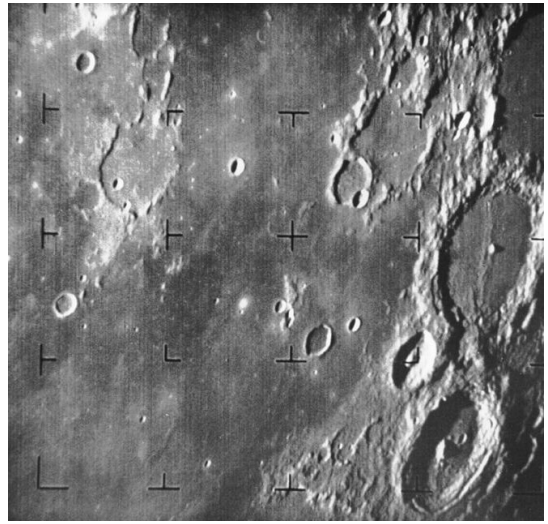
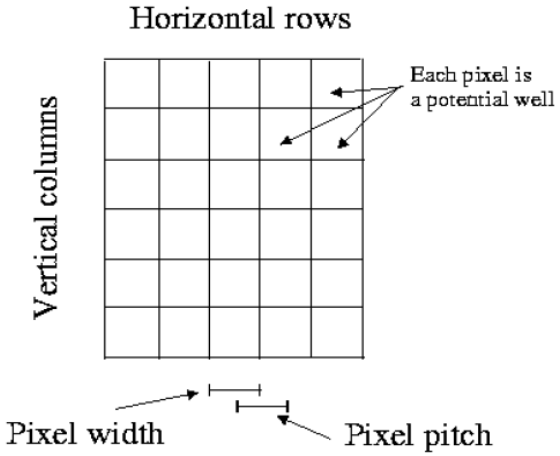
Lesson 9: Digital Data Analysis

What Is Data?

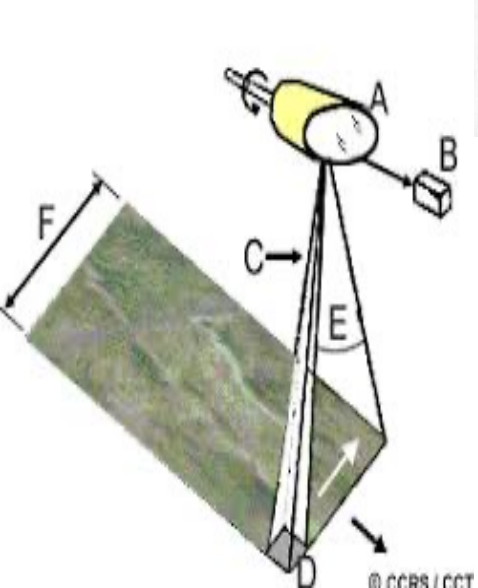
- Pixel represents is the smallest element of a feature with discreet values of brightness
- Pixel has both (location and value)

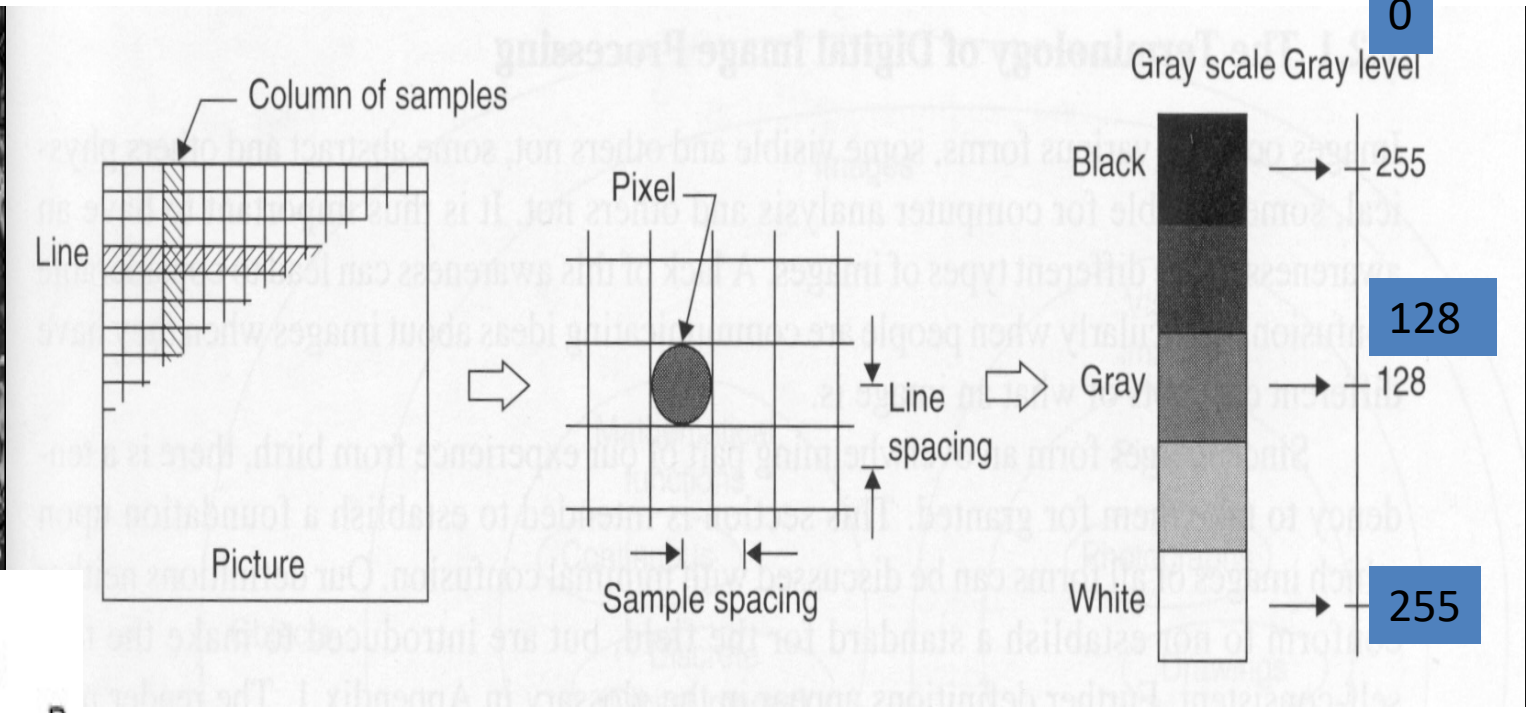


y

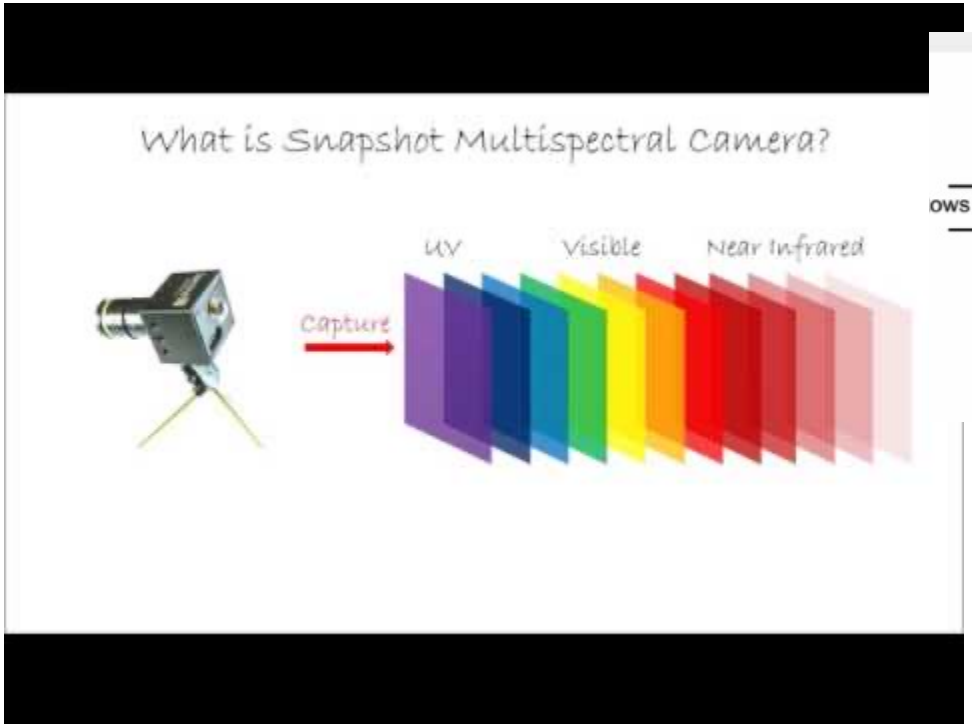


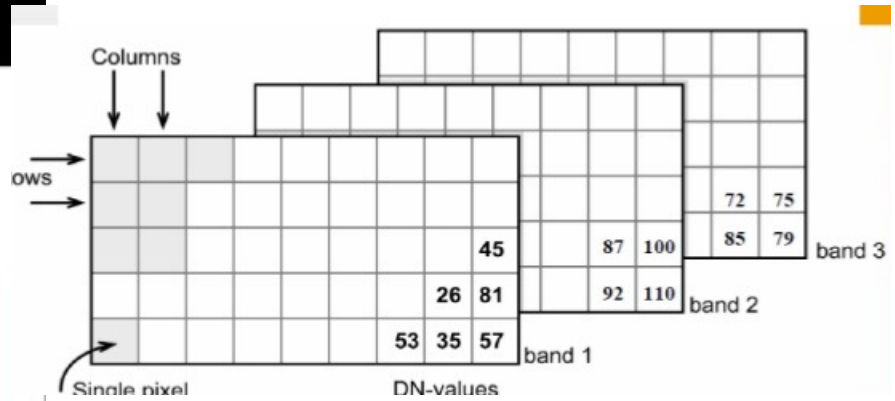
WHAT IS DATA



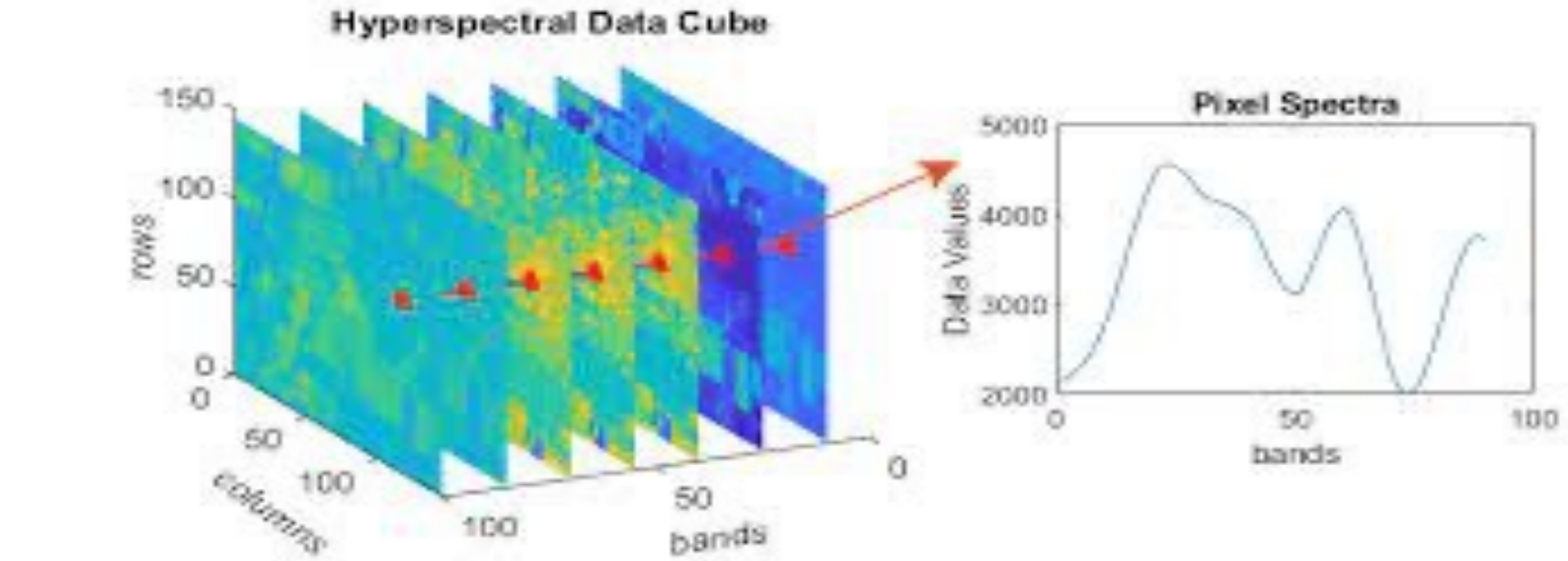


Each pixel is characterized
by some single value of radiation
(e.g., reflectance) impinging on
a detector that is converted electrical signal and detected





MULTISPECTRAL DATA



DIGITAL IMAGE PROCESSING:

Pre processing:

Correct for the errors during image acquisitions
(errors like geometry, radiometry, atmosphere)

Enhancement:

Improve the interpretability of image

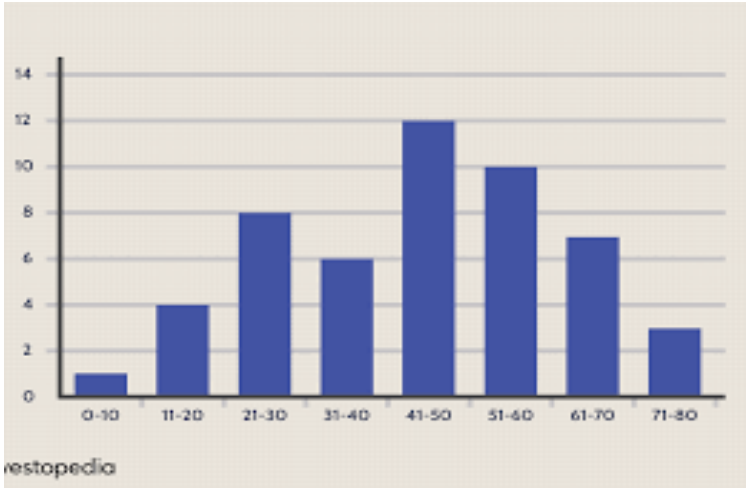
Image transormation :

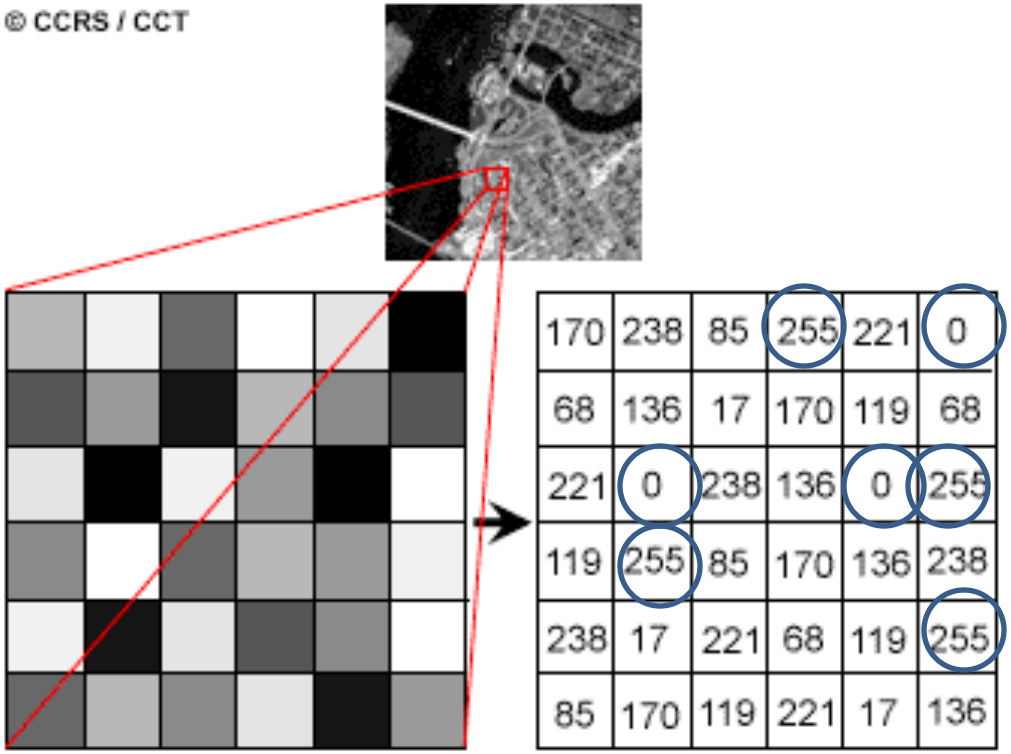
Manipulation of multiband data to generate new image , which highlights particualr feature
e.g vegetaion index

Image classification:

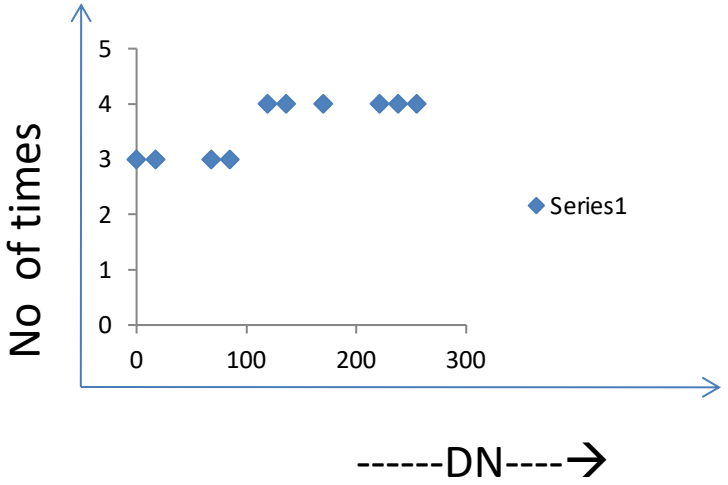
Process of sorting pixels into finite number of individual classes.

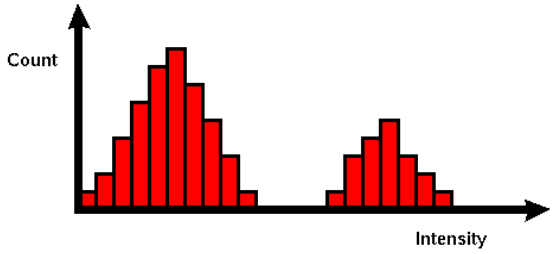
A **histogram** is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.





A **histogram** shows the statistical frequency of data distribution in a dataset. In the case of **remote sensing**, the dataset is an image, the **data distribution is the frequency of the pixels in the range of 0 to 255,** which is the range of the 8-byte numbers used to store image information on computers.





HISTOGRAM USES:

Overall statistics of the data (mean, sd, range and distribution)

contrast stretching: which involves altering the distribution and range of DN values,

1. Purpose: Improve the quality/ interpretability of image , Image stretching

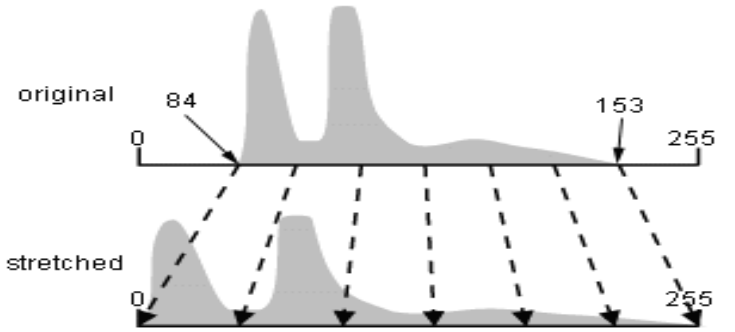
2. Classification:

Categorize the pixels into number of classes

Image Enhancement

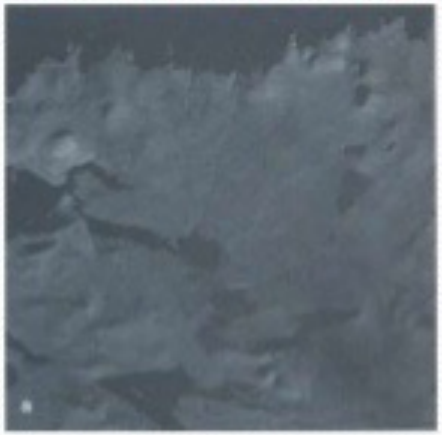
- ***Contrast Stretching:*** Quite often the useful data in a digital image populates only a small portion of the available range of digital values (commonly 8 bits or 256 levels). Contrast enhancement involves changing the original values so that more of the available range is used, this then increases the contrast between features and their backgrounds. There are several types of ***contrast enhancements*** which can be subdivided into ***Linear*** and ***Non-Linear procedures***.

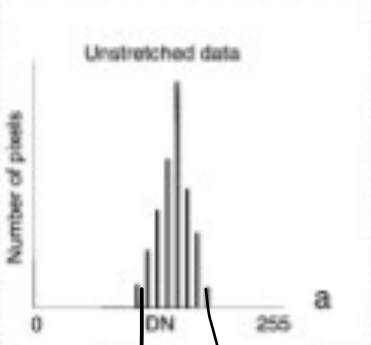
$$g(x,y) = \frac{f(x,y)-fmin}{fmax-fmin} * 2^{bpp}$$



Linear Contrast Stretching…

Landsat TM Band-5 image and Histogram





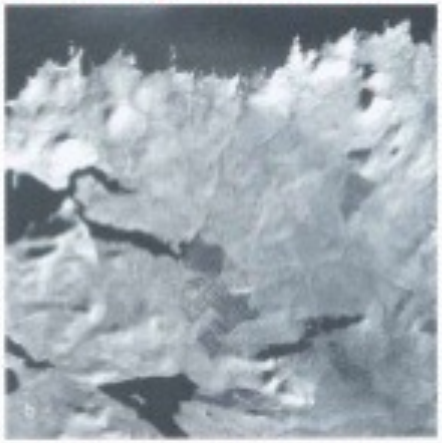
Displayed in an 8-bit system

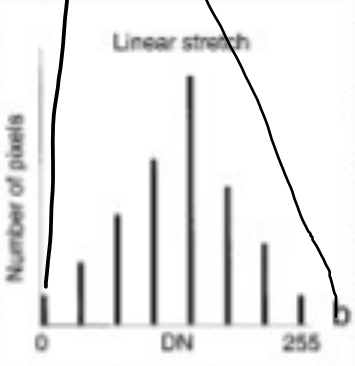
Image is vague

DN values range from 60-158

0-59 and 159-255 are not utilized

Linearly stretched Landsat TM Band-5 image and Histogram





DN values are stretched to 0-255

Contrast is improved

Light tones appear lighter

Dark tones appear darker

Contrast stretching

151	151	151	152	152
150	151	151	151	152
150	150	150	151	151
149	149	149	150	151
148	147	148	150	151
148	148	148	150	150

204	204	204	255	255
153	204	204	204	255
153	153	153	204	204
102	102	102	153	204
51	0	51	153	204
51	51	51	153	153

A simple equation shows the procedure for scaling the digital values is the image:

[(gray value – MIN gray value)/(MAX gray value – MIN gray value)]* 255

[(147-147)/(152-147)]*255 = 0 or [(152-147)/152-147)]*255=255

The scaling of the gray level values in an image to make hidden information visible is called *contrast stretching*. The idea behind contrast stretching is to increase the dynamic range of the gray levels in the image

HISTOGRAM USE

- Thresholding: DNs are divided to two groups
- DNs less than threshold $\rightarrow$ 0
- DNs more that threshold $\rightarrow$ 1
- E.g. separate water areas from land areas

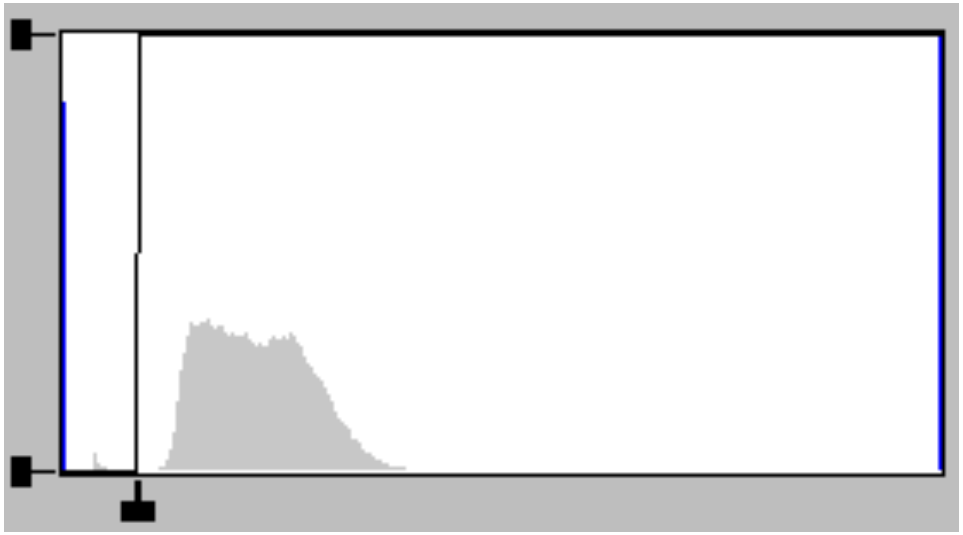


Image transformation involves manipulation of multiple bands of data, from a single multispectral image or from two or more images of the same area acquired at different times.

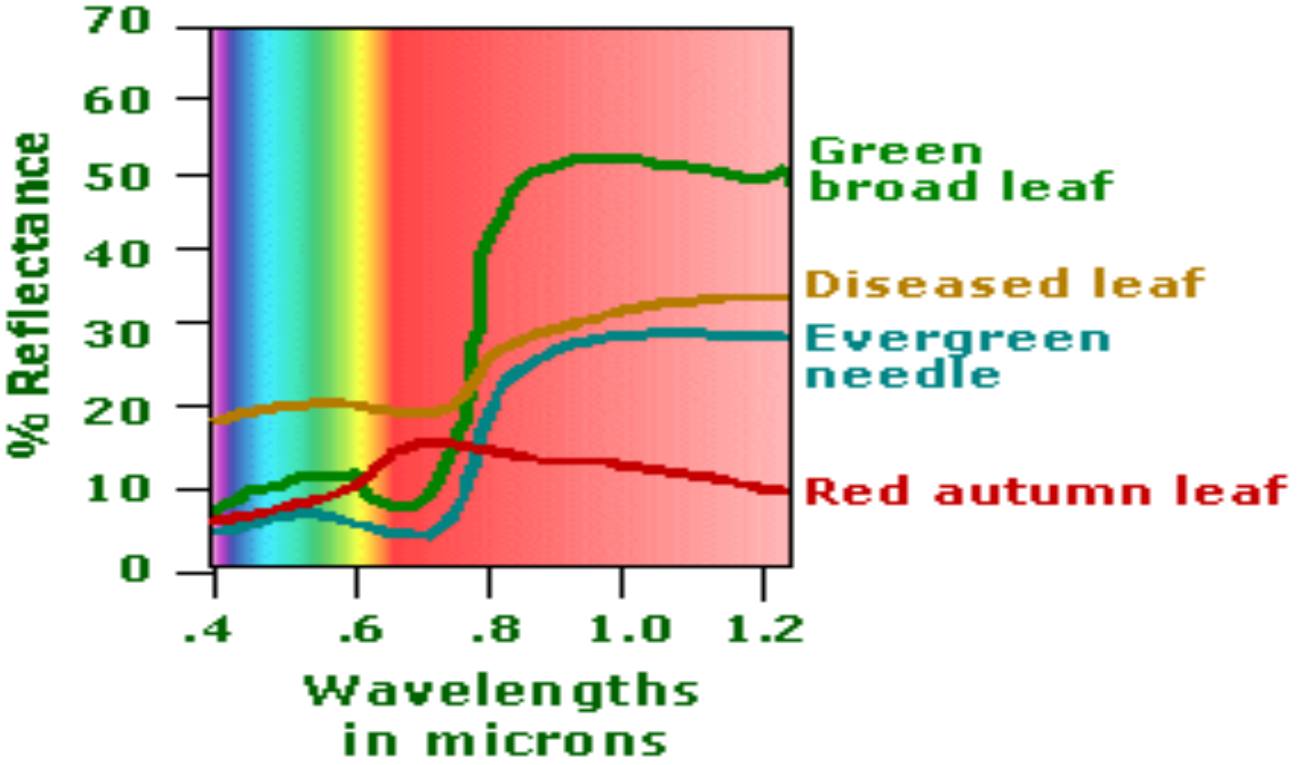
In this process, a raw image or several set of images undergo some mathematical treatment to achieve a new imagery.

As a consequence, new resultant transformed image generated from two or more sources highlights particular features or properties of interest better than original input images.

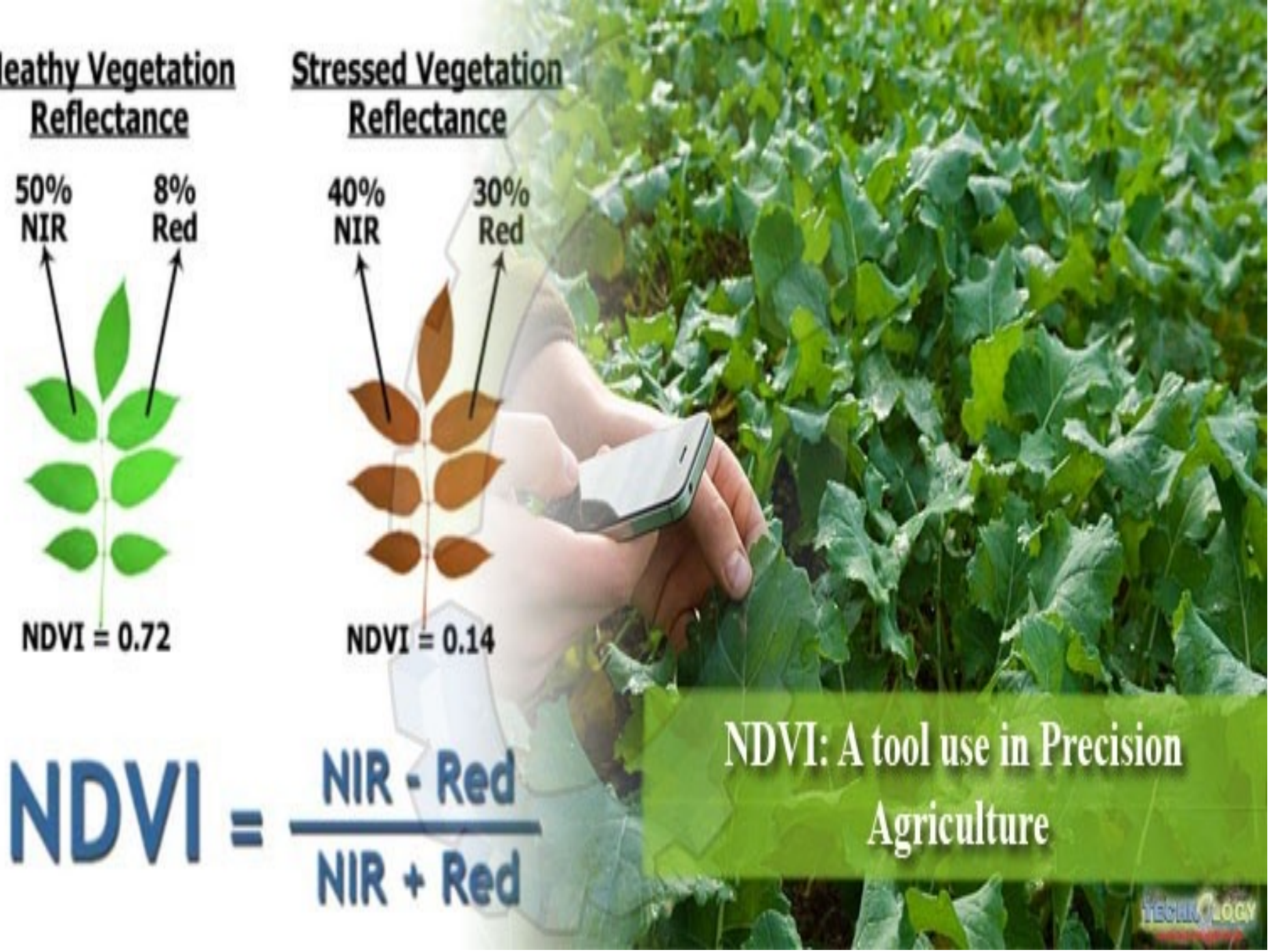
Hence, the transformed image may have properties that make it more suited to a particular purpose than original input images.

- **Chlorophyll in healthy vegetation absorbs most of visible red and blue for photosynthesis.**

- **Amount of near infrared energy reflected is a function of**
  - **internal structure**
  - **amount of moisture**



1×10^{-6} metre


leathy Vegetation
Reflectance
50%
NIR
8%
Red
NDVI = 0.72
Stressed Vegetation
Reflectance
40%
NIR
30%
Red
NDVI = 0.14
NDVI = NIR - Red / NIR + Red
NDVI: A tool use in Precision Agriculture

VEGETATION INDEX

NDVI

Normalized Difference Vegetation Index (NDVI) describes the vegetation density and assessing changes in plant health. NDVI is calculated as a ratio between the **red (R)** and **near-infrared (NIR).** In this tutorial learn how to apply the **NDVI** formula and calculate vegetation patterns.

NDVI Formula

(NIR – R) / (NIR + R)

Landsat 4-7, NVI = (Band 4 – Band 3) / (Band 4 + Band 3).
The Landsat 8, NDVI = (Band 5 – Band 4) / (Band 5 + Band 4).
IRS Liss-III, NDVI = (Band 3 – Band 2) / (Band 3 + Band 2).
NDVI always ranges from -1 to +1.

Normalized Difference Water Index

The NDWI index is designed for water body mapping. A water body has strong absorbability and low radiation in the visible to infrared range of wavelengths. Consequently, the green and Near Infra-red bands from remote sensing images are used in the calculation of the index. The NDWI can usually enhance the detection of water bodies

NDWI is calculated as:

$$NDWI = \frac{(NIR - SWIR1)}{(NIR + SWIR1)}$$

The NDWI index for assessing risk of fire is used to determine the presence of moisture in vegetation cover. Higher NDWI values indicate sufficient moisture, while a low value indicates water stress.

The NDWI product is dimensionless and varies from -1 to +1, depending on the hardwood content, as well as the type of vegetation and cover.

The high NDWI values) correspond to high plant water content and coating of high plant fraction.

Low NDWI values correspond to low vegetation content and cover with low vegetation. During periods of water stress the NDWI rate will decrease.

https://vedas.sac.gov.in/vstatic/vegetation_monitoring/index.html

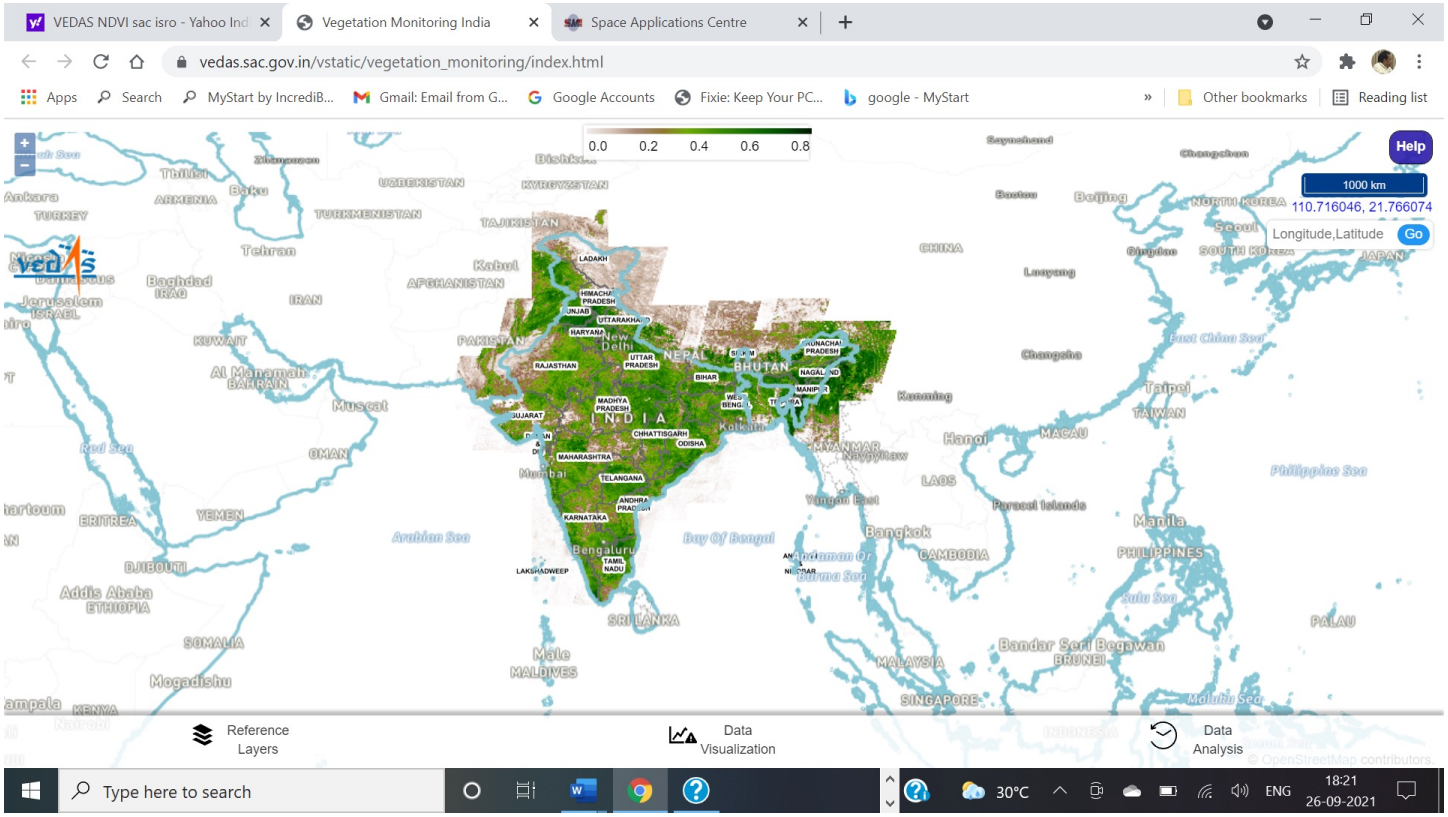
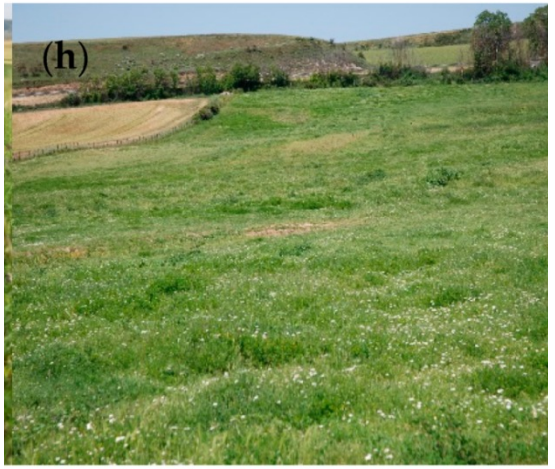


IMAGE CLASSIFICATION

SPECTRAL : It is based on the fact that the different classes of the image **have different combinations of digital values** in each band due to its reflectance or emittance.

Spatial : It is based on analyzing the relationship between neighboring pixels, considering aspects such as **texture, proximity, size, shape, repetition, etc**

Temporal : It is based on analyzing the relationship between neighboring pixels, considering aspects such as texture, proximity, size, shape, repetition, etc.



CLASSIFICATION :Supervised vs. Unsupervised Approaches

- Supervised - image analyst "supervises" the selection of spectral classes that represent patterns or land cover features that the analyst can recognize

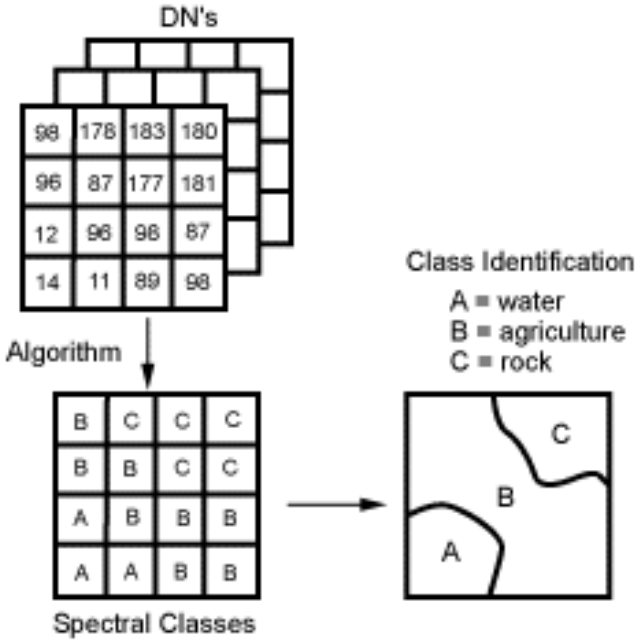
  Prior Decision

- Unsupervised - statistical "clustering" algorithms used to select spectral classes inherent to the data, more computer-automated

- **Unsupervised Image Classification**
- **The process requires a minimal amount of initial input from the analyst**
- **A numeric operation searches for natural grouping of the spectral properties of pixels**
- **The analyst determines the information class for each spectral class after the spectral classes are formed**

- ***Unsupervised Classifications***

  this is a computerized method without direction from the analyst in which pixels with similar digital numbers are grouped together into ***spectral classes*** using statistical procedures such as ***nearest neighbour*** and ***cluster analysis***. The resulting image may then be interpreted by comparing the clusters produced with maps, airphotos, and other materials related to the image site.



Natural clustering: Spectral

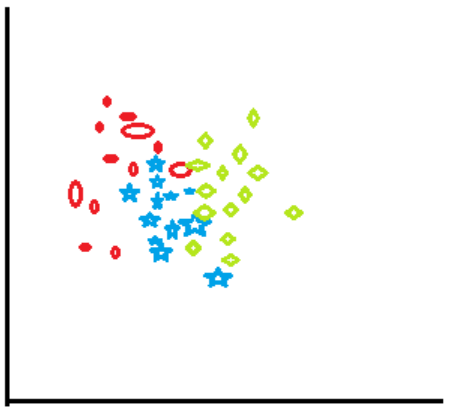


fig 1: before applying k-means clustering

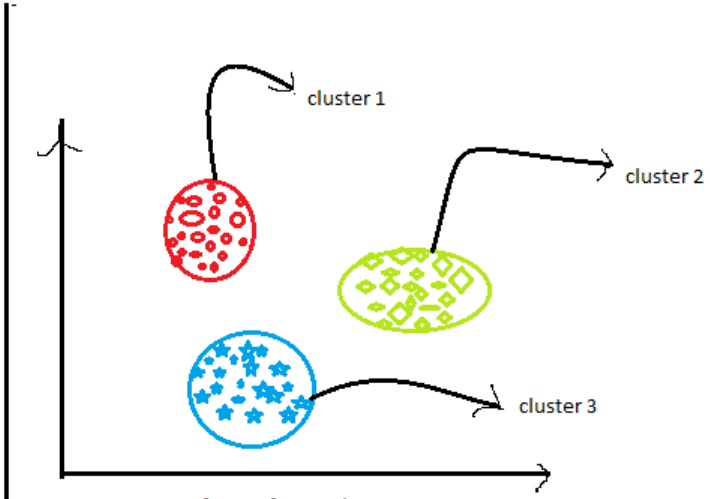


fig 2: After applying K-means clustering

What is K-means?

K-means clustering is one in every of the only and most common unsupervised machine learning algorithms.

Typically, unsupervised algorithms create inferences from datasets mistreatment solely input vectors while not bearing on famed, or tagged, outcomes.A cluster refers to a collection of data points aggregated together because of certain similarities.

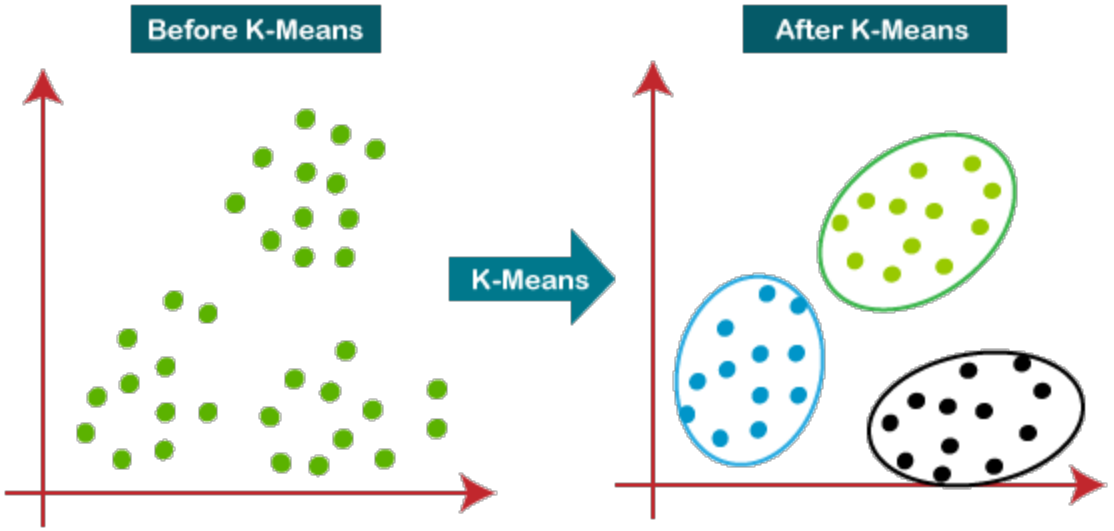
K-means Overview

- An unsupervised clustering algorithm
- **"*K*" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate *K***
- *K*-means algorithm is iterative in nature
- Works only for numerical data
- Easy to implement

The steps involved are:

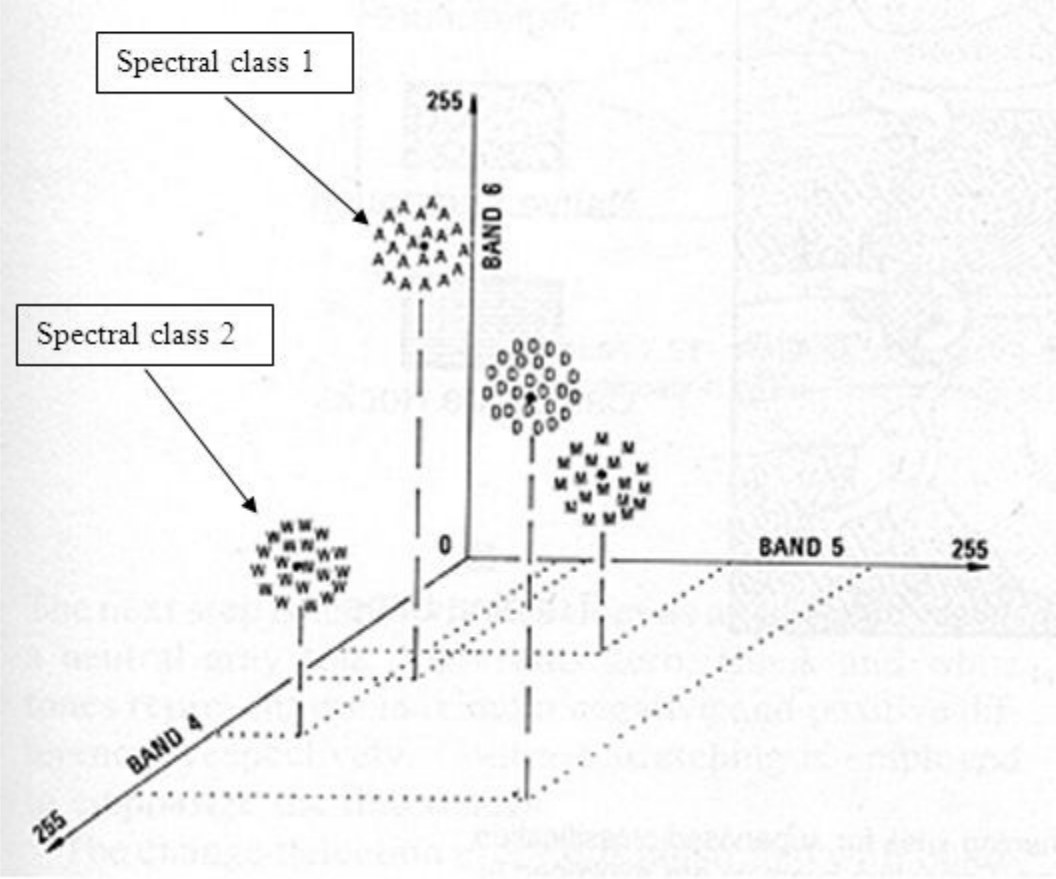
- **Specify the number of clusters as k**
- **Random initialization of centroids & allocation of data points to the nearest centroid**
- **Compute centroids by averaging data points within a cluster**
- **Re-assign data points to their closest centroid**
- **Re-compute centroids**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



Unsupervised Classification

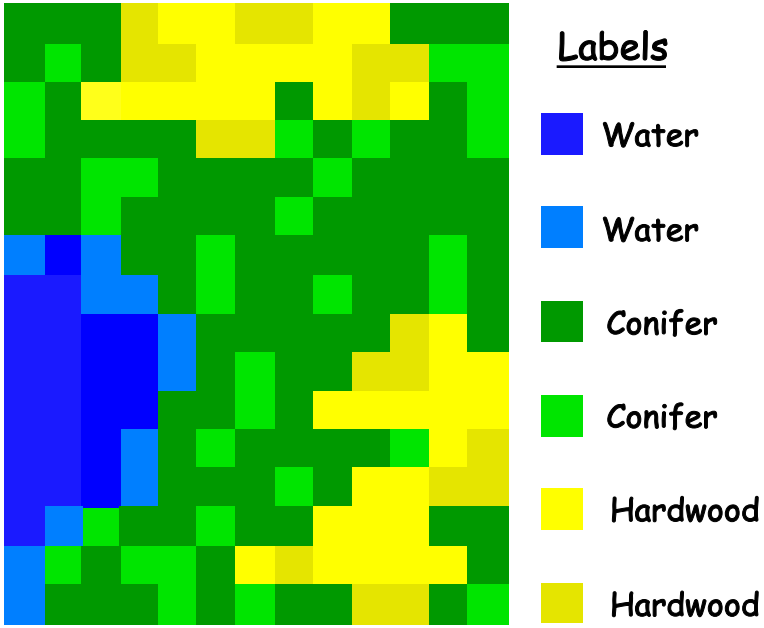
- Assumes no prior knowledge
- Computer groups all pixels according to their spectral relationships and looks for natural clusterings
- Assumes that data in different cover class will not belong to same grouping
- Once created, the analyst assesses their utility and can adjust clustering parameters



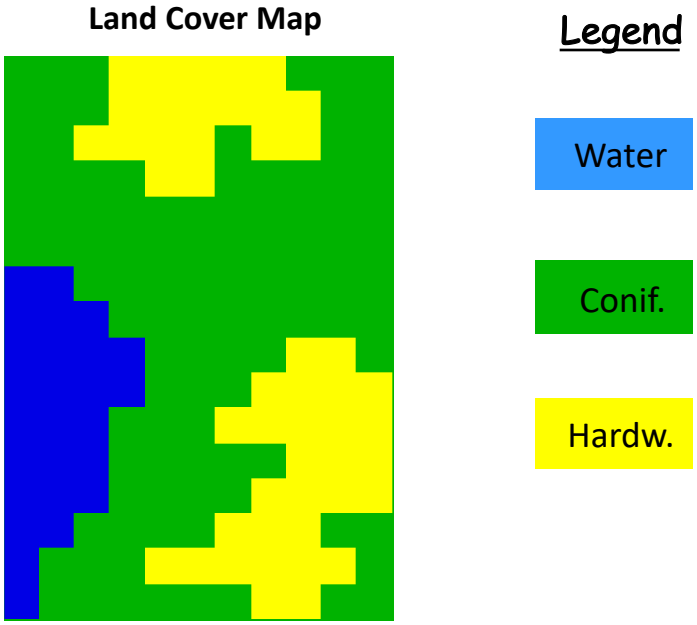
Source: F.F. Sabins, Jr., 1987, Remote Sensing: Principles and Interpretation.

Unsupervised Classification

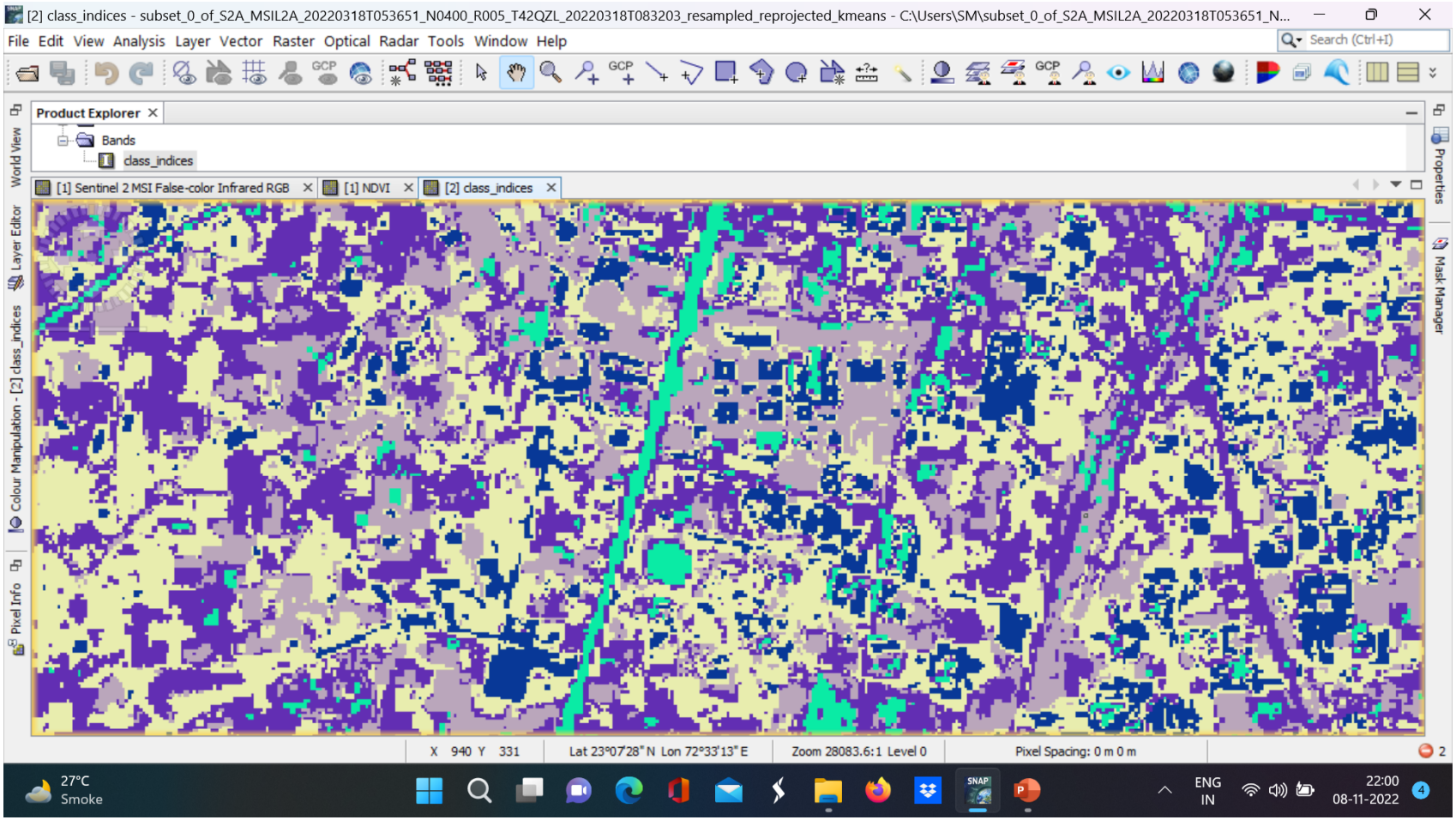
It is a simple process to regroup (recode) the clusters into meaningful information classes (the legend).



The result is essentially the same as that of the supervised classification:



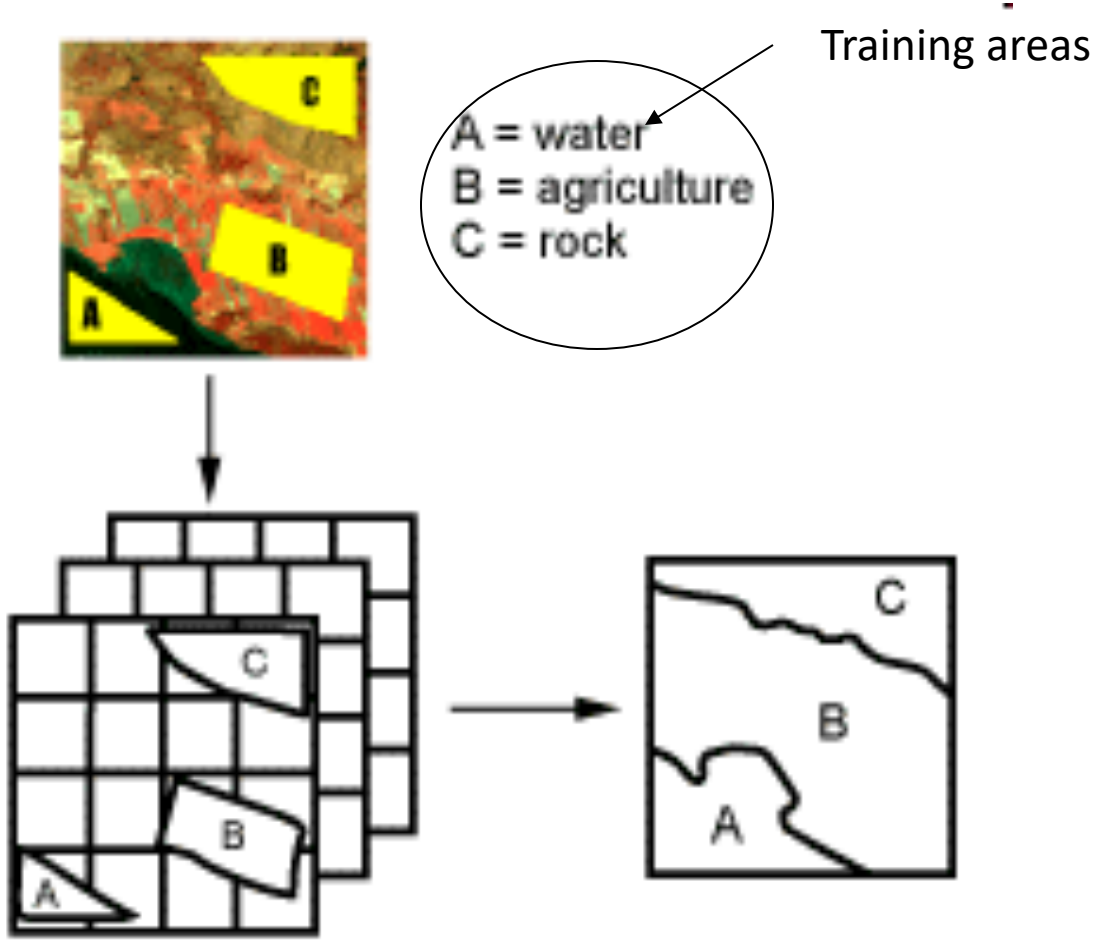
Unsupervised classifier
k=5, seed 200, iteration 10



Supervised classification

- Operator controlled
- There are three distinct stages:
  - Training
  - Classification
  - Output

- ***Supervised Classification*:**

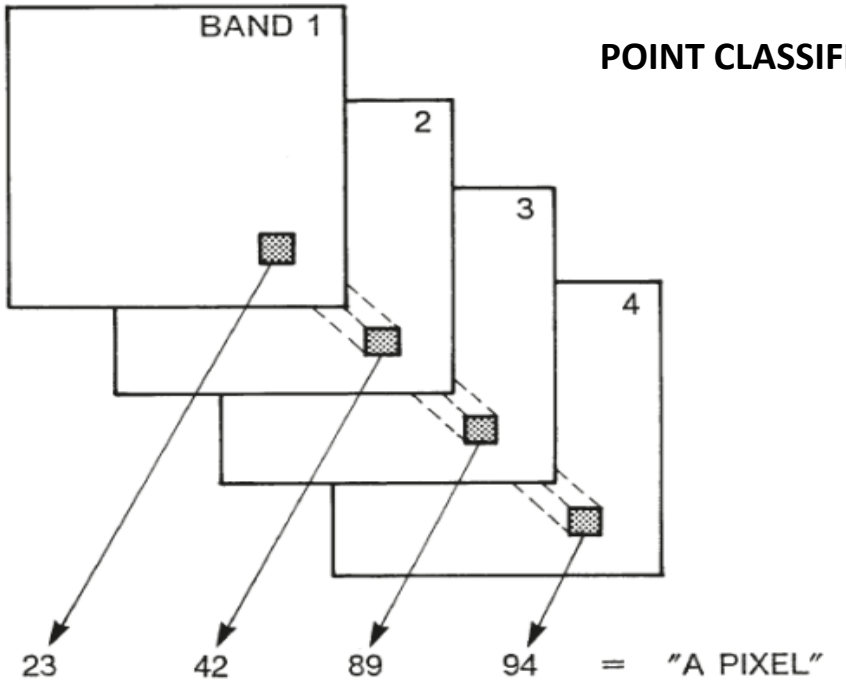


INPUT: TRAINING SAMPLE

using samples of known identity (i.e., pixels already assigned to informational classes) to classify pixels of unknown identity (i.e., to assign unclassified pixels to one of several classes).

Samples of known identity are those pixels located within training areas, or training fields. The analyst defines training areas by identifying regions on the image that can be clearly matched to areas of known identity on the image. Such areas should typify spectral properties of the categories they represent and, of course, must be homogeneous in respect to the informational category to be classified.

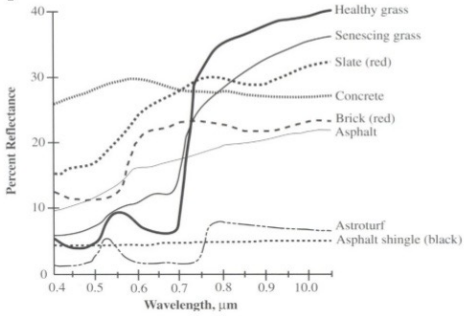
training samples used to guide the classification algorithm to assign specific spectral values to appropriate informational classes. Clearly, the selection of these training data is a key step in supervised classification



POINT CLASSIFIER

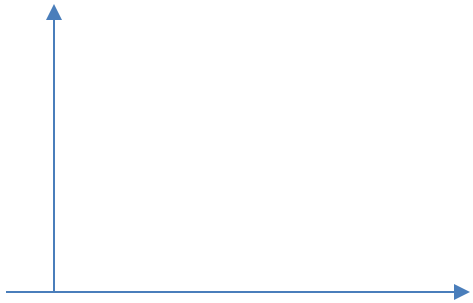
A. The nature of urban areas affecting image interpretation

(i) Spectral characteristics of common urban materials



33

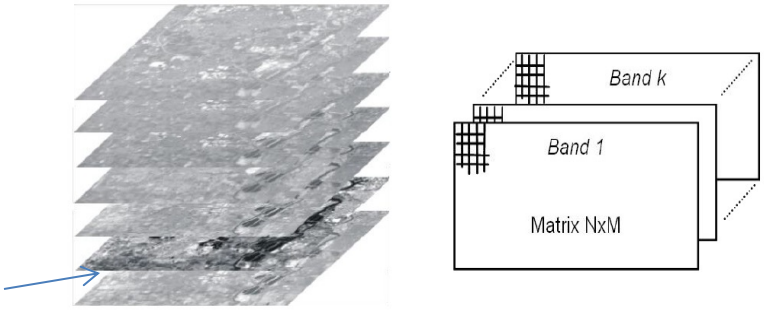
Source: Remote Sensing of the environment by J.R. Jensen, 2000



EASY IN IMPLEMENTATION

DO NOT CONSIDER SPATIAL RELATIONSHIP WITH NEIGHBOURING PIXEL

Uses the spectral information represented by the digital numbers in one or more spectral bands.

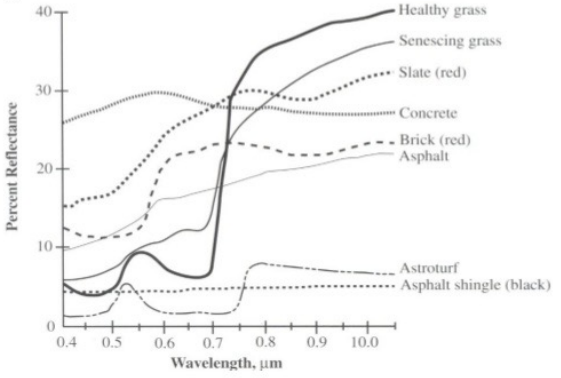


Classifies each individual pixel based on this spectral information.
Also known as spectral pattern recognition.

The objective: To assign all pixels in the image to particular classes or themes (e.g., water, coniferous forest deciduous forest crops bare soil etc)

A. The nature of urban areas affecting image interpretation

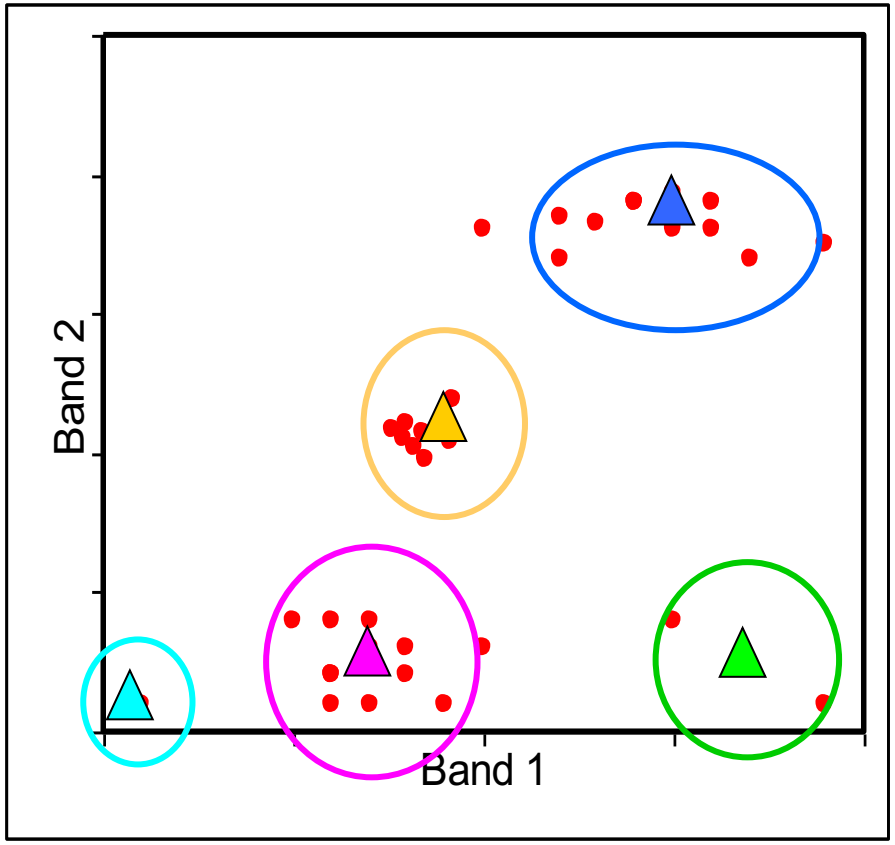
(i) Spectral characteristics of common urban materials

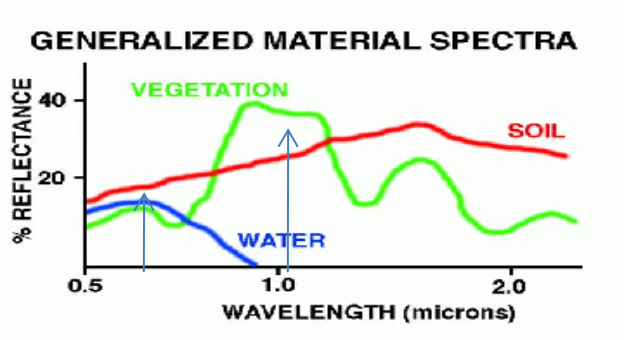


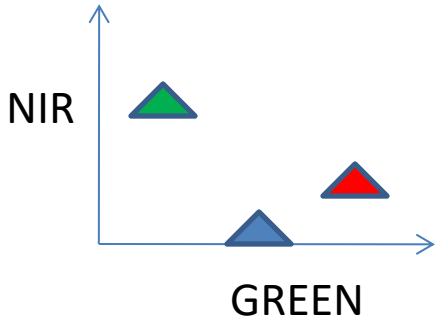
Source: Remote Sensing of the environment by J.R. Jensen, 2000

Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

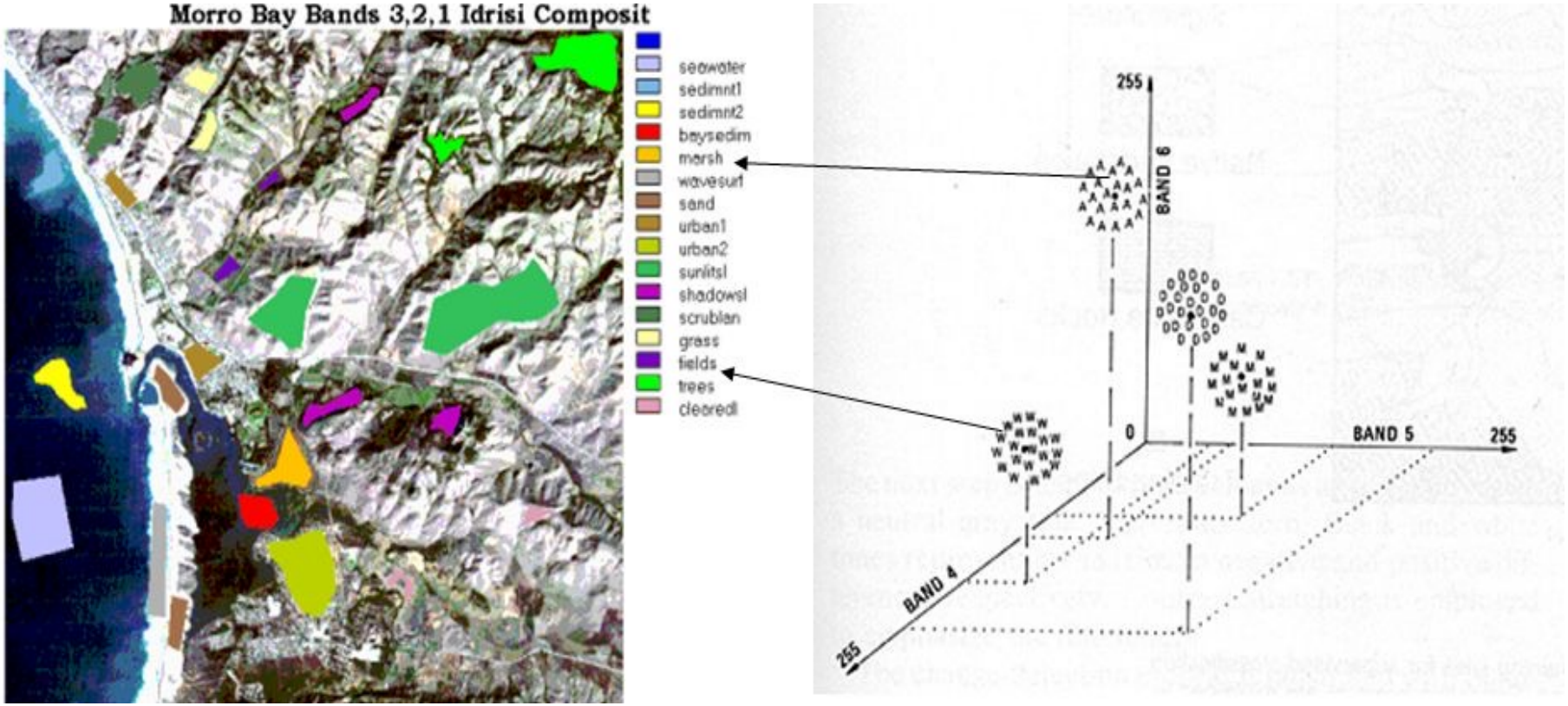
The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.



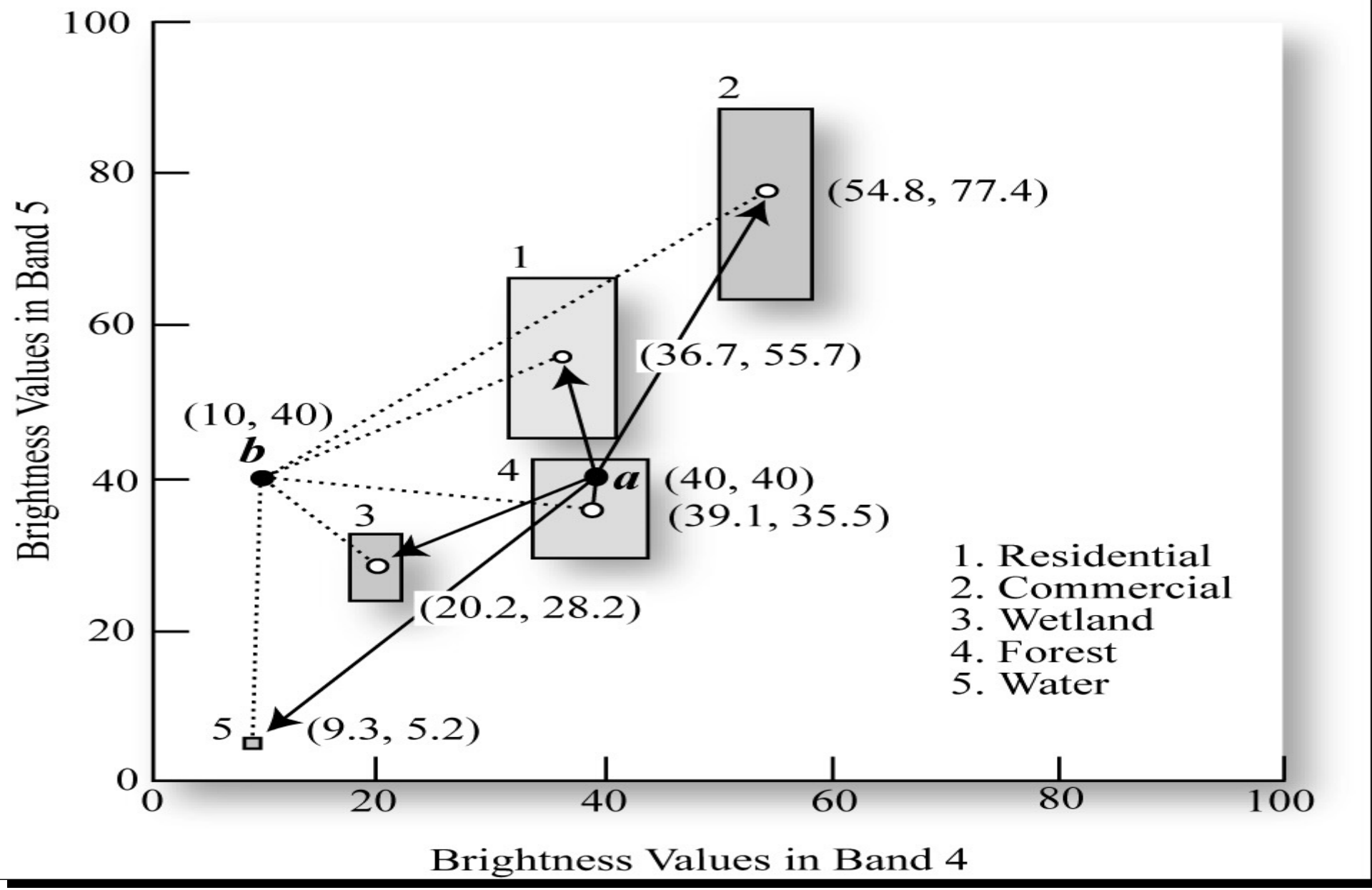




TRAINING


Morro Bay Bands 3,2,1 Idrisi Composit
seawater
sedimnt1
sedimnt2
baysedim
marsh
wavesurf
sand
urban1
urban2
sunlitsl
shadowsl
scrublan
grass
fields
trees
clearedl
255
BAND 6
0
BAND 5
255
BAND 4
255

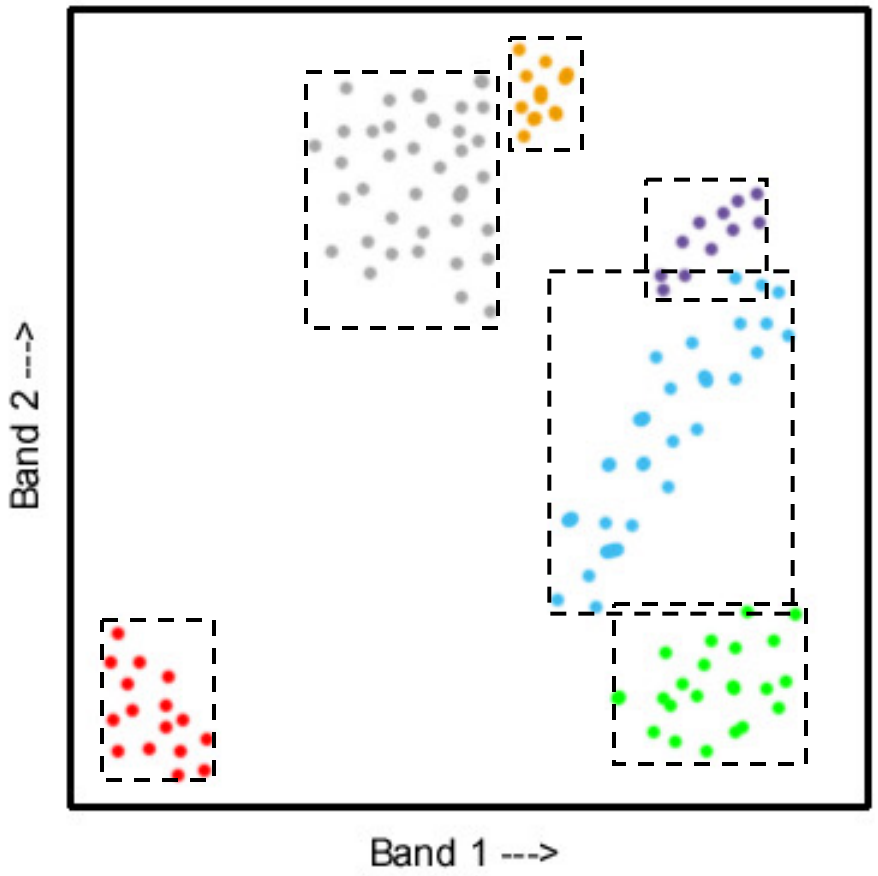
- Algorithms include
ALGORITHM:
  - Minimum distance to means classification (Chain Method)
  - Gaussian Maximum likelihood classification
  - Parallelpiped classification
- Each will give a slightly different result
- The simplest method is "minimum distance" in which a theoretical center point of point cloud is plotted, based on mean values, and an unknown point is assigned to the nearest of these. That point is then assigned that cover class.



$$\mu_{ck} - \sigma_{ck} \leq BV_{ijk} \leq \mu_{ck} + \sigma_{ck}$$

Parallelepiped example continued

- Each class type defines a spectral box
- Note that some boxes overlap even though the classes are spatially separable.
- This is due to band correlation in some classes.
- Can be overcome by customising boxes.



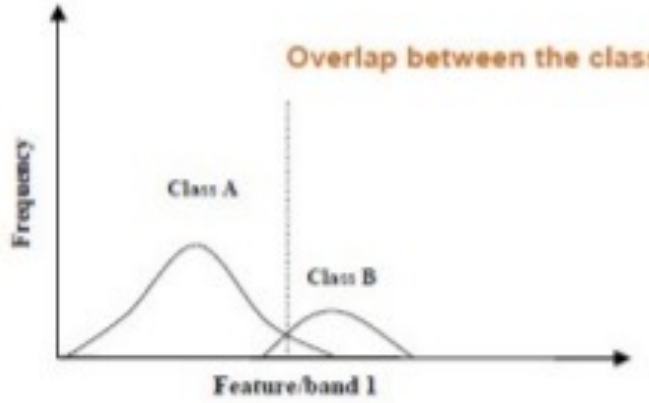
Maximum likelihood

- Instead based on training class multispectral *distance* measurements, the *maximum likelihood decision rule* is based on probability.
- The maximum likelihood procedure assumes that each training class in each band are *normally distributed* (Gaussian). Training data with bi- or *n*-modal histograms in a single band are not ideal. In such cases the individual modes probably represent unique classes that should be trained upon individually and labeled as separate training classes.
- **the probability of a pixel belonging to each of a predefined** set of *m* classes is calculated based on a normal probability density function, and the pixel is then assigned to the class for which the probability is the highest.

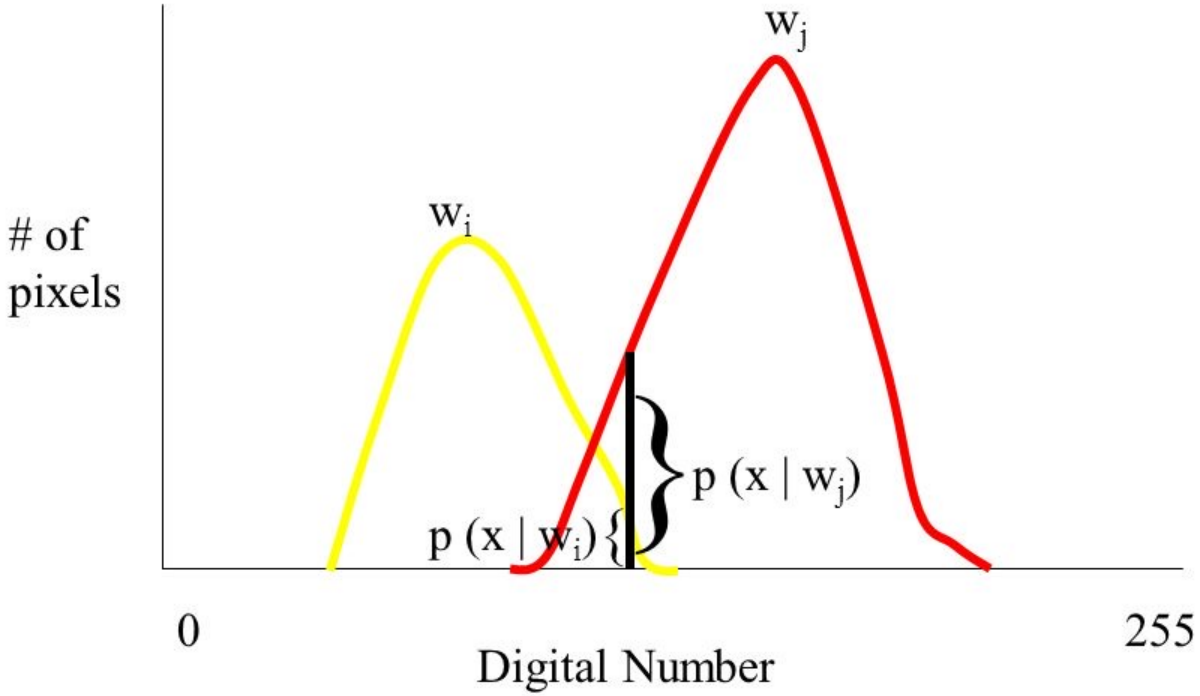
Boundaries between clusters

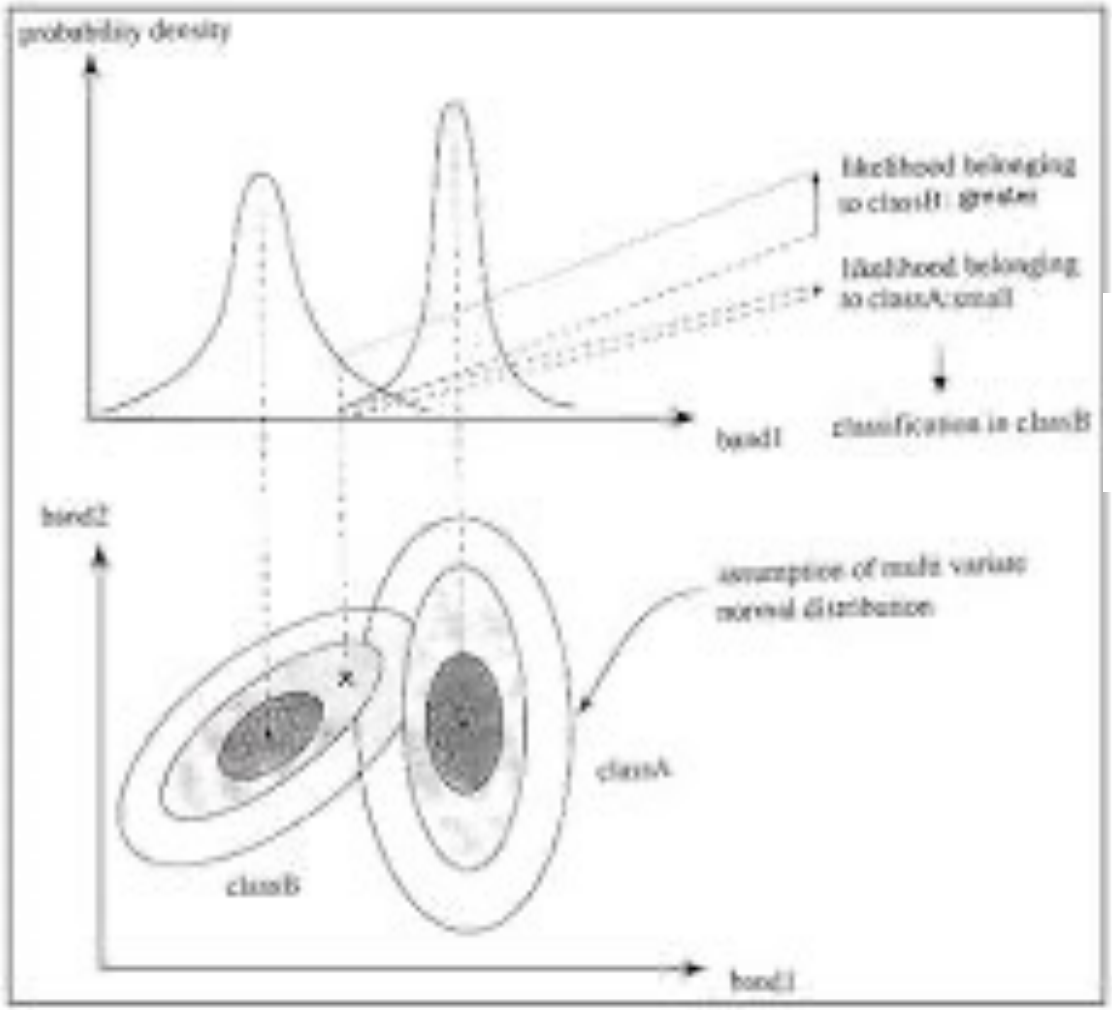
Clear distinction of class A and B

Overlap between the classes



Probabilities used in likelihood ratio





$$d_e(m_{ik}) = \sqrt{\sum_{j=1}^{nb} (m_{ij} - k_j)^2}$$ Numerical form

$$= [(C_i - k)^T (C_i - k)]^{1/2}$$ Matrix form

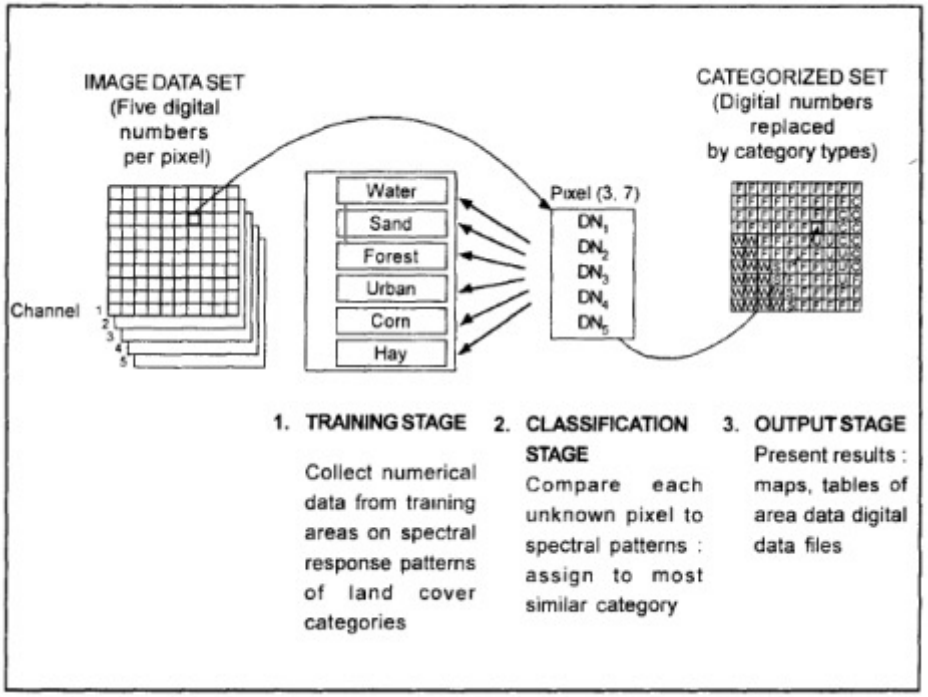
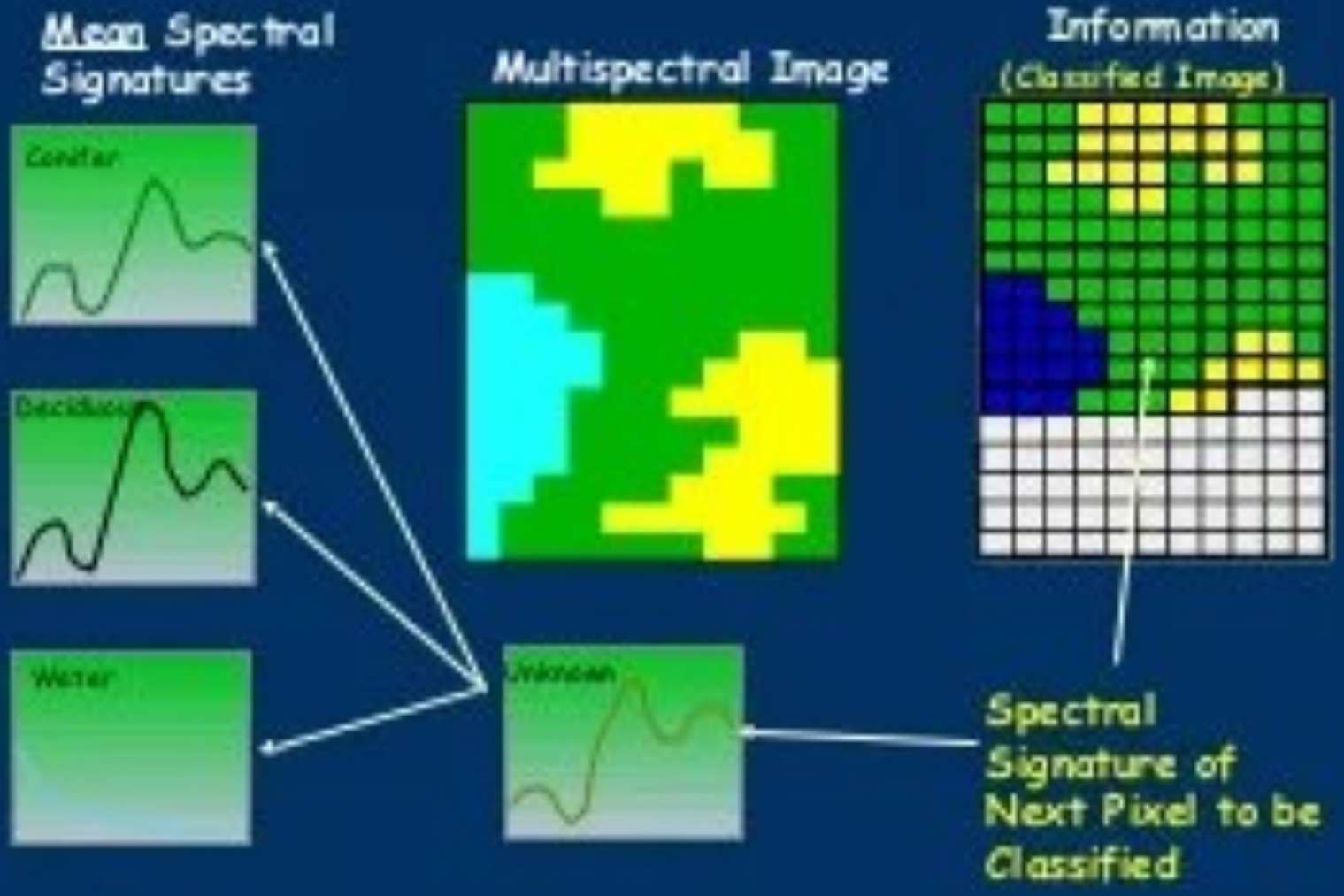


IMAGE DATA SET
(Five digital numbers per pixel)
CATEGORIZED SET
(Digital numbers replaced by category types)
Channel
1
2
3
4
5
Water
Sand
Forest
Urban
Corn
Hay
Pixel (3, 7)
DN_1
DN_2
DN_3
DN_4
DN_5
1. TRAINING STAGE
Collect numerical data from training areas on spectral response patterns of land cover 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

Supervised Classification



How do we do accuracy assessment?

- Collect reference data, i.e., “ground truth”

  determining class types at specific locations
- Compare a map with the reference to compute accuracy measures
- Interpretation of the results

Accuracy assessment

- Collect reference data, i.e., “ground truth” determining class types at specific locations
- **Compare a map with the reference to compute accuracy measures**
- Interpretation of the results

Overall Accuracy

- Of all of the reference sites, what proportion were mapped correctly?

- Easiest to understand but least amount c information for map users and map prod (us).

Overall Accuracy

Reference data	Classified image			
	Water	Forest	Urban	Total
Water	21	5	7	33
Forest	6	31	2	39
Urban	0	1	22	23
Total	27	37	31	95

Correctly classified:
21 + 31 + 22 = 74

Total number reference sites = 95

Overall accuracy
= 74 / 95 = **77.9%**

Example of confusion matrix

	Training data (known cover types)						
Classification results	Water	Concrete	High buildings	Bare soils	Grass slopes	Forest	Row total
Water	**93**	0	2	1	0	0	96
Concrete	0	**65**	4	6	0	0	75
High buildings	2	3	**124**	5	9	12	155
Bare soils	2	3	21	**165**	24	12	227
Grass slopes	0	0	6	16	**201**	45	268
Forest	0	0	8	9	76	**512**	605
Column total	97	71	165	202	310	581	1426

Producer's accuracy

W = 93/97 = 96% B = 165/202 = 82%

C = 65/71 = 92% G = 201/310 = 65%

H = 124/165 = 75% F = 512/581 = 88%

User's accuracy

W = 93/96 = 97% B = 165/227 = 73%

C = 65/75 = 87% G = 201/268 = 75%

H = 124/155 = 80% F = 512/605 = 85%

Overall accuracy = (93 + 65 + 124 + 165 + 201 + 512)/ 1426 = 81%

κ= (1160 - 365.11) / (1426 - 365.11) = 0.749

- 124 sample points have been correctly classified as high buildings
- But 2 genuine high building samples have been classified as water,
- and 2 water samples have been classified as high buildings

Accuracy is how close a value is to its true value. An example is how close an arrow gets to the bull's-eye center.

Precision is how repeatable a measurement is. An example is how close a second arrow is to the first one (regardless of whether either is near the mark)

Precision is how close two or more measurements are to each other. If you consistently measure your height as 5'0" with a yardstick, your measurements are precise.

Accurate and Precise | Not Accurate, Precise | Not Accurate, Not Precise

Kappa Coefficient

reference data \ classified image	forest	shrubland	grassland	urban	*totals*
forest	150	5	15	10	*180*
shrubland	15	55	5	5	*80*
grassland	10	20	105	5	*140*
urban	25	20	5	50	*100*
totals	*200*	*100*	*130*	*70*	*500*

$$\hat{K} = \frac{M\sum_{i=j=1}^{r} n_{ij} - \sum_{i=j=1}^{r} n_i n_j}{M^2 - \sum_{i=j=1}^{r} n_i n_j}$$

$$= \frac{(500\times 360) - [(180\times 200) + (80\times 100) + (140\times 130) + (100\times 70)]}{500^2 - [(180\times 200) + (80\times 100) + (140\times 130) + (100\times 70)]}$$

$$= \frac{180{,}000 - 69{,}200}{250{,}000 - 69{,}200} = \frac{110{,}800}{180{,}800}$$

$$= 0.613$$

Further reading:
Image processing and analysis
https://crisp.nus.edu.sg/~research/tutorial/process.htm

Image Classification Techniques in Remote Sensing
https://gisgeography.com/image-classification-techniques-remote-sensing/

Question BANK:

1. **What is image histogram?**
2. **What are the uses of histogram.**
3. **Explain linear contrast stretching.**
4. **In 8 bit remote sensing data of M*N matrix, minimum and maximum digital**
5. **number values are: 152 and 147.**

What would be the transformed values of DN=150,151,147,152

6. **Explain following spectral indices:NDVI and NDWI. What is use of these indices.**
7. **Following reflectivity measurements are done in two agricultural fields:**

	NIR Band	Red band	NDVI
Crop 1	0.5	0.08	
Crop 2	0.4	0.3	

Calculate NDVI of both the crop.

Which crop is healthy

8.What is image classification?

9. Explain spectral, spatial and temporal image classification

10.What do you mean by unsupervised and supervised image classification.

11. Explain advantages and disadvantages of unsupervised and supervised image classification techniques.

12. Explain “K” means clustering. What are the steps involved in K-means Unsupervised clustering

13. Explain steps involved in supervised classification technique

14. What do you mean by classification accuracy?.

15. Explain accuracy and precision.

16: An experiment resulted in following accuracy table.
What is overall classification accuracy?

Reference data	Classified image			
	Water	Forest	Urban	Total
Water	21	5	7	33
Forest	6	31	2	39
Urban	0	1	22	23
Total	27	37	31	95