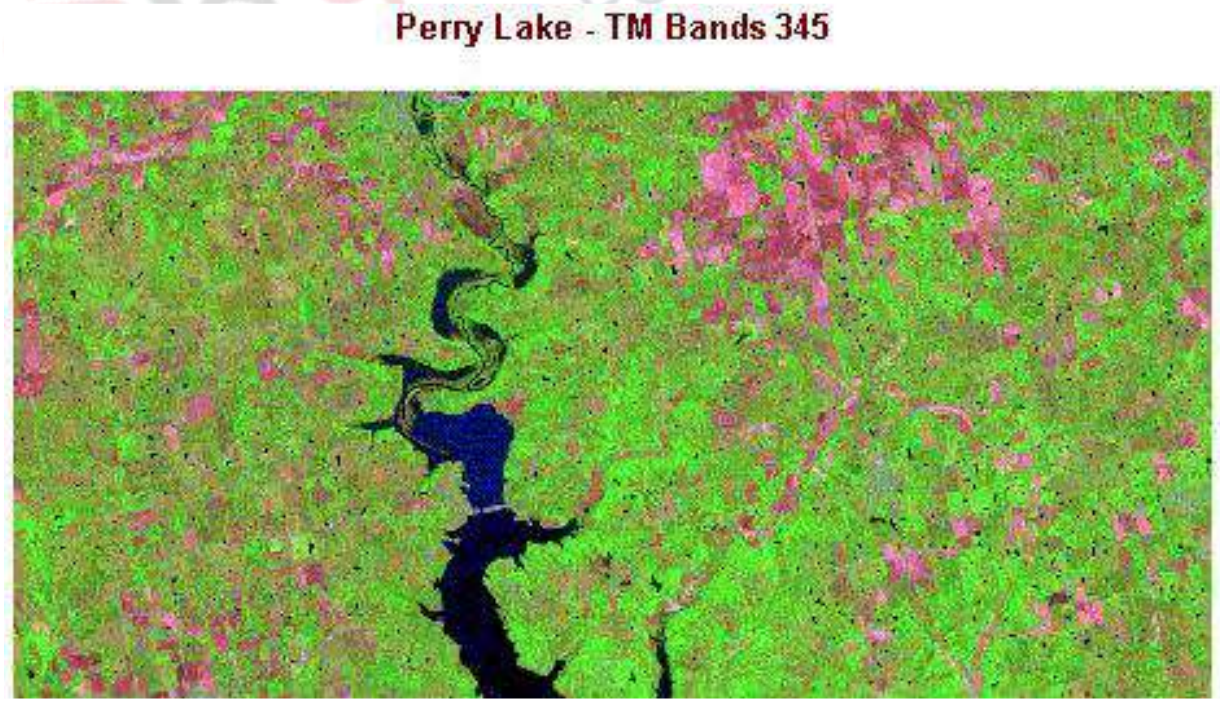

Fig no. 2 Perry lake displayed in BGR

This composition is from the same Landsat scene shown using bands 3, 4, and 5 (displayed in BGR respectively). In this composition vegetation appears in bright green while some of the agricultural fields, still with a good portion of bare soil, are displayed in a pink color. Also, there is a considerable amount of grassland in this scene that is displayed with a combination of the green and pink colors.

By examining the second composition, the land use patterns can be seen quite well. A good portion of the bare soil is located along the Kansas River in the bottom part of the image. There is also the presence of tilled fields located throughout the image with larger concentrations in the northwest corner. From the enlarged image numerous small lakes and ponds can be identified (dark blue). Most of these small ponds can be seen in the grassland areas. These ponds are most likely located in pastures and used as watering holes by cattle. Other important land features include the forested areas as well as the dam and outflow of Perry Reservoir. The forest sprawls outward from the reservoir and follows small drainages, while the man made dam lies at the south end of Perry Reservoir and releases water into the Kansas River. In order to get this image into a more usable format several attempts were made to classify the land uses into separate categories.

Unsupervised classification: Idea

An attempt made to classify the various land uses in Idrisi was done using unsupervised classification techniques. Unsupervised classification techniques do not require the user to specify any information about the features contained in the images. This example was conducted using the ISOCLUST module in Idrisi. With ISOCLUST, the user simply identifies which bands Idrisi should use to create the classifications, and how many classes to categorize the land cover features into. Again Landsat TM bands 1-5 and 7 were used. The resulting classified image is seen below in Fig no. 3.

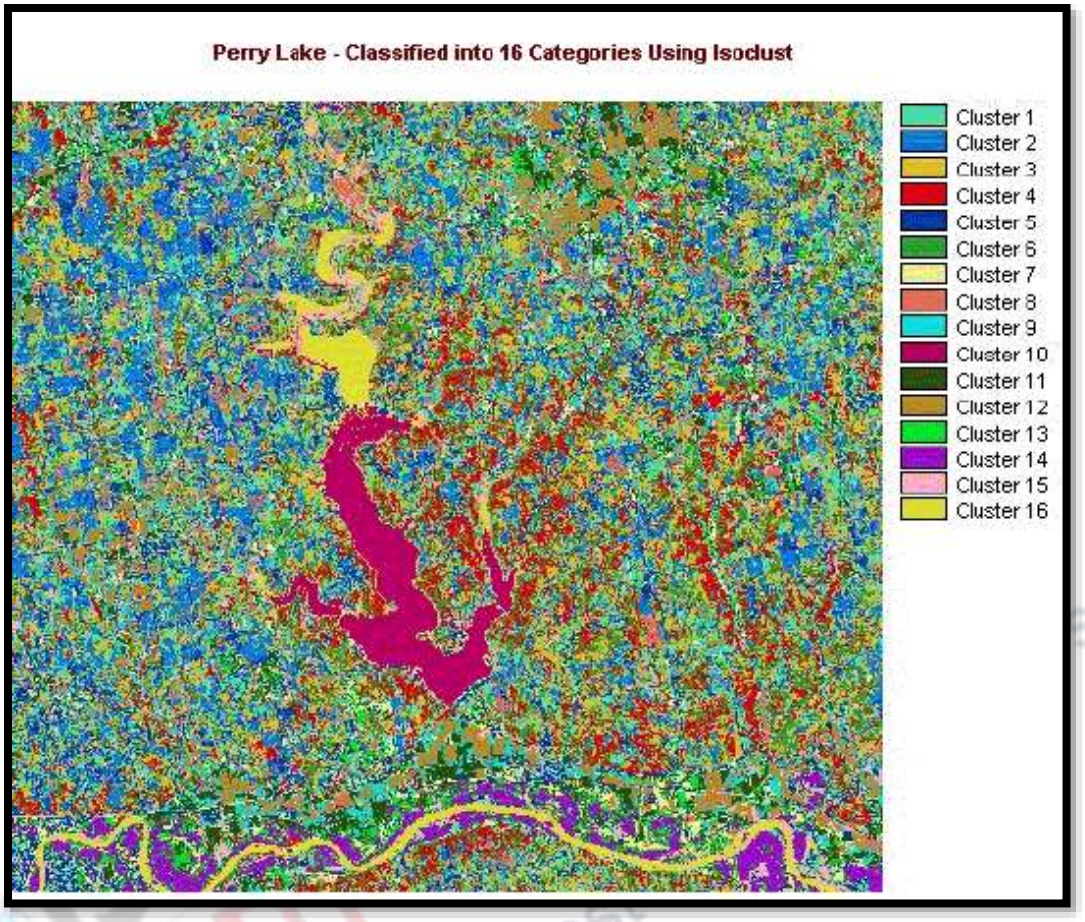

Fig no.3 ISOCLUST of Perry Lake

Source: http://academic.emporia.edu/aberjame/student/banman5/perry3.html

At this point, the image is difficult to interpret. Looking to the image and classified output, decisions need to be made to convert spectral groups into feature or thematic classes. To make these decisions, other materials and knowledge of the area are useful. Ground truthing what is seen in the digital image with what was actually present at the time the image was recorded makes this task more efficient and more accurate. If this knowledge is not available, scientific reasoning may be used to group the various categories together into land use categories. In the demonstartion example Six land cover types were identified from the original 16 categories as shown in the image. (Fig no. 4).

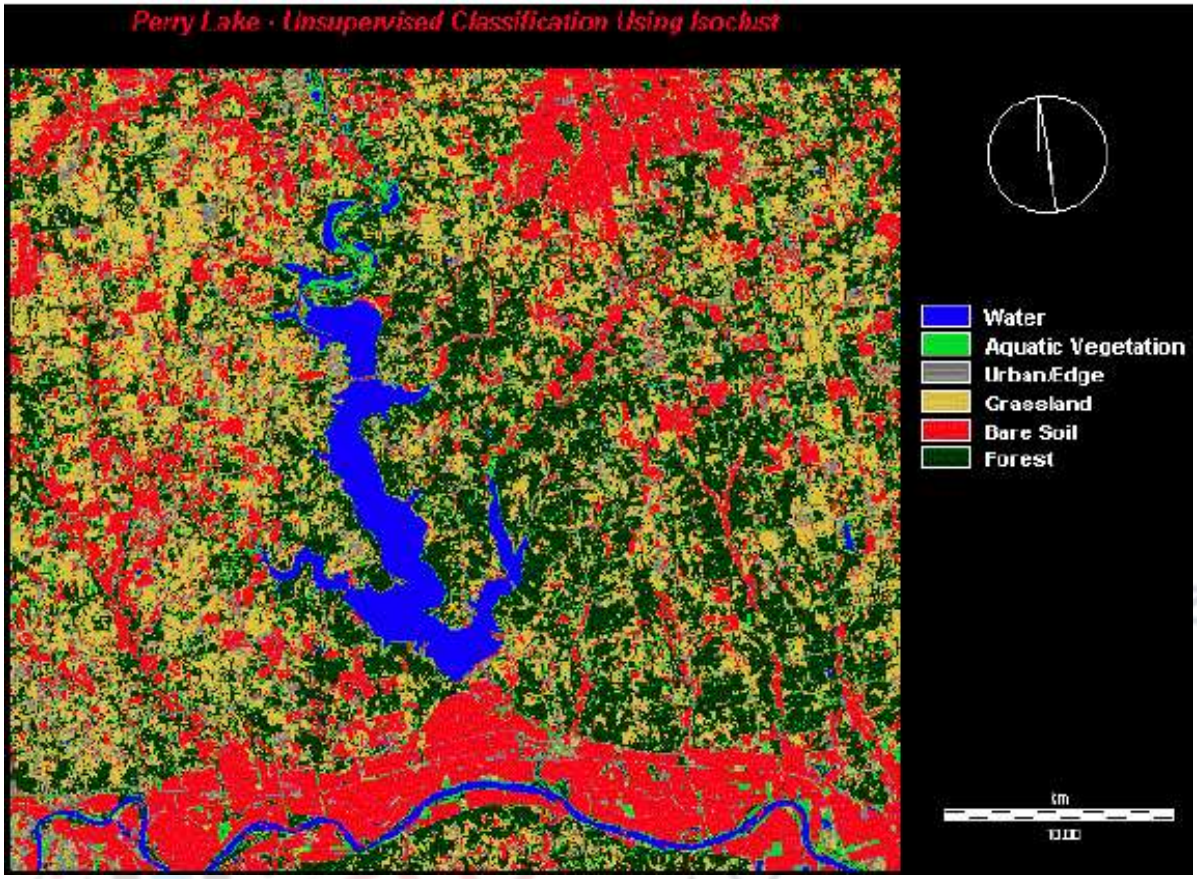

Fig. No. 4 Classified image with 6 categories

Source: http://academic.emporia.edu/aberjame/student/banman5/perry3.html

Area of each category can be calculated through software as shown in Table no 2:

Table No. 2

Sl. No.	Area of category (Hectares)	Land Use
1	6167	water
2	6410	Aquatic vegetation
3	10946	Urban/Edge
4	32976	Grassland
5	29926	Bare Soil

6	38651	Forest
	125076	Total

The image below shows a close-up of the dam and outflow area of Perry Reservoir (Fig No. 5). Again it can be seen that areas classified as Agricultural fall in areas where agriculture is not likely present. Similar examples can be found throughout the image.

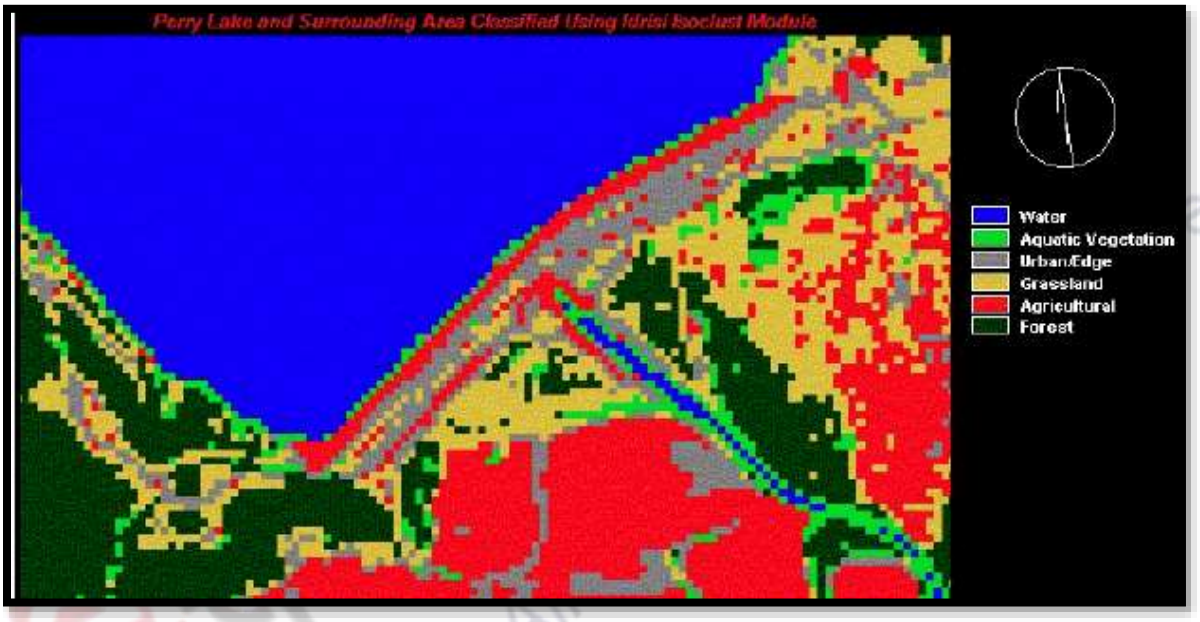

Fig. No. 5 Close-up of the dam and outflow area of Perry Reservoir

Source: http://academic.emporia.edu/aberjame/student/banman5/perry3.html

There are two major methods of unsupervised classification.

Clustering

1. K-means clustering
2. Isodata clustering

1. K-means:

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function as in this equation.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of *n* data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.K-means is a simple algorithm that has been adapted to many problem domains.

There are two schemes to use k-means to classify data

Scheme-1

One method used is to separate the data according to class labels and apply k-means to every class separately. If we have two classes, we would perform k-means twice, once for each group of data. At the end, we acquire a set of prototypes for each class. When we have a new data point, we put all of the prototypes together and find which one is closest to the new data point. This prototype is associated with a class because the prototypes are created by clustering each class of data individually. The class of this prototype is taken as the class of the new data point.

Scheme1of using k-means clustering to the training data in each class separately, using R prototypes per class.

- Assign a class label to each of the $K \times R$ prototypes.
- Classify a new feature vector x to the class of the closest prototype.

Below is a result from the textbook using this scheme (Figure No. 6). There are three classes green, red, and blue. k-means is applied using 5 prototypes for each class. We can see below that for each class, the 5 prototypes chosen are shown by filled circles.

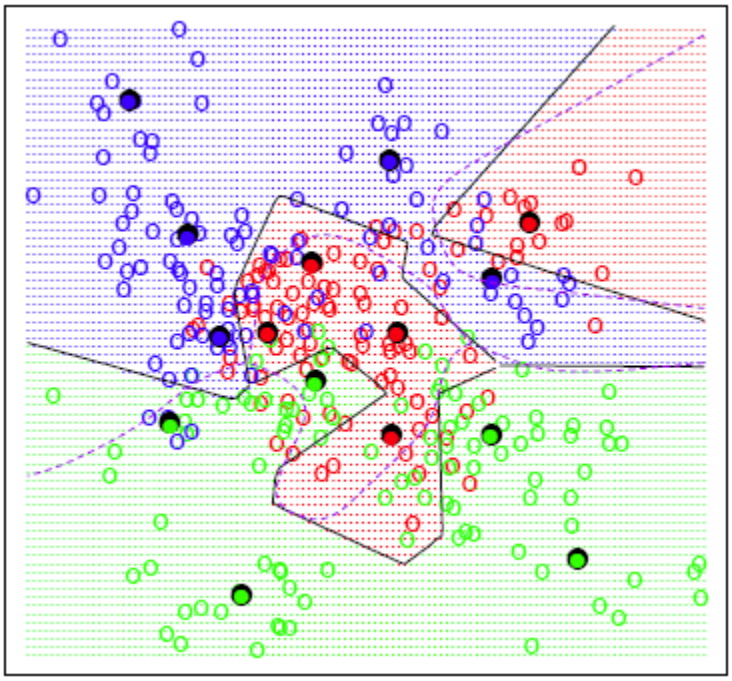

Figure No. 6 K-means clustering

Source: https://onlinecourses.science.psu.edu/stat857/node/113

According to the classification scheme, for any new point, among these 15 prototypes, we would find the one closest to this new point. Then, depending on the color code of that prototype, the corresponding class will be assigned to the new point.

The black lines are classification boundaries generated by the k-means algorithm. Specifically, these are the classification boundaries induced by the set of prototypes based on the nearest neighbor. The decision boundary between any two prototypes based on the nearest neighbor rule is linear. Every prototype occupies some region in the space. The region around each prototype is sometimes called the voronoi region and is bounded by hyperplanes. Because we have more than one prototype for each class, the classification boundary between the two classes is connected segments of the straight lines, which gives a zigzag look

Scheme- 2

The second scheme of classification by k-means is to put all the data points together and perform k-means once. There's no guarantee that points in the same group are of the same class because

we conduct k-means on the class blended data. To associate a prototype with a class, we count the number of data points in each class that are assigned to this prototype. The dominant class with the most data points is associated with the prototype. During the classification of a new data point, the procedure then goes in the same way as Scheme 1.

Steps of Scheme 2:

- Apply k-means clustering to the entire training data, using *M* prototypes.
- For each prototype, count the number of samples from each class that are assigned to this prototype. Associate the prototype with the class that has the highest count.
- Classify a new feature *x* to the class of the closest prototype.

2. ISODATA Clustering:

The Iterative Self-Organizing Data Analysis Technique (ISODATA) represents a comprehensive set of heuristic procedures that have been incorporated into an iterative classification algorithm. Many of the steps incorporated into the algorithm are a result of experience gained through experimentation.

The ISODATA algorithm is a modification of the k-means clustering algorithm, which includes a) merging clusters if their separation distance in multispectral feature space is below a user-specified threshold, and

b) rules for splitting a single cluster into two clusters.

• ISODATA is iterative because it makes a large number of passes through the remote sensing dataset until specified results are obtained, instead of just two passes.

• ISODATA does not allocate its initial mean vectors based on the analysis of pixels in the first line of data the way the two-pass chain algorithm does. Rather, an initial arbitrary assignment of all Cmax clusters takes place along an n-dimensional vector that runs between very specific points in feature space. The region in feature space is defined using the mean and standard deviation of each band in the analysis. This method of automatically seeding the original Cmax vectors makes sure that the first few lines of data do not bias the creation of clusters.

ISODATA is self-organizing because it requires relatively little human input. Typical ISODATA algorithms normally require the analyst to specify the following criteria:

• **Cmax**: the maximum number of clusters to be identified by the algorithm (e.g., 20 clusters). However, it is not uncommon for fewer to be found in the final classification map after splitting and merging take place.

• **T**: the maximum percentage of pixels whose class values are allowed to be unchanged between iterations. When this number is reached, the ISODATA algorithm terminates. Some datasets may never reach the desired percentage unchanged. If this happens, it is necessary to interrupt processing and edit the parameter.

M: the maximum number of times ISODATA is to classify pixels and recalculate cluster mean vectors. The ISODATA algorithm terminates when this number is reached. Minimum members in a cluster (%): If a cluster contains less than the minimum percentage of members, it is deleted and the members are assigned to an alternative cluster. This also affects whether a class is going to be split (see maximum standard deviation). The default minimum percentage of members is often set to 0.01.

Maximum standard deviation (σmax): When the standard deviation for a cluster exceeds the specified maximum standard deviation and the number of members in the class is greater than twice the specified minimum members in a class, the cluster is split into two clusters. The mean vectors for the two new clusters are the old class centers ±1 of standard deviation (σ). Maximum standard deviation values between 4.5 and 7 are typical.

Split separation value: If this value is changed from 0.0, it takes the place of the standard deviation in determining the locations of the new mean vectors plus and minus the split separation value.

Minimum distance between cluster means (C): Clusters with a weighted distance less than this value are merged. A default of 3.0 is often used.

The below figure(7) shows the schemetic of iso-data clustering:

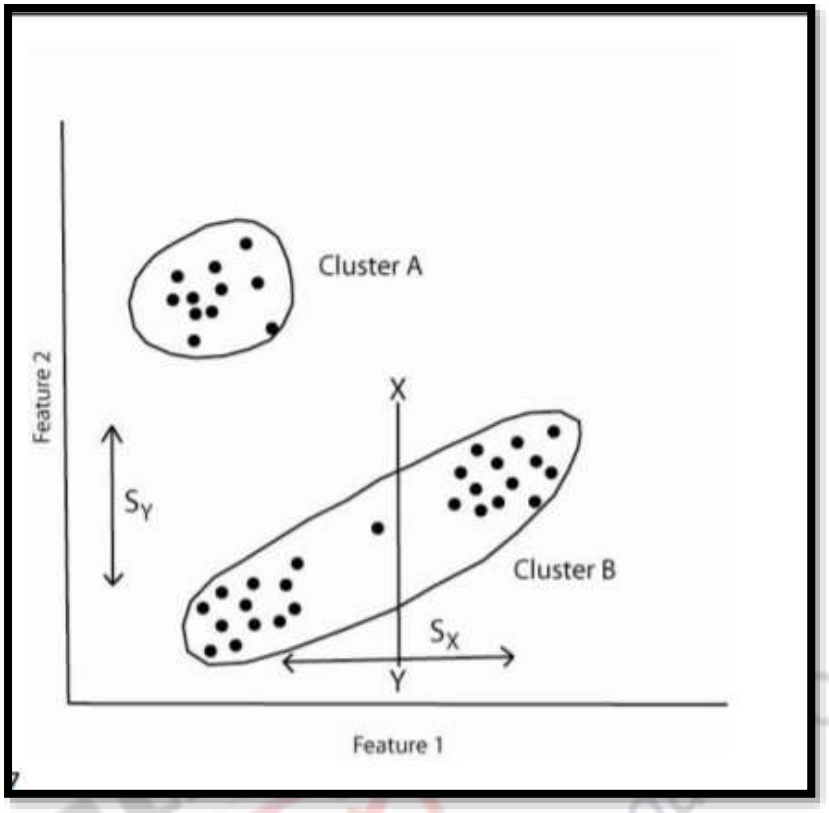

Figure No. 7 Isodata clustering

Source: https://thesai.org/Downloads/Volume7No4/Paper_25 Hyperspectral_Image_Classification_Using_Unsupervised_Algorithms.pdf