GNR602
Advanced Methods in Satellite Image Processing

Instructor: Prof. B. Krishna Mohan

CSRE, IIT Bombay

bkmohan@csre.iitb.ac.in

Slot 13

Lecture 16-17 Hyperspectral Image Analysis

Contents of the Lecture

Concept of High Dimensional Image Analysis

Architecture of Hyperspectral Image Analysis System

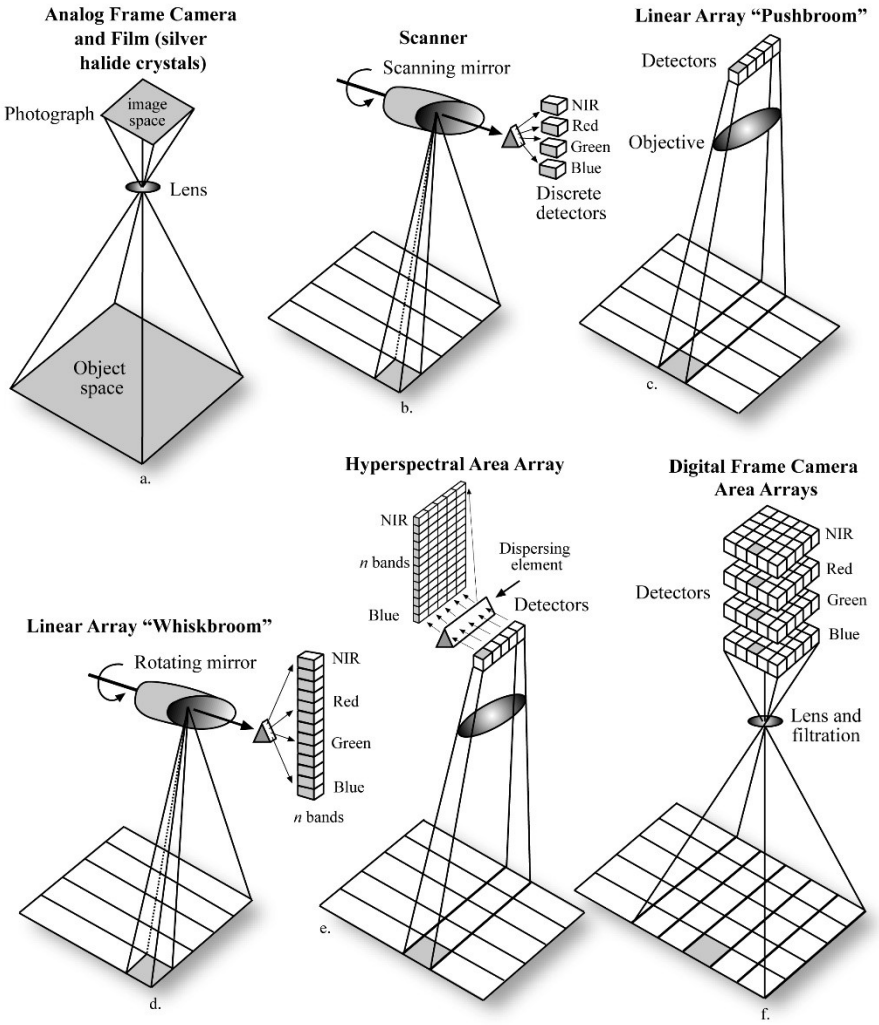
Steps in hyperspectral image analysis

Dimensionality reduction methods

Hyperspectral Image Analysis

- Hyperspectral images are very high dimensional data sets
- Number of bands in a typical hyperspectral image > 200
- A new perspective on image analysis
- New sensor technology
- Spatial resolution relatively coarse

Reproduced with permission from the lecture notes of Prof. John Jensen, University of South Carolina



Scanning Mechanism

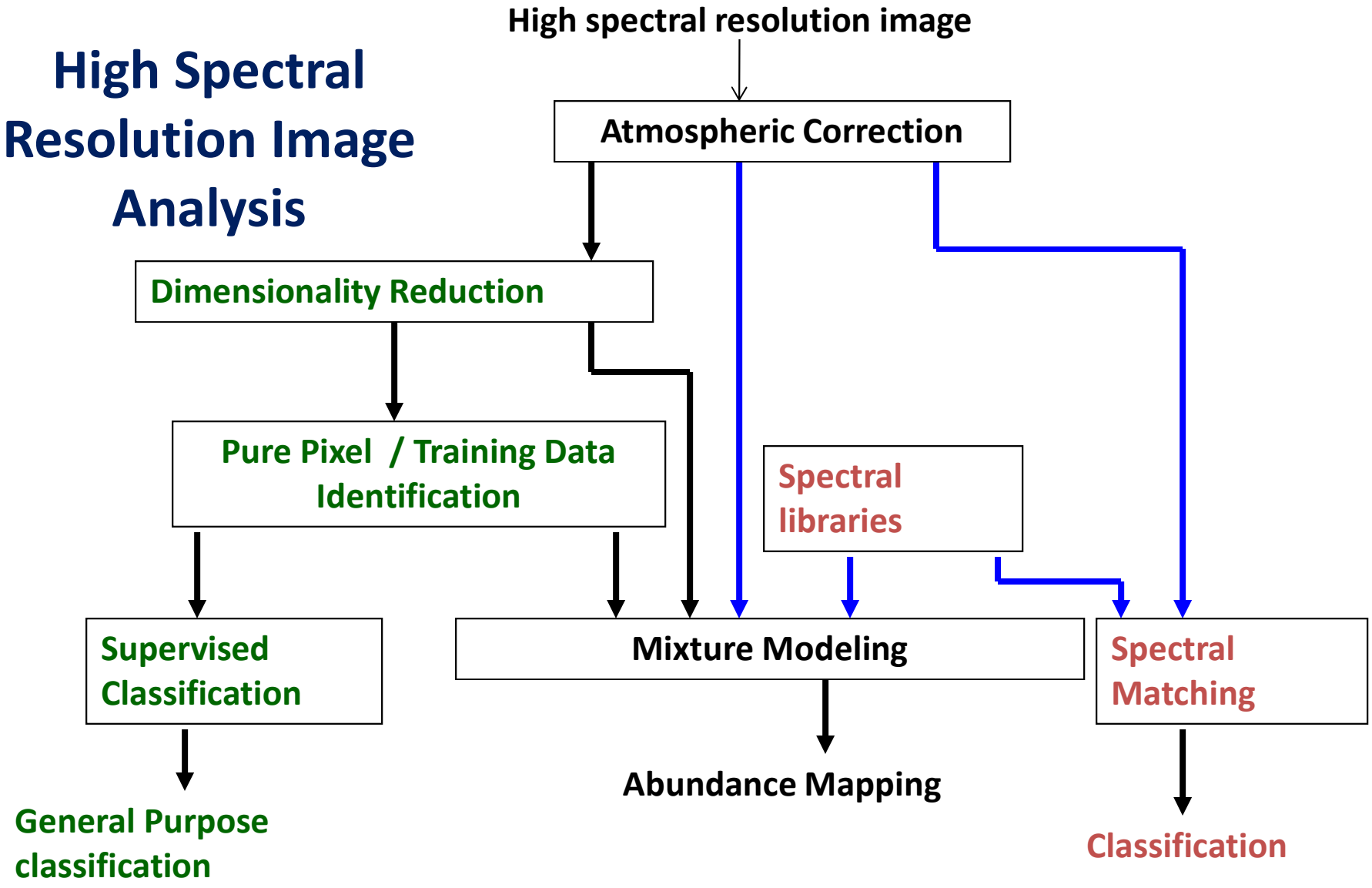
Hyperspectral Image Analysis

- Strategy for analyzing hyperspectral imagery
- Digital Image Analysis approach
  - Correct images
  - Reduce dimensionality
  - Supervised classification
  - Accuracy assessment
-

Hyperspectral Image Analysis

- Strategy for analyzing hyperspectral imagery
- Imaging Spectroscopy approach
  - Obtain reference spectra of materials of interest
  - Correct images
  - Match pixel spectra with reference spectra
  - Model mixing of classes
  - Generate map of proportion of each pure class with a pixel

High Spectral Resolution Image Analysis

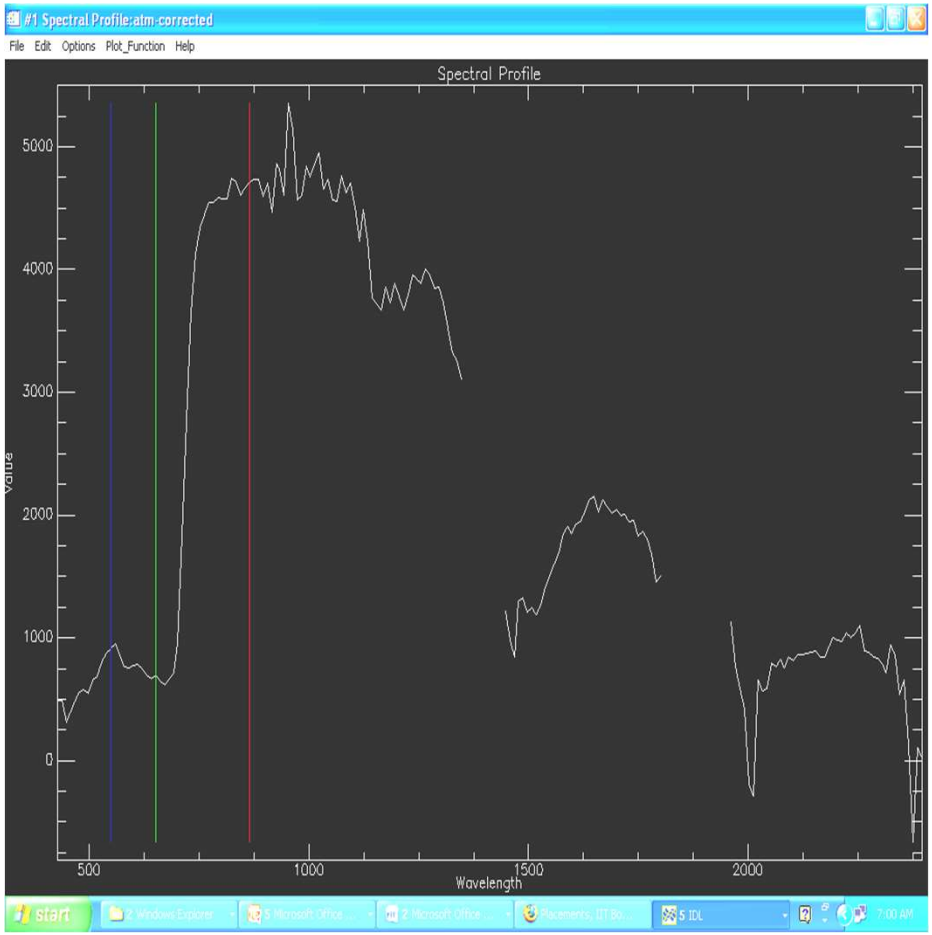


Hyperspectral Image Analysis

- Hyperspectral images are very high dimensional data sets
- Number of bands in a typical hyperspectral image > 200
- A new perspective on image analysis
- New sensor technology
- Spatial resolution relatively coarse

Why do we get spectra ?

- Spectrum is the representation of e.m energy at various wavelengths
- Examining the maxima and minima of spectral reflectance curves – minima are caused by absorption
- Difference in absorption and scattering for different wavelengths can be used to identify the features



Hyperspectral Image Analysis

- Hyperspectral images are very high dimensional data sets
- Number of bands in a typical hyperspectral image > 200
- A new perspective on image analysis
- New sensor technology
- Spatial resolution relatively coarse

Radiance, Irradiance and Reflectance

- Radiance describes the amount of light that passes through or is emitted from a particular area and falls within a given solid angle in a specified direction
- Units : W / Sr m^2

- Irradiance: radiometric measure when em radiation is incident on the surface
- Units : W/m^2

- Reflectance is the ratio of amount of light leaving a target to the amount of light striking the target

Hyperspectral Remote Sensing

- Sensor with hundreds of spectral channels with each channel covering a narrow and contiguous portion of the light spectrum
- It is not the number of measured wavelengths that defines a sensor as hyperspectral, rather it is the narrowness and contiguous nature of measurements.

  Ex: 20 bands with 10 nm wide – Hyperspectral

  20 bands, 100 nm wide – Non hyperspectral

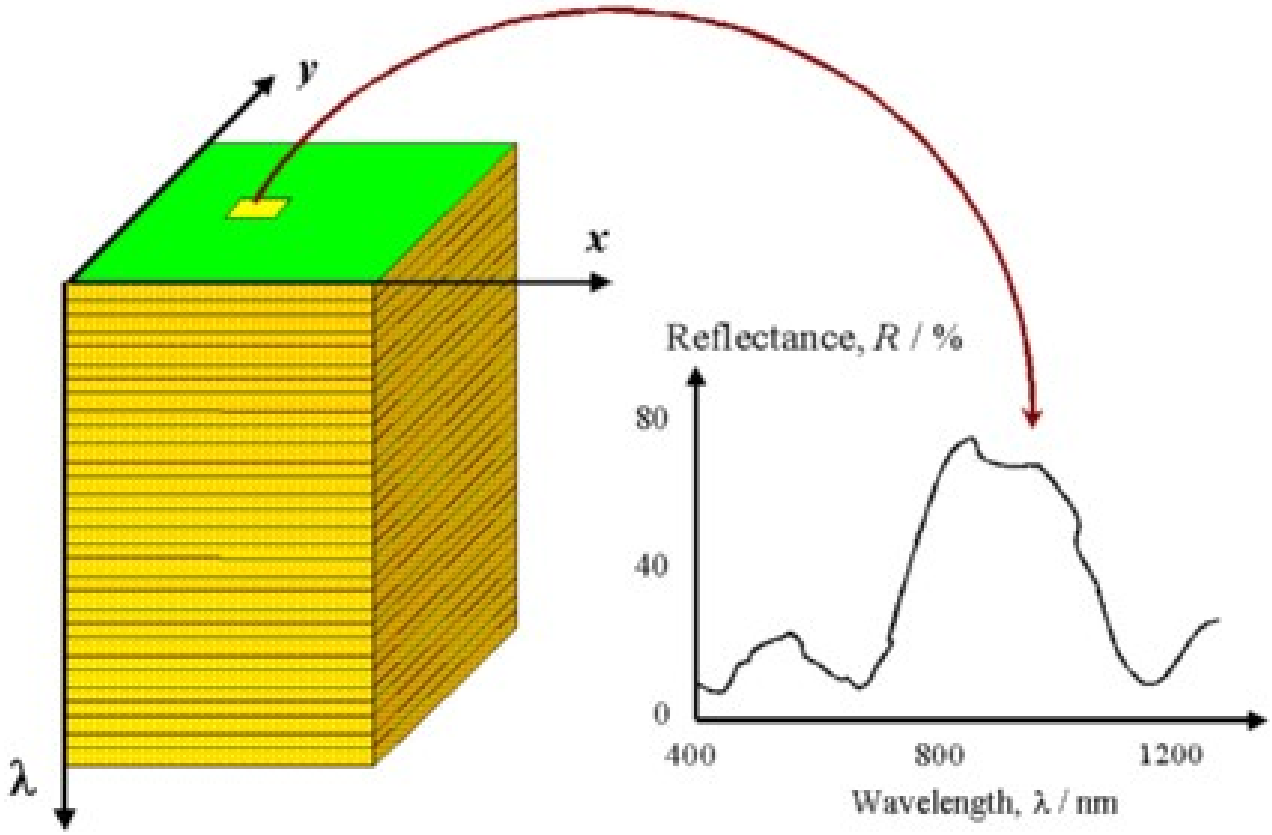
Hyperspectral Remote Sensing

- Sensor with hundreds of spectral channels with each channel covering a narrow and contiguous portion of the light spectrum
- It is not the number of measured wavelengths that defines a sensor as hyperspectral, rather it is the narrowness and contiguous nature of measurements.

  Ex: 20 bands with 10 nm wide – Hyperspectral

  20 bands, 100 nm wide – Non hyperspectral

Hyperspectral Remote Sensing



Comparison with Multispectral Remote Sensing

Multispectral

- Separated spectral bands
- Wider bandwidths
- Coarse representation of the spectral signature
- Smaller image size

Hyperspectral

- No gaps
- Narrow bandwidths
- Complete representation of the spectral signature
- Ability to detect subtle spectral features
- Large image size
- Radiometric and spectral calibration are time consuming

Challenges to Interpretation

- Data Volume
  - LandSat and Hyperion data

    Landsat TM : Number of Bands-7, Radiometric resolution-8 bit

    Hyperion : Number of Bands-242, Radiometric resolution-16 bit

    Ratio of increase in data volume is Landsat : Hyperion = 1:70
- Redundancy
  - Lot of redundancy (overlap) between adjacent bands
  - Information content of one band can be fully or partly predicted from the other bands in the data
  - Dimensionality reduction techniques prominent tools in data processing

Selected Hyperspectral Sensors

HYDICE

- HYDICE: Hyperspectral Digital Imagery Collection Experiment
- It was one of the first (1994) airborne hyperspectral instruments to be operated from a relatively low altitude thereby achieving a high spatial resolution.
- Spectral Resolution of 400 – 2500 nm with spectral channel bandwidth 3 – 15 nm
- It combined a high SNR and significantly good spatial and spectral resolution and radiometric accuracy

AVIRIS

- JPL developed the Airborne Visible/Infrared Imaging Spectrometer in 1983
- First imaging spectrometer to measure the solar reflected spectrum from 400 nm to 2500 nm
- 224 contiguous bands of 10 nm width
- Spatial resolution of 20m and size of one scene is 11 Km X 800 Km

HyMAP

- HyMap Airborne Imaging Spectrometer is operated by HyVista Corporation
- It provides 126 bands across the wavelength region of 400 – 2500 nm except in atmospheric water vapour bands
- Bandwidths are between 11 – 21 nm
- Spatial configuration of Sensor is
  - IFOV 2.5 mrad along track, 2 mrad across track, FOV is 61.3 degrees (512 pixels)
  - Swath width = 2.56 km
- High SNR > 500:1

AIMS – ISRO's Imaging Spectrometer

- SAC, ISRO developed AIMS in 1997
- Specifications
  - IFOV (µrad) : 660, 2mX2m from 3km altitude
  - Swath width (degrees) : 14.5, 770m from 3 km
  - Spectral Range : 450 - 880 (nm)
  - Encoding bits/pixel : 10
  - Number of Spectral Bands : 143
  - Spectral Bandwidth : 3 nm

Hyperion Spaceborne Sensor

- First spaceborne HSI to acquire both VNIR and SWIR through two spectrometers and a single telescope
- 220 unique spectral channels collected with a complete spectrum covering from 355 nm to 2577 nm with 10 nm bandwidth
- 30 m spatial resolution
- Instrument can image 96 km by 7.5 km land area per image
- Radiometric resolution : 16 bit signed integer
- Hyperion VNIR sensor has 70 bands (355.589 – 851.92 nm) and the SWIR has 172 bands (1057.36 – 2577.07 nm)
- SNR varies from 190 to 40 as wavelengths increases

CHRIS/PROBA
Compact High-Resolution Imaging Spectrometer

- Launched from Sriharikota, India on Oct 22, 2001
- Orbit is Sun-synchronous, at an altitude of 550 km – 670 km
- Spectral range : 400 nm – 1050 nm with 18 spectral channels
- Typical image is of 13 x 13 km size at 17m spatial resolution
- At 34 m spatial resolution the sensor can capture in 62

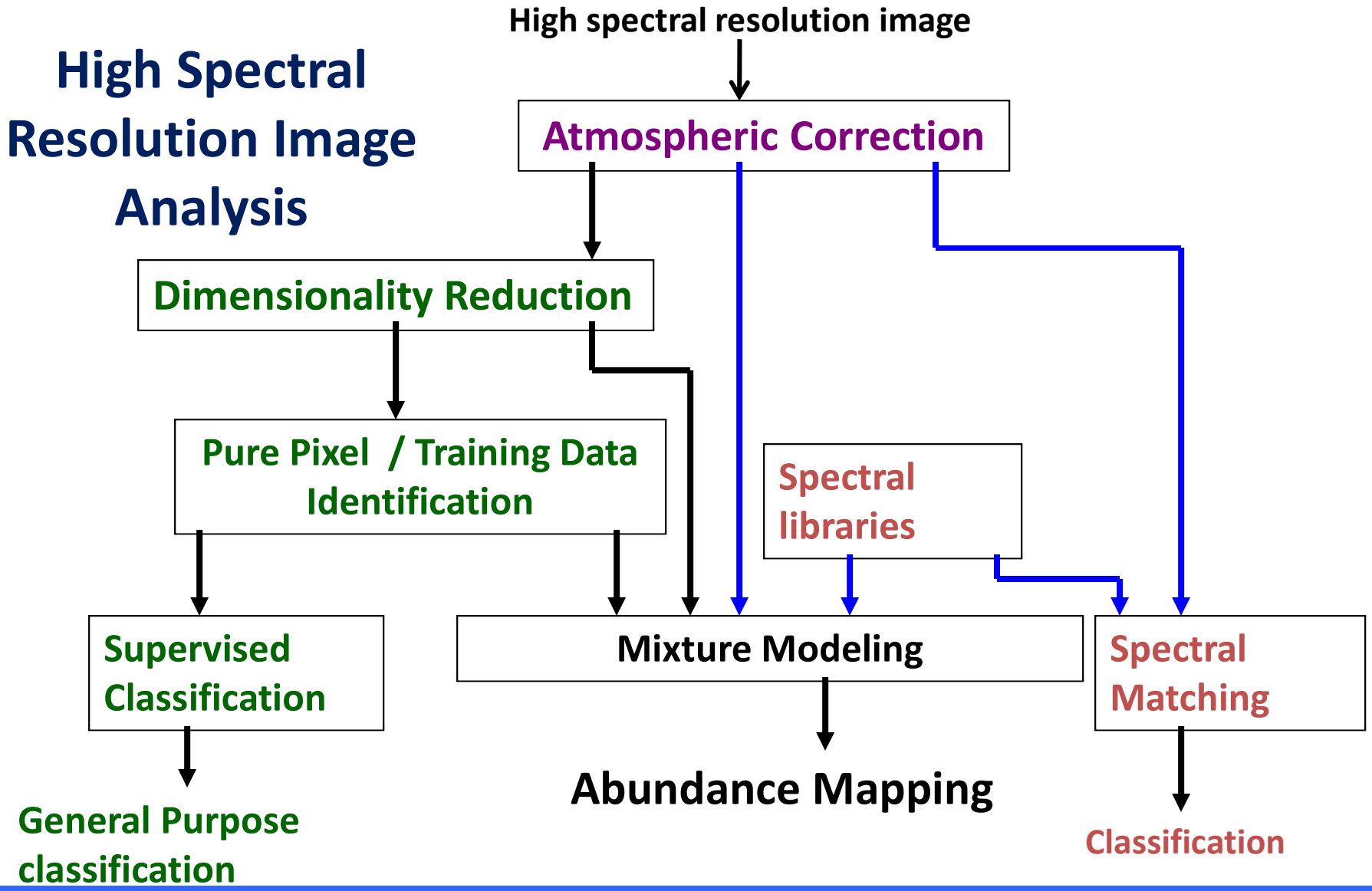
HySI – Hyperspectral Imager onboard Chandrayaan - 1

- On-board Chandrayaan-1, India's first mission to moon
- Specifications
  - Spectral range : 400 – 950 nm
  - Spectral resolution < 15 nm
  - Spatial resolution : 80 m
  - Swath width : 20 km

Hyperion Preprocessing

- Stored in BIL format - 256 cols and 3232 rows
- Level 1 radiometric product has a total of 242 bands but only 198 bands are calibrated
- Because of an overlap between the VNIR and SWIR focal planes, there are only 196 unique channels
- Overlapped bands are 56, 57 and 77, 78
- Calibrated Channels are : 8 – 57 for VNIR and 77 – 224 for SWIR
- Reason for not calibrating all 242 channels is mainly due to detectors low responsivity
- The bands that are not calibrated are set to zero for those channels

High Spectral Resolution Image Analysis



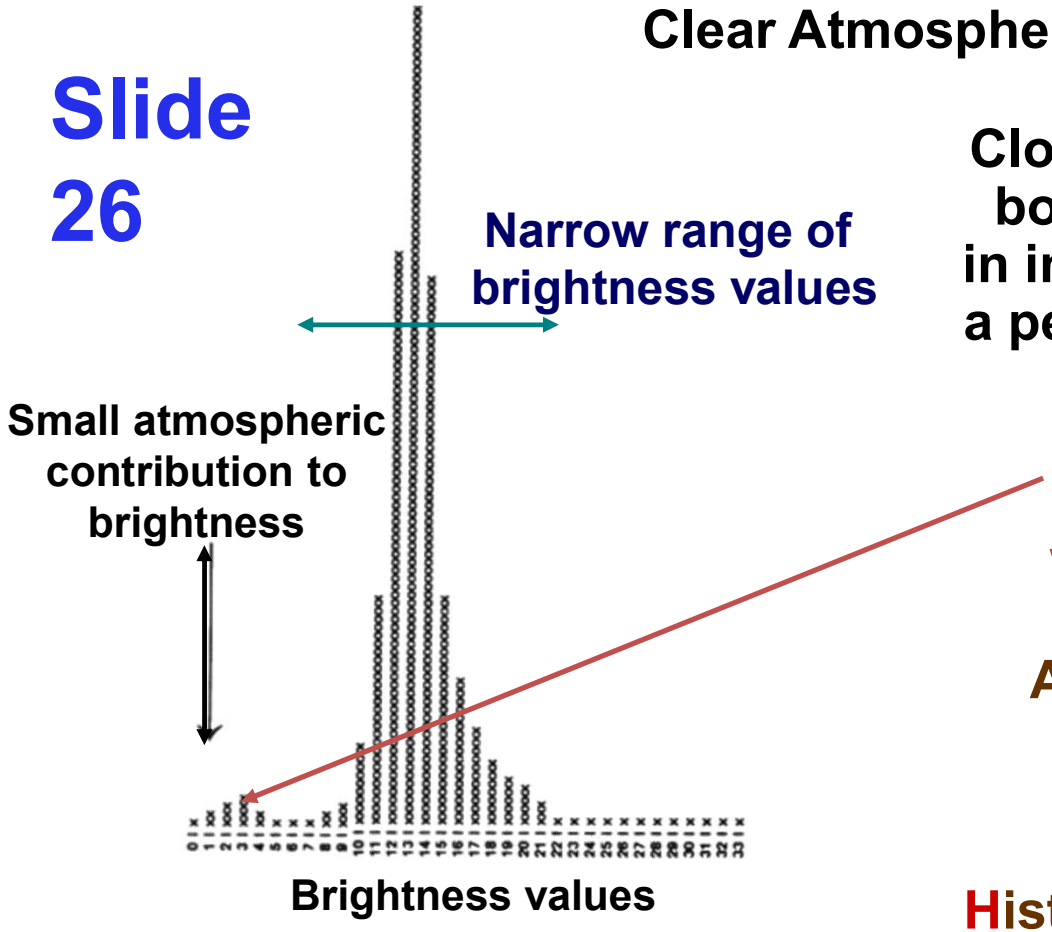
Need for Atmospheric Correction

- Atmospheric Correction
  - Radiation affected by the absorption and scattering in the atmosphere
  - Water Absorption features at 820nm, 940nm, 1135nm
  - Scattering:- Rayleigh Scattering and Mie scattering

Simple Atmospheric Corrections – Histogram Adjustment

Slide 26

Clear Atmosphere



Cloud shadowed region and water bodies have very low reflectance in infrared bands. This should give a peak near zero on the histogram.

The shifted peak is due to the low reflectance regions with atmospheric scattering.

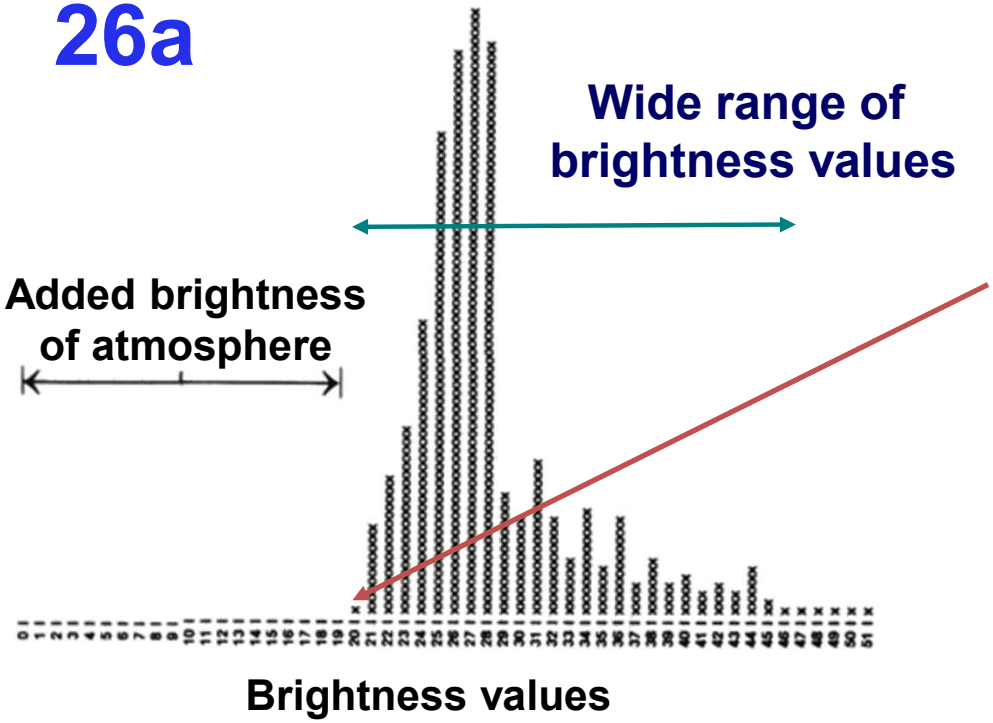
A correction can be obtain by removing this value from all pixels.
This method is called the Histogram Minimum Method (HMM)

Campbell 10.4

Simple Atmospheric Corrections – Histogram Adjustment

Slide 26a

Hazy Atmosphere



In this case, the minimum value is higher, and the histogram shape has changed

Darkest values far from zero

Campbell 10.4

Atmospheric Correction Models

Physical models simulate the physical process of scattering at the level of individual particles and molecules

Slide 27

Absorption by gases
scattering by aerosols

LOWTRAN 7
MODTRAN
CAM5S, 6S

Complex models that need many meteorological data as input.
The data may not always be available

Campbell 10.4

Atmospheric Correction Models

Second Simulation of the Satellite Signal in the Solar Spectrum: 6S

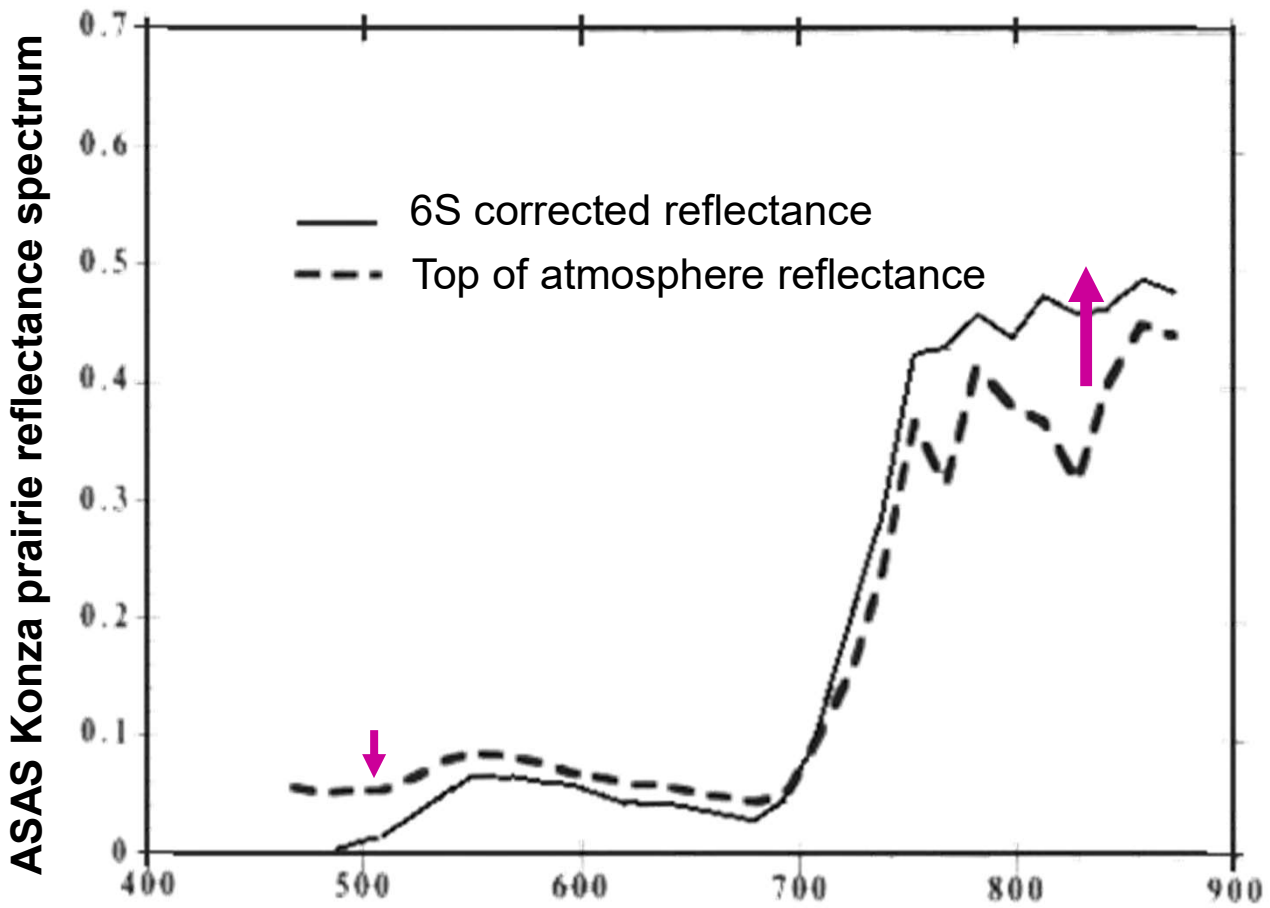
Slide 27a

Input file example (Saskatchewan study site; Landsat imagery):

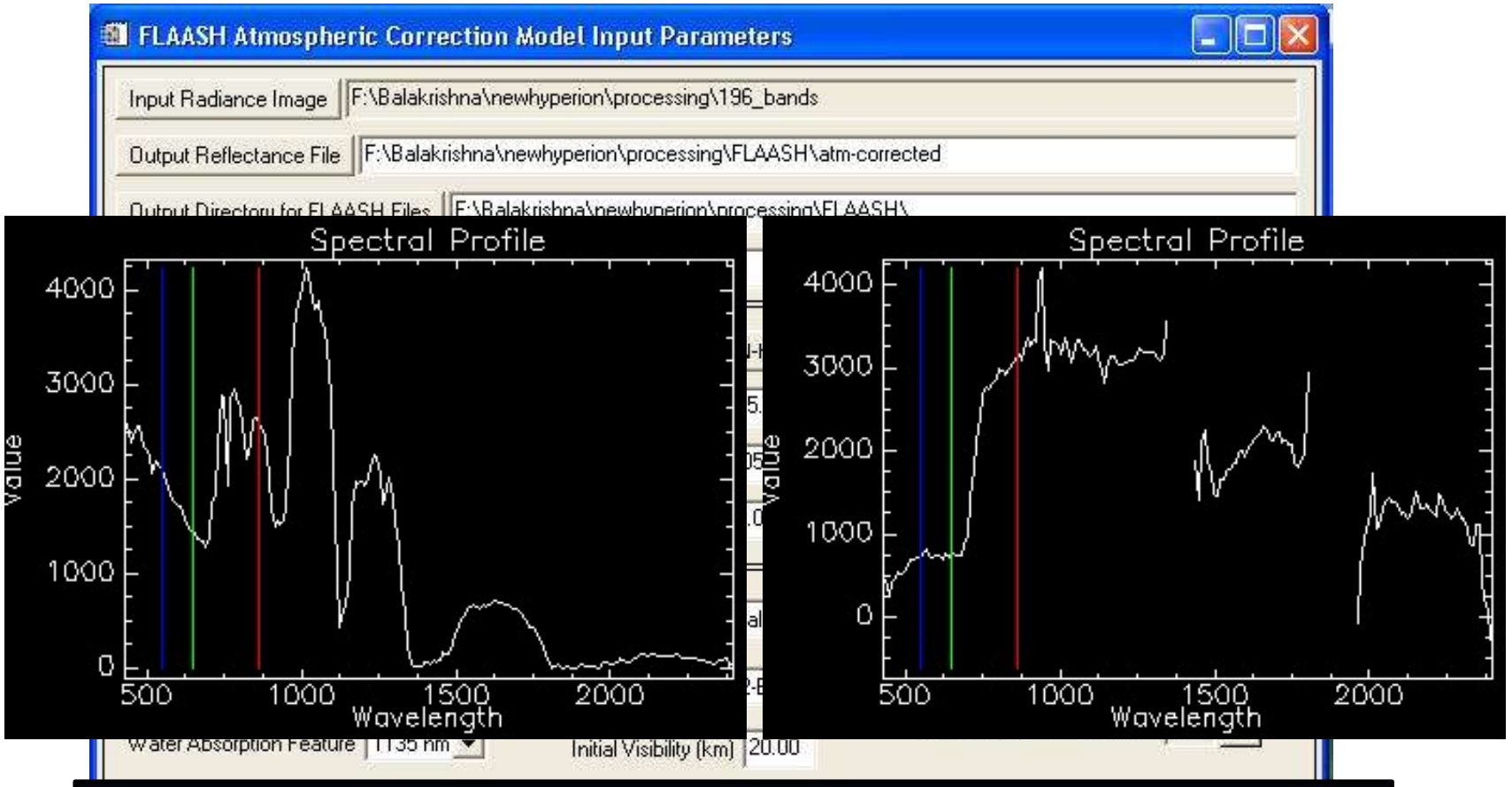
7	(landsat TM)
9 02 17.14 -105.22 53.85	(month,day,hour,long,lat)
2	(mid lat summer)
1	(continental)
30	(visibility, km)
-0.59	(TARGET ALTITUDE IN KM)
-1000	(SATELLITE CASE)
29	(Landsat band 1)
0	(HOMOGENEOUS CASE)
0	(NO BRDF effect)
1	(uniform target = vegetation)
-2.0	(no atm. correction)

Atmospheric Correction Models

Slide 27b

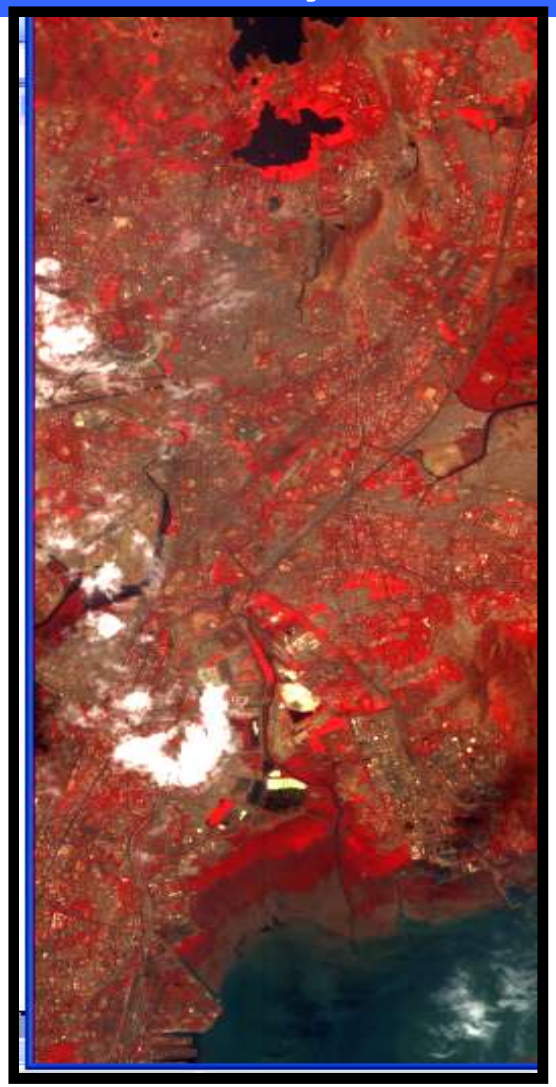


Vermote *et al.*, 1997

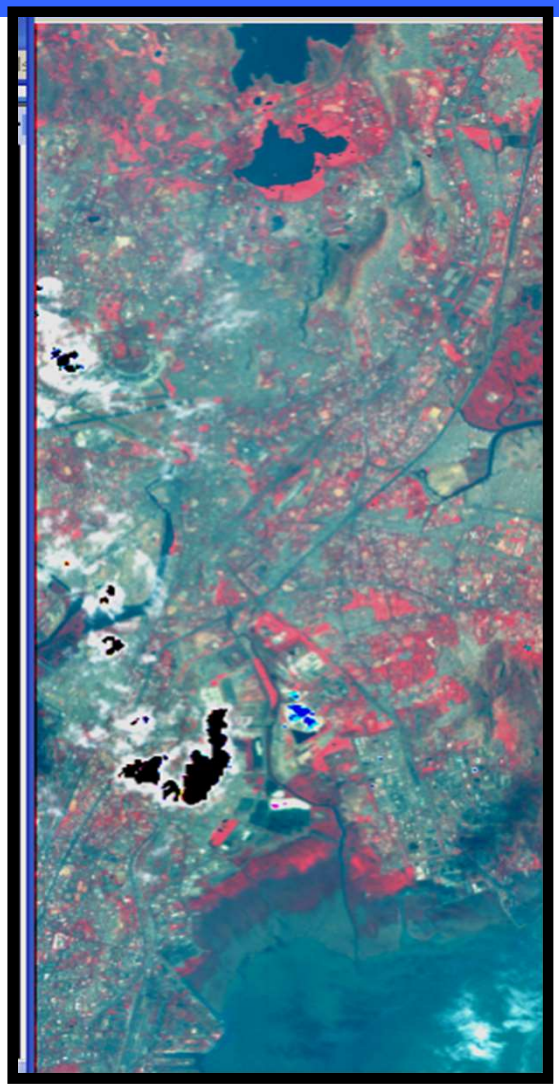


Z profile spectrum shown over vegetation area before atmospheric correction and after correction

Illustration of FLAASH



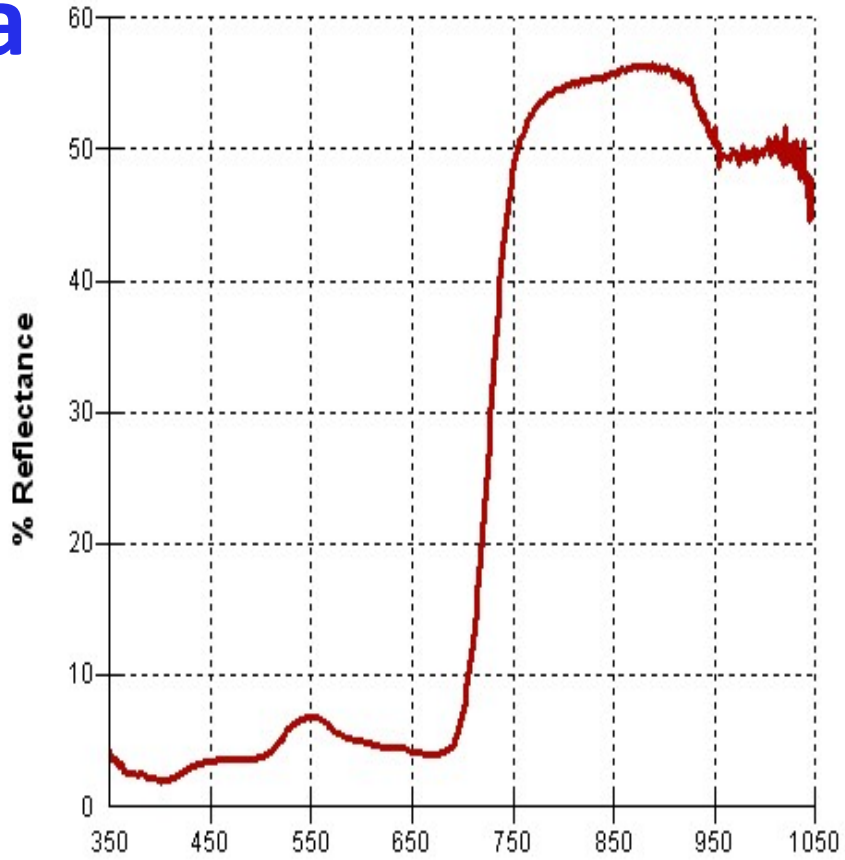
Example of FLAASH



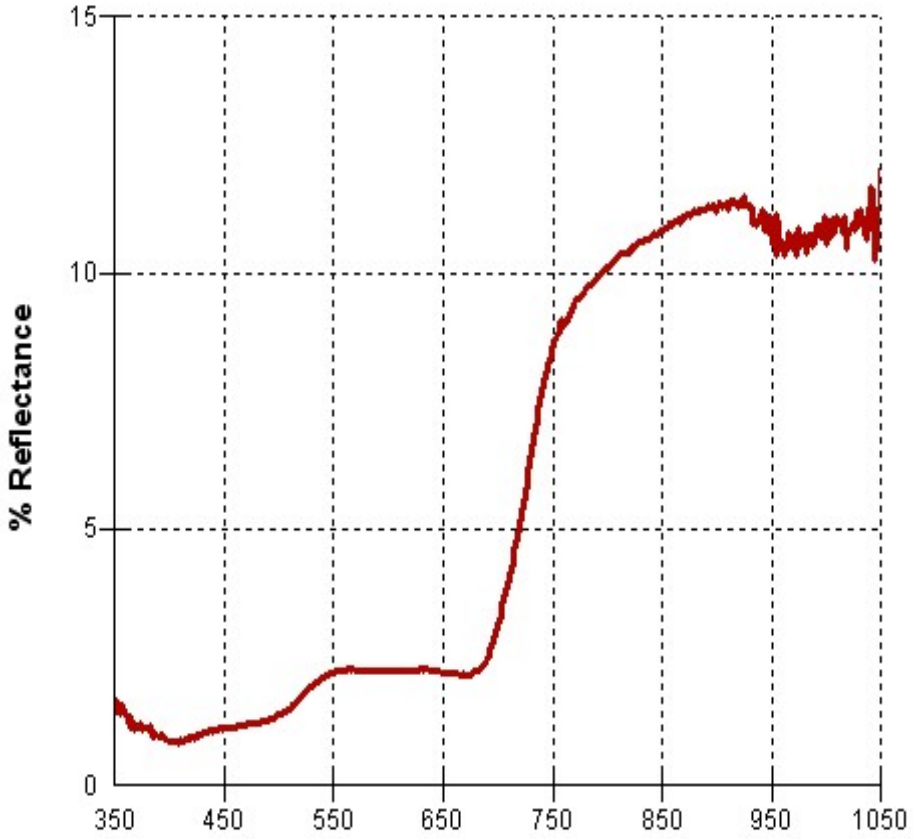
Field Collection Data

- Spectroradiometer used : GER 1500
- Spectral Range : 350 nm – 1050 nm with 512 spectral bands
- Signature of Crops

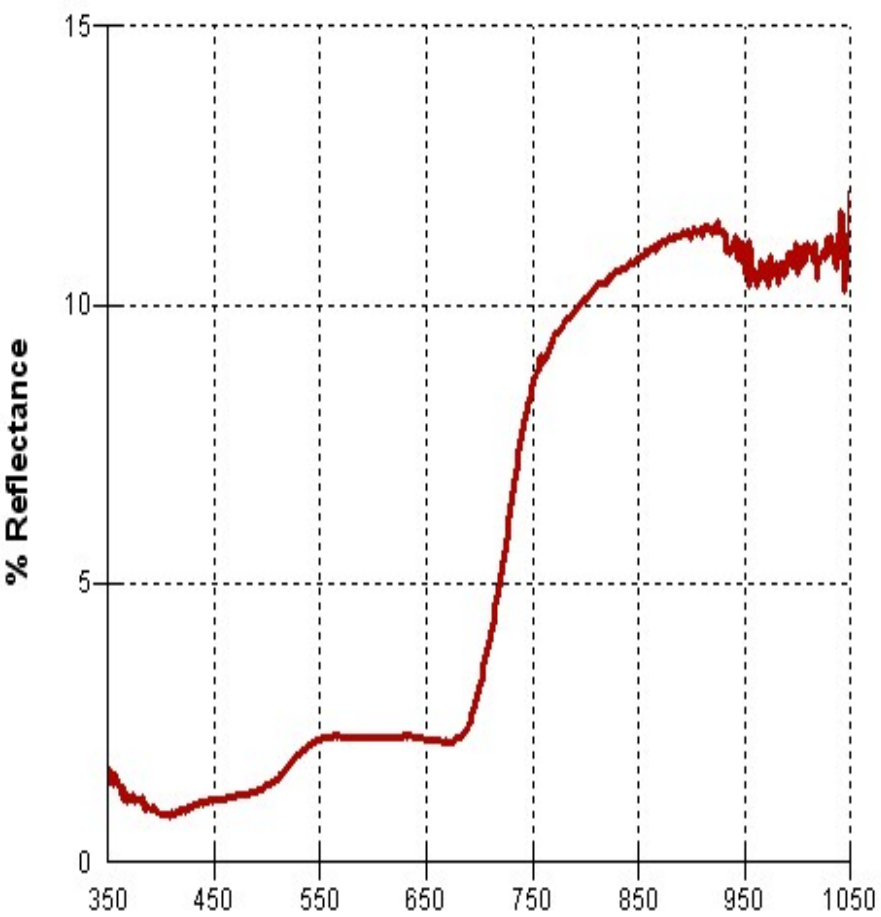
Signature of Chana



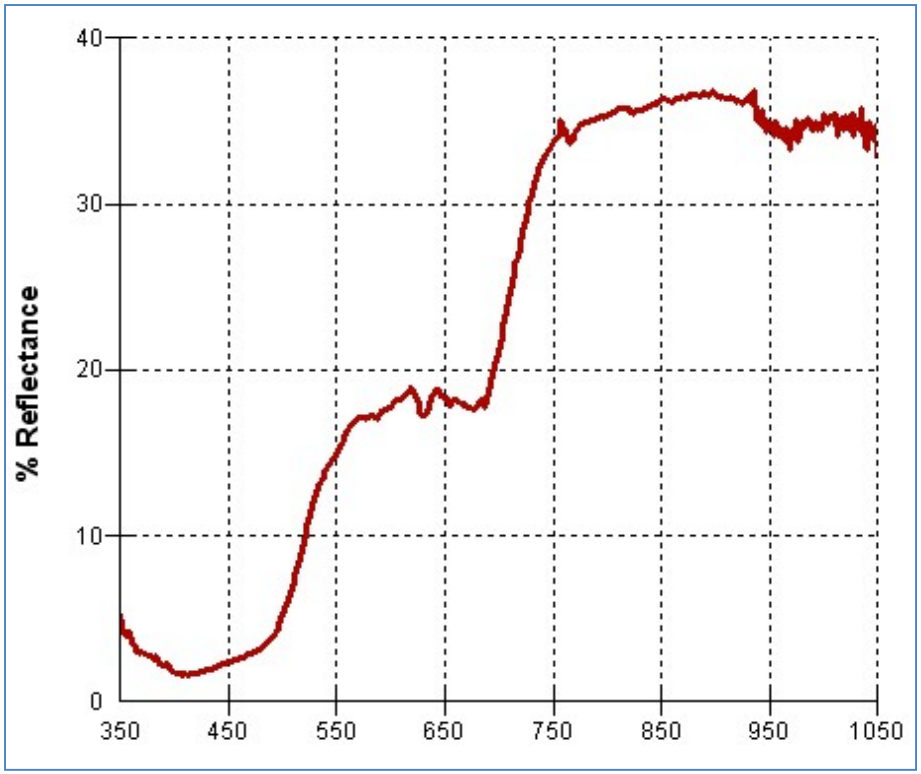
Signature of Wheat



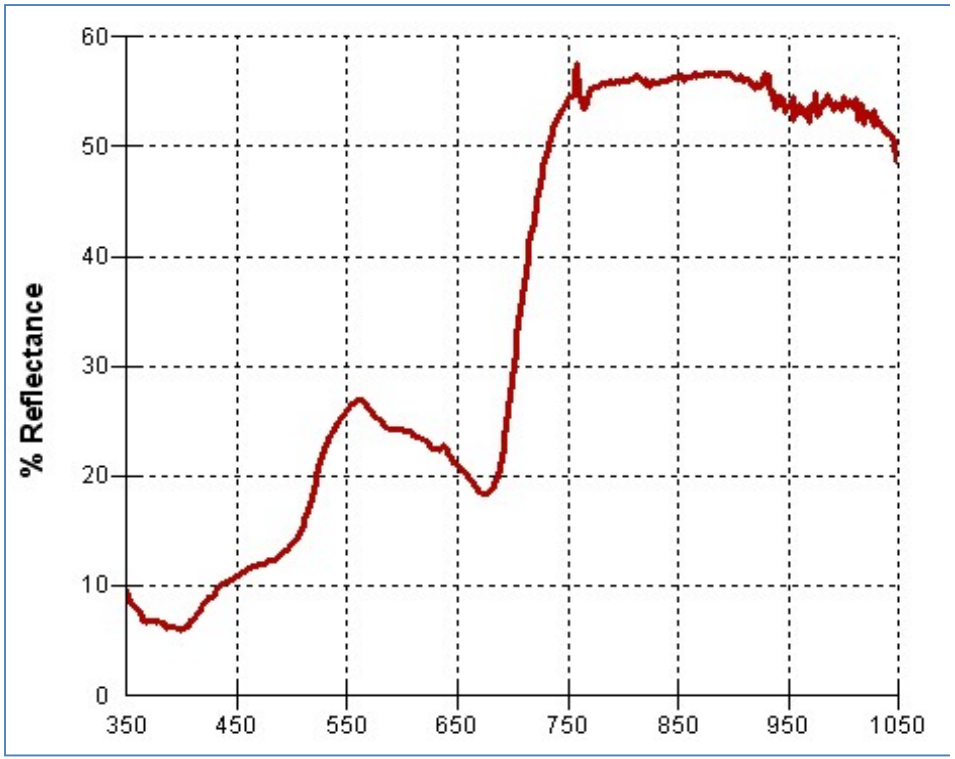
Signature of Bajra



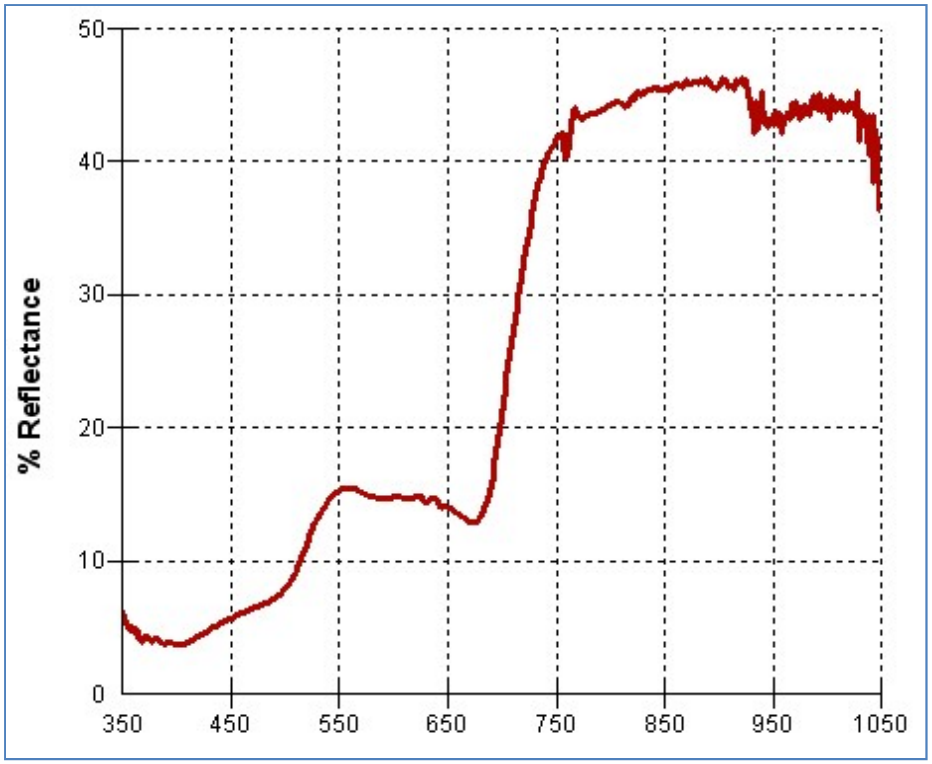
Signature of Kardi



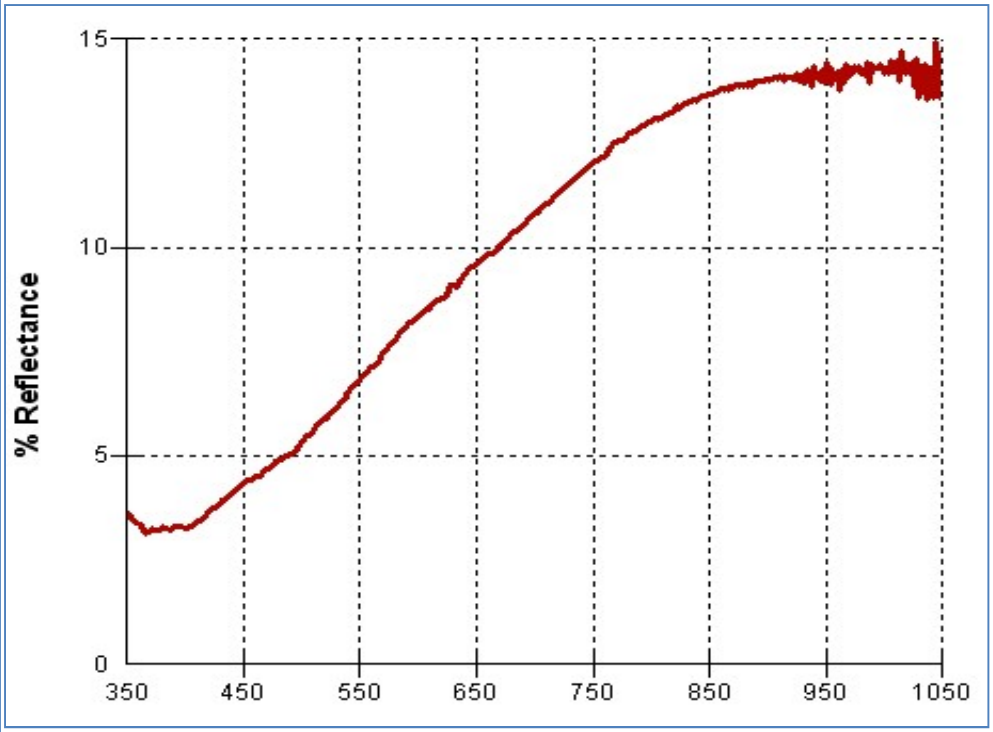
Signature of Jowar



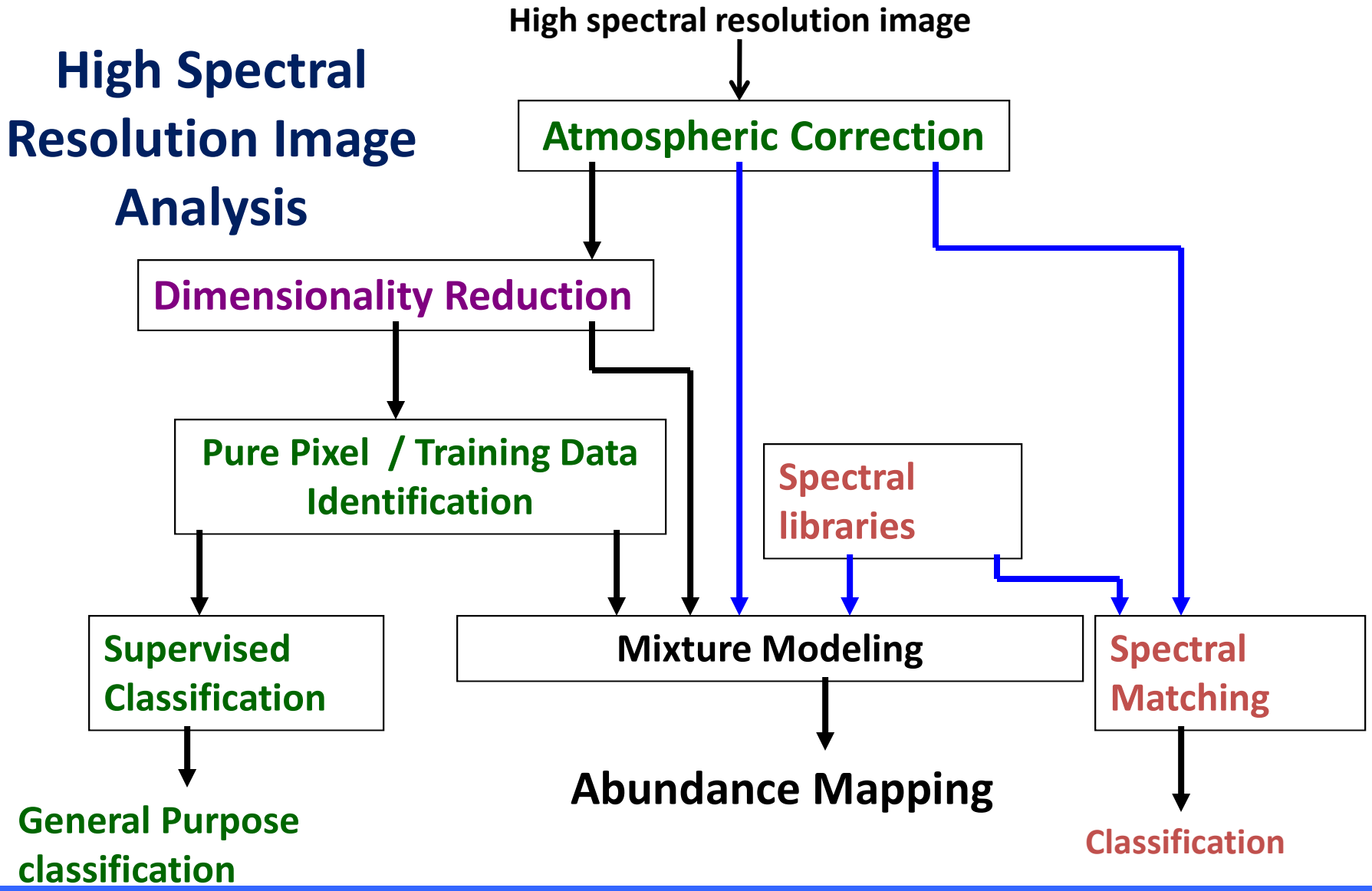
Signature of Tur Dal



Harvested Land



High Spectral Resolution Image Analysis



Dimensionality Reduction Algorithms

- Issues with high dimensionality
  - The class parameters are estimated by statistics computed using training samples
  - For high dimensional data the number of unknowns in the statistics are very large
  - Demand for larger training sample size grows with dimensionality of the input feature set
  - Difficult with coarse resolution hyperspectral imagery

Dimensionality Reduction Algorithms

- Issues with high dimensionality
  - For n-band image, the covariance matrix contains
  - n.(n+1)/2 unique elements to be determined
  - Need large number of samples for large n.
  - Thumb rules suggest about 10p to 30p samples for p dimensional data

Dimensionality Reduction Algorithms

- Hughes Phenomenon

The minimum achievable error in a classification problem is the Bayes error.

A decision rule that assigns a sample to the class that has the ***M***aximum ***A*** *posteriori* ***P***robability **(MAP** classifier) achieves the Bayes error. In order to design such a classifier, knowledge of the posterior probabilities and thus, the class conditional probability density functions is required.

Hughes Phenomenon

If knowledge of class conditional probability density functions is available then by increasing the dimensionality one would expect to enhance the performance.

In other words, the Bayes error is a decreasing function of the dimensionality of the data. In practice, however, class conditional probability density functions (pdfs) need to be estimated from a set of training samples.

When these estimates are used in place of the true values of the pdfs the resulting decision rule is sub-optimal and hence has a higher probability of error.

Hughes Phenomenon

The expected value of the probability of error, taken over all training sample sets of a particular size is, therefore, larger than the Bayes error.

When a new feature is added to the data the Bayes error decreases, but at the same time the bias of the classification error increases.

This increase is due to the fact that more parameters need to be estimated from the same number of samples.

If the increase in the bias of the classification error is more than the decrease in the Bayes error, then the use of the additional features degrades the performance of the decision rule.

This is called the Hughes phenomenon.

Dimensionality Reduction Algorithms

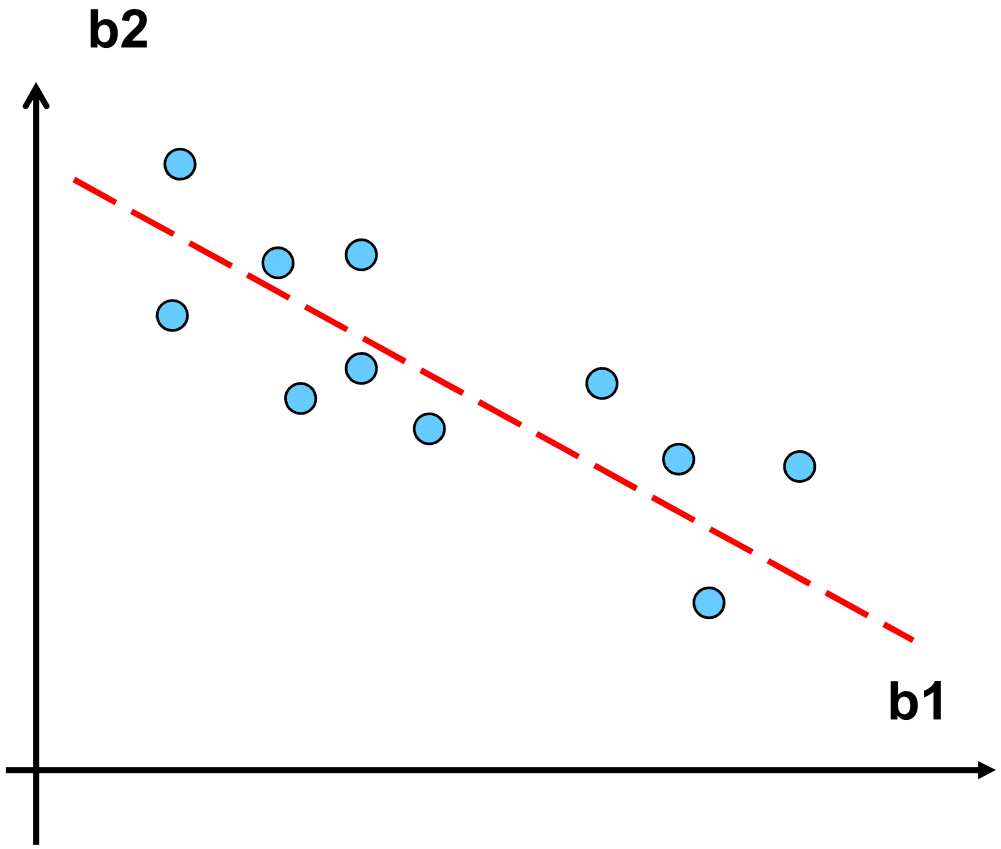
- Principal Component Analysis
- Maximum Noise Fraction Analysis
- Independent Component Analysis
- Genetic Algorithm Based Search
- …

Principal Component Transform

- Highlights the ***redundancy*** in the data sets due to similar response in *some* of the wavelengths
- Original bands variables represented along different coordinate axes, redundancy implies variables are ***correlated***, not independent
- Gray level in a band at a pixel can be predicted from the knowledge of the pixel gray level in other bands

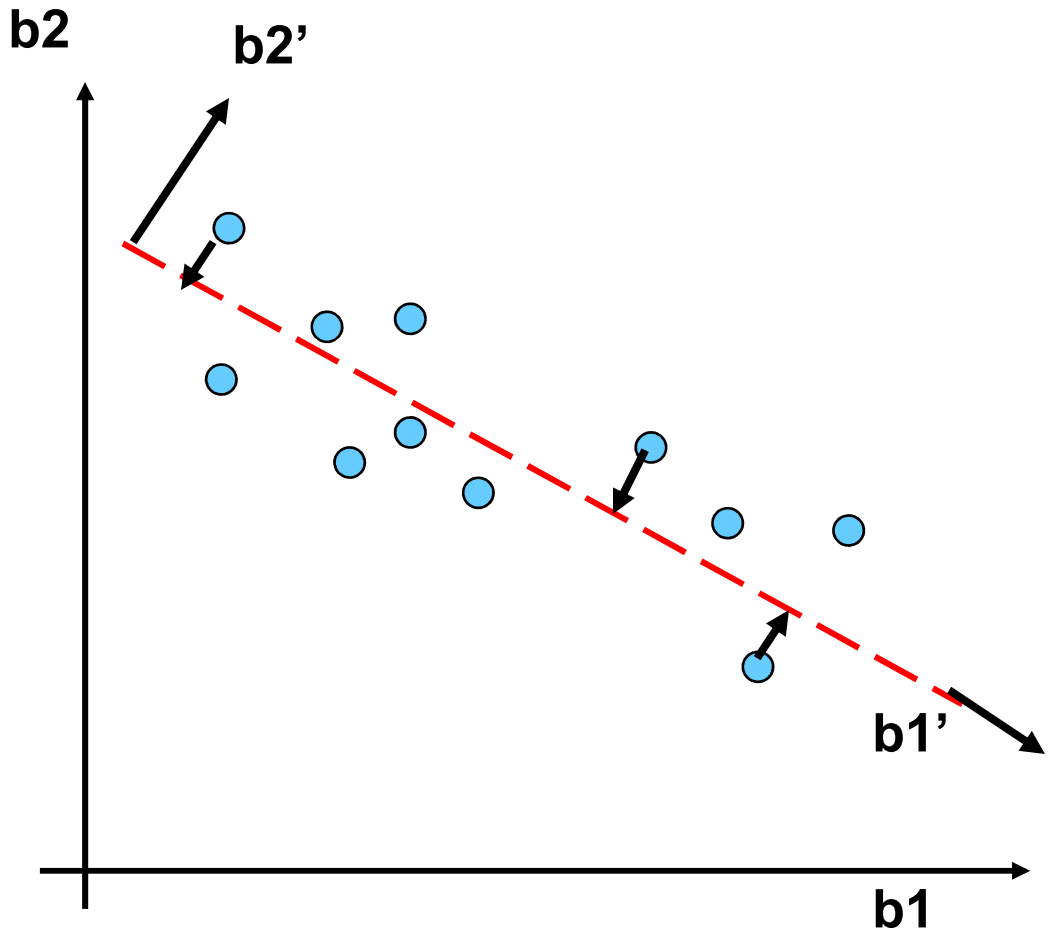
Example of Redundancy in Data

- **Example:** Highly correlated data
- Values along band b1 leads to knowledge along band b2 of the data element
- Linear variation (nearly) between b1 and b2
- Often true in case of visible bands



Example of Redundancy in Data

- Points projected onto the line ➔a small error in the position of the point.
- Points represented by only one coordinate b1' ➔ half data reduced
- For highly correlated data, this error will be minimal



Decorrelating Multispectral Remotely Sensed Data

- How do we identify the **optimum axes** along which the remotely sensed data should be projected so that the transformed data would be uncorrelated?
- What should be the way to **rank the new axes** so that we can discard the least important dimensions of the transformed data?
- **Invertibility** of the transformation?

Useful band statistics

Mean → $\dfrac{\sum_{i=1}^{M}\sum_{j=1}^{N} g_{ij}^{k}}{M \ . N}$

Variance → $\left[\dfrac{\sum_{i=1}^{M}\sum_{j=1}^{N}(g_{ij}^{k}-\mu_k)^2}{M.N}\right]$

$\dfrac{\sum_{i=1}^{M}\sum_{j=1}^{N}(g_{ij}^{k}-\mu_k)(g_{ij}^{l}-\mu_l)}{M.N}$ **← Covariance**

Covariance Matrix

- $C = \{C_{kl} \mid k = 1, \ldots, K, l = 1, \ldots, K\}$
- K is the number of bands in which the multispectral dataset was generated
- C is a symmetric matrix
- $C_{kl} = C_{lk}$
- Diagonal elements of C are the intra-band variances
- Off-diagonal elements are the inter-band covariances

Relation between correlation and covariance

- Correlation R_{kl} = $$\frac{\sum_{i=1}^{M}\sum_{j=1}^{N} g_{ij}^{k}\, g_{ij}^{l}}{M \,.\, N}$$
- It can be shown that $R_{kl} = C_{kl} + \mu_k \mu_l$
- For data with zero-mean, correlation and co-variance will be equal

Principal Component Transformation

Problem to solve:

- Find a transformation to be applied to the input multispectral image such that the covariance matrix of the result is reduced to a diagonal matrix
- Further, we should find an axis $\underline{\boldsymbol{v}}$ such that the variance of the projected coordinates ($z_k = \underline{\boldsymbol{v}}_{\underline{k}}^t \underline{\mathbf{x}}$) is maximum.

Solution

Given the transformed vector

$$z_k = \underline{\boldsymbol{v}_k}^t \underline{\mathbf{x}}$$

The variance $\sigma_z^2 =$ $$\frac{\sum_{i=1}^{M}\sum_{j=1}^{N} v^t (x_{ij} - \mu_k)(x_{ij} - \mu_l)^t v}{M.N}$$

This simplifies to $\sigma_z^2 = \underline{\boldsymbol{v}}^t C \underline{v}$

C, the covariance matrix is a positive, semi-definite, real symmetric matrix.

Finding vector $\underline{v}$

- To maximize the projected variance σ_z^2, find a $\underline{\boldsymbol{v}}$ such that $\underline{\boldsymbol{v}}^t C \underline{v}$ is maximum, subject to the constraint $\underline{\boldsymbol{v}}^t \underline{v} = 1$. Combining the maximization function with the constraint, we can write
- $\underline{\boldsymbol{v}}^t C \underline{v} - \lambda(\underline{\boldsymbol{v}}^t \underline{v} - 1)$ = maximum
- Differentiating w.r.t. $\underline{\boldsymbol{v}}$,

$$\frac{\partial}{\partial v}\left[\mathbf{v}^t C\mathbf{v} - \lambda(\mathbf{v}^t\mathbf{v} - 1)\right] = 0$$

Finding v

The derivative results in

$$C\underline{\boldsymbol{v}} = \lambda\underline{\boldsymbol{v}} \quad \text{(Verify!)}$$

Therefore, $\underline{\boldsymbol{v}}$ is an *eigenvector of* C

$$\underline{\boldsymbol{v}}^t C\underline{v} = \underline{\boldsymbol{v}}^t(\lambda\underline{\boldsymbol{v}}) = \lambda\underline{\boldsymbol{v}}^t\underline{\boldsymbol{v}} = \lambda$$

This implies that $\underline{\boldsymbol{v}}$ is the eigenvector of C with the largest eigenvalue

Therefore all the eigenvectors with decreasing eigenvalues lead to axes with decreasing variance along them.

Sample Eigenvectors and Eigenvalues

Covariance Matrix

34.89	55.62	52.87	22.71
55.62	105.95	99.58	43.33
52.87	99.58	104.02	45.80
22.71	43.33	45.80	21.35

Eigenvalues

253.44	7.91	3.96	0.89

Eigenvectors

0.34	−0.61	0.71	−0.06
0.64	−0.40	−0.65	−0.06
0.63	0.57	0.22	0.48
0.28	0.38	0.11	−0.88

Sample Eigenvectors and Eigenvalues

$$\Sigma_x = \begin{bmatrix} 874.98 & 550.56 & 698.00 & 335.54 & 858.15 & 551.21 \\ 550.56 & 363.82 & 454.79 & 230.30 & 558.88 & 358.38 \\ 689.00 & 454.79 & 580.63 & 288.11 & 747.97 & 471.72 \\ 335.54 & 230.30 & 288.11 & 722.46 & 742.35 & 387.61 \\ 858.15 & 558.88 & 747.97 & 742.35 & 1544.70 & 871.29 \\ 551.21 & 358.38 & 471.72 & 387.61 & 871.29 & 514.18 \end{bmatrix}$$

eigenvalues	3727.35	613.34	226.14	23.52	8.16	2.25
eigenvectors	first	second	third	fourth	fifth	sixth
	0.433	0.485	−0.307	−0.684	−0.089	0.088
	0.282	0.294	−0.218	0.369	0.094	−0.801
	0.364	0.347	−0.127	0.627	−0.153	0.561
	0.303	−0.673	-0.671	0.018	0.042	0.056
	0.615	−0.322	0.562	−0.052	−0.429	−0.129
	0.362	−0.047	0.275	−0.026	0.880	0.127

Transformation

New component value = dot product of eigenvector and pixel vector

(i,j) ➔ pixel position

n eigenvectors for n principal components

1st principal component ➔ dot product of pixel vector with eigenvector corresponding to largest eigenvalue

Principal Components

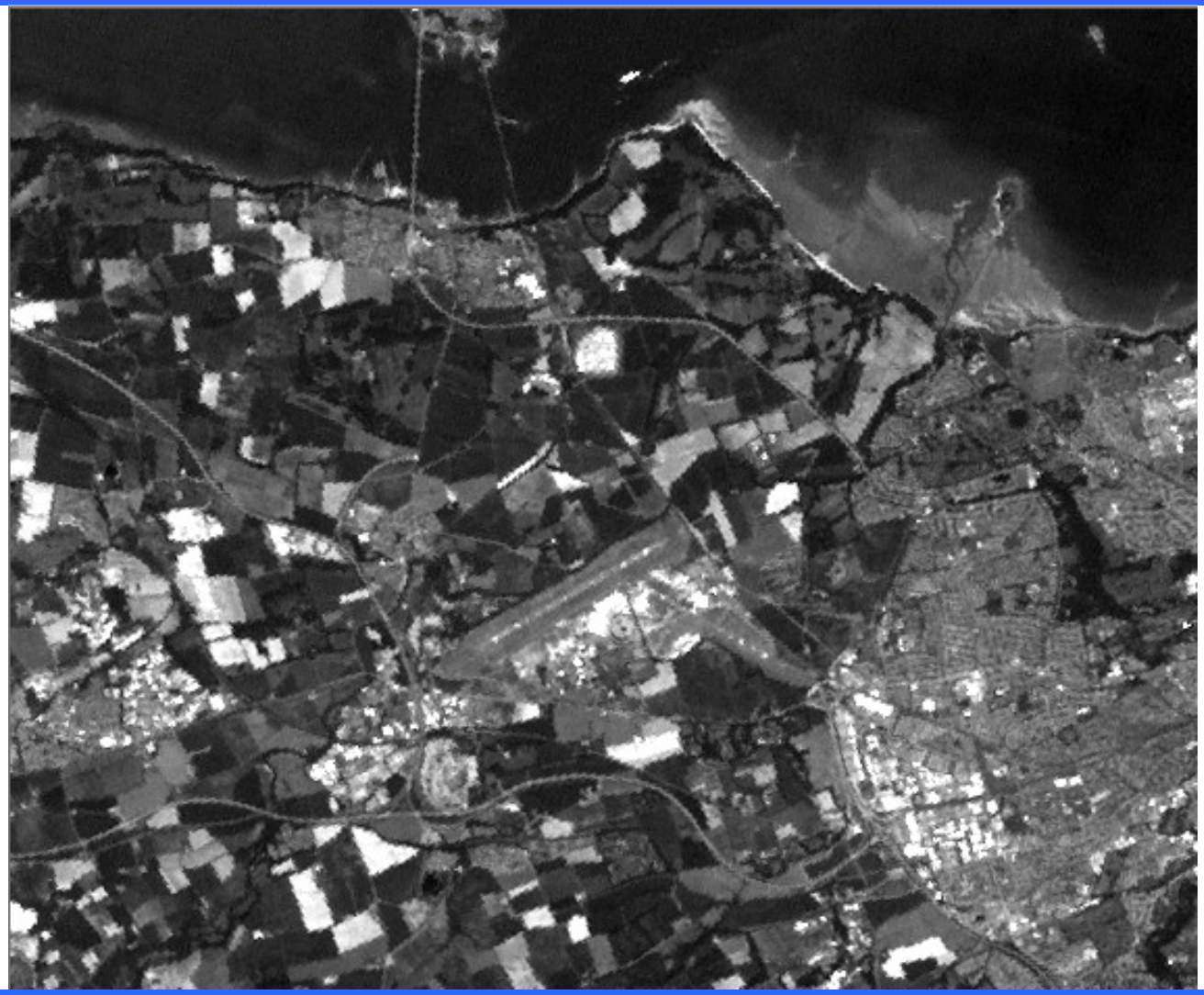
For n input bands, n principal components are computed

The utility of the principal components gradually decreases from 1st towards the last

e.g., For Landsat TM, last three PCs are generally of very little value

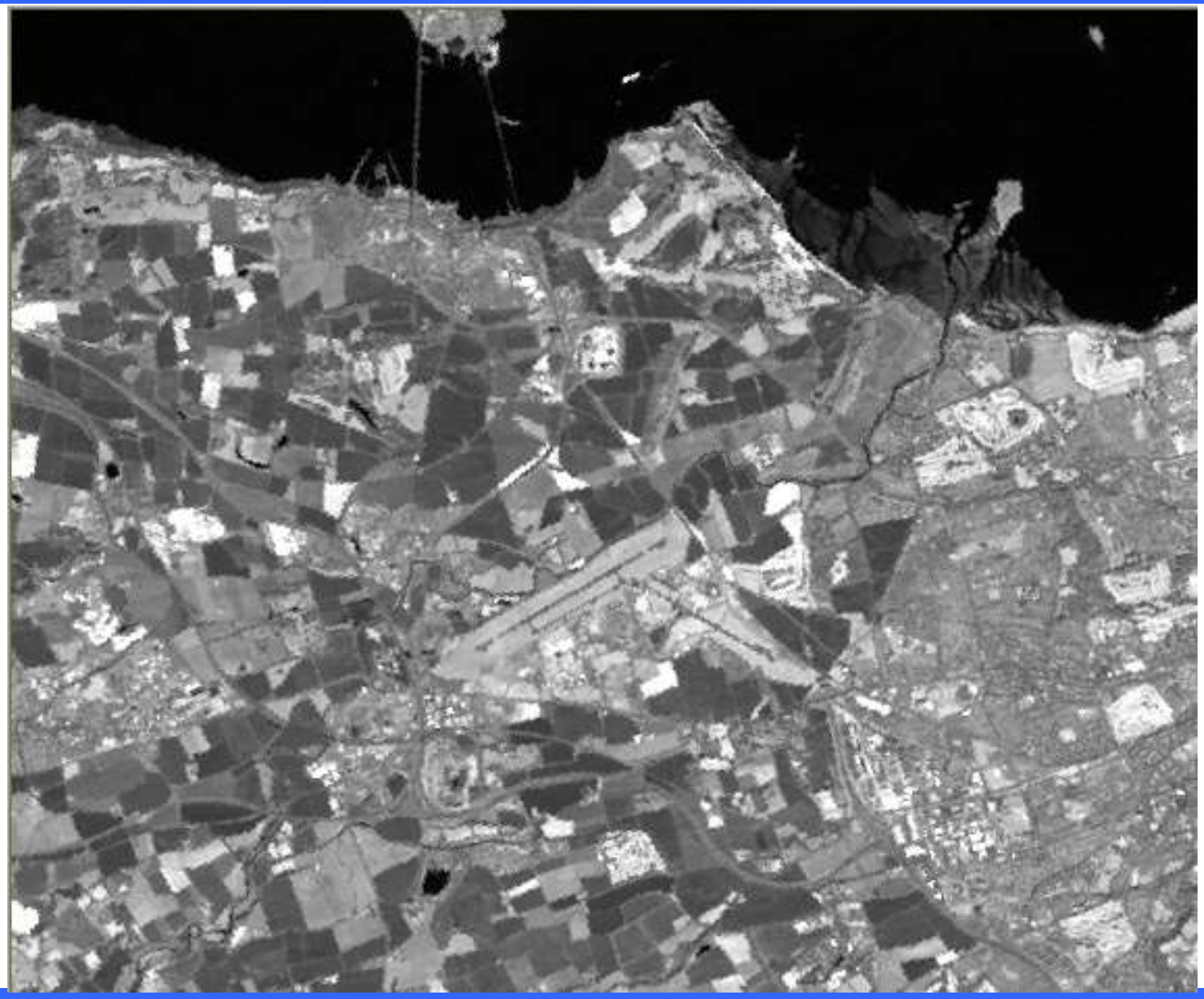
Band 1
(Blue)

Band 2
(Green)

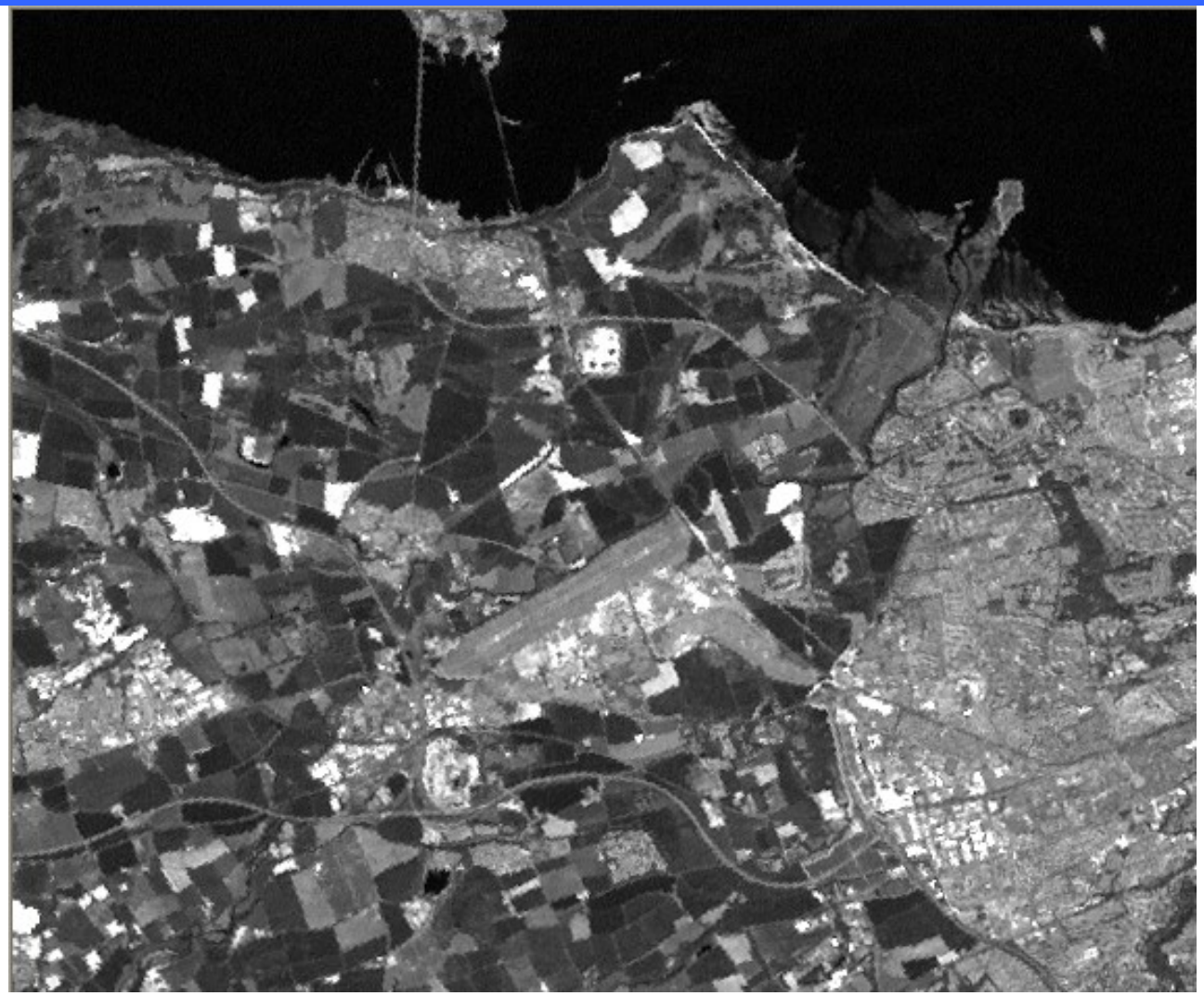


Band 3
(Red)

Band 4
(NIR)



Band 5
(SWIR)

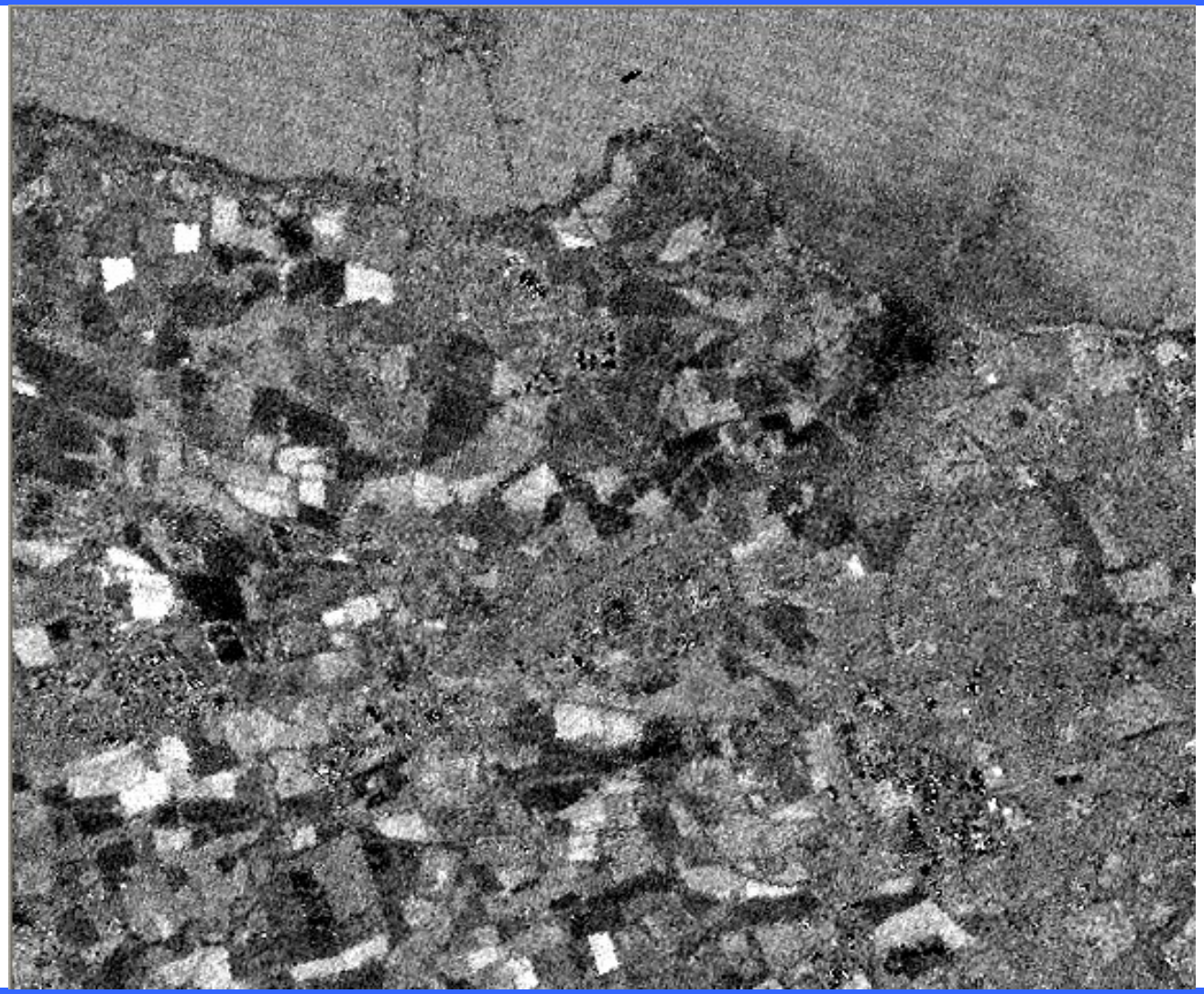


Band 7
(SWIR)

PC1

PC2

PC3



PC6

Input Image FCC

Decorrelation Stretch

Limitation of Principal Component Analysis

- Simple principal component analysis affected by noise in data
- Noise disturbs the interband correlation
- Structure of covariance matrix does not reflect the actual redundancy in the data
- Noise problem is severe due to small signal strength available at any detector due to very narrow spectral bands

Maximum Noise Fraction Analysis

- Alternative proposed is the Maximum Noise Fraction method
- Decorrelates the noise, reduces it to zero mean and unit standard deviation
- On the noise adjusted data, standard principal component analysis is applied
- Results superior to simple principal component analysis

Maximum Noise Fraction Analysis

- Ranks the transformed components in increasing order of noise variance.
- Linear transformation which is essentially two cascaded Principal Components Transforms.
- Let Σ_x and Σ_y be the covariances of original and transformed coordinates .
  - Then the noise fraction which has to be minimized is

$$\Sigma_x = \Sigma_x^n + \Sigma_x^s$$

$$\gamma = \frac{d^t \Sigma_x^n d}{d^t \Sigma_x d}$$

See full derivation in Richards and Jia's book 4th ed., pp. 154-156

Maximum Noise Fraction Analysis

- Minimizing it we get $(\sum_x^n \sum_x^{-1} - \gamma I)d = 0$
- If the noise covariance is scaled to identity the above equation changes as $\left(\sum_x^{-1} - \gamma I\right)d = 0$
- If both sides above are multiplied by Σ_x the above can be rewritten as $\left(\sum_x - \nu I\right)d = 0$

$$V = \gamma^{-1}$$

- This is a standard equation for PCT.

Maximum Noise Fraction Analysis

- To estimate noise covariance first find

$$\Delta X(i,:) = X(i,:) - X(i+1,:)$$

The justification for this step is that normally adjacent pixels will have nearly same values, any difference is due to noise and then make the assumption

$$Cov(\Delta X) = Cov(N) = \Sigma_X^n$$

- $\sum_x^n$ can be diagonalized by standard similarity transformation

$$\Lambda = E^{-1} \sum_x^n E$$

Maximum Noise Fraction Analysis

- $\sum_x^n$ can be rescaled to identity matrix by the transformation

$$I = \Lambda^{(-1/2)t} E^{-1} \sum_x^n E \Lambda^{-1/2} \text{ where } \Lambda \text{ is noise variance}$$

- E is eigenvector matrix whose columns are eigenvectors of $\sum_x^n$
- If $F = E\Lambda^{-1/2}$ is defined then

$$y = F^T x$$

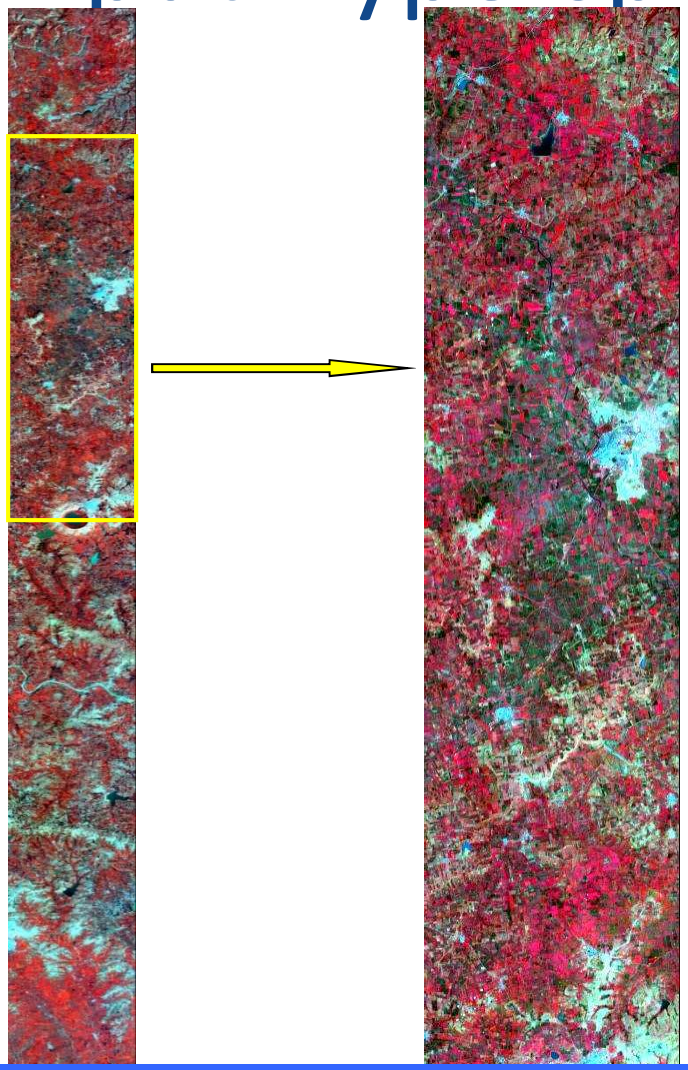
results in a new image with noise covariance matrix being identity matrix (noise is uncorrelated with the image)

- **Now the standard PCA can be applied on y.**

Maximum Noise Fraction Analysis

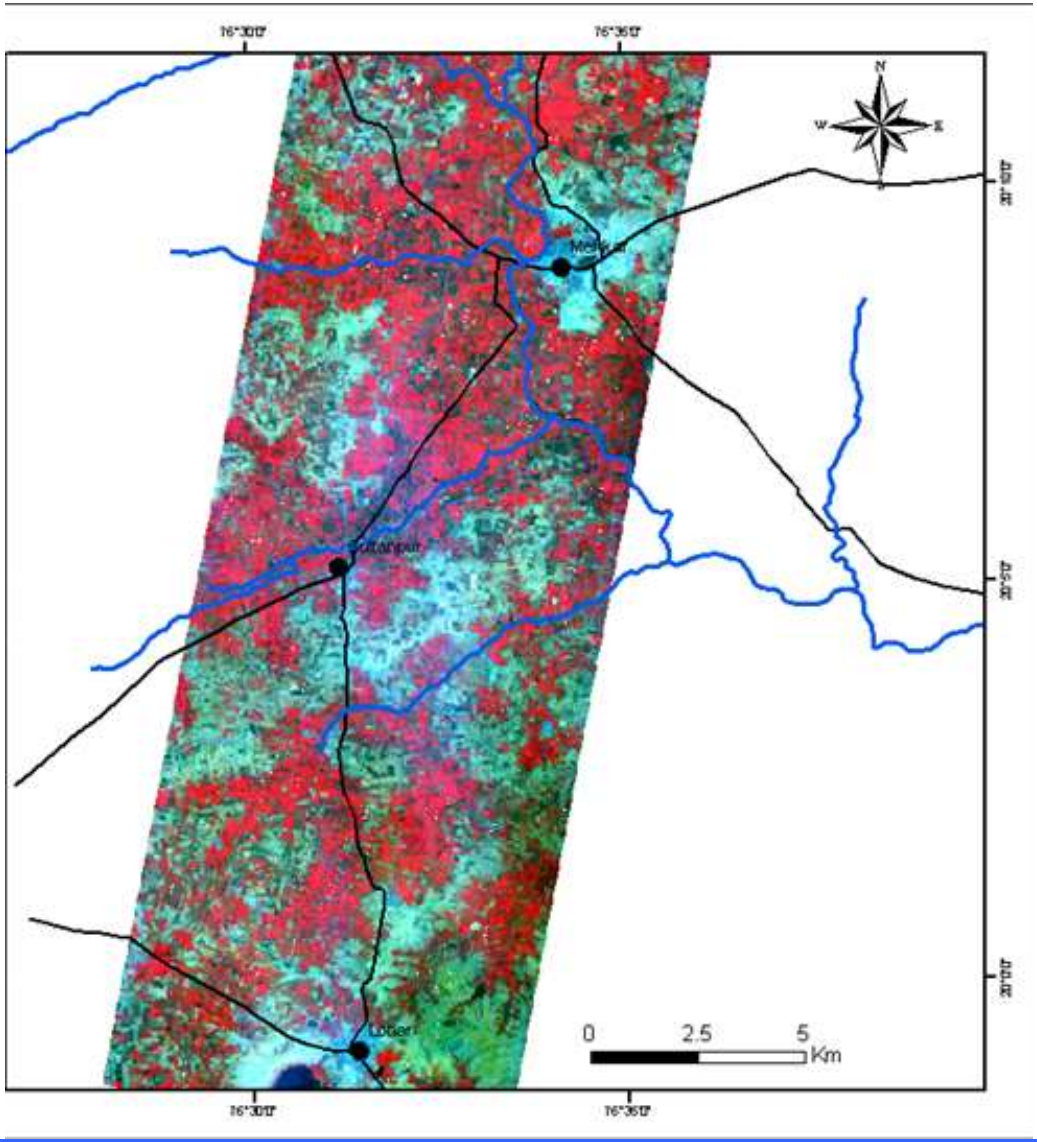
- Since noise variance is forced to be 1, all principal components corresponding to and close to eigenvalues of unit magnitude correspond to noise
- Principal components with large eigenvalues correspond to noise-free part of data

Input Hyperspectral Data



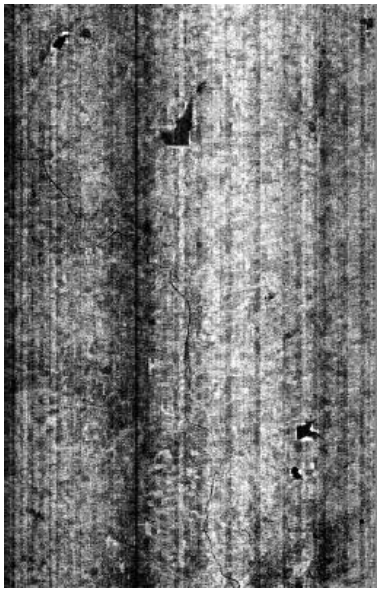
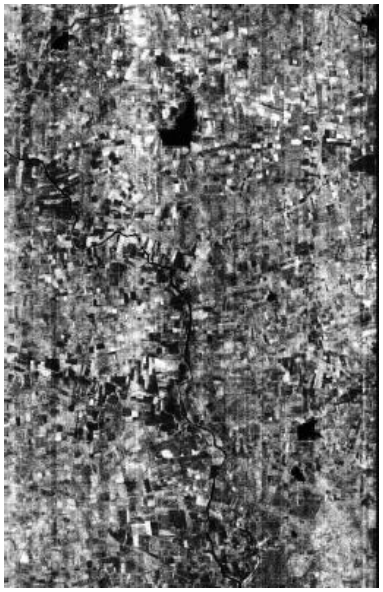
- Number of rows = 1400
- Number of columns = 256
- Number of bands = 242

Study Area



Maximum Noise Fraction Transform

Maximum Noise Fraction Transform


Definition of ICA Problem

– Consider n random variables x_i and n independent components s_i

$$x_1 = a_{11}s_1 + a_{12}s_2 + \ldots\ldots\ldots + a_{1n}s_n$$
$$x_2 = a_{21}s_1 + a_{22}s_2 + \ldots\ldots\ldots + a_{2n}s_n$$
$$x_3 = a_{31}s_1 + a_{32}s_2 + \ldots\ldots\ldots + a_{3n}s_n$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots$$
$$x_n = a_{n1}s_1 + a_{n2}s_2 + \ldots\ldots\ldots + a_{nn}s_n$$

– In matrix form $X = AS$, such that $S = A^{-1}X$, where A is the mixing matrix.

– We have to estimate a W such that $S = WX$

Challenge

- Sources S are unknown
- Mixing matrix A is unknown
- All that is available is recorded observations x.
- Need to estimate $A^{-1} = W$, and then Wx = estimate of S

Definition and Fundamental Properties

Statistical Independence:

- Two variables y1 and y2 are said to be statistically independent if information on the value of y1 does not give any information on the value of y2, and vice versa.

- If $p(y_1)$ and $p(y_2)$ are the marginal pdf's of y_1 and y_2 and $p(y_1, y_2)$ their joint probability density then we have

$$p(y_1) = \int p(y_1, y_2) dy_2 \qquad p(y_2) = \int p(y_1, y_2) dy_1$$

Definition and Fundamental Properties

- y_1 and y_2 are statistically independent if and only if their joint probability density is factorizable as

$$p(y_1, y_2) = p(y_1).p(y_2)$$

- Another fundamental property of independent random variables is : Given the random variables y_1 and y_2 ,

$$E\{h_1(y_1)h_2(y_2)\} = E\{h_1(y_1)\}.E\{h_2(y_2)\} \quad (1)$$

Definition and Fundamental Properties

- Uncorrelated variables are only partly independent.
  - Two variables are uncorrelated if

$$E\{(y_1 y_2)\} - E\{y_1\}.E\{y_2\} = 0$$

  - If the variables are independent, they are uncorrelated which is clear from eq. above by substituting $h_1(y_1)$ by y_1 and $h_2(y_2)$ by y_2.

Example

- Suppose (y_1, y_2) are discrete valued and have a distribution with probability 1/4 equal to any of the following values: (0,1), (0,-1), (1,0), (-1,0).
- Then y_1 and y_2 are uncorrelated, as can be simply calculated.
- On the other hand,

$$E\left\{y_1^2 y_2^2\right\}(here\ \ 0) \neq E\left\{y_1^2\right\}E\left\{y_2^2\right\}(here\ \ \frac{1}{4})$$

Therefore the condition for independence is violated and hence the variables cannot be independent

Method cannot work with Gaussian variables

- The fundamental restriction in ICA is that the independent components must be non-Gaussian for ICA to be possible.
- Assume that the mixing matrix is orthogonal and the independent components s_i are Gaussian.

$$p(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{s_1^2 + s_2^2}{2} \right)$$

- The figure shows that density is completely symmetric. Therefore, it does not contain any information on the directions of the columns of the mixing matrix A.

Multivariate distribution of two independent Gaussian variables

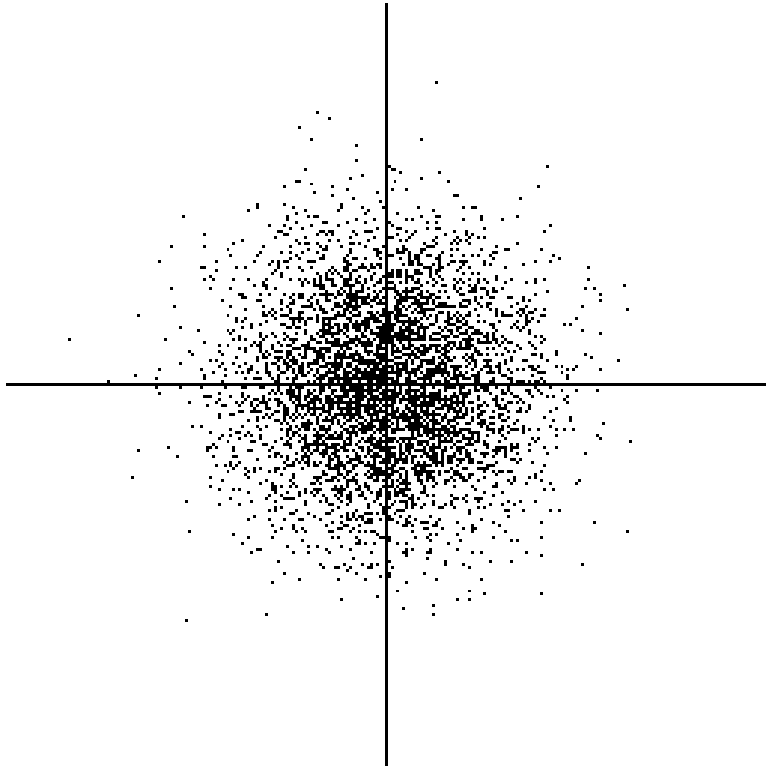
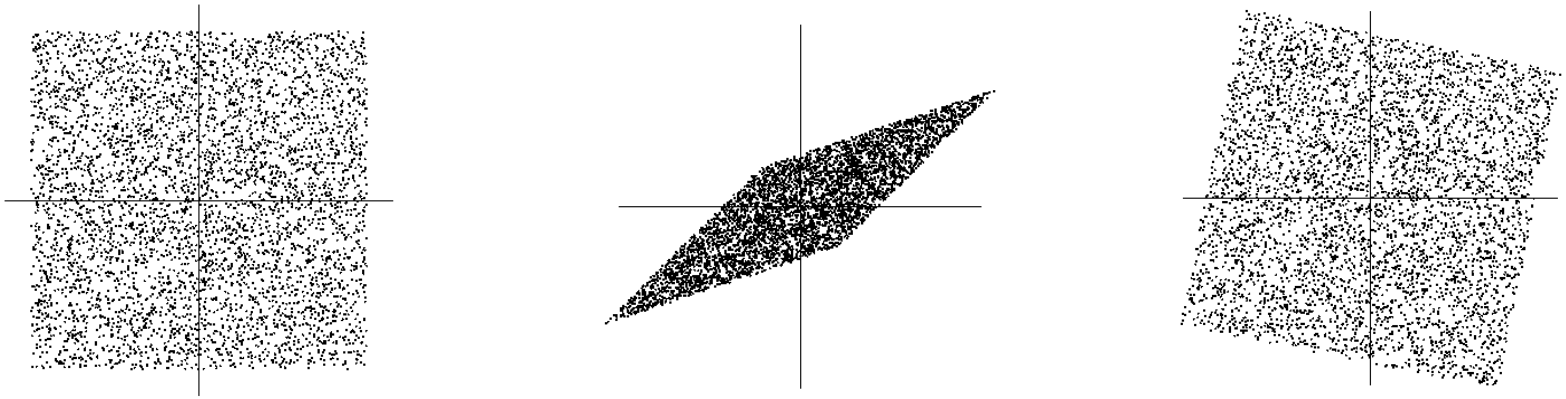


Illustration of Independent Components, Mixtures and Retrieved Components

Two ICs with uniform distributions:



Original variables, observed mixtures, whitened mixtures.
Cummulative frequency gaussian density: symmetric in all directions.

Nongaussianity is independence

- **Central limit theorem: A random variable generated by sum of n independent random variables will have a gaussian probability density function as n tends to infinity**
  - **A sum of even two independent random variables usually has a distribution that is closer to gaussian than any of the two original random variables.**
- Consider a linear combination of x, $y = w^T x$

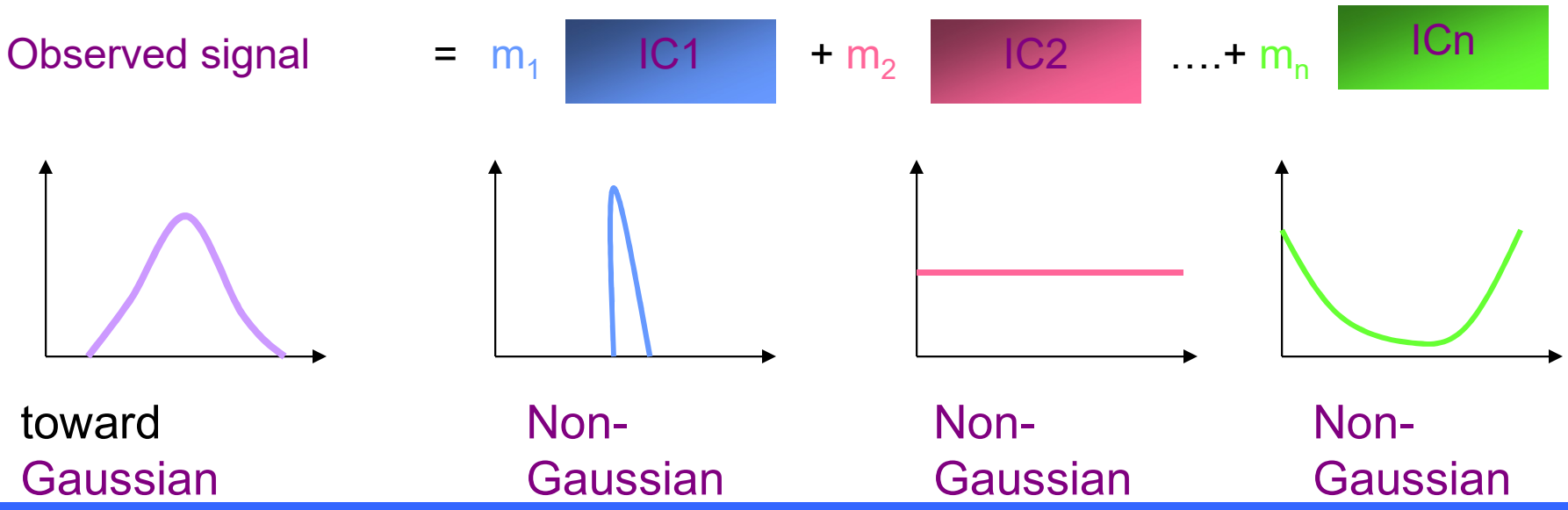
$$y = w^T As = (A^T w)^T s$$

$$= z^T s$$

Central limit theorem

- The distribution of a sum of independent random variables tends toward a Gaussian distribution

Source:
www.ym.edu.tw/~ytwu/ica.**ppt**



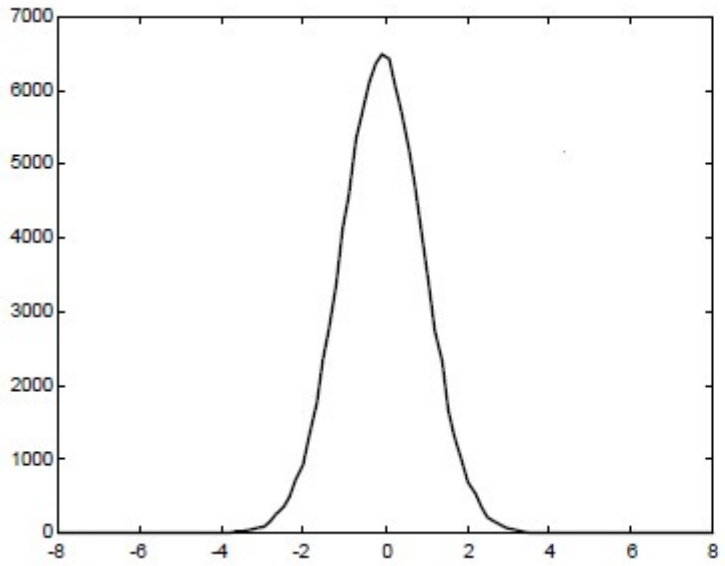
Non-Gaussianity is independence

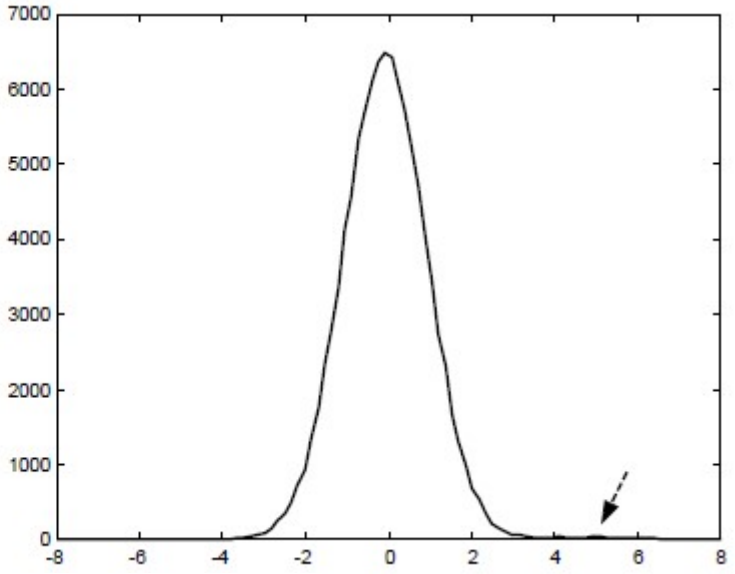
- This restriction provides an indirect measure of independence i.e. non-Gaussianity.
- Maximizing non-Gaussianity is maximizing independence of the components.
- Measures of non-Gaussianity
  - ***Kurtosis***
  - ***Negentrropy***

Kurtosis

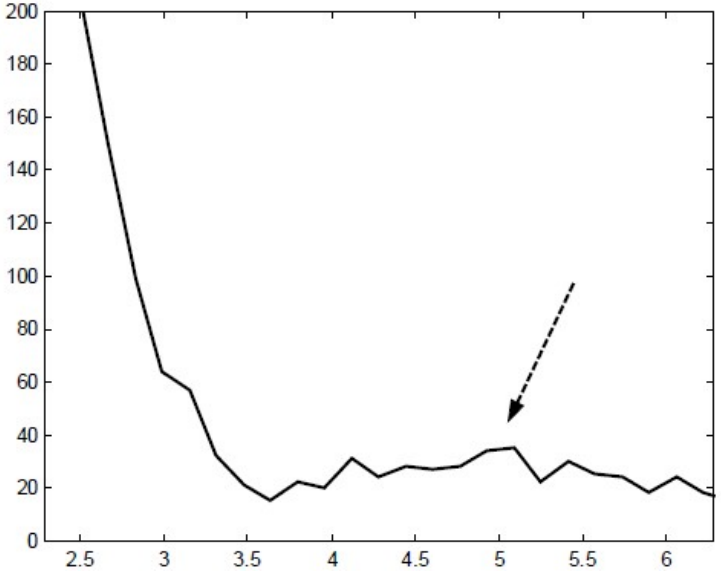
$$kurt\ (y) = E\{y^4\} - 3\left(E\{y^2\}\right)^2$$

- For Gaussian distribution, $kurt(y) = 0$
- And for $kurt(y) > 0$, the random variable is called supergaussian or leptokurtic.
- While for $kurt(y) < 0$, the random variable is called subgaussian or platykurtic.
- The main drawback in using kurtosis is that it is very sensitive to outliers. Its not a robust measure of nongaussianity.





Kurtosis 0.0022



Kurtosis 2.5639

Source: Stefan Robila's PhD thesis

Sensitivity of Kurtosis

Negentropy

- Negentropy is based on the information theoretic quantity of (differential) entropy.
- The entropy of a random variable is interpreted as the average information that the observation of the variable gives.

Negentropy

- The more "random", i.e. unpredictable and unstructured the variable is, the larger is its entropy.

$$H(Y) = -\sum_t P(Y_t)\log P(Y_t)$$

- For continuous variables the entropy is termed as differential entropy and is given by

$$H(y) = -\int f(y)\log f(y)dy$$

Negentropy

- A fundamental property is that a Gaussian variable has highest entropy among all the random variables of equal variance.

- A slightly modified version of differential entropy is 'negentropy' which is zero for Gaussian variable and is always nonnegative. Negentropy is defined as

$$J(y) = H(y_{gauss}) - H(y)$$

where y_{Gauss} is a Gaussian random variable of the same covariance matrix as y

Approximation of negentropy:

- The best approximation of negentropy for a random variable y is given by

$$J(y) = \left[E\{G_t(y)\} - E\{G_t(v)\}\right]^2$$

- Where v is a Gaussian variable of zero mean and unit variance, y is assumed to be of zero mean and unit variance, and the functions G_t are some nonlinear and nonquadratic functions.
- Generalization of (square of) kurtosis when $G(y) = y^4$

Approximation of negentropy:

- Following choices of G have proved very useful
  - $G(u) = (1/a_1)*\log \cosh a_1 u$
  - $G(u) = - \exp (-u^2/2)$

Mutual Information

$$I(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} H(y_i) - H(y)$$

- A natural measure of dependence between random variables.
- Always non-negative.
- Zero if and only if the variables are statistically independent.
- Takes into account the whole dependence structure of the variables.

Mutual Information

- Mutual information in terms of negentropy

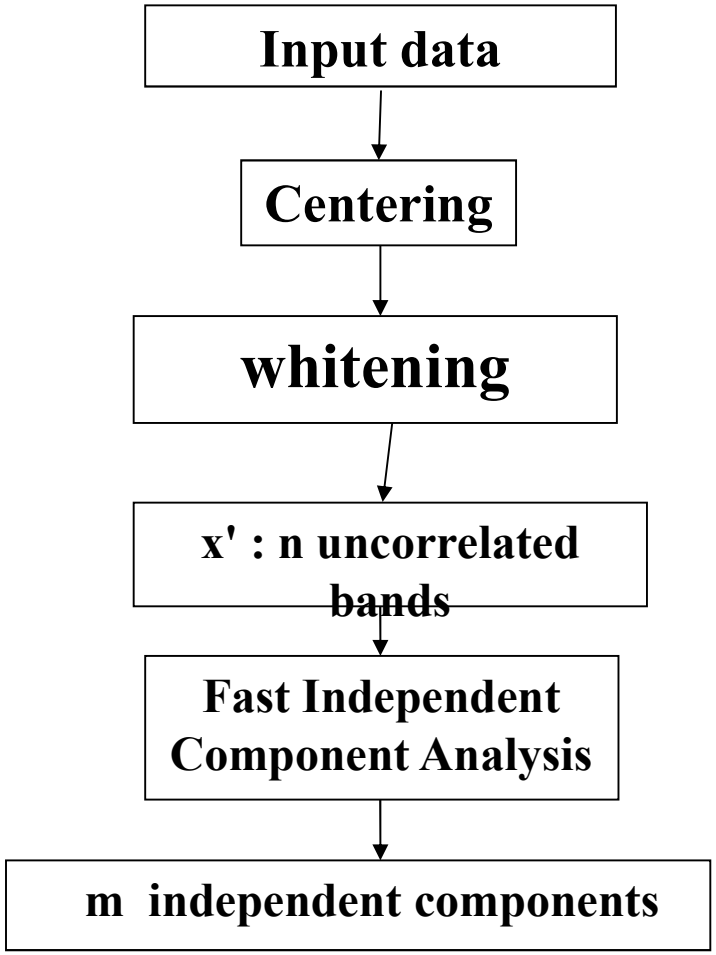
$$I(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} H(y_i) - H(y)$$

$$I(y_1, y_2, \ldots, y_n) = C - \sum_i J(y_i)$$

- Use the transform $y = w^T x$ such that w minimizes the mutual information.

- Minimization of mutual information is roughly equivalent to finding directions in which negentropy is maximized.

Preprocessing for ICA



Block Diagram showing various stages in Fast Independent Component Analysis.

Centering

- The most basic and necessary preprocessing is to center the data X, i.e. subtract its mean vector m=E{x} so as to make x a zero mean variable.

- After estimating the mixing matrix A, we can complete the estimation by adding the mean vector of s back to the centered estimates of s.

- The mean vector of s is given by $A^{-1}m$, where m is the mean that was subtracted in the preprocessing.

Whitening

- Before applying ICA algorithm it is required to transform the vector x linearly so that we obtain a vector x which is white, i.e. its components are uncorrelated and their variances equal unity.
- This can be done using Principal Component Transform or Maximum noise fraction Transform.
- This step is also useful in reducing the dimensionality of the data, by discarding those components whose eigenvalues are too small.
- This step reduces the noise and prevents the ICA from overlearning, which can sometimes be observed in ICA.

Algorithm for Fast ICA

1) Center the data to make its mean zero.
2) Choose m, the number of independent components to

   estimate from PCA/MNF.
3) Whiten the data to give Z.

Algorithm for Fast ICA

4) Choose a random unmixing matrix W.

5) $W = (WW^T)^{-1/2}W$. Symmetric decorrelation

6) Let $W_1 = E\left\{Z g\left(W^T Z\right)\right\} - E\left\{g'\left(W^T Z\right)\right\} W$

7) Where g is defined as g(y) = tanh(y) or

$g(y) = y^3$ or

$g(y) = \sin(y^3)$

8) $W_1 = (W_1 W_1^{T})^{-1/2} W_1$.

9) If not converged, go back to step 6.

10) $W_2 \leftarrow W_1 / \|W_1\|$

11) For second ICA go to step 6.

12) Repeat for i = 1, 2, 3…m.

Algorithm for Fast ICA

Convergence criteria is that the old and new values of W point in same direction. i.e, their dot product ≈ 1.0

Algorithm for Fast ICA

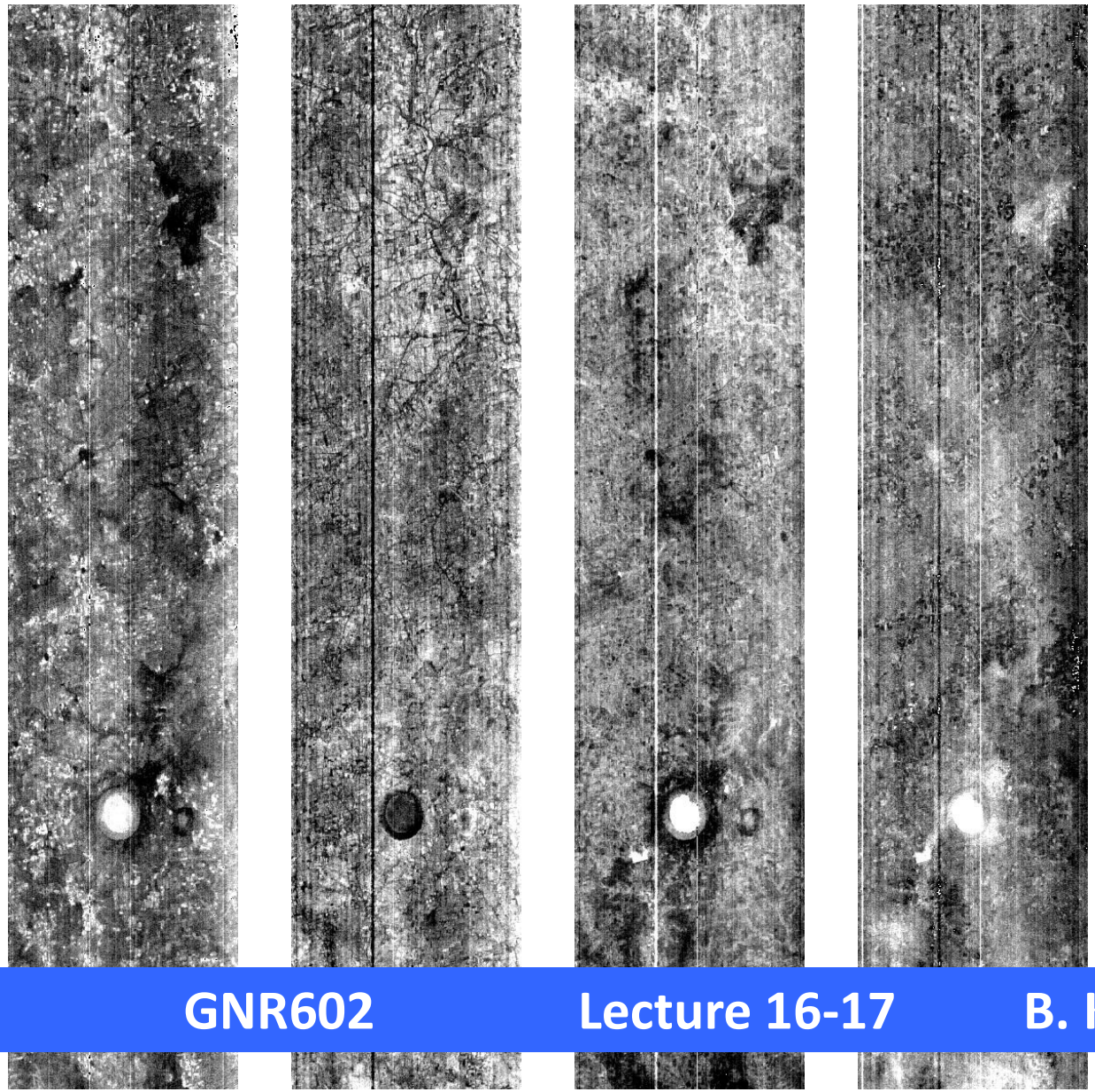
- Once p independent components or p vectors $\mathbf{w_1}$, ...,$\mathbf{w_p}$ are obtained, we run algorithm for w_{p+1} , and after every iteration subtract from w_{p+1} the projections $w_{p+1}w_jw_j$ $j = 1,2,\ldots p$. and then normalize it

$$w_{p+1} = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T w_j w_j$$

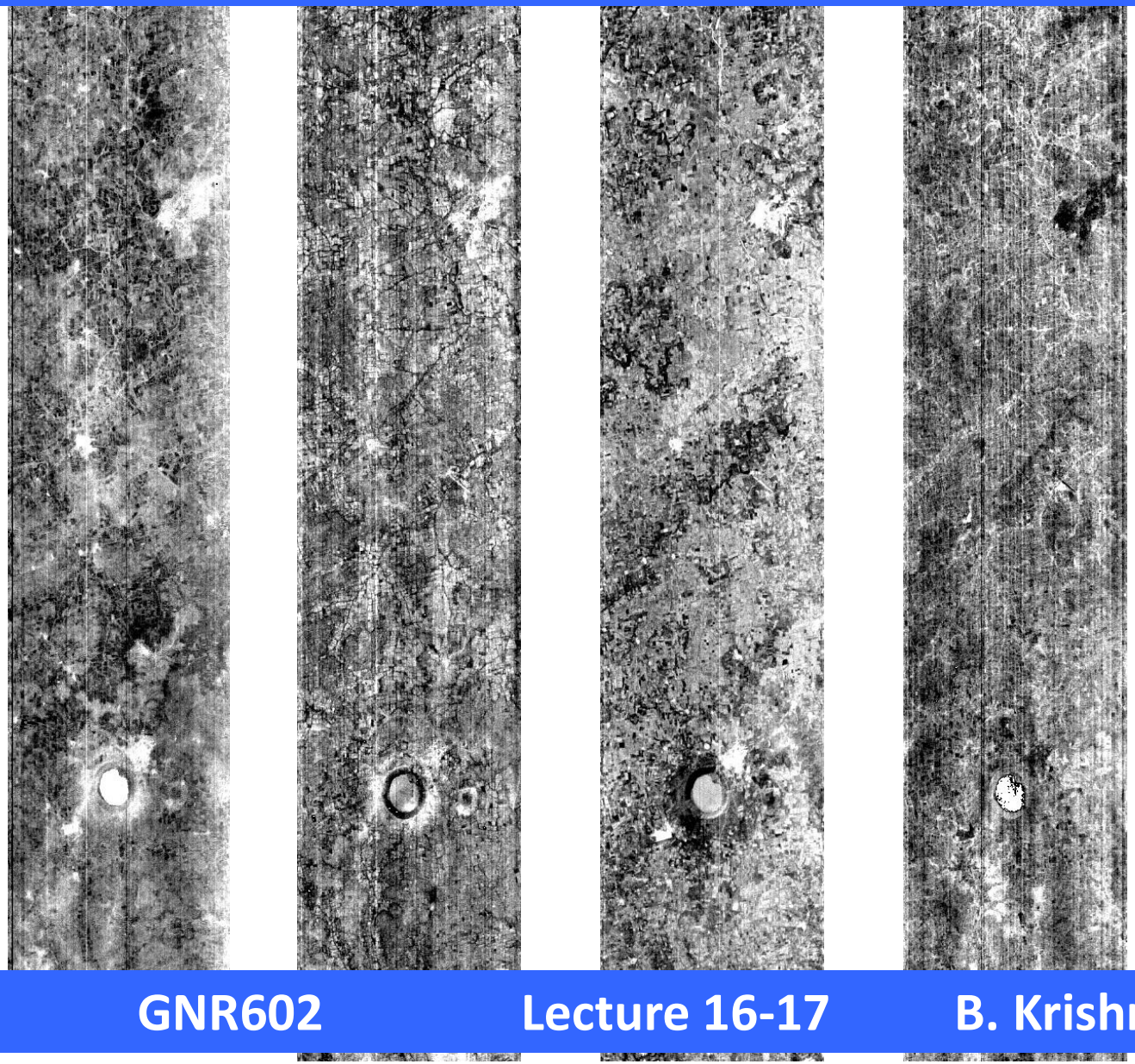
$$w_{p+1} = w_{p+1} \Big/ \sqrt{w_{p+1}^T w_{p+1}}$$

Case Study: Input Data

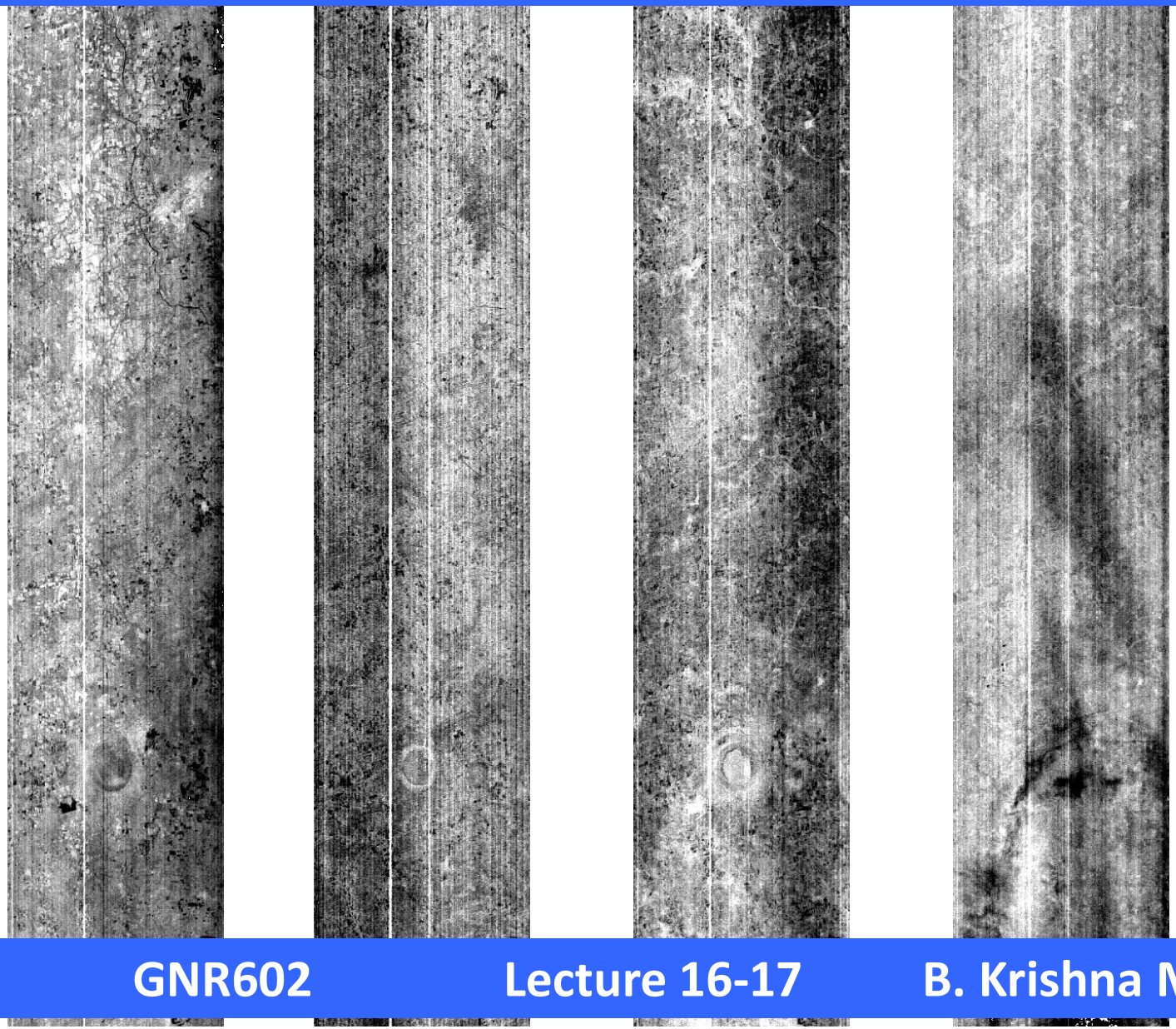
- Area - Jalna
- Sensor – Hyperion
- Original Scene specifications
  - Number of lines in the file=3248
  - Number of samples in the file = 256
  - Number of bands in the file =242
- Dimensions after atmospheric Correction
  - Number of lines in the file=1200
  - Number of samples in the file = 254
  - Number of bands in the file =175



Fast ICA Output

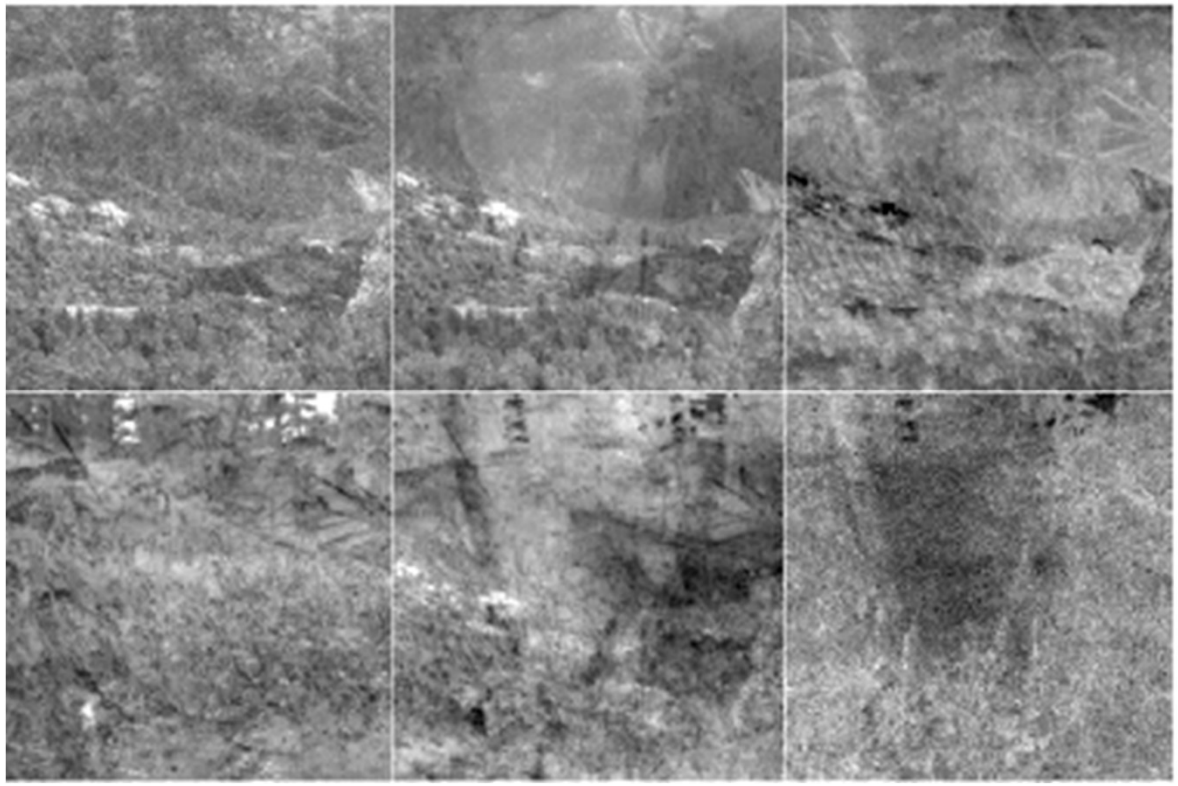


Fast ICA Output

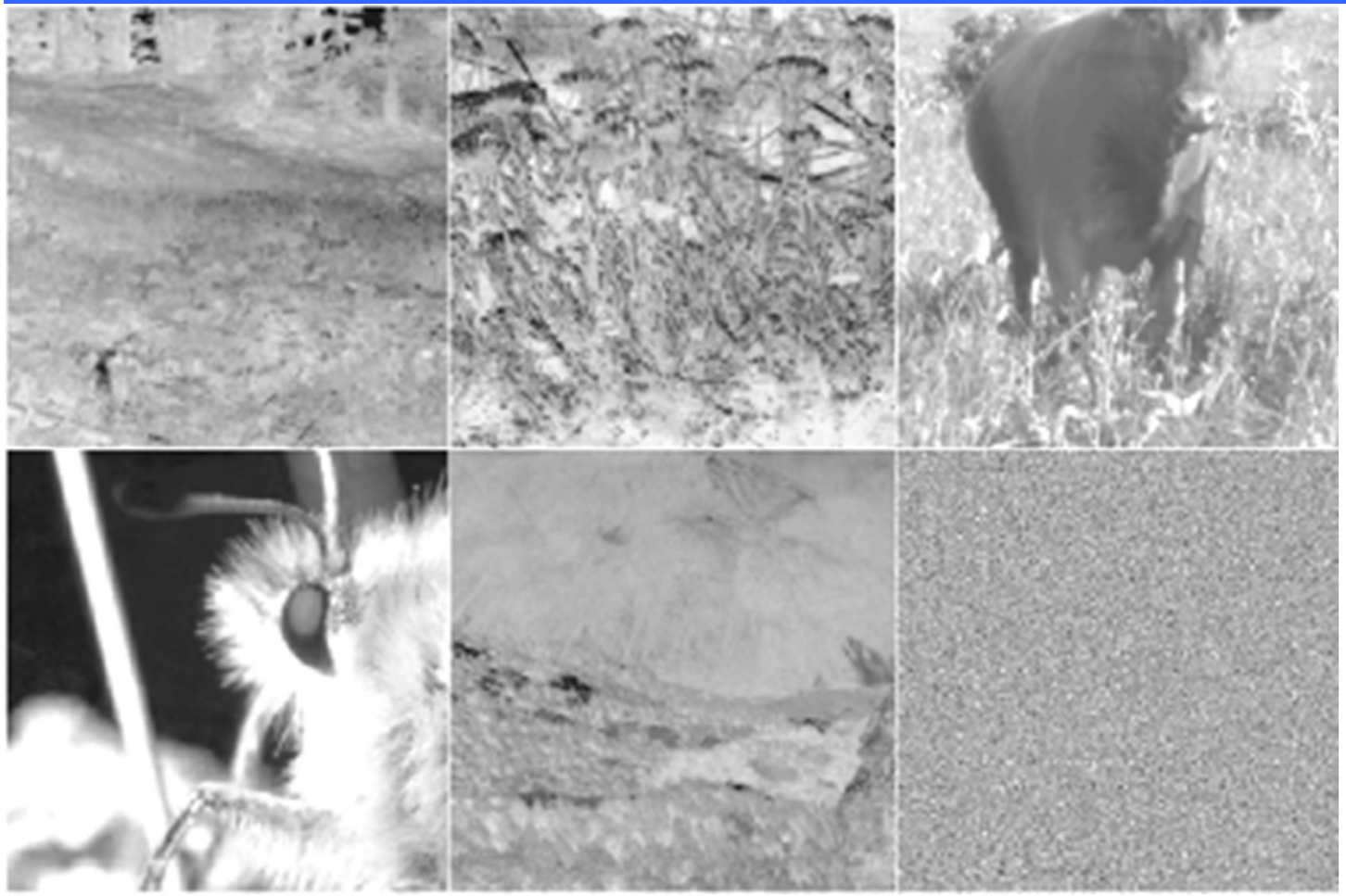


Fast
ICA
Output

Example 2

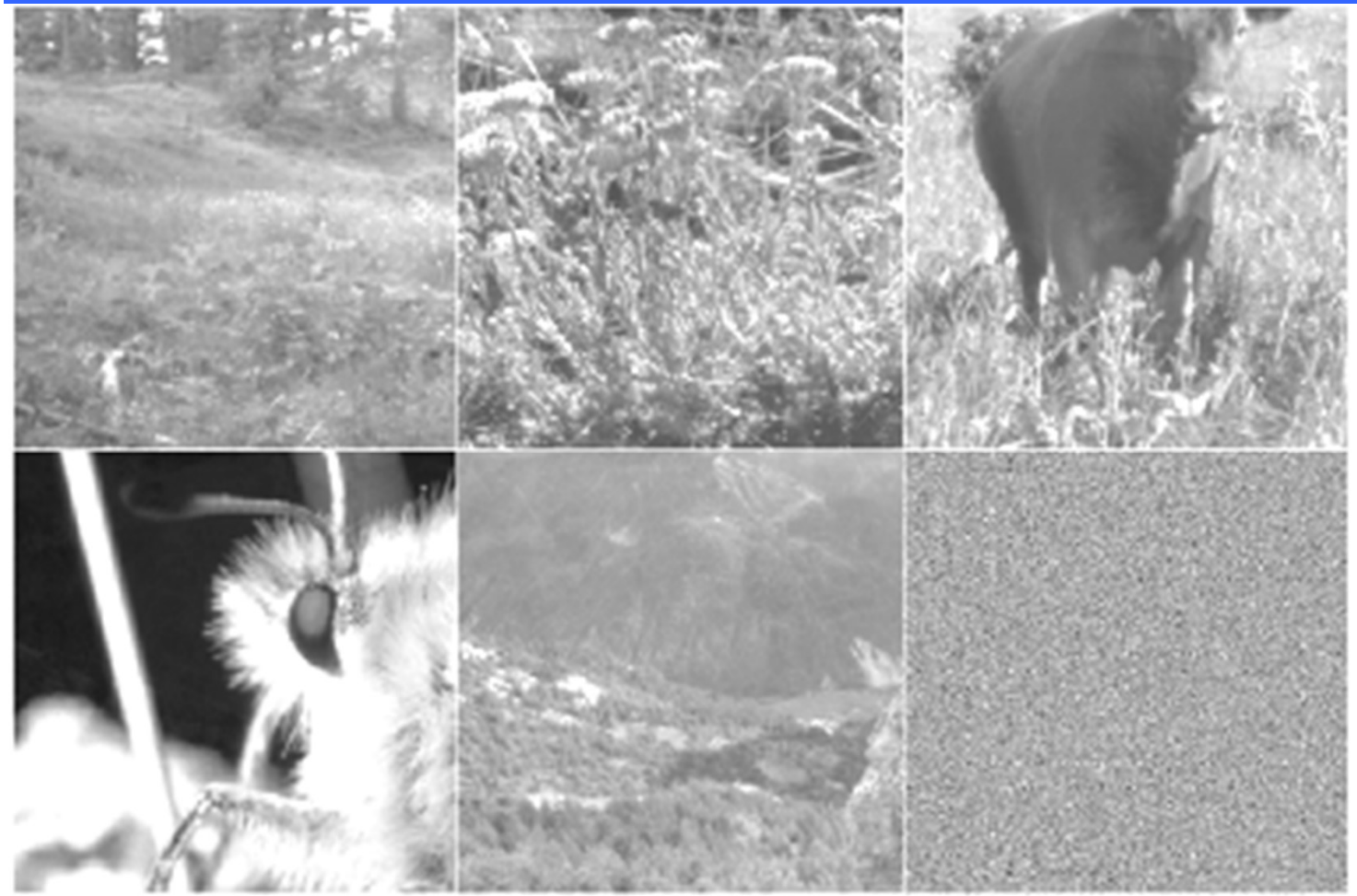


- 6 images
- Linear mixture of six originals
- Determine the originals



From the ICA Tutorial by Hyvarinen and Oja

6 independent components



From the ICA Tutorial by Hyvarinen and Oja

Independent latent (hidden) variables

Properties of the FastICA Algorithm

- The convergence is cubic (or at least quadratic), under the assumption of the ICA data model
- This is in contrast to ordinary ICA algorithms based on (stochastic) gradient descent methods, where the convergence is only linear.
- Contrary to gradient-based algorithms, there are no step size parameters to choose.

Properties of the FastICA Algorithm

- The performance of the method can be optimized by choosing a suitable nonlinearity g.

- The independent components can be estimated one by one. This is useful in decreasing the computational load of the method in cases where only some of the independent components need to be estimated

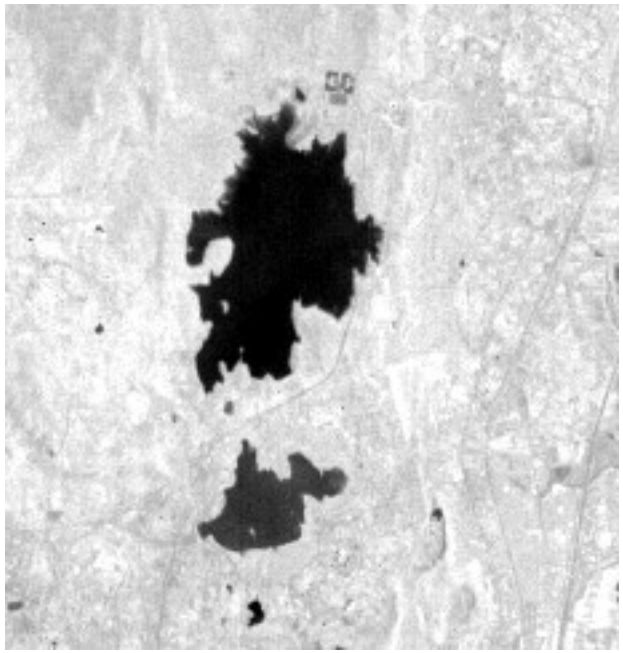
Applications

- Audio noise cancelling (cocktail party problem)
- Dimensionality reduction and mixture modeling
- Reducing Noise in Natural Images.
- Telecommunications.

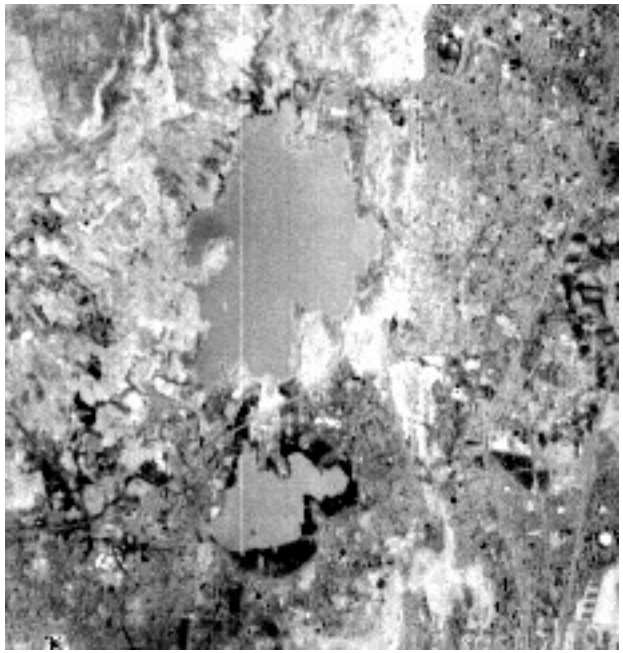
Summary of ICA

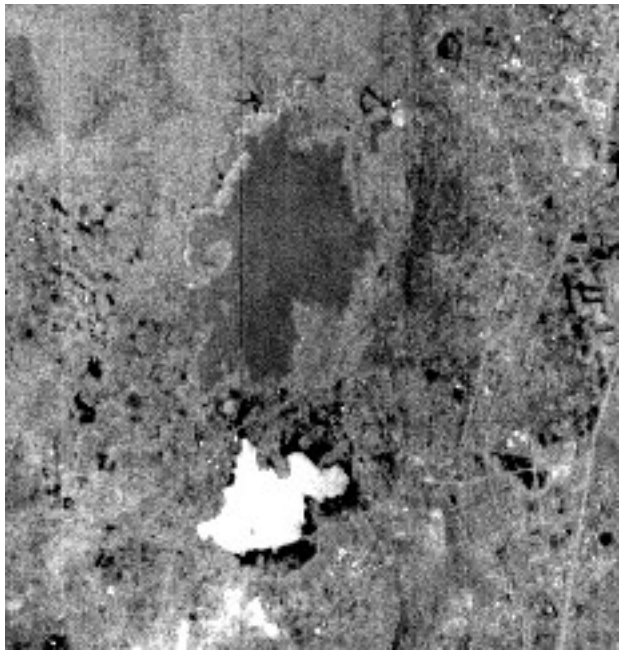
- ICA is a very simple model
- Simplicity implies wide applicability
- A non-gaussian alternative to PCA
- Decorrelation or whitening is only half of ICA
- The other half uses higher order statistics of non-gaussian variables
- Basic principle is to find maximally non-gaussian directions

IC's 1-3

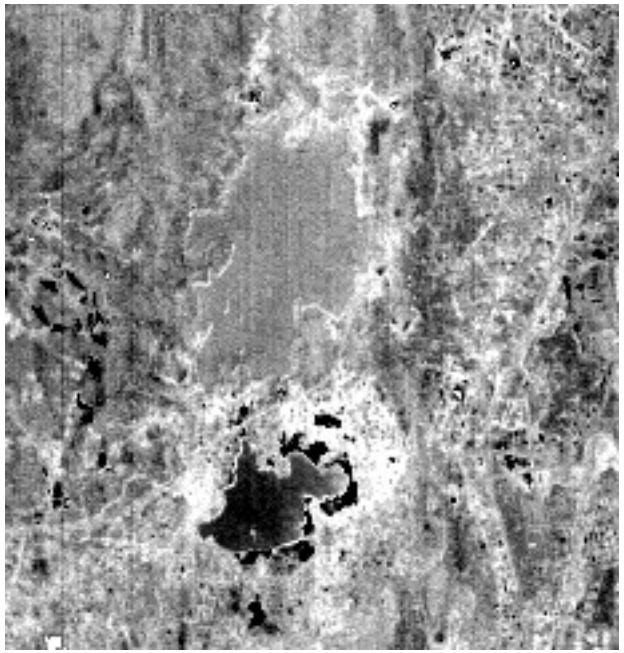


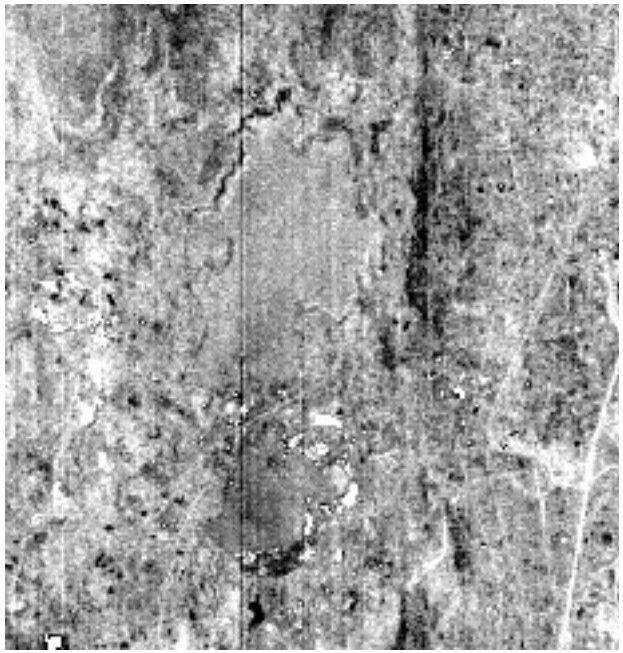
Components 4-6





Components 7-9





Selected References for ICA

- Wang, J., and Chang, C.-I (2006), Independent Component Analysis-Based Dimensionality Reduction with Applications in Hyperspectral Image Analysis, IEEE Transaction on Geoscience and Remote Sensing, vol. 44(6), pp. 1586-1600.
- Varshney, P.K. and Arora, M.K. (2004), Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data, Springer, chapter 3, 4, 8, pp. 89-132, 199-216.
- Hyvarninen, A. and Oja, E. (2000) Independent Component Analysis: Algorithms and Applications, Neural Networks, vol. 13(4-5), pp. 411-430.
- Fodor, I.K. (2002), A Survey of Dimensionality Reduction Techniques, Scientific Report UCRL-ID-148494, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Liverpool.

Genetic Algorithm Based Search

- Genetic algorithm based search procedures search for subsets of input bands where the class separability is high and classification accuracy is also high
- Fitness function is dependent on pair-wise class divergence and classification accuracy. Higher fitness function candidates prosper.

Fitness Function

- Fitness associated with a band subset =

  Accuracy of classification for the entire set of test samples + Separability of classes

- High accuracy + High separability = High fitness

GA Based Feature Selection

- **Algorithm for GA based feature selection and classification**
  1. Choose a proper data structure for encoding (integer encoding) the chromosome to represent the band numbers of subset of bands.
  2. Create and initialize the population.
  3. Evaluate fitness of the members of the population using accuracy of SAM classification on training data, mean Euclidean distance between class means, mean inter-band correlation.

Fitness = α(accuracy) + β(avg_dis) + (1 – α – β) (avg_cor)

GA Based Feature Selection

Apply crossover on the population considering the crossover probability.

Apply mutation on the population considering the mutation probability.

Evaluate fitness of the new population.

Use proper selection method to choose population for next generation from old and new population.

Repeat the steps from 3 to 6 until fixed number of iterations are completed or convergence criteria are reached.

Once the optimal subset of bands is obtained, classify the band subset using SAM classifier and calculate the confusion matrix and kappa coefficient using test-data.

Map the output class pixel values to color.

Example: Comparison of MNF and GA results

- Class separabilities of MNF components

AVE	MIN	Class Pairs:													
		1: 2	1: 3	1: 4	1: 5	1: 6	2: 3	2: 4	2: 5	2: 6	3: 4	3: 5	3: 6	4: 5	4: 6
		5: 6													
		Best Average Separability													
1400	171	1271	1134	6531	316	838	423	1281	220	203	2811	407	171	2240	2752
		395													

Example: Comparison of MNF and GA results

- Class separabilities of GA components

AVE	MIN	Class Pairs:						
		1: 2	1: 3	1: 4	1: 5	1: 6	2: 3	2: 4
		2: 5	2: 6	3: 4	3: 5	3: 6	4: 5	4: 6
		5: 6						
		Best Average Separability						
2639	337	2166	992	4767	517	1088	812	7314
		3170	602	6178	509	337	4188	5270
		1676						

Endmember Selection

Issue with Pixel Spectra

- Hyperspectral images, due to their very high spectral resolution, have much coarser spatial resolution compared to the high spatial resolution sensors like Cartosat-2A, Quickbird, Ikonos, etc.
- Footprints of most pixels contain more than one class
- The resultant spectrum recorded at such pixels is a mixture of the individual ***pure*** spectra

Pure Pixels or End Members

- Endmembers in an image are those that largely contain a single class
- Most often high spectral resolution images are of coarse spatial resolution
- Since pixel footprint on ground is large, more than one category may be present within each pixel
- Majority of pixels may be ***mixed pixels***

What are Endmembers?

- Endmembers are pure pixels that contain typically more than 75% of a single material
- The advantage of finding the endmembers is that one can classify all the pixels based on their relative similarity to the endmembers
- The relative proportion of each endmember can be estimated in each pixel, resulting in ***abundance map***

Pixel Purity Index (PPI)

Used to extract Endmembers

The PPI algorithm is characteristic in its supervised nature;

it consists of the following steps. First, a noise-whitening and dimensionality reduction step is performed by using the MNF transform

Then, a pixel purity score is calculated for each point in the image cube by randomly generating co-ordinate axes in the space comprising the MNF-transformed data.

All the points in that space are projected onto the axes, and the ones falling at the extremes of each line are counted.

After many repeated projections to different random axes, those pixels that count above a certain cutoff threshold are declared “pure.”

Principle Behind Pixel Purity Index (PPI)

An endmember is a feature vector representing a class, and is supposed to be different from other feature vectors

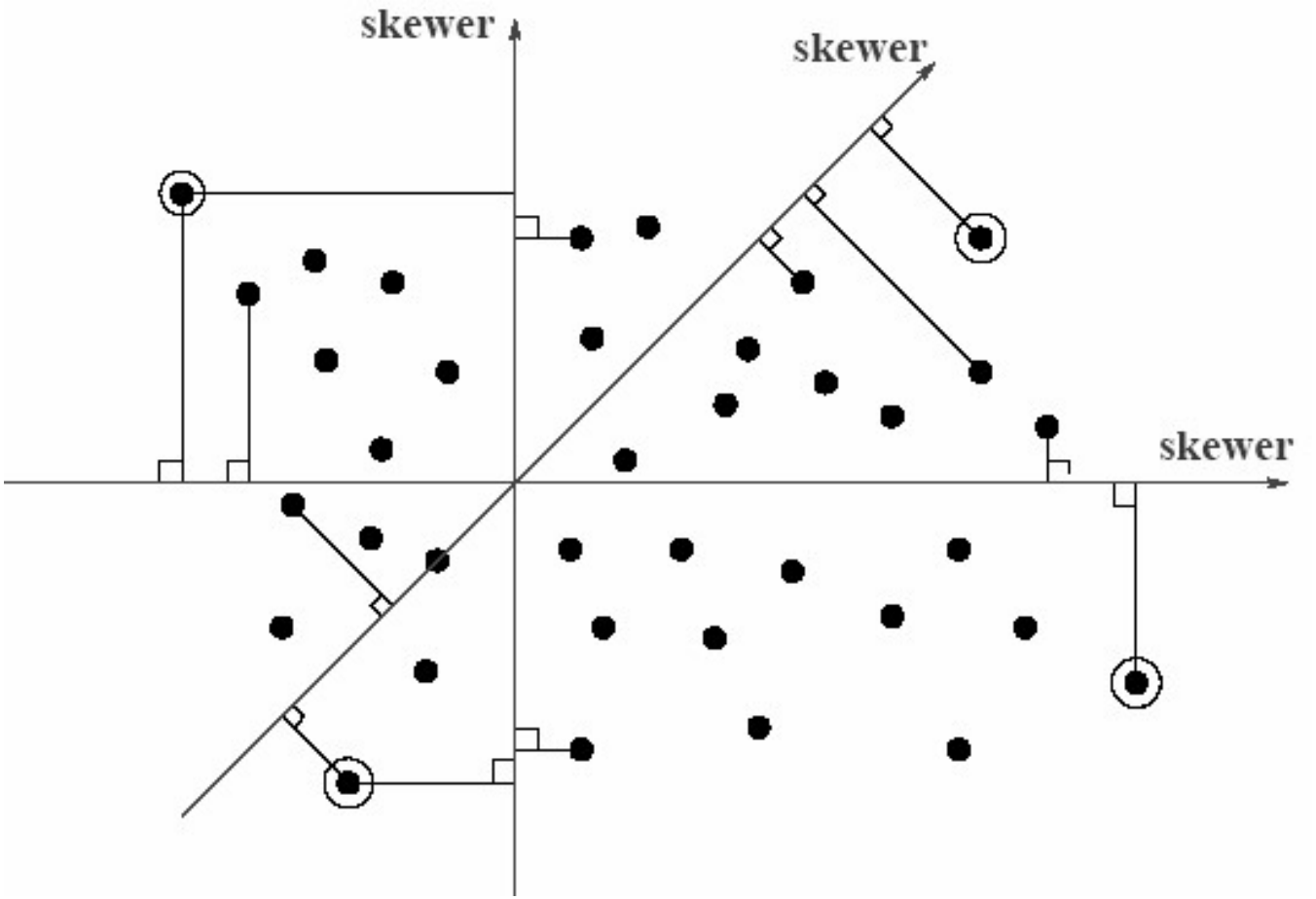
Therefore, when projected onto any coordinate axis, the endmember is expected to get projected onto the left or right extreme of the axis.

Feature vectors that are mixtures of endmembers get projected towards the middle of the axis.

Pure Pixel Extraction

- Pixel purity index (implemented in ENVI)
  - Project the data onto a large number of random axes
  - Count the pixels that fall on the extremes of the axes
  - Pixels with the highest count are taken as pure pixels

Pixel Purity Index



Principle Behind Pixel Purity Index (PPI)

The points corresponding to extremes, for each *skewer* direction, are stored.

A cumulative account records the number of times each pixel (*i.e.*, a given spectral vector) is found to be an extreme. The pixels with the highest scores are the purest ones.

PPI was, firstly, conceived as supervised tool to identify endmembers.

Therefore, the purest pixels are loaded into a multi-dimensional visualization tool and then endmembers are identified visually as the extreme pixels in the data cloud.

For this purpose, a threshold is specified on the counts of the pixels that were on the extremes of the skewers

Linear Mixture Modeling

- If the mixing is rather large, then the mixing of the signatures can be represented as a linear model. In other words, if large classes get mixed inside a pixel's footprint then a linear mixing model will suffice. However, if several small objects are contained in the pixel, a linear model is strictly not valid.

Linear Mixture Modeling

- Consider a pixel whose reflectance spectrum is R_i that is made up of linear combination of all the endmember spectra Re_j

$$R_i = \sum_{j=1}^{n} f_j \, \mathrm{Re}_{ij} + \varepsilon_i$$

$$0 \le \sum_{j=1}^{n} f_j \le 1 \qquad \sum_{j=1}^{n} f_j = 1$$

- Each component of R_i is given by a linear combination of the corresponding components of ALL the endmembers with f_j being the weights

Linear Mixture Modeling

$$R_i = \sum_{j=1}^{n} f_j \, \mathrm{Re}_{ij} + \varepsilon_i \quad 0 \le \sum_{j=1}^{n} f_j \le 1 \quad \sum_{j=1}^{n} f_j = 1$$

- R_i represents the observed reflectance or gray level value in wavelength i.
- This can be represented as a matrix-vector product when all wavelengths are included in the hyperspectral data
- **[Re]f** + ε = **R**, where **Re** is a matrix of endmember spectra, whose size = DxK; D is the dimensionality of the data, K is the number of classes. Each column of **[Re]** is endmember of a class. Each element in the column is the reflectance or gray level in a band.

Linear Mixture Modeling

$$R_i = \sum_{j=1}^{n} f_j \, \mathrm{Re}_{ij} + \varepsilon_i \quad 0 \le \sum_{j=1}^{n} f_j \le 1 \quad \sum_{j=1}^{n} f_j = 1$$

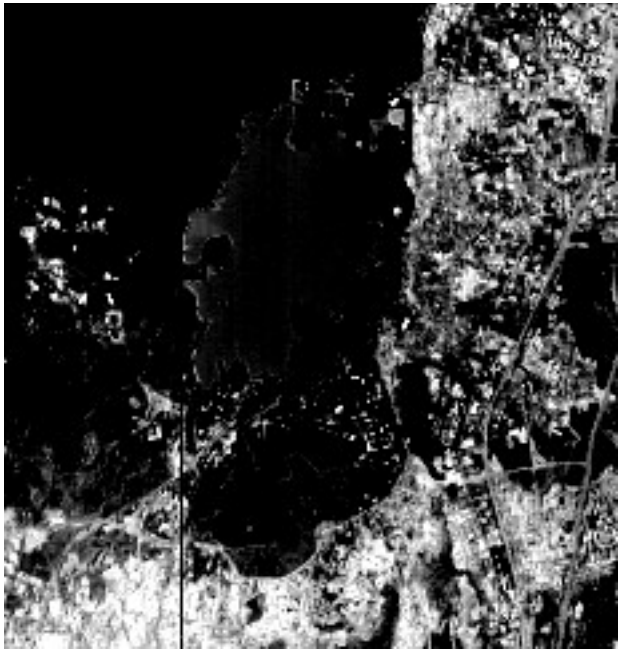
- ε is noise vector of the same dimension as the feature vector (=D)
- **f** is the vector denoting the proportion of each endmember in a pixel, to be determined.
- The nonlinear (over-specified) system of equations is solved to find **f**
- $\mathbf{f} = [\mathbf{Re}]^{+} \mathbf{R} - [\mathbf{Re}]^{+} \varepsilon$
- $[\mathbf{Re}]^{+}$ is called the pseudo-inverse of the rectangular matrix **[Fe]**

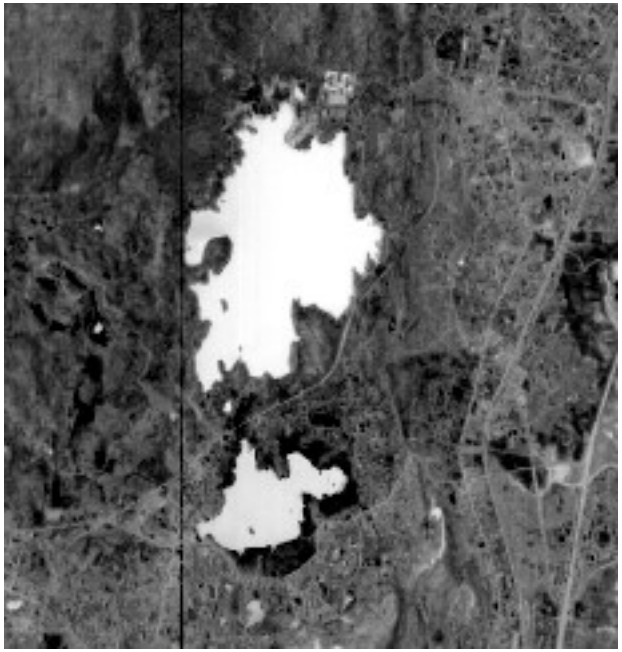
Linear Mixture Modeling

$$R_i = \sum_{j=1}^{n} f_j \, \mathrm{Re}_{ij} + \varepsilon_i \qquad 0 \le \sum_{j=1}^{n} f_j \le 1 \qquad \sum_{j=1}^{n} f_j = 1$$

- **f = [Re]$^+$ R - [Re]$^+$ ε**
- The above solution does not guarantee the non-negativity and sum-to-unity requirements of the end-member proportions.
- Imposing these constraints makes the solution a linear programming problem with equality and inequality constraints.

Endmember Abundances using the Endmembers Selected

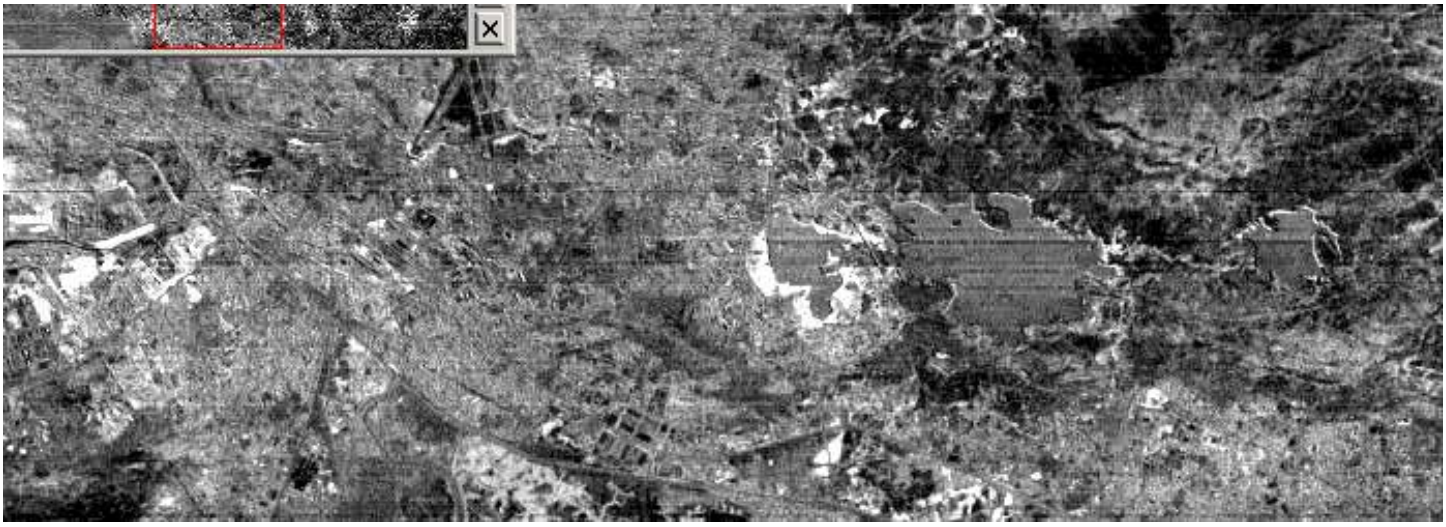


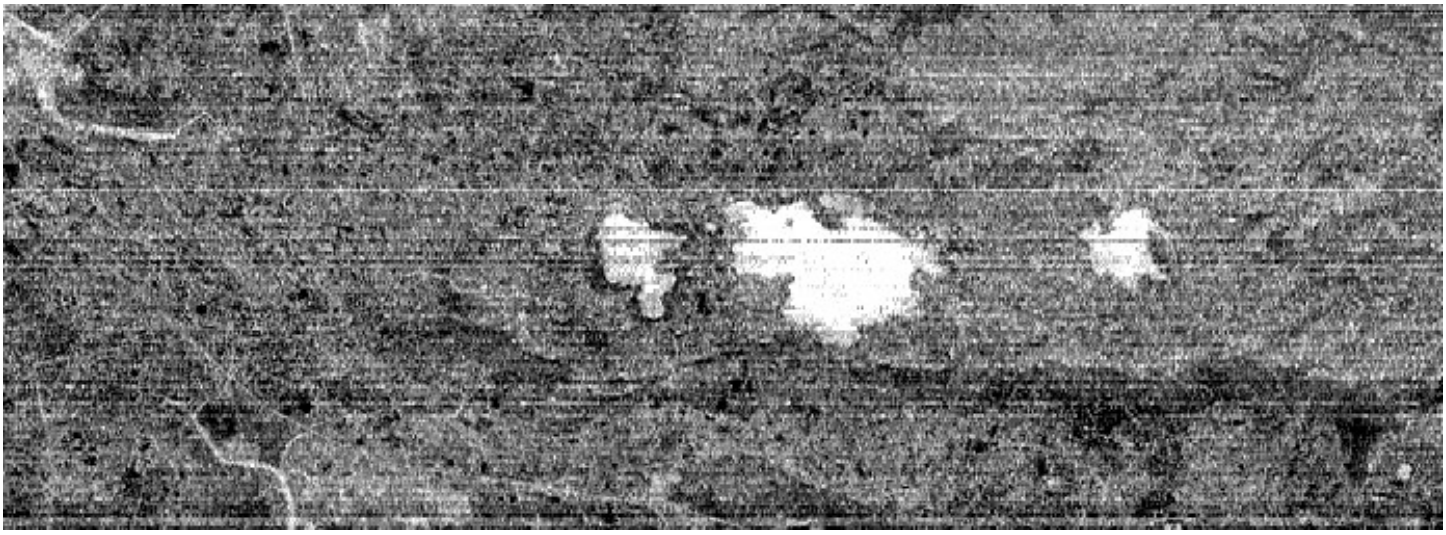


End Member Abundances

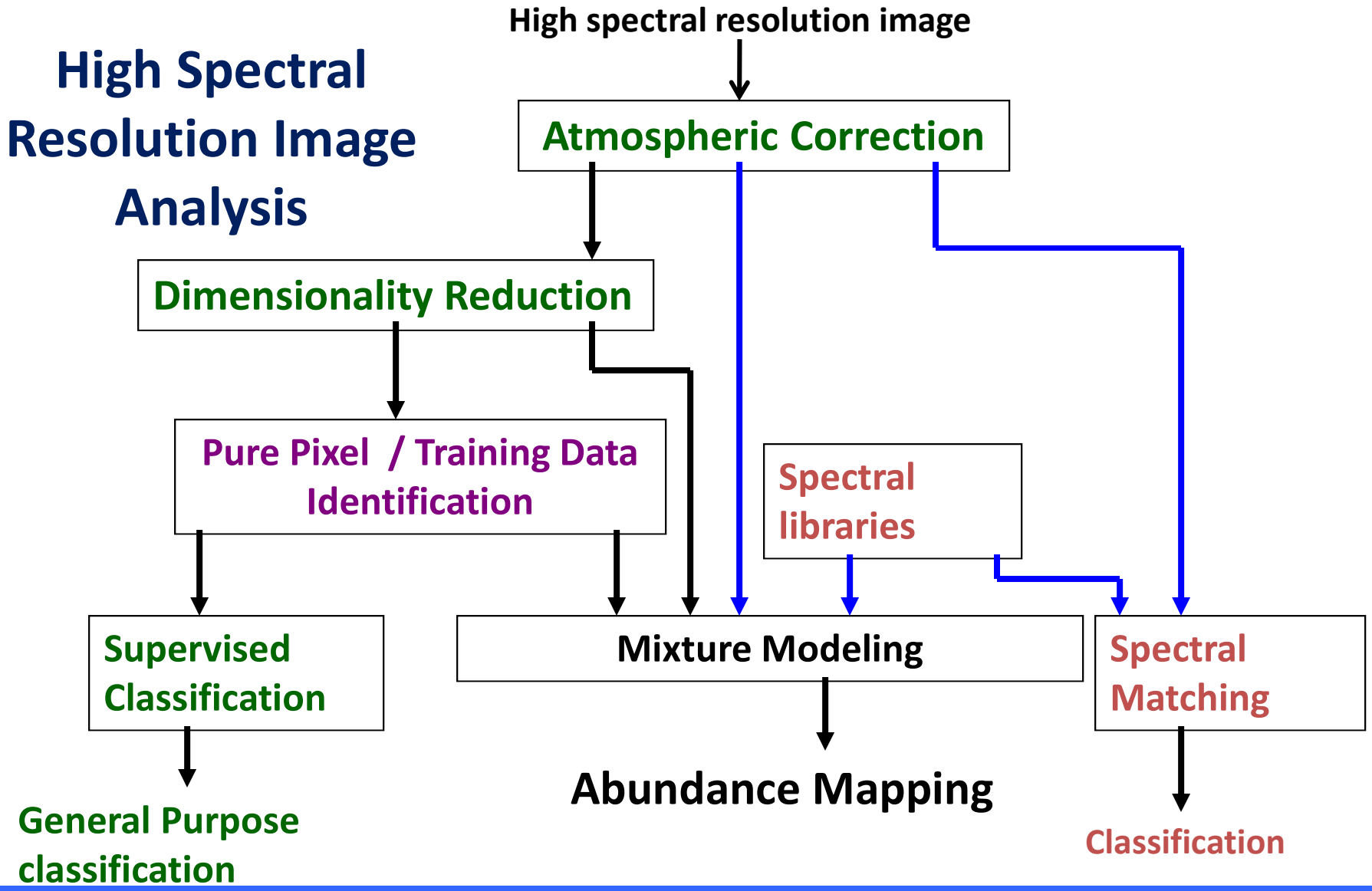
Example with Manually Selected Endmembers

- From the Hyperion image of Mumbai city, endmembers are manually selected
- The major classes are:
  - Two types of vegetation found near the lakes
  - Water
  - Roads





High Spectral Resolution Image Analysis



Spectral Angle Mapper

- Given a large dimensional data set, computing the covariance matrix, its inverse, and the distance for each pixel
- $(X - \mu)^T \Sigma^{-1} (X - \mu)$ is highly time consuming and if the covariance matrix is close to singular then its inverse can be unstable, leading to erroneous results
- In such cases, alternate methods can be applied, such as Spectral Angle Mapper

S.A.M. Principle

- If each class is represented by a vector $\mathbf{v}_i$, then the angle between the class vector and the pixel feature vector $\mathbf{x}$ is given by
- $\cos\theta = [\mathbf{v}_i.\mathbf{x}] / [|\mathbf{v}_i||\mathbf{x}|]$
- For small values of θ, the value of $\cos\theta$ is large
- The likelihood of $\mathbf{x}$ to belong to different classes can be ranked according to the value of $\cos\theta$.

S.A.M. Advantage

- The value of the vector would not be greatly affected by minor changes in $\mathbf{v}_i$ or $\mathbf{x}$.
- The computation is simpler compared to the Mahalanobis distance computation involved in ML method

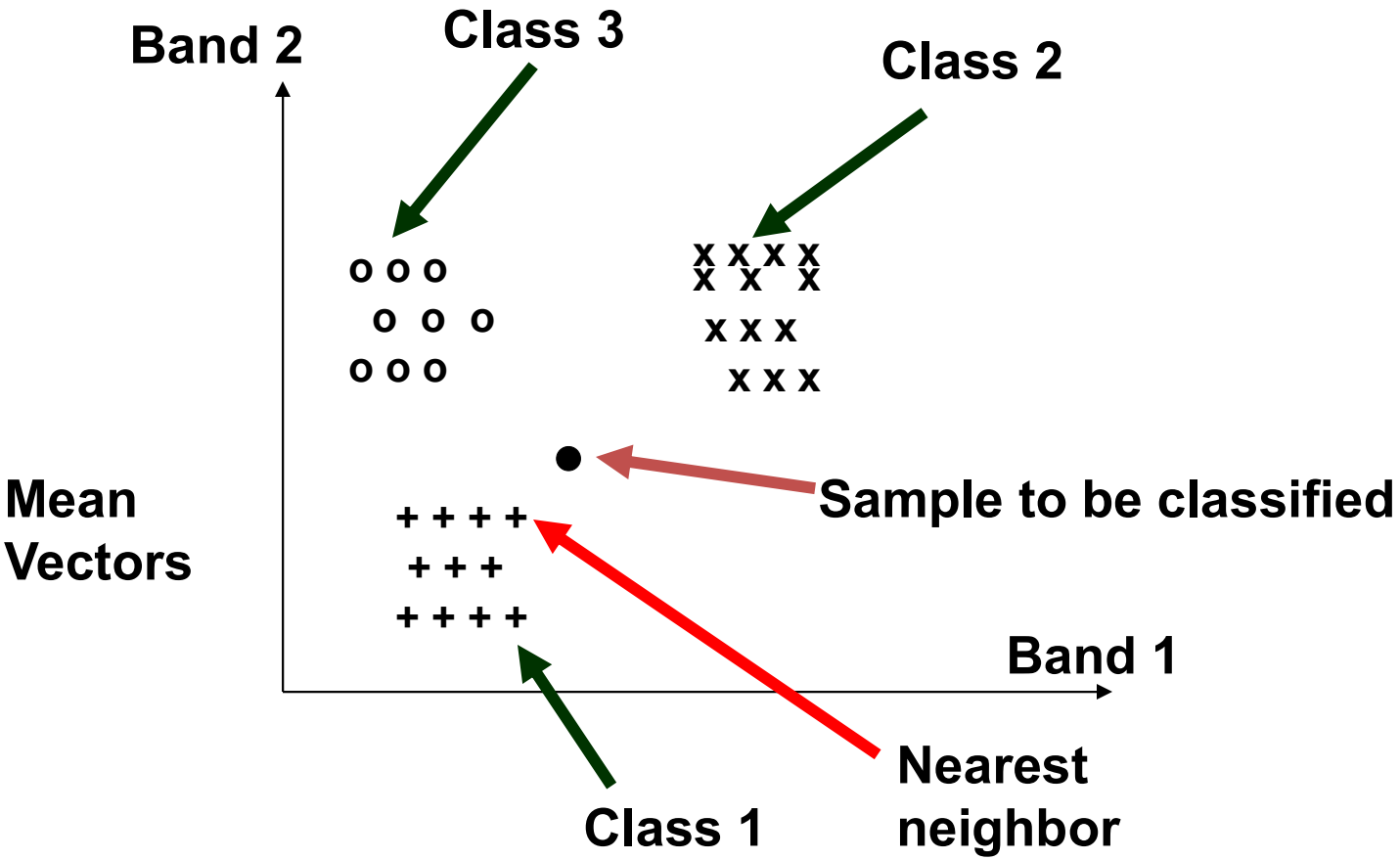
Simple Classification Techniques for Hyperspectral Images

Nearest-Neighbor Classifier

- This is one of the non-parametric classifiers
- The algorithm is:
  - Find the distance of given feature vector **x** from ALL the training samples
  - **x is assigned to the class of the nearest training sample (in the feature space)**
  - **This method does not depend on the class statistics like mean and covariance.**

Concept of Nearest Neighbor Classification

- 2-D Feature Space



The nearest sample to the pixel to be classified is from class 1. Hence this pixel is assigned to class 1

Principle of Nearest-Neighbor Classifier

- The ***nearest neighbor*** classifier is a general technique developed by Pattern Recognition practitioners. Image processing community adopted it.
- The nearest neighbor referred to in this context is the training sample with the most similar feature vector, i.e., at the smallest distance from the given pixel's feature vector. This has no relevance to where that pixel is in the image. Any pixel with the same feature vector will be identically classified, irrespective of its location.

K-NN Classifier

- K-nearest neighbour classifier
- Simple in concept, time consuming to implement
- For a pixel to be classified, find the K closest training samples (in terms of feature vector similarity or smallest feature vector distance)
- Among the K samples, find the most frequently occurring class C_m
- Assign the pixel to class C_m

K-NN Classifier

- Let k_i be number of samples for class C_i (out of K closest samples), i=1,2,…,L (number of classes)
- Note that

$$\sum_i k_i = K$$

- The discriminant for K-NN classifier is
- $g_i(x) = k_i$
- The classifier rule is
- *Assign x to class C_m if $g_m(x) > g_i(x)$, for all i, $i \neq m$*

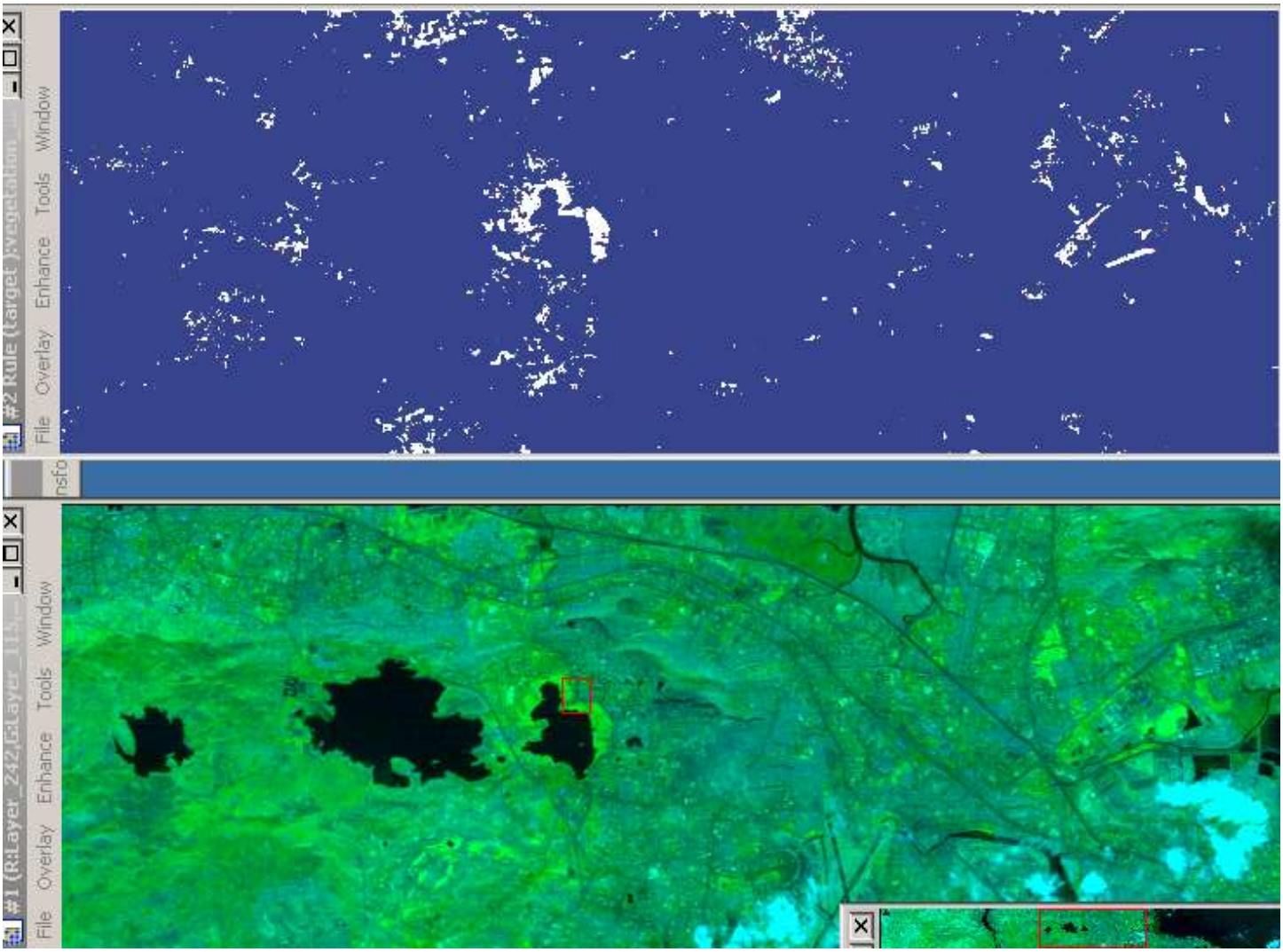
S.A.M. Principle

- If each class is represented by a vector $\mathbf{v}_i$, then the angle between the class vector and the pixel feature vector **x** is given by
- $\cos\theta = [\mathbf{v}_i.\mathbf{x}] / [|\mathbf{v}_i||\mathbf{x}|]$
- For small values of θ, the value of $\cos\theta$ is large
- The likelihood of **x** to belong to different classes can be ranked according to the value of $\cos\theta$.

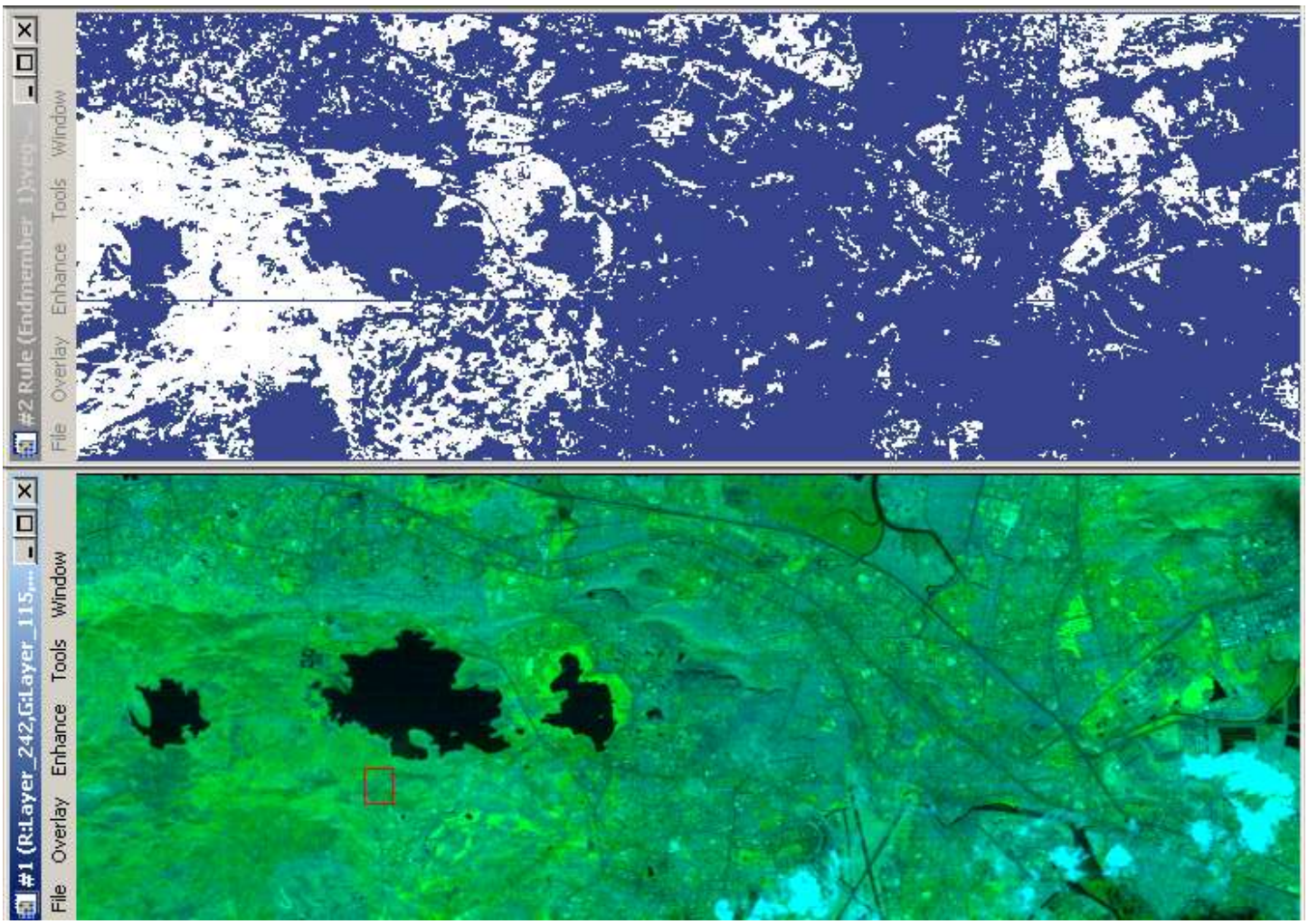
S.A.M. Advantage

- The value of the vector would not be greatly affected by minor changes in $\mathbf{v}_i$ or **x**.
- The computation is simpler compared to the Mahalanobis distance computation involved in case of multispectral classifiers

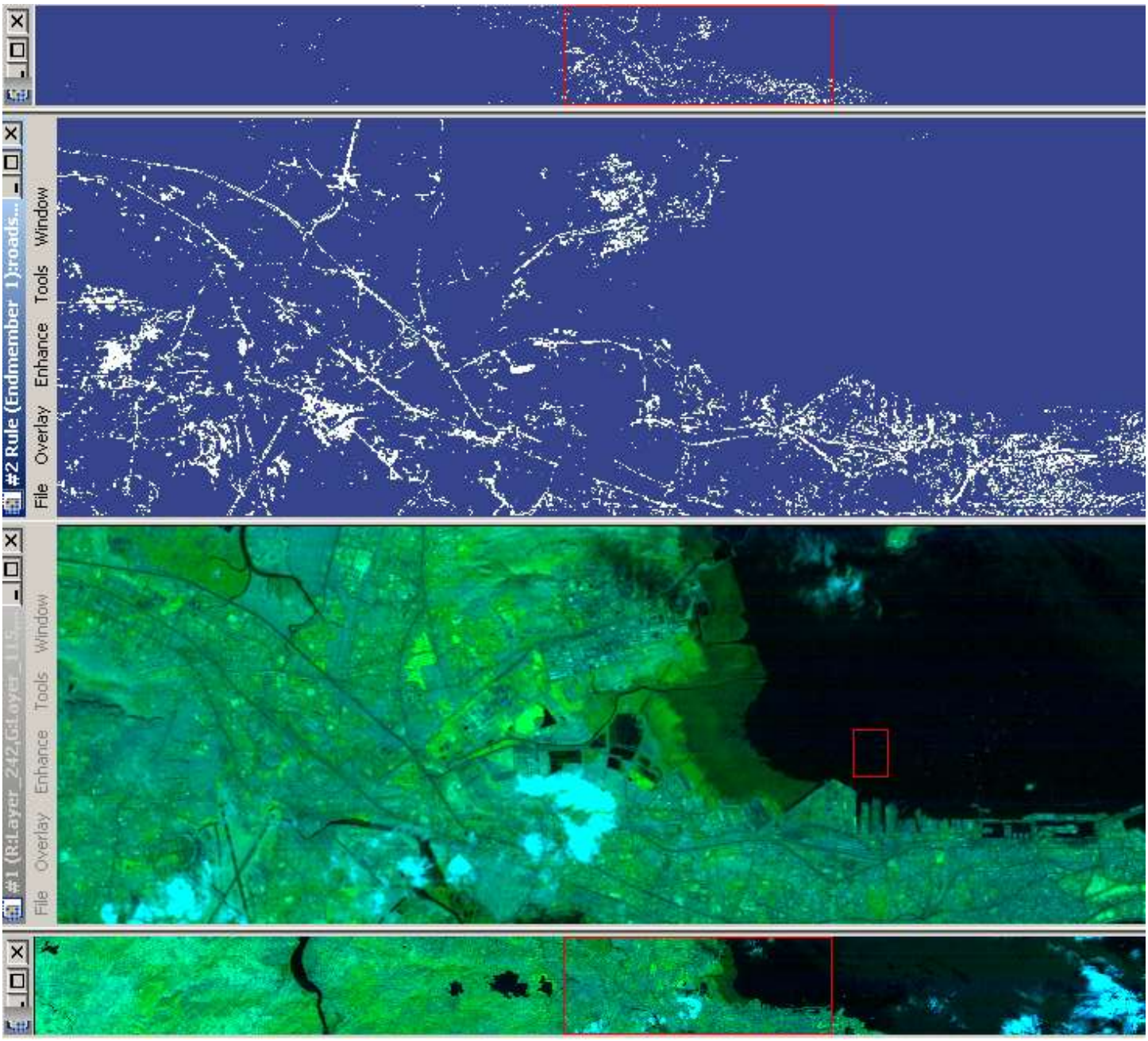
Classification by SAM



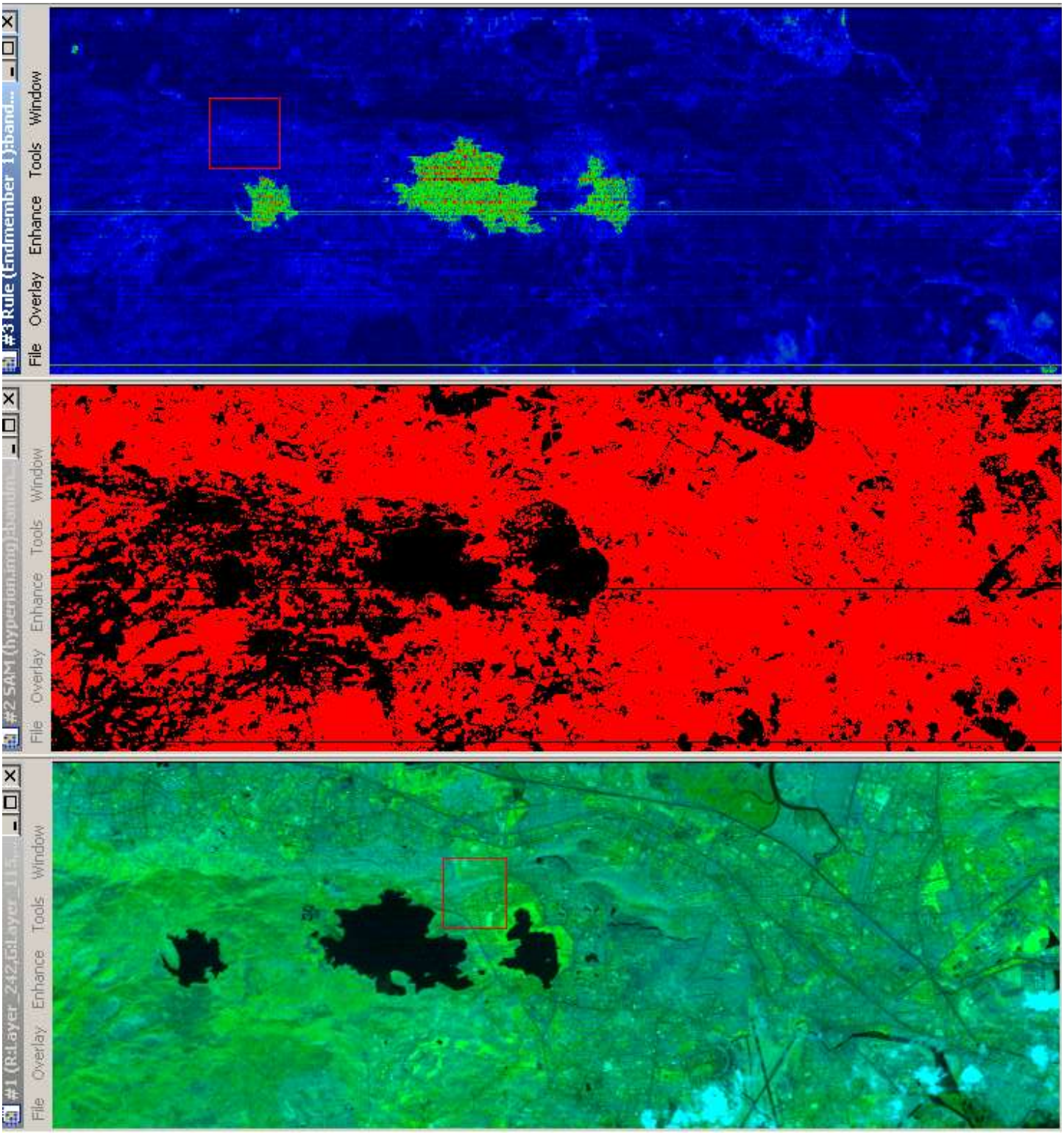
Classification by SAM



Classification by SAM



Classification Map for Water by SAM



Contd…