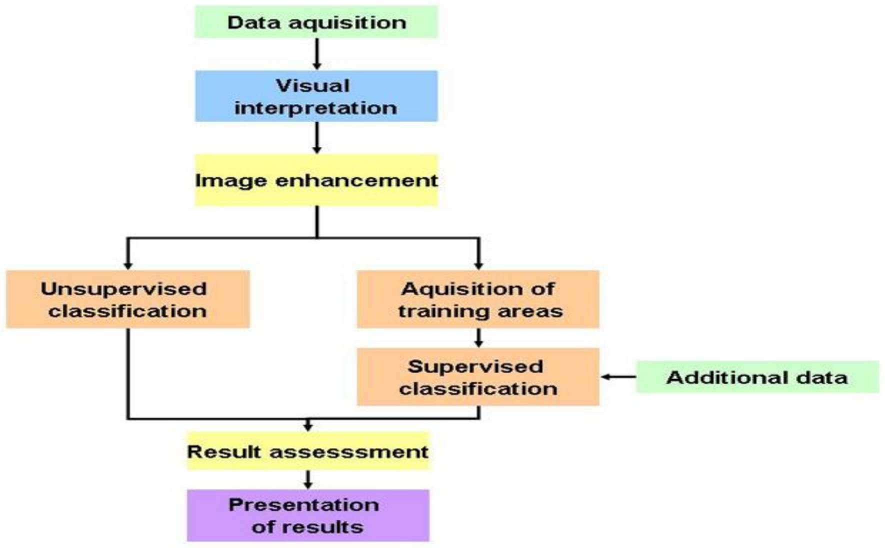
IMAGE CLASSIFICATION

# WHAT IS IMAGE CLASSIFICATION?

Image classification refers to a process in computer vision that can classify an image according to its visual content. For example, an image classification algorithm may be designed to tell if an image contains a human figure or not. While detecting an object is trivial for humans, robust image classification is still a challenge in computer vision applications. Image classification is the process of assigning land cover classes to pixels. For example, classes include water, urban, forest, agriculture and grassland.

Image classification refers to the task of extracting information classes from a multiband raster image. The resulting raster from image classification can be used to create thematic maps. Depending on the interaction between the analyst and the computer during classification, there are two types of classification: supervised and unsupervised.

Image classification is assigning pixels in the image to categories or classes of interest Examples: built-up areas, water body, green vegetation, bare soil, rocky areas, cloud, shadow etc. in order to classify a set of data into different classes or categories, the relationship between the data and the classes into which they are classified must be well understood. To achieve this by computer, the computer must be (i) trained Training is key to the success of classification, (ii) Classification techniques were originally developed (iii) Out of research in Pattern Recognition field



Digital Image Processing and Result Assessment

Computer classification of remotely sensed images involves the process of the computer program learning the relationship between the data and the information classes.

Image classification is a procedure to automatically categorize all pixels in an Image of a terrain into land cover classes. Normally, multispectral data are used to Perform the classification of the spectral pattern present within the data for each pixel is used as the numerical basis for categorization. This concept is dealt under the Broad subject, namely, Pattern Recognition. Spectral pattern recognition refers to the Family of classification procedures that utilizes this pixel-by-pixel spectral information as the basis for automated land cover classification. Spatial pattern recognition involves the categorization of image pixels on the basis of the spatial relationship with pixels surrounding them. Image classification techniques are grouped into two types, namely supervised and unsupervised. The classification process may also include features, Such as, land surface elevation and the soil type that are not derived from the image.

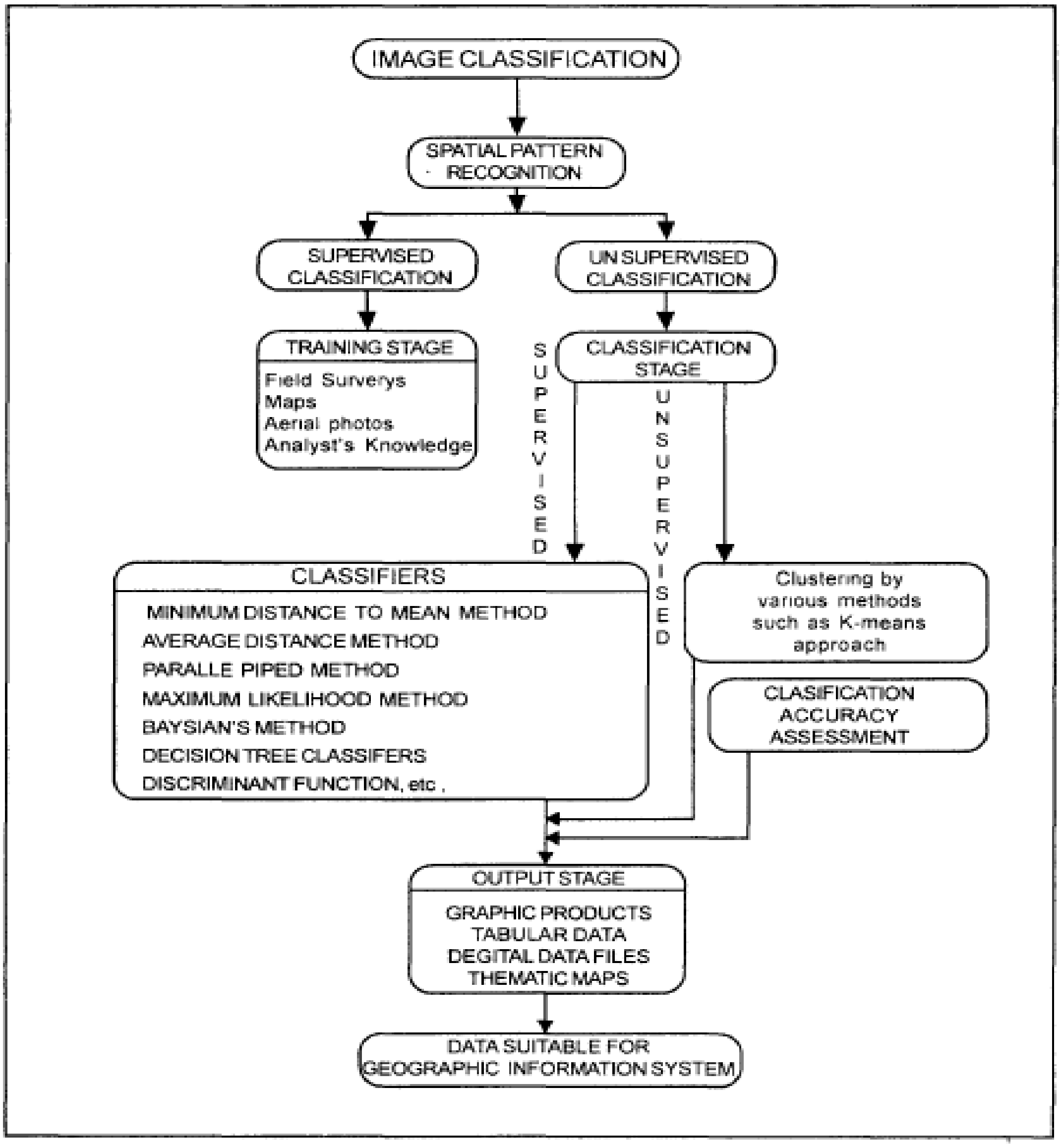


IMAGE CLASSIFICATION
SPATIAL PATTERN RECOGNITION
SUPERVISED CLASSIFICATION
UN SUPERVISED CLASSIFICATION
TRAINING STAGE
Field Surverys
Maps
Aerial photos
Analyst's Knowledge
SUPERVISED
CLASSIFICATION STAGE
UNSUPERVISED
CLASSIFIERS
MINIMUM DISTANCE TO MEAN METHOD
AVERAGE DISTANCE METHOD
PARALLE PIPED METHOD
MAXIMUM LIKELIHOOD METHOD
BAYSIAN'S METHOD
DECISION TREE CLASSIFERS
DISCRIMINANT FUNCTION, etc ,
Clustering by various methods such as K-means approach
CLASIFICATION ACCURACY ASSESSMENT
OUTPUT STAGE
GRAPHIC PRODUCTS
TABULAR DATA
DEGITAL DATA FILES
THEMATIC MAPS
DATA SUITABLE FOR GEOGRAPHIC INFORMATION SYSTEM

Flow Chart showing Image Classification

With the ArcGIS Spatial Analyst extension, there is a full suite of tools in the Multivariate toolset to perform supervised and unsupervised classification. The classification process is a multi-step workflow; therefore, the Image Classification toolbar has been developed to provide an integrated environment to perform classifications with the tools. Not only does the toolbar help with the workflow for performing unsupervised and supervised classification, it also contains additional functionality for analyzing input data, creating training samples and signature files, and determining the quality of the training samples and signature files. The recommended way to perform classification and multivariate analysis is through the Image Classification toolbar.

The amount of image data that is received from satellite is constantly increasing. For example, nearly 3 terabytes of data are being sent to Earth by NASA's satellites every day. Advances in satellite technology and computing power have enabled the study of multi-modal, multi-spectral, multi-resolution and multi-temporal data sets for applications such as urban land use monitoring and management, Geographic Information System (GIS) and mapping, environmental change, site suitability, agricultural and ecological studies. Automatic content extraction, classification and contentbased retrieval have become highly desired goals for developing intelligent systems for effective and efficient processing of remotely sensed data sets. Gong and Howarth (1992) discussed the classification of remote sensing data is a complex process and requires consideration of many factors. The user's need, scale of the study area, economic condition, and analyst's skills are important factors influencing the selection of remotely sensed data, the design of the classification procedure and the quality of the classification results. The major steps of image classification may include image preprocessing, feature extraction, selection of training samples, selection of suitable classification approaches, post-classification processing, and accuracy assessment.

Phinn et al (2000) described classification of remote sensing data is used to assign corresponding labels with respect to homogeneous characteristics of groups. The main aim of classification is to discriminate multiple objects from each other within the image. Classification will be executed on the base of spectral or spectrally defined features, such as density, texture etc., in the feature space. It can be said that classification divides the feature space into several classes based on a decision rule. In many cases, classification will be undertaken using a computer, with the use of mathematical classification techniques.

# STEPS IN IMAGE CLASSIFICATION:

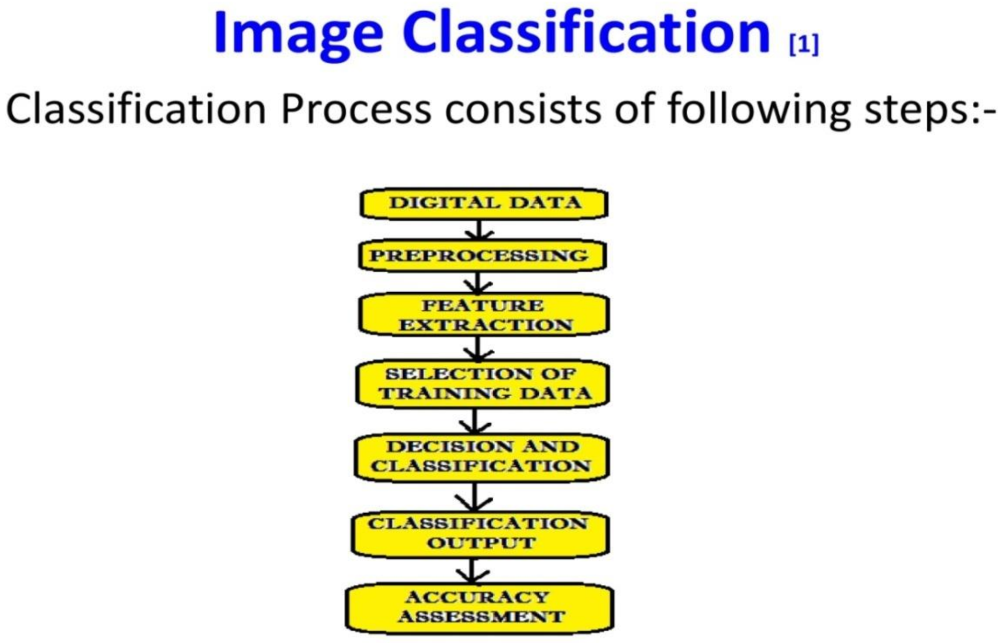


Image Classification [1]

Classification Process consists of following steps:-

DIGITAL DATA

PREPROCESSING

FEATURE EXTRACTION

SELECTION OF TRAINING DATA

DECISION AND CLASSIFICATION

CLASSIFICATION OUTPUT

ACCURACY ASSESSMENT

Step 1: Definition of Classification Classes Depending on the objective and the characteristics of the image data, the classification classes should be clearly defined.

Step 2: Selection of Features to discriminate between the classes should be established using multi-spectral or multi-temporal characteristics, colour, textures etc.

Step 3: Sampling of Training Data Training data should be sampled in order to determine appropriate decision rules. Classification techniques such as supervised or unsupervised learning will then be selected on the basis of the training data sets. 1

Step 4: Finding of proper decision rule Various classification techniques will be compared with the training data, so that an appropriate decision rule is selected for subsequent classification.

Step 5: Classification depending upon the decision rule, all pixels are classified in a single class. There are two methods of pixel by pixel classification and per-field classification, with respect to segmented areas.

Step 6: Verification of Results The classified results should be checked and verified for their accuracy and reliability.

# IMAGE CLASSIFICATION TECHNIQUES:

The learning algorithms are broadly classified into supervised and unsupervised learning techniques. The distinction is drawn from how the learner classifies data. In supervised learning, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by a human. In practice, a certain classes of data will be labeled with these classifications. Althausen (2002) reviewed the classes are then evaluated based on their predictive capacity in relation to measures of variance in the data itself. Some of the examples of supervised classification techniques are Back Propagation Network (BPN), Learning Vector Quantization (LVQ), Self Organizing Map (SOM), Support Vector Machine (SVM), etc., The basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether that can be characterized as forming a group. These groups are termed clusters. Unsupervised classification, often called as clustering, the system is not informed how the pixels are grouped. The task of clustering is to arrive at some grouping of the data. One of the very common of cluster analysis is K-means clustering.

The intent of the classification process is to categorize all pixels in a digital image into one of several land cover classes, or *"themes"*. This categorized data may then be used to produce thematic maps of the land cover present in an image. Normally, multispectral data are used to perform the classification and, indeed, the spectral pattern present within the data for each pixel is used as the numerical basis for categorization (Lillesand and Kiefer, 1994). The objective of image classification is to identify and portray, as a unique gray level (or color), the features occurring in an image in terms of the object or type of land cover these features actually represent on the ground.

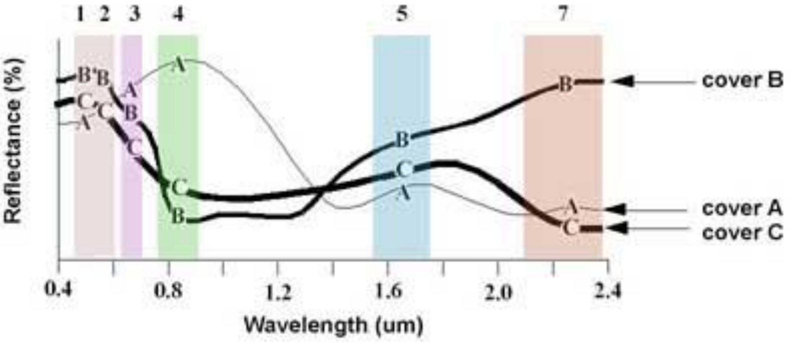


Figure: Spectral Reflectance curve of 3 land covers

Image classification is perhaps the most important part of digital image analysis. It is very nice to have a *"pretty picture"* or an image, showing a magnitude of colors illustrating various features of the underlying terrain, but it is quite useless unless to know what the colors mean. (PCI, 1997). Two main classification methods in GIS are *Supervised Classification* and *Unsupervised Classification.* Two major categories of image classification techniques include **unsupervised** (calculated by software) and **supervised** (human-guided) classification.

The 2 main image classification techniques in remote sensing are:

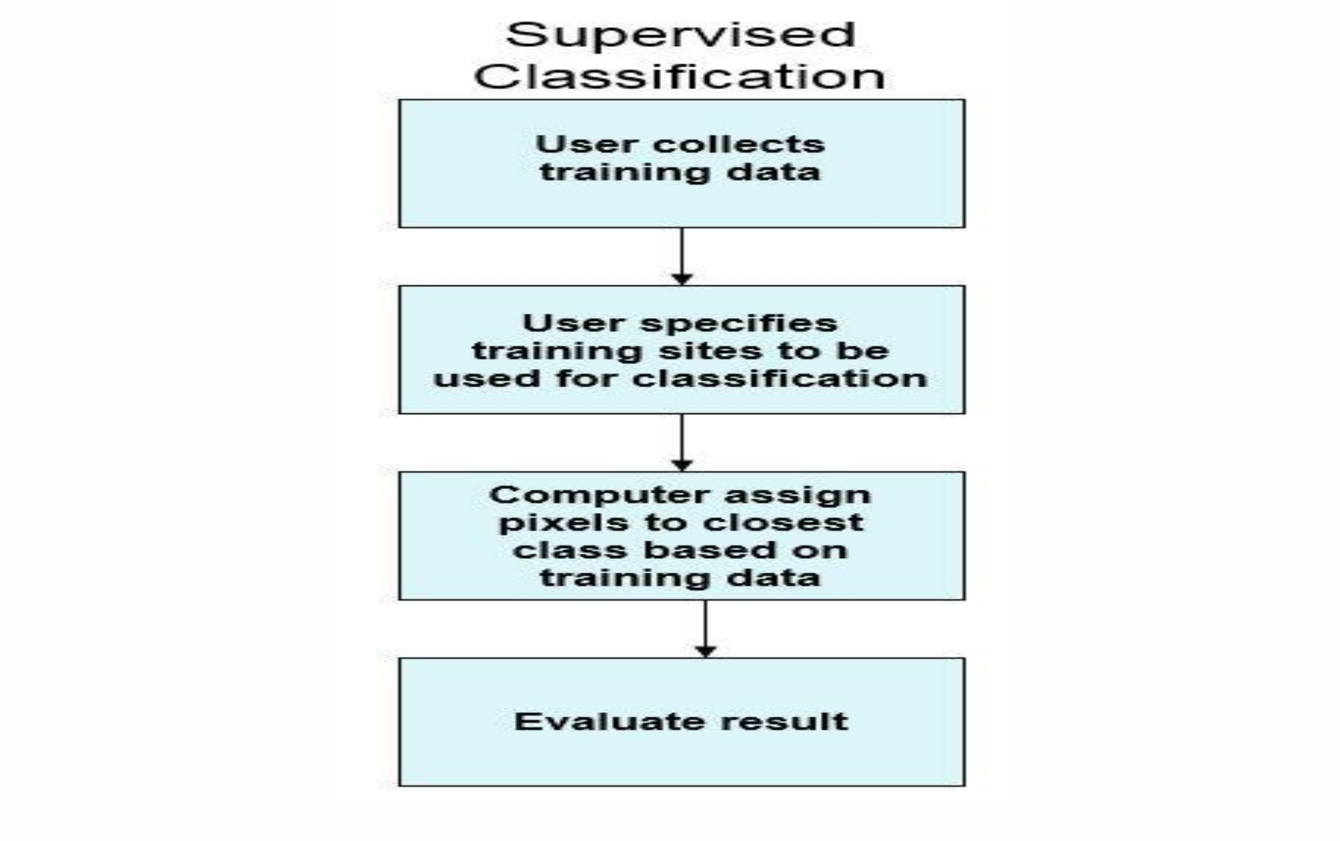
- **Unsupervised image classification**
- **Supervised image classification**

Unsupervised and supervised image classification is the two most common approaches. However, object-based classification has gained more popularity because it’s useful for high-resolution data.

1. Supervised classification

Supervised classification uses the spectral signatures obtained from training samples to classify an image. With the assistance of the Image Classification toolbar, you can easily create training samples to represent the classes you want to extract. You can also easily create a signature file from the training samples, which is then used by the multivariate classification tools to classify the image.

The classifier has the advantage of an analyst or domain knowledge using which the classifier can be guided to learn the relationship between the data and the classes. The number of classes, prototype pixels for each class can be identified using this prior knowledge. When prior knowledge is available for some classes, and not for others,⁄ For some dates and not for others in a multitemporal⁄ dataset, Combination of supervised and unsupervised methods can be employed for partially supervised classification of images.


Supervised Classification
User collects training data
User specifies training sites to be used for classification
Computer assign pixels to closest class based on training data
Evaluate result

Supervised classification is based on the idea that a user can select sample pixels in an image that are representative of specific classes and then direct the image processing software to use these training sites as references for the classification of all other pixels in the image. Training sites (also known as testing sets or input classes) are selected based on the knowledge of the user. The user also sets the bounds for how similar other pixels must be to group them together. These bounds are often set based on the spectral characteristics of the training area, plus or minus a certain increment (often based on "brightness" or strength of reflection in specific spectral bands). The user also designates the number of classes that the image is classified into. Many analysts use a combination of supervised and unsupervised classification processes to develop final output analysis and classified maps. In supervised classification the user or image analyst "supervises" the pixel classification process. The user specifies the various pixels values or spectral signatures that should be associated with each class. This is done by selecting representative sample sites of a known cover type called **Training Sites or Areas**. The computer algorithm then uses the spectral signatures from these training areas to classify the whole image. Ideally, the classes should not overlap or should only minimally overlap with other classes.

Huang et al (2009) presented supervised classification identifies the specific area of interest i.e., water body and non-water body in the image. The supervised classification is much more accurate for mapping classes, but depends heavily on the cognition and skills of the image specialist. The strategy is simple: the specialist must recognize conventional classes or meaningful classes in a scene from prior knowledge, such as personal experience with what's present in the scene, or more generally, the region it's located in, by experience with thematic maps, or by on-site visits. This familiarity allows the individual making the classification to choose and set up discrete classes and then, assign them category names. As a rule, the classifying person also locates specific training sites on the image to identify the classes. The resulting training sites are areas representing each known land cover category that appear fairly homogeneous on the image. For each class thus outlined, mean values and variances of the each band used to classify them are calculated from all the pixels enclosed in each site. More than one polygon is usually drawn for any class. The classification program then acts to classify the data representing each class. When the classification for a class is plotted as a function of the band the result is a spectral signature or spectral response curve for that class. The multiple spectral signatures so obtained are for all of the materials within the site that interact with the incoming radiation.

Classification now proceeds by statistical processing in which every pixel is compared with the various signatures and assigned to the class whose signature comes closest. Many of the classes for the satellite images are almost self-evident in portraying ocean water, waves, beach, marsh, shadows. For example, difference between ocean and bay waters, but their gross similarities in spectral properties would probably make separation difficult. Some classes are broad-based, representing two or more related surface materials that might be separable at high resolution but are inexactly expressed in the image. Thus, the supervised classification has an edge over the unsupervised methodology. Learning Vector Quantization (LVQ) and the Support Vector Machines (SVM) are the two supervised classification methodologies implemented to perform the analysis for the water body and non-water body classification.

Steps involved in Supervised Classification:

A supervised classification algorithm requires a training sample for each class, that is, a collection of data points known to have come from the class of interest. The classification is thus based on how “close” a point to be classified is to each training sample. We shall not attempt to define the word “close” other than to say that both Geometric and statistical distance measures are used in practical pattern recognition algorithms. The training samples are representative of the known classes of interest to the analyst. Classification methods that relay on use of training patterns are called supervised classification methods. In supervised classification, you select representative samples for each land cover class. The software then uses these “training sites” and applies them to the entire image. The three basic steps involved in a typical supervised classification procedure are as follows:

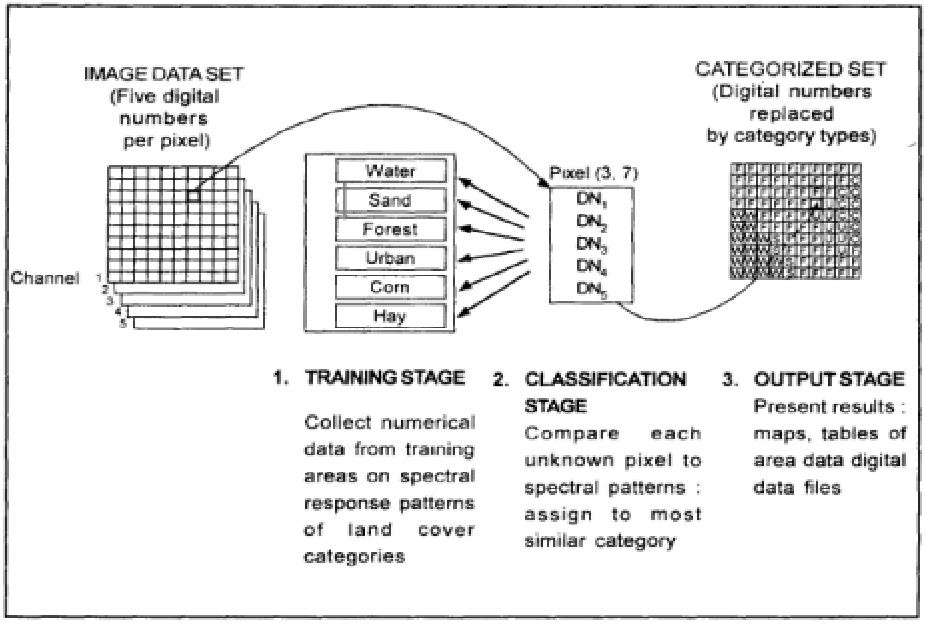


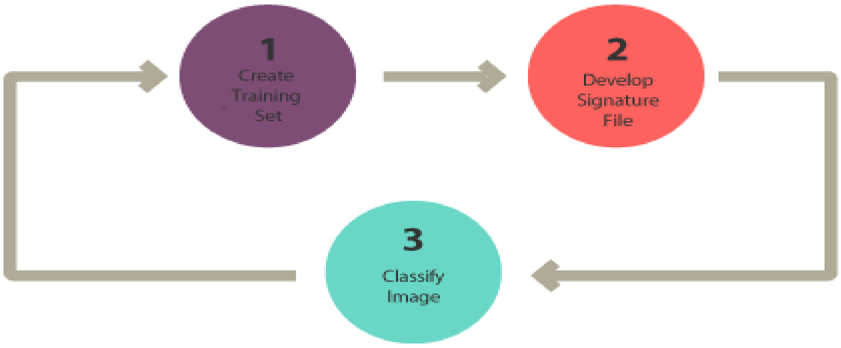
IMAGE DATA SET
(Five digital numbers per pixel)
Channel
1
2
3
4
5
Water
Sand
Forest
Urban
Corn
Hay
Pixel (3, 7)
DN_1
DN_2
DN_3
DN_4
DN_5
CATEGORIZED SET
(Digital numbers replaced by category types)
1. TRAINING STAGE
Collect numerical data from training areas on spectral response patterns of land cover categories
2. CLASSIFICATION STAGE
Compare each unknown pixel to spectral patterns : assign to most similar category
3. OUTPUT STAGE
Present results : maps, tables of area data digital data files

Basic steps of supervised classification

(i) Training stage: The analyst identifies representative training areas and develops numerical descriptions of the spectral signatures of each land cover type of interest in the scene. Training sites are areas that are known to be representative of a particular land cover type. The computer determines the spectral signature of the pixels within each training area, and uses this information to define the statistics, including the mean and variance of each of the classes. Preferably the location of the training sites should be based on field collected data or high resolution reference imagery. It is important to choose training sites that cover the full range of variability within each class to allow the software to accurately classify the rest of the image. If the training areas are not representative of the range of variability found within a particular land cover type, the classification may be much less accurate. Multiple, small training sites should be selected for each class. The more time and effort spent in collecting and selecting training site the better the classification results.

(ii) The classification stag (Decision Rule) or Generate signature file: Each pixel in the image data set IS categorized into the land cover class it most closely resembles. If the pixel is insufficiently similar to any training data set it is usually labeled ‘Unknown’.

(iii) The output stage or Classify: The results may be used in a number of different ways. Three typical forms of output products are thematic maps, tables and digital data files which become input data for GIS. The output of image classification becomes input for GIS for spatial analysis of the terrain. Fig. 2 depicts the flow of operations to be performed during image classification of remotely sensed data of an area which ultimately leads to create database as an input for GIS. Plate 6 shows the land use/ land cover color coded image, which is an output of image



For supervised image classification, you first create training samples. For example, you mark urban areas by marking them in the image. Then, you would continue adding training sites representative in the entire image.

For each land over class, you continue creating training samples until you have representative samples for each class. In turn, this would generate a signature file, which stores all training samples spectral information.

Supervised Classification Principles:

The classifier learns the characteristics of different thematic classes – forest, marshy vegetation, agricultural land, turbid water, clear water, open soils, manmade objects, desert etc. This happens by means of analyzing the statistics of a small sets of pixels in each class that are reliably selected by a human analyst through experience or with the help of a map of the area

With supervised classification, we identify examples of the Information classes (i.e., land cover type) of interest in the image. These are called *"training sites"*. The image processing software system is then used to develop a statistical characterization of the reflectance for each information class. This stage is often called "*signature analysis"* and may involve developing a characterization as simple as the mean or the rage of reflectance on each bands, or as complex as detailed analyses of the mean, variances and covariance over all bands. Once a statistical characterization has been achieved for each information class, the image is then classified by examining the reflectance for each pixel and making a decision about which of the signatures it resembles most. (Eastman, 1995)

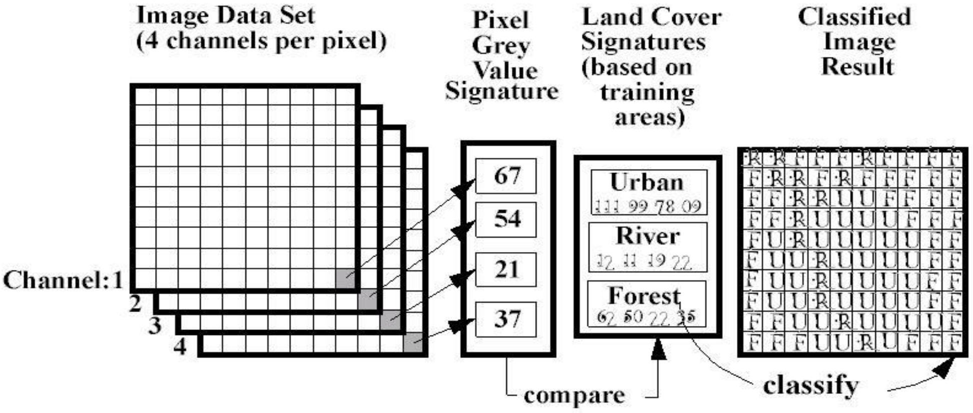


Figure: Steps in Supervised classification

Supervised Classification Algorithms:

Finally, the last step would be to use the signature file to run a classification. From here, you would have to pick a classification algorithms such as:

- Maximum likelihood
- Minimum-distance
- Principal components
- Support vector machine (SVM)
- Iso cluster
- Artificial Neural Networks (ANN)
- Parallel piped
- **Mahalanobis Distance**

As shown in several studies, **SVM is one of the best classification algorithms** in remote sensing. But each option has its own advantages, which you can test for yourself.

(i) Maximum likelihood Classification

Maximum likelihood Classification is a statistical decision criterion to assist in the classification of overlapping signatures; pixels are assigned to the class of highest probability. Assumes that the statistics for each class in each band are normally distributed and calculates the probability that a given pixel belongs to a specific class. Each pixel is assigned to the class that has the highest probability (that is, the maximum likelihood). This is the default.

The maximum likelihood classifier is considered to give more accurate results than parallelepiped classification however it is much slower due to extra computations. We put the word *`accurate'* in quotes because this assumes that classes in the input data have a Gaussian distribution and that signatures were well selected; this is not always a safe assumption.

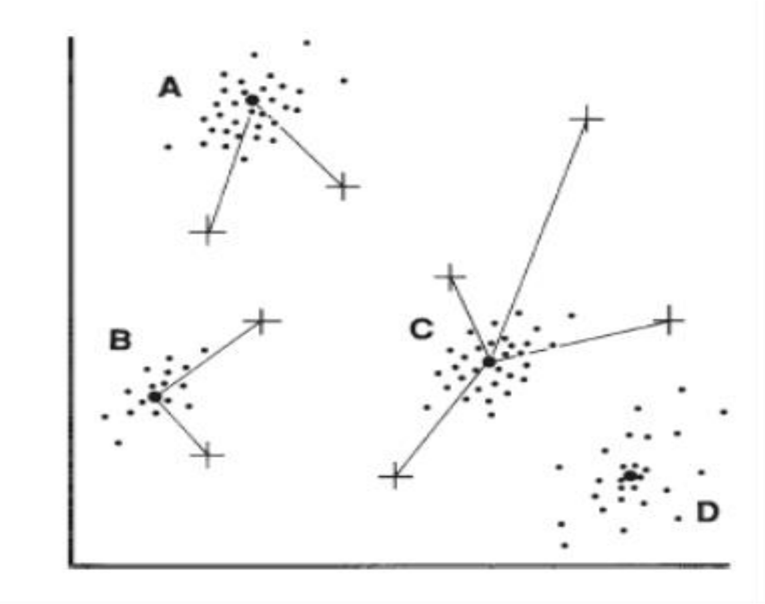
In nature the classes that we classify exhibit natural variation in their spectral patterns. Further variability is added by the effects of haze, topographic shadowing, system noise, and the effects of mixed pixels. As a result, remote sensing images seldom record spectrally pure classes; more typically, they display a range of brightness’s in each band. The classification strategies considered thus far do not consider variation that may be present within spectral categories and do not address problems that arise when frequency distributions of spectral values from separate categories overlap. The maximum likelihood (ML) procedure is the most common supervised method used with remote sensing. It can be described as a statistical approach to pattern recognition where the probability of a pixel belonging to each of a predefined set of classes is calculated; hence the pixel is assigned to the class with the highest probability MLC is based on the Bayesian probability formula.

(ii) Minimum distance Classification

Minimum distance classifies image data on a database file using a set of 256 possible class signature segments as specified by signature parameter. Each segment specified in signature, for example, stores signature data pertaining to a particular class. Only the mean vector in each class signature segment is used. Other data, such as standard deviations and covariance matrices, are ignored (though the maximum likelihood classifier uses this). Uses the mean vectors for each class and calculates the

Euclidean distance from each unknown pixel to the mean vector for each class. The pixels are classified to the nearest class.

Minimum Distance Classification for supervised classification, these groups are formed by values of pixels within the training fields defined by the analyst.Each cluster can be represented by its centroid, often defined as its mean value. As unassigned pixels are considered for assignment to one of the several classes, the multidimensional distance to each cluster centroid is calculated, and the pixel is then assigned to the closest cluster. Thus the classification proceeds by always using the "minimum distance" from a given pixel to a cluster centroid defined by the training data as the spectral manifestation of an informational class. Minimum distance classifiers are direct in concept and in implementation but are not widely used in remote sensing work. In its simplest form, minimum distance classification is not always accurate; there is no provision for accommodating differences in variability of classes, and some classes may overlap at their edges. It is possible to devise more sophisticated versions of the basic approach just outlined by using different distance measures and different methods of defining cluster centroids.



Minimum distance classifier

The result of the classification is a theme map directed to a specified database image channel. A theme map encodes each class with a unique gray level. The gray-level value used to encode a class is specified when the class signature is created. If the theme map is later transferred to the display, then a pseudo-color table should be loaded so that each class is represented by a different color.

(iii) Mahalanobis Distance: A direction-sensitive distance classifier that uses statistics for each class. It is similar to maximum likelihood classification, but it assumes all class covariances are equal, and therefore is a faster method. All pixels are classified to the closest training data.

Mahalanobis Distance is similar to Minimum Distance, except that the covariance matrix is used in the equation. Mahalanobis distance is a well-known statistical distance function. Here, a measure of variability can be incorporated into the distance metric directly. Mahalanobis distance is a distance measure between two points in the space defined by two or more correlated variables. That is to say, Mahalanobis distance takes the correlations within a data set between the variable into consideration. If there are two non-correlated variables, the Mahalanobis distance between the points of the variable in a

2D scatter plot is same as Euclidean distance. In mathematical terms, the Mahalanobis distance is equal to the Euclidean distance when the covariance matrix is the unit matrix. This is exactly the case then if the two columns of the standardized data matrix are orthogonal. The Mahalanobis distance depends on the covariance matrix of the attribute and adequately accounts for the correlations. Here, the covariance matrix is utilized to correct the effects of cross-covariance between two components of random variable.

$D=(X-M_c)^T (COV_c)^{-1}(X-M_c)$ (2)

where

D = Mahalanobis Distance, c = a particular class, X = measurement vector of the candidate pixel Mc = mean vector of the signature of class c, Covc = covariance matrix of the pixels in the signature of class c, Covc-1 = inverse of Covc, T = transposition function.

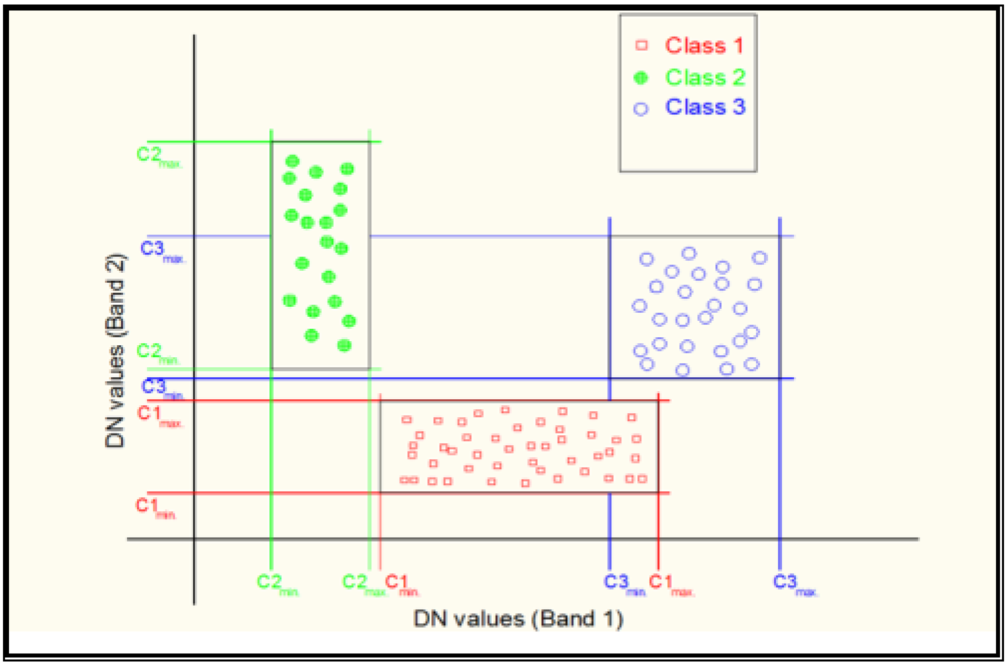
(iv) Parallelepiped Classification

The parallelepiped classifier uses the class limits and stored in each class signature to determine if a given pixel falls within the class or not. The class limits specify the dimensions (in standard deviation units) of each side of a parallelepiped surrounding the mean of the class in feature space.

If the pixel falls inside the parallelepiped, it is assigned to the class. However, if the pixel falls within more than one class, it is put in the overlap class (code 255). If the pixel does not fall inside any class, it is assigned to the null class (code 0).

The parallelepiped classifier is typically used when speed is required. The drawback is (in many cases) poor accuracy and a large number of pixels classified as ties (or overlap, class 255).

Parallelepiped classification, sometimes also known as box decision rule, or level-slice procedures, are based on the ranges of values within the training data to define regions within a multidimensional data space. The spectral values of unclassified pixels are projected into data space; those that fall within the regions defined by the training data are assigned to the appropriate categories. In this method a parallelepiped-like (i.e., hyper-rectangle) subspace is defined for each class. Using the training data for each class the limits of the parallelepiped subspace can be defined either by the minimum and maximum pixel values in the given class, or by a certain number of standard deviations on either side of the mean of the training data for the given class. The pixels lying inside the parallelepipeds are tagged to this class. Figure depicts this criterion in cases of two-dimensional feature space.



Implementation of the parallelepiped classification method for three classes using two spectral bands

Comparison of supervised classification techniques:

One of the most important keys to classify land use or land cover using suitable techniques the table showed advantages and disadvantages of each techniques

techniques	advantage	disadvantage
Parallelepiped	Fast and simple, calculations are made, thus cutting processing Not dependent on normal distributions.	Since parallelepipeds have corners, pixels that are actually quite far, spectrally, from the mean of the signature may be classified
Minimum Distance Classification	Since every pixel is spectrally closer to either one sample mean or another, there are no unclassified pixels. Fastest decision rule to compute, except for parallelepiped	Pixels that should be unclassified,, this problem is alleviated by thresholding out the pixels that are farthest from the means of their classes. Does not consider class variability
Mahalanobis Distance	Takes the variability of classes into account, unlike Minimum Distance or Parallelepiped	Tends to overclassify signatures with relatively large values in the covariance matrix. Slower to compute than Parallelepiped or Minimum Distance
Maximum Likelihood	Most accurate of the classifiers In classification. Takes the variability of classes into account by using the covariance matrix, as does Mahalanobis Distance	An extensive equation that takes a long time to compute Maximum Likelihood is parametric, meaning that it relies heavily on anormal distribution of the data in each input band

Advantages and Disadvantages of Supervised Classification:

In supervised classification the majority of the effort is done prior to the actual classification process. Once the classification is run the output is a thematic image with classes that are labeled and correspond to information classes or land cover types. Supervised classification can be much more accurate than unsupervised classification, but depends heavily on the training sites, the skill of the individual processing the image, and the spectral distinctness of the classes. If two or more classes are very similar to each other in terms of their spectral reflectance (e.g., annual-dominated grasslands vs. perennial grasslands), mis-classifications will tend to be high. Supervised classification requires close attention to the development of training data. If the training data is poor or not representative the classification results will also be poor. Therefore supervised classification generally requires more times and money compared to unsupervised.

2. UNSUPERVISED CLASSIFICATION:

Unsupervised classification is where the outcomes (groupings of pixels with common characteristics) are based on the software analysis of an image without the user providing sample classes. The computer uses techniques to determine which pixels are related and groups them into classes. The user can specify which algorism the software will use and the desired number of output classes but otherwise does not aid in the classification process. However, the user must have knowledge of the area being classified when the groupings of pixels with common characteristics produced by the computer have to be related to actual features on the ground (such as wetlands, developed areas, coniferous forests, etc.).

The steps for running an unsupervised classification are:

1. Generate clusters
2. Assign classes

Generate clusters

In this step, the software clusters pixels into a set number of classes. So, the first step is to assign the number of classes you want it to generate. In addition, you have to identify which bands you want it to use.

If you’re using Landsat, here is a list of **Landsat bands**. For Sentinel, here are **Sentinel-2 bands**. We also have a **handy guide on spectral signatures** which explains which spectral bands are useful for classifying different classes.

In ArcGIS, the steps for generating clusters are:

- First, you have to activate the spatial analyst extension (Customize ‣ Extensions ‣ Spatial Analyst).
- In this unsupervised classification example, we use Iso-clusters (Spatial Analysis Tools ‣ Multivariate ‣ Iso clusters).

INPUT: The image you want to classify.

NUMBER OF CLASSES: The number of classes you want to generate during the unsupervised classification. For example, if you are working with **multispectral imagery** (red, green, blue and NIR bands), then the number here will be 40 (4 classes x 10).

MINIMUM CLASS SIZE: This is the number of pixels to make a unique class.

When you click OK, it creates clusters based on your input parameters. But you still need identify which land cover classes each cluster belongs to.

Assign classes

Now that you have clusters, the last step is to identify each class from the iso-clusters output. Here are some tips to make this step easier:

- In general, it helps to select colours for each class. For example, set water as blue for each class.
- After setting each one of your classes, we can merge the classes by using the reclassify tool.

If land cover appears in 2 classes, you will need to make some manual edits. For example, if vegetation was mistakenly classified as water (perhaps algae in the water), you will have to manually edit the polygon.

In most cases, it helps to convert the raster to vector and use the editing toolbar. You can split polygons to help properly identify them.

The unsupervised clustering is a kind of clustering which takes place with minimum input from the operator; no training sample is available and part of the feature space is achieved by identifying natural groupings of the given data. In unsupervised clustering technique, an individual pixel is compared to each cluster to see the closest pixel. Finally, a map of all pixels in the image, classified as to different clusters, each pixel is most likely to belong, is produced. This then must be interpreted by the user as to what the colour patterns may mean in terms of classes, etc. that are actually present in the real world scene; this requires some knowledge of the scene's feature or cluster content from general experience or personal familiarity with the area imaged. The objective here is to group multi-band spectral response patterns into clusters that are statistically separable. The cluster numbers are initially set and the total number can be varied arbitrarily. Generally, in an area within a SAR image, multiple pixels in the same cluster correspond to some ground feature or cluster so that patterns of gray levels result in a new image depicting the spatial distribution of the clusters. These levels can then be assigned to produce a cluster map. This can be done by either being adequately familiar with the major classes expected in the scene, or, where feasible, by visiting ground truth and visually correlating map patterns to their ground counterparts. Since the classes are not selected beforehand, this method is termed as unsupervised classification.

Unsupervised Classification Techniques:

There are three unsupervised classification techniques namely K-means clustering, PCA based K-means clustering and Fuzzy C-Means clustering (FCM).

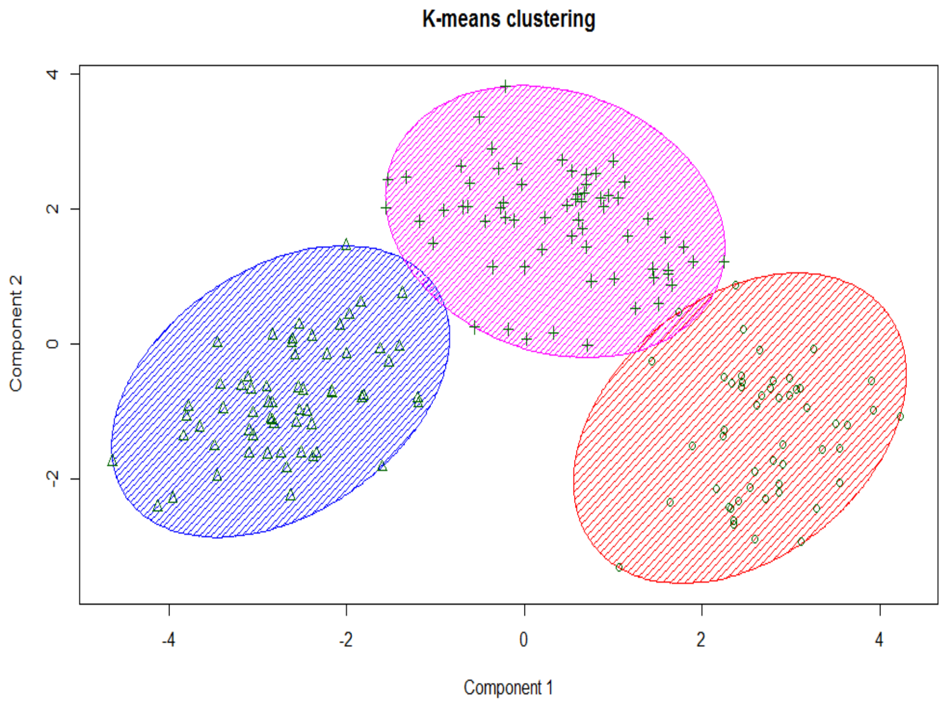
K- means clustering: K-means algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach kmeans follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a breakdown of how we can solve it mathematically



Unsupervised classification finds spectral classes (or clusters) in a multiband image without the analyst's intervention. The Image Classification toolbar aids in unsupervised classification by providing access to the tools to create the clusters, capability to analyze the quality of the clusters, and access to classification tools. Unsupervised classification is a method which examines a large number of unknown pixels and divides into a number of classed based on natural groupings present in the image values. Unlike supervised classification, unsupervised classification does not require analyst-specified training data. The basic premise is that values within a given cover type should be close together in the measurement space (i.e. have similar gray levels), whereas data in different classes should be comparatively well separated (i.e. have very different gray levels) (PCI, 1997; Lillesand and Kiefer, 1994; Eastman, 1995)

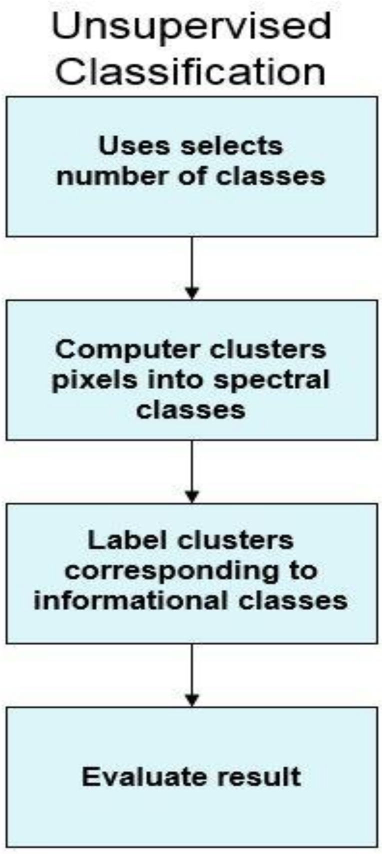
The classes that result from unsupervised classification are spectral classed which based on natural groupings of the image values, the identity of the spectral class will not be initially known, must compare classified data to some form of reference data (such as larger scale imagery, maps, or site visits) to determine the identity and informational values of the spectral classes. Thus, in the supervised approach, to define useful information categories and then examine their spectral separability; in the unsupervised approach the computer determines spectrally separable class, and then define their information value. (PCI, 1997; Lillesand and Kiefer, 1994)

Unsupervised classification is becoming increasingly popular in agencies involved in long term GIS database maintenance. The reason is that there are now systems that use clustering procedures that are extremely fast and require little in the nature of operational parameters. Thus it is becoming possible to train GIS analysis with only a general familiarity with remote sensing to undertake classifications that meet typical map accuracy standards. With suitable ground truth accuracy assessment procedures, this tool can provide a remarkably rapid means of producing quality land cover data on a continuing basis.

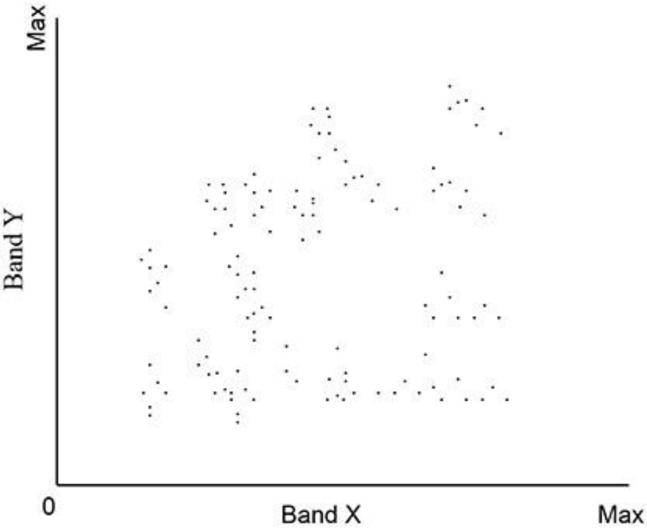
When access to domain knowledge or the experience of an analyst is missing, the data can still be analyzed by numerical exploration, whereby the data are grouped into subsets or clusters based on statistical similarity

Using the Image Classification toolbar and Training Sample Manager, it was determined the training samples were representative for the area and statistically separate. Therefore, a maximum likelihood classification was performed from the toolbar.

Steps of Unsupervised Classification:


Unsupervised Classification
Uses selects number of classes
Computer clusters pixels into spectral classes
Label clusters corresponding to informational classes
Evaluate result

Unsupervised classification is a form of pixel based classification and is essentially computer automated classification. The user specifies the number of classes and the spectral classes are created solely based on the numerical information in the data (i.e. the pixel values for each of the bands or indices). Clustering algorithms are used to determine the natural, statistical grouping of the data. The pixels are grouped together into based on their spectral similarity. The computer uses feature space to analyze and group the data into classes. Roll over the below image to see how the computer might use feature space to group the data into ten classes.



While the process is basically automated, the user has control over certain inputs. This includes the Number of Classes, the Maximum Iterations, (which is how many times the classification algorithm runs) and the Change Threshold %, which specifies when to end the classification procedure. After the data has been classified the user has to interpret, label and color code the classes accordingly.

Advantages of Unsupervised Classification:

Unsupervised classification is fairly quick and easy to run. There is no extensive prior knowledge of area required, but you must be able to identify and label classes after the classification. The classes are created purely based on spectral information; therefore they are not as subjective as manual visual interpretation.

Disadvantages of Unsupervised Classification:

One of the disadvantages is that the spectral classes do not always correspond to informational classes. The user also has to spend time interpreting and label the classes following the classification. Spectral properties of classes can also change over time, so you can’t always use the same class information when moving from one image to another.

Supervised classification generally performs better than unsupervised classification IF good quality training data is available unsupervised classifiers are used to carry out preliminary analysis of data prior to supervised classification