Carleton UNIVERSITY

GEOM 3002 – Lecture 9

Semi-automated Classification of Remote Sensing Images for Thematic Mapping: Supervised Classification

Readings

- As given for Lecture 8

Lillesand et al. 2015

- Supervised Classification 7.8-7.10
- Accuracy Assessment 7.17

Jensen 2015

- Supervised Classification pp. 376-401
- Accuracy Assessment pp. 557-571

Olofsson et al. 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42-57.

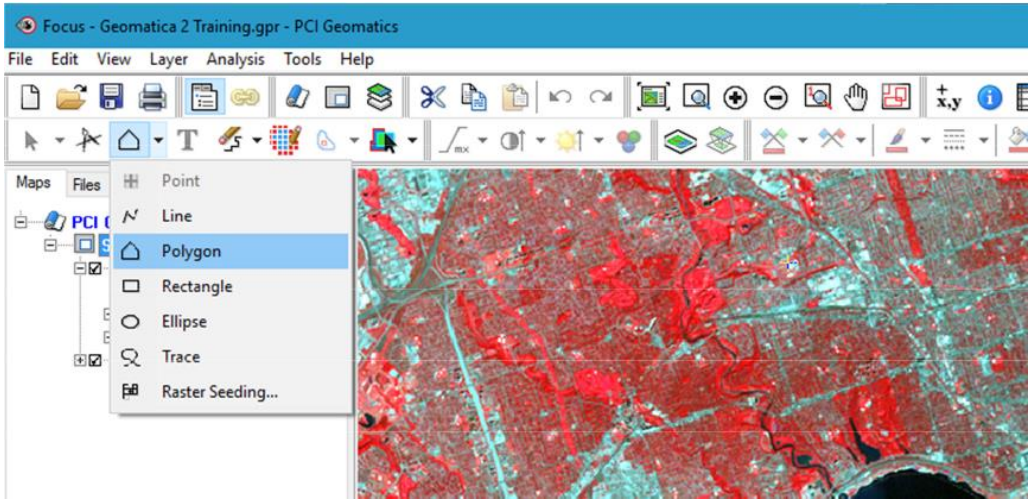
Supervised Classification

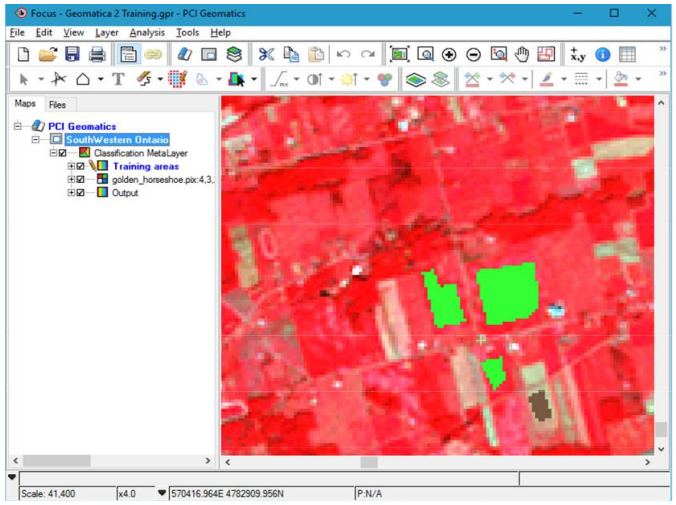
As previously mentionned:

- The analyst selects sample pixels (“Training Data”) of known classes
- The algorithm determines the data characteristics / statistics of these samples for each class (e.g., average brightness (reflectance) or other data values (elevation, etc.) if used as input.
- The algorithm then assigns each image pixel (or segmented object) to the class whose data it most resembles
  - Or for subpixel classifiers, it determines the classes the pixel most resembles and gives them weightings that are often related to area proportions of each class within the pixel

Supervised Classification: Training Data Collection

- Often ≥1 polygon for each class is delineated manually
  - Sometimes people use a computer seeding algorithm (pick a pixel, the algorithm grows a polygon around it of similar pixel values or with given relative variance)
  - Data for pixels in all polygons for a given class are combined

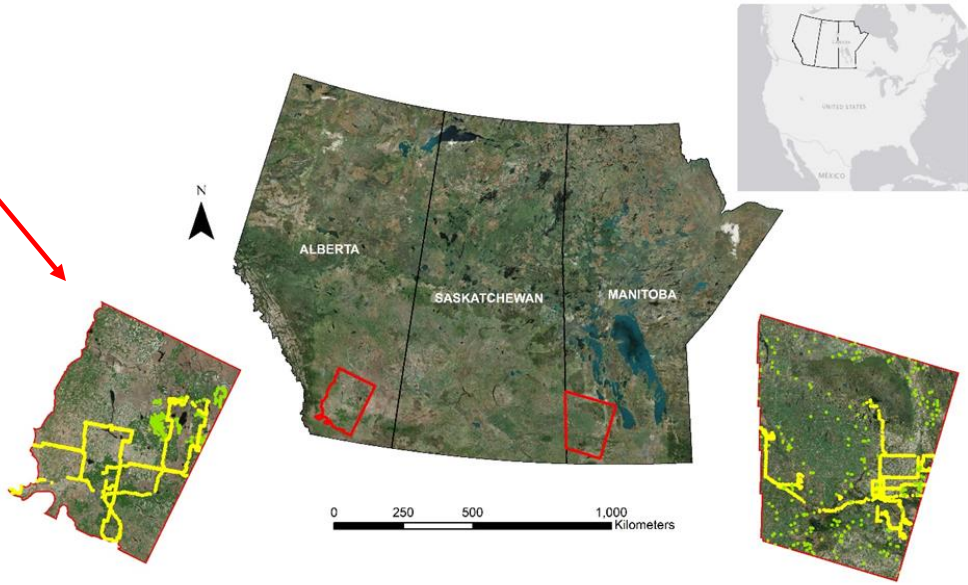




PCI Geomatics Inc

Supervised Classification: Training Data Collection

- Often, use field observations taken at GPS'd locations
- Generally use locations where at least 3x3 pixels could be used as training.
  - E.g. For Landsat, select a location where the given class extends over an area of at least 90x90 m.
- Find the pixel in the geo-referenced image with the same coordinates as the given field location; draw a 3x3 pixel (or larger) square centred on that location.



Vehicle routes (yellow) for crop/rangeland/forage class reference information collection and additional reference points (green) from provincial crop insurance datasets. Lindsay, King, Davidson, Daneshfar. 2018. J. Rangeland Ecology and Management. In Press.

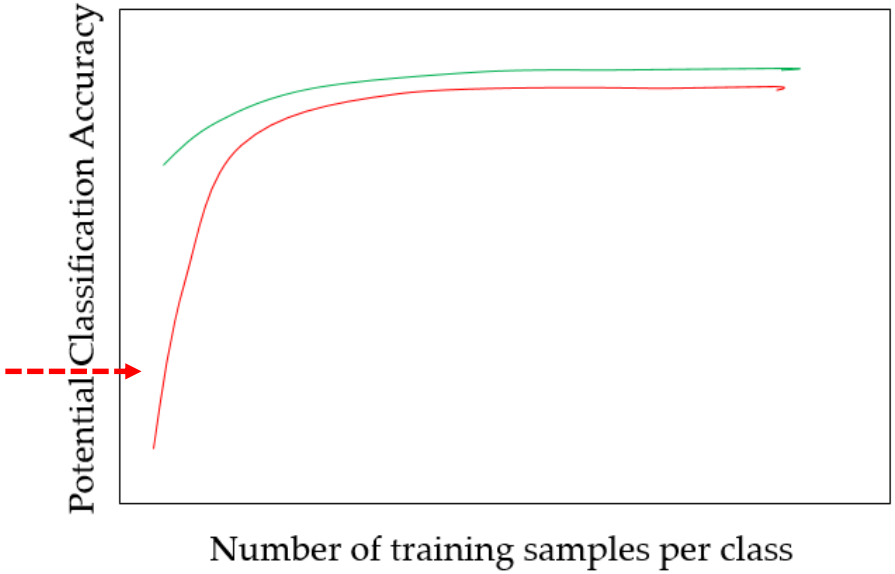
Supervised Classification: Training Data Collection

- Selection of training pixels is critical:
  - They must represent the variance of the class so should be selected throughout the area in which the class exists
  - Don't want too much variance, because training data distributions will overlap between classes more, and greater chance of misclassification
  - Don't want too little variance, because pixels outside the training data distributions will be mis-classified or not classified at all (left as a blank 'null' class)
  - Example on right shows variability in "rangeland" land use class in a Manitoba Landsat scene (see previous slide)

Rangeland class: E) upland native grassland; F) lowland native grassland (foreground) with meadow (middle ground); and G) native tall grass prairie under conservation with shrub encroachment. E. Lindsay, MSc

Number of Training Pixels per Class

- How many training pixels are needed for each class?
  - Many rules proposed. E.g. For 'n' bands (data variables) need >10n pixels (Jensen, 2015) for each class to compute a representative variance–covariance matrix required by parametric classification algorithms such as Maximum Likelihood.
    - With fewer pixels, mean and variance of the class may not be well represented and accuracy may be low
  - Some non parametric classifiers (e.g. Support Vector Machines), can achieve high accuracy with few training pixels (green line in graph).



Common Steps in Collecting and Analyzing Training Data

1. Assemble all information needed for the region to be mapped.
   - E.g. maps, aerial photographs.
2. Conduct field studies to acquire reference information over study area.
   - Determine # sample locations, perhaps using statistical guidelines
   - Decide on field logistics (travel routes, equipment, etc.) accounting for time and cost.
   - In field, note class type and more detailed observations, take photos, GPS location.
3. Delineate training areas using the previous guidelines.
4. Display and inspect frequency histograms of each training polygon.
5. Modify training polygons to eliminate bimodal frequency distributions, outliers and mixed pixels
   - Outline new training areas if necessary.
6. Assess the separability of classes (using separability metrics – later) and merge classes with low separability (poor potential for accurate classification), or edit training data to improve separability.
   - Note: It is easy to obtain high class separability by selecting training pixels only within the central more uniform areas of a given class, or by not sampling throughout the image. This results in narrow training data distributions that are unrealistic. Avoid this. It is best to select pixels that represent the variance of each class well.
7. When satisfied that training data are representative, proceed with the classification process.

Training Data Distributions

In 1 spectral band (as before)

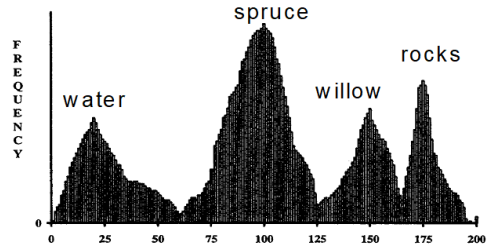
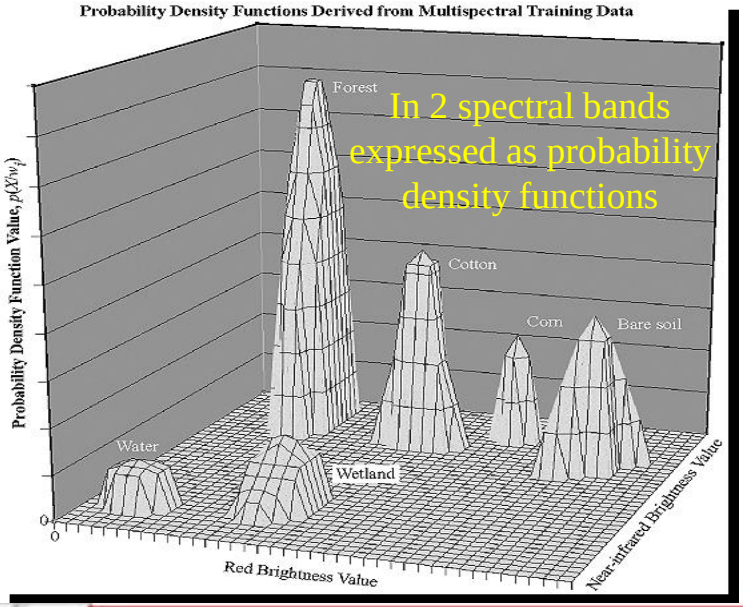
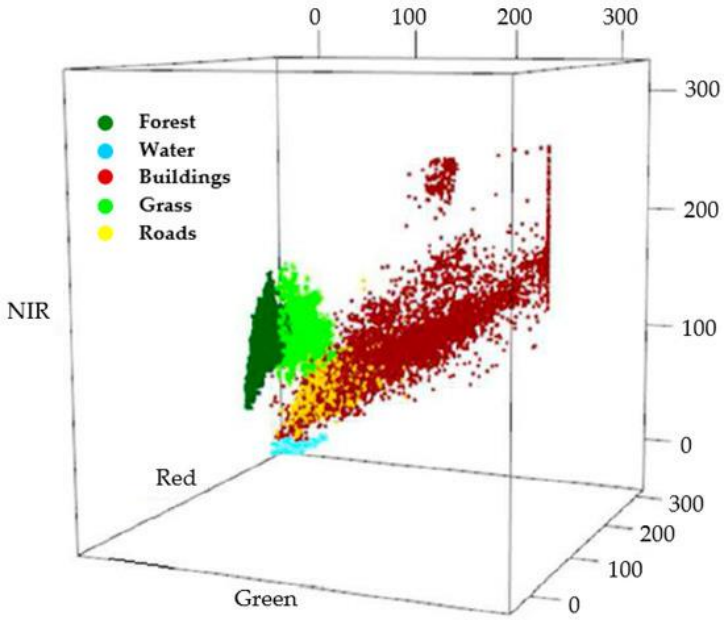


Figure 6.1. Histogram of near-infrared digital numbers from hypothetical image.



Jensen, 2015

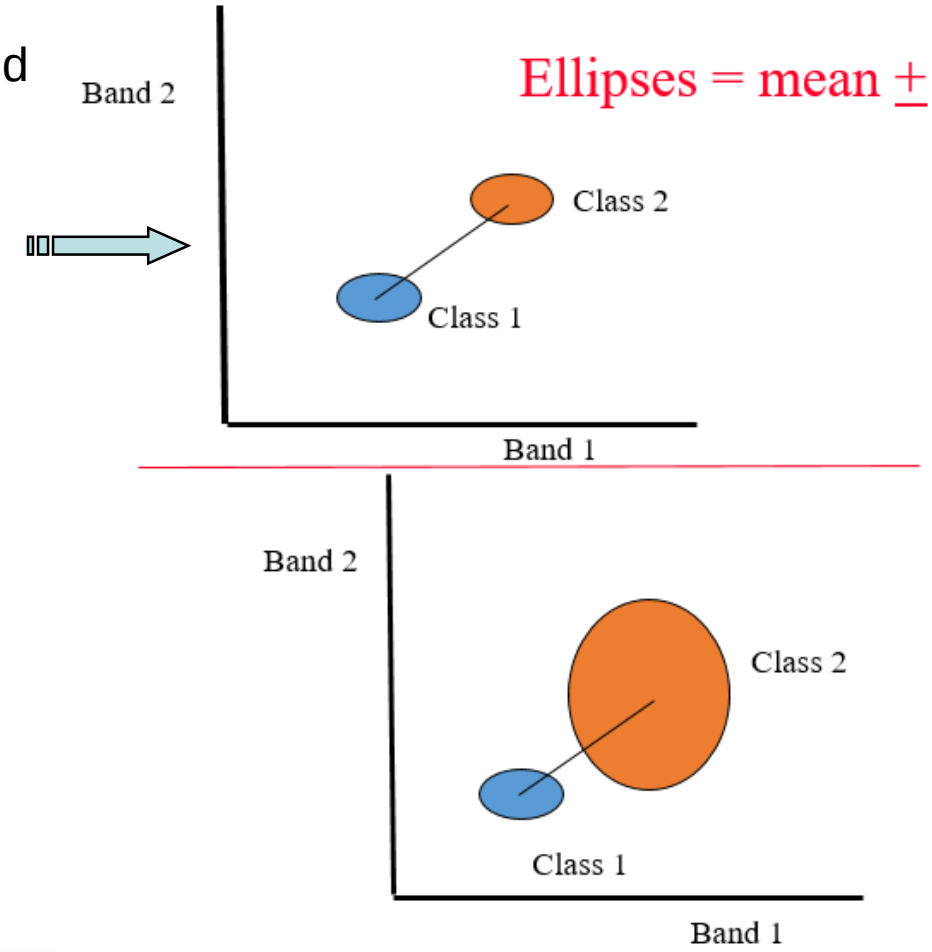
In 3 spectral bands



Hsiao and Cheng, 2016. Remote Sensing 8(9), 705

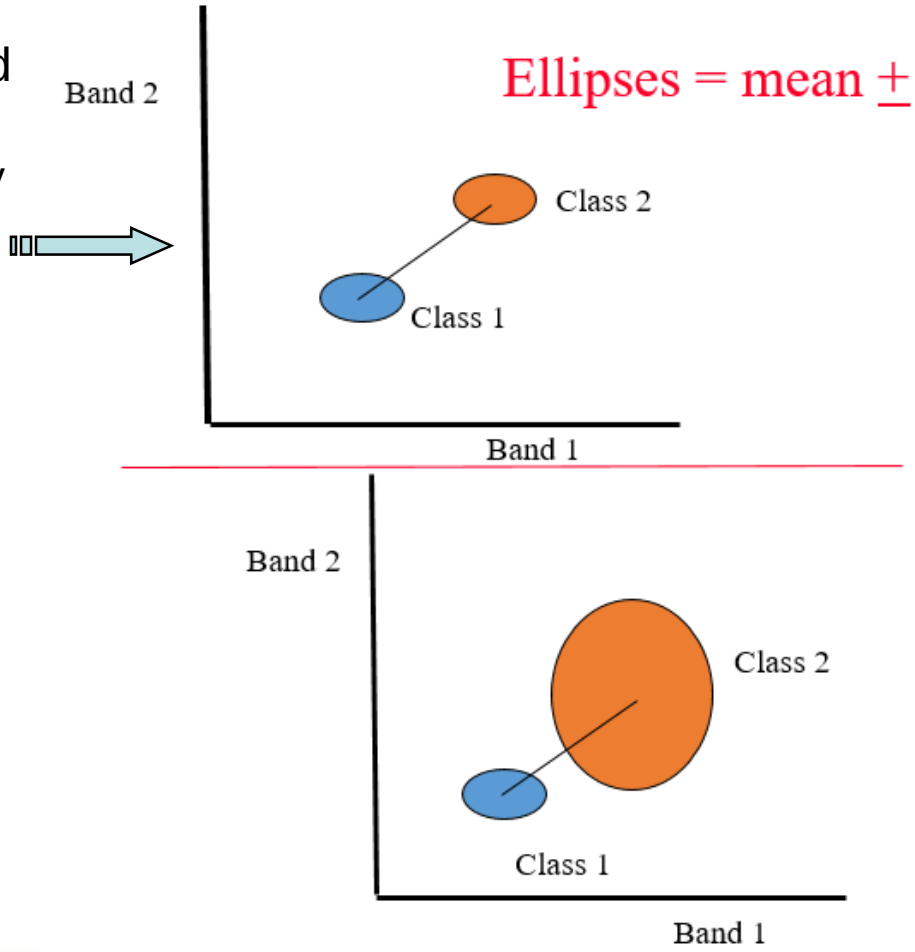
Analyzing Potential Classification Accuracy: Training Data Separability Analysis

- Which graph shows greater separability between Class 1 and Class 2?



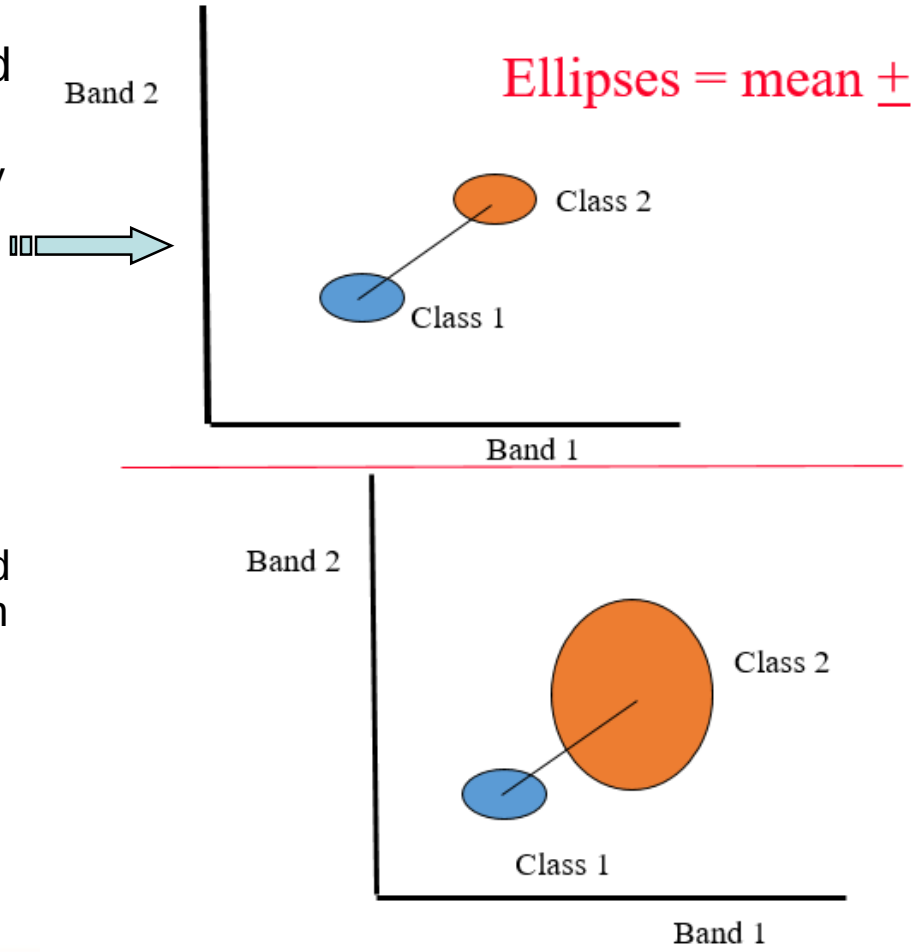
Analyzing Potential Classification Accuracy: Training Data Separability Analysis

- Which graph shows greater separability between Class 1 and Class 2?
  - Class 1 + 2 means separated by same Euclidian (straight line) distance in both graphs



Analyzing Potential Classification Accuracy: Training Data Separability Analysis

- Which graph shows greater separability between Class 1 and Class 2?
  - Class 1 + 2 means separated by same Euclidian (straight line) distance in both graphs
  - In bottom graph, we might think the classes are more similar because their data distributions are closer (less separable)
  - The bottom graph shows we should account for variance (and covariance) of the two classes in Bands 1 and 2.
  - What kind of measure can do that?



Analyzing Potential Classification Accuracy: Training Data Separability Analysis

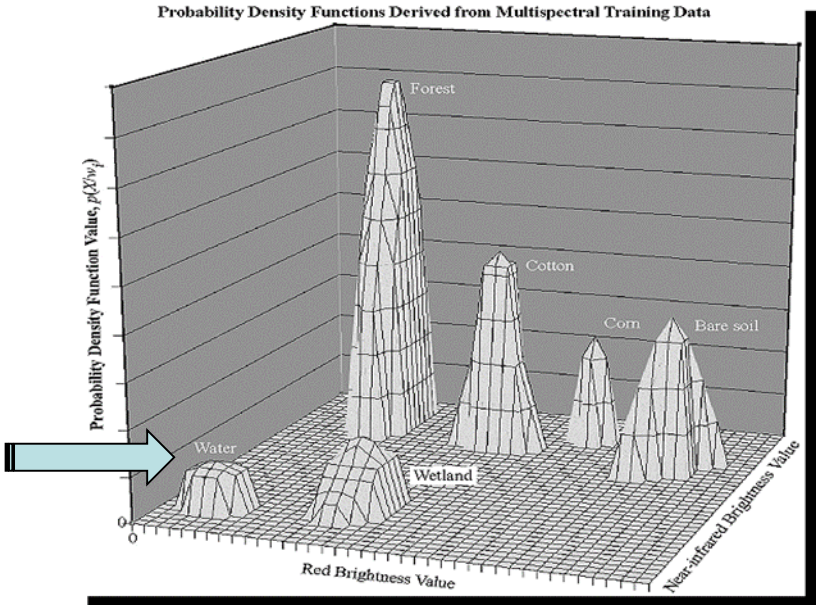
- Statistical 'separability' of a class with all other classes
  - Use a distance measure that is inversely weighted by the variance-covariance of the classes in all bands
    - i.e. when the class training data variance and covariance is high for two classes the separability distance value should be lower
  - E.g., Bhattacharyya distance (ranges from 0 (complete overlap in training data distributions) to 2.0 (training data distributions completely separated - no overlap))
    - >1.9 is good separability and indicates potentially high class accuracy
    - 1.7-1.9 indicates potential for moderate class accuracy (e.g. >70%)
    - < 1.7 indicates the classes will likely have low accuracy (e.g. <70%)

$$Bhat_{cd} = \frac{1}{8}(M_c - M_d)^T \left(\frac{V_c + V_d}{2}\right)^{-1} (M_c - M_d) + \frac{1}{2}\log_e \left[\frac{\left|\frac{V_c + V_d}{2}\right|}{\sqrt{|V_c| \cdot |V_d|}}\right]$$

Analyzing Potential Classification Accuracy: Training Data Separability Analysis

Once all training data collected:

- Check all class pair separabilities
  - Some algorithms produce a separability matrix
- Merge classes that have low separability
  - E.g., BHAT < 1.5-1.7
- Or, edit training pixels to remove non-representative pixels until separability is high
  - But as before, make sure they are representative of the class variance. The training data distributions in this example are probably too narrow as vegetated classes would normally have overlapping distributions.



Supervised Classification: The Classification Process

- Each pixel or segmented object in image is assigned to one of the classes based on rule(s).

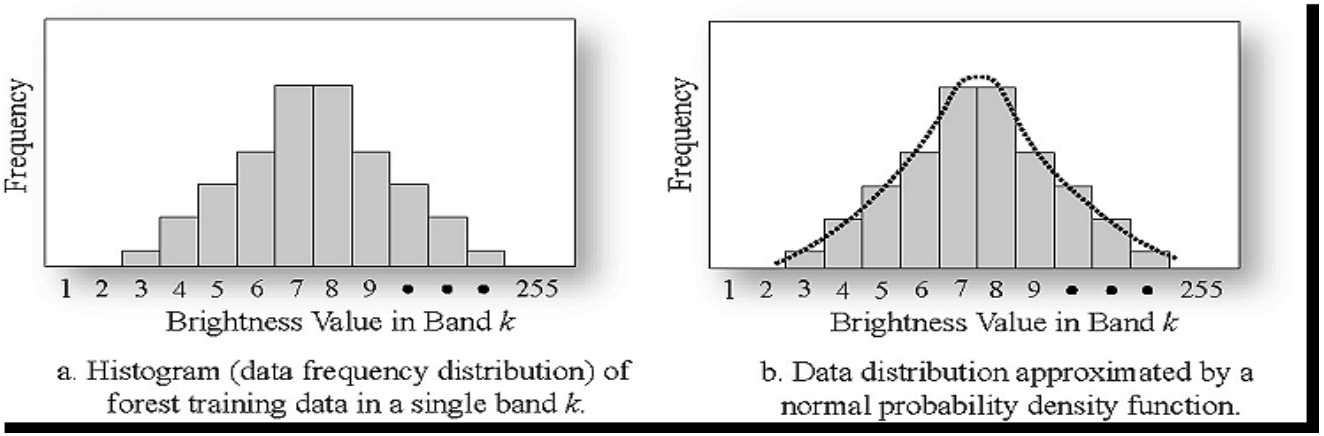
As on a previous slide:

- *Parametric classifiers*: The distribution of training pixel values for a land cover type is assumed to be normal (Gaussian) and can be represented by the mean and var-covariance matrix.
  - We will start with this type of classifier
- *Non-parametric classifiers:* have no assumptions
  - We will briefly look at decision trees and other non parametric classifiers.

Statistical Background for Parametric Classifiers based on Normal Distribution

- For *parametric classifiers* based on the normal distribution, training data in n bands are used to determine the probability density function (PDF) for each class in n-dimensional spectral space (or including other data types).

e.g. PDF for **1 band** from histogram of training data



a. Histogram (data frequency distribution) of forest training data in a single band *k*.

b. Data distribution approximated by a normal probability density function.

- PDFs must be normally distributed
  - They can then be represented by just their mean and variance
- For N bands, the multi-dimensional PDF is calculated using the variance - covariance matrix

Statistical Background for Parametric Classifiers based on Normal Distribution

- What is covariance?
  - How data variables (e.g. spectral bands) vary with each other.
    - As one band increases, does the other also increase by a predictable amount?

$$SP_{kl} = \sum_{i=1}^{n} (BV_{ik} - \mu_k)(BV_{il} - \mu_l)$$

$$\text{cov}_{kl} = \frac{SP_{kl}}{n-1}$$

- Is it related to correlation (which we saw with scatterplots)?
  - Yes, correlation is the covariance standardized by the product of the standard deviations of each variable
    - Covariance is not bounded
    - Correlation is standardized to a range of -1 to +1

$$r_{kl} = \frac{\text{cov}_{kl}}{s_k s_l}$$

Statistical Background for Parametric Classifiers based on Normal Distribution

Computation of Variance-Covariance Between Bands 1 and 2 of the Sample Data

	Band 1	(Band 1 x Band 2)	Band 2
Pixel 1	130	7,410	57
Pixel 2	165	5,775	35
Pixel 3	100	2,500	25
Pixel 4	135	6,750	50
Pixel 5	145	9,425	65
Total	675	31,860	232

Note: SP formula is a condensed form of formula on previous slide

$$SP_{12} = (31{,}860) - \frac{(675)(232)}{5} = 540$$

$$\text{cov}_{12} = \frac{540}{4} = 135$$

Jensen, 2015

Statistical Background for Parametric Classifiers based on Normal Distribution

- A variance-covariance matrix is the same as a correlation matrix, except:
  - Off-diagonal cells are the covariance between the two variables (bands) instead of the correlation between them
  - The diagonals are the variance, or often just the numerator; i.e., the sum of squares of the [training pixel values minus the mean] in each band.

	Band 1 (green)	Band 2 (red)	Band 3 (near-infrared)	Band 4 (near-infrared)
Band 1	SS_1	$cov_{1,2}$	$cov_{1,3}$	$cov_{1,4}$
Band 2	$cov_{2,1}$	SS_2	$cov_{2,3}$	$cov_{2,4}$
Band 3	$cov_{3,1}$	$cov_{3,2}$	SS_3	$cov_{3,4}$
Band 4	$cov_{4,1}$	$cov_{4,2}$	$cov_{4,3}$	SS_4

Jensen, 2015

$$\text{var}_k = \frac{\sum_{i=1}^{n} (BV_{ik} - \mu_k)^2}{n} \rightarrow SS$$

Supervised Classification: Maximum Likelihood Classifier

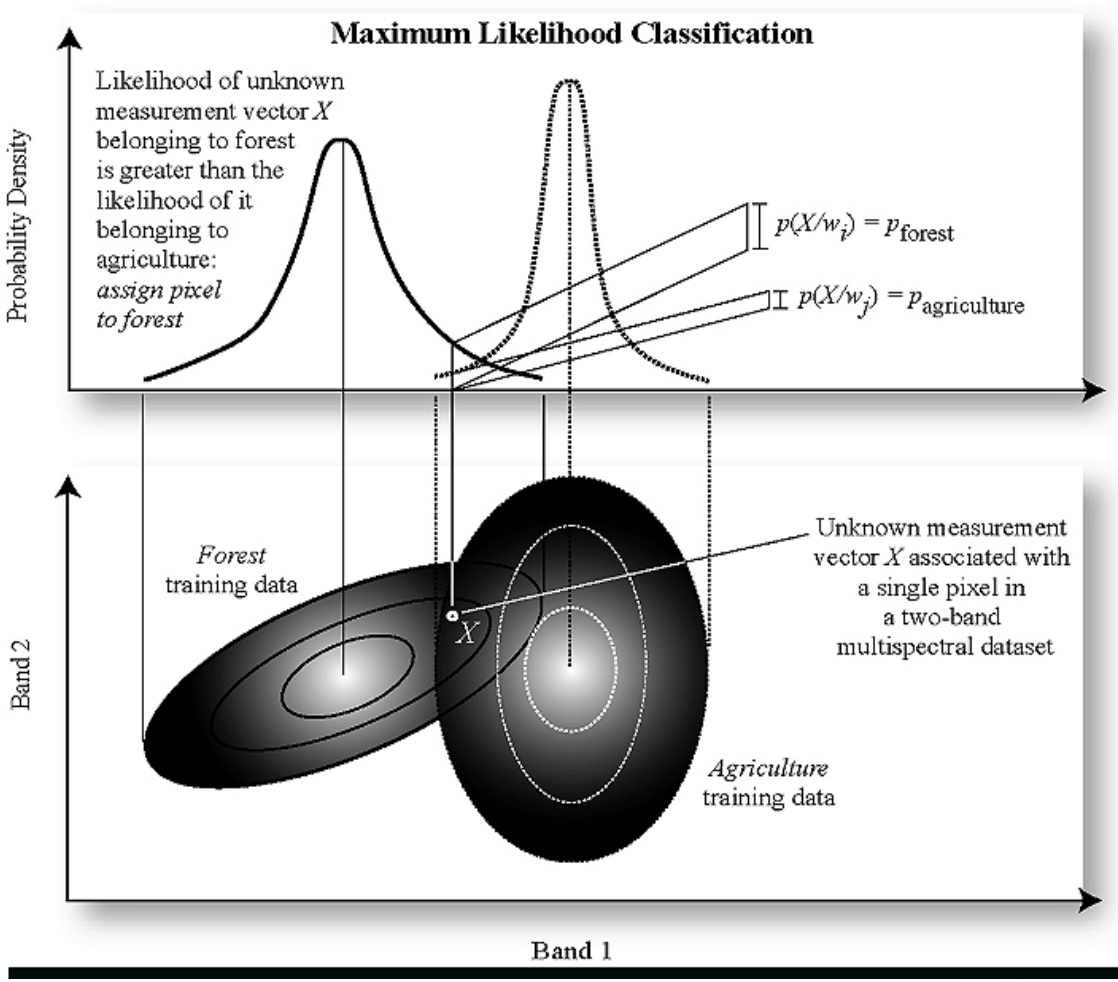
- For 1 band, the estimated probability of a pixel with value x being in class w_i (e.g., forest) is:

$$\hat{p}(x \mid w_i) = \frac{1}{(2\pi)^{\frac{1}{2}} s_i} \exp\left[-\frac{1}{2} \frac{(x - \bar{x}_i)^2}{s_i^2} \right]$$

where x is the brightness value (reflectance) of the unclassified pixel, and the mean ($\bar{x}_i$) and variance (s^2) of the class are known.

Maximum Likelihood Classifier

- Assign unknown pixel to class that it has the highest probability (maximum likelihood) of belonging to.
- Here, assign it to forest because probability is greater for agriculture



Maximum Likelihood Classifier

- Maximum Likelihood Decision Rule:
  - An unknown pixel's values are compared to each class probability density function and the probability that it belongs to each class is determined.
  - The pixel is then assigned to the class for which the probability is the highest

$$p(X \mid w_i) \cdot p(w_i) \geq p(X \mid w_j) \cdot p(w_j)$$

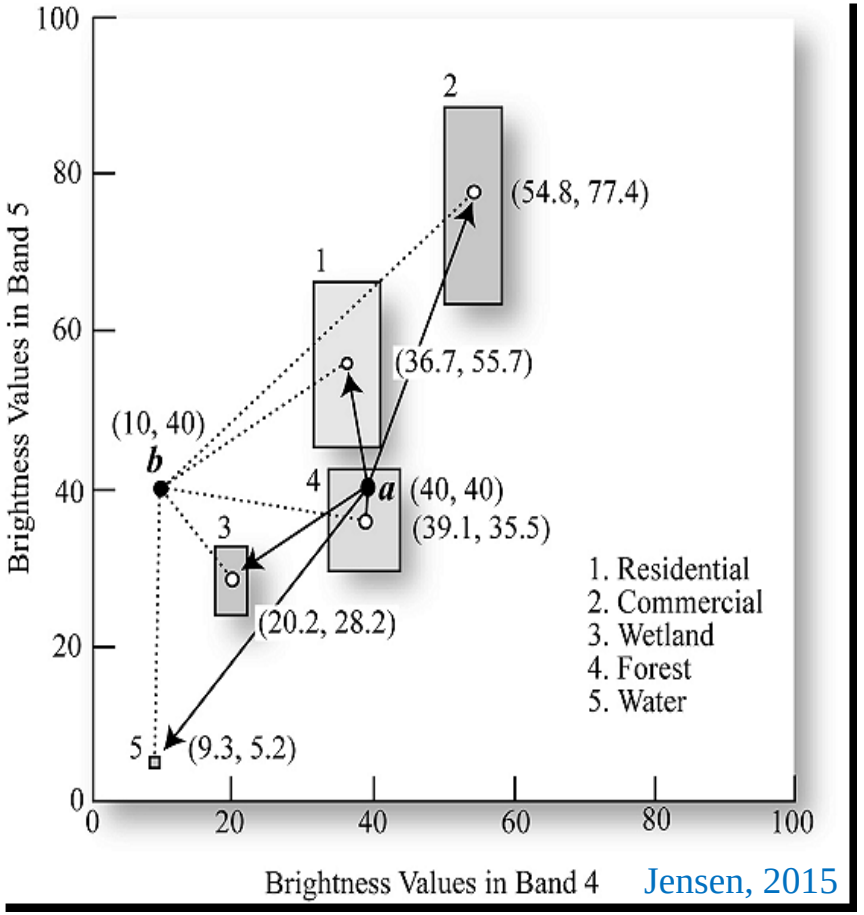
  - i.e. For a pixel to be assigned to class w_i, the probability of belonging to class w_i *times the probability that class w_i exists* (its proportional coverage in the area = the *a priori* probability of class w_i) must be greater than that for all other classes
  - If *a priori* probabilities [($p(w_i)$, $p(w_j)$…] are not known, assume all are equal and just compare probabilities of belonging to each class.
    - Sometimes a previous land cover map can be used to estimate the proportions covered by each class, i.e. the a priori probabilities
  - Pixels beyond the distributions of all classes may be assigned to an 'unclassified' (null) class, or all pixels can be classified

Supervised Classification: Non-parametric Classifiers

- Common *non-parametric* classifiers:
  - Parallelepiped
    - Possibly with bounds based on cut-off in data distributions (e.g. 1s, 1.5s)
    - Pixel a assigned to?
    - Pixel b assigned to?
  - Minimum distance to means
    - Straight line distance as for unsupervised clustering

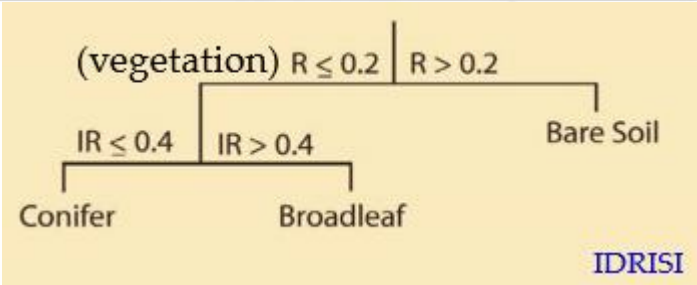
$$Dist = \sqrt{\left(BV_{ijk} - \mu_{ck}\right)^2 + \left(BV_{ijl} - \mu_{cl}\right)^2}$$

    - Pixel a assigned to?
    - Pixel b assigned to?



Jensen, 2015

Non-parametric Classifiers (cont'd)



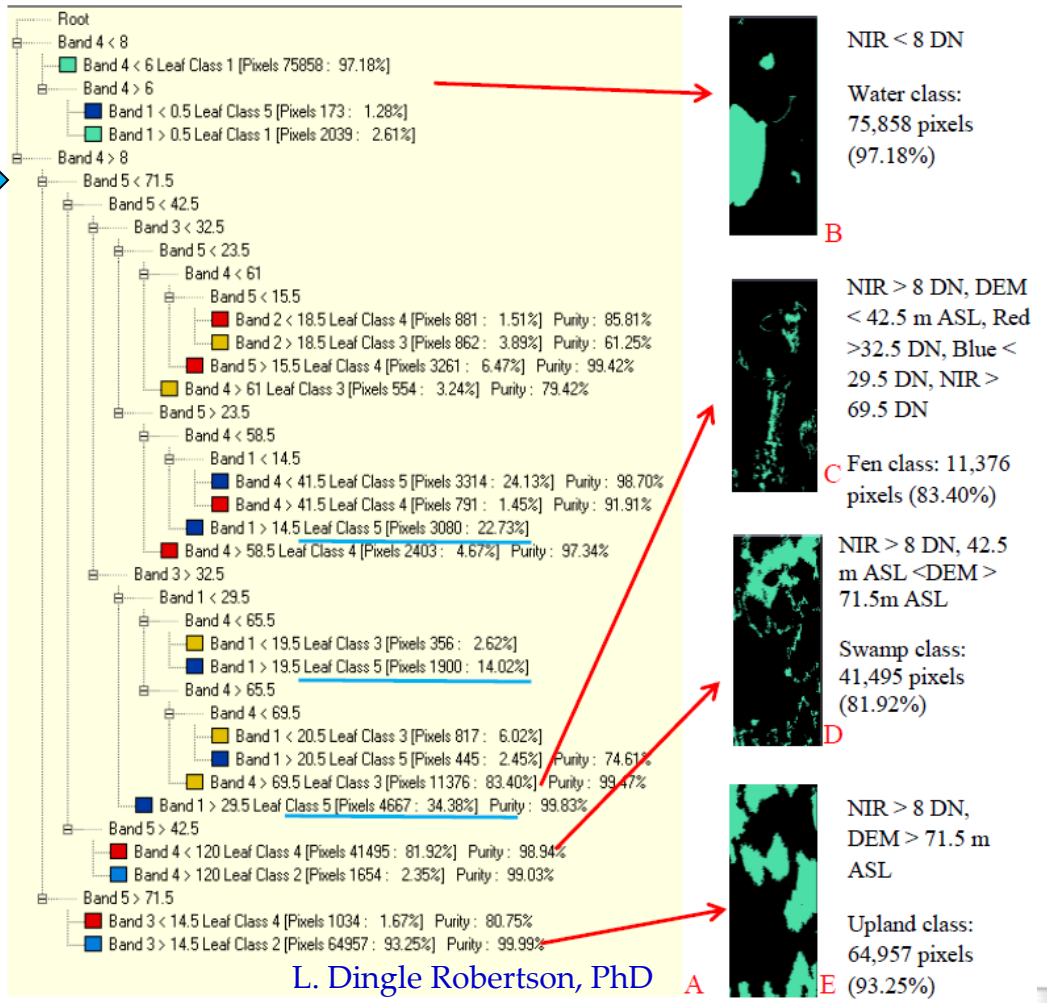
Decision Trees

- Type of machine learning classifier
- Set of hierarchical rules
- Start with all data and find the variable and variable value that best splits the data into 2 groups
  - All values of the training data for all variables are assessed
  - The variable and value selected provide the best improvement in "purity" over the original combined data.
    - i.e. the two subgroups are more "pure" (homogeneous) than any other two subgroups that could be derived using other data values.
  - Purity is often based on entropy. The variable and value that best decreases the entropy in the two subsets over the combined parent set is selected. Various metrics representing entropy decrease have been developed for such decision tree algorithms.
- In each group repeat the above process. Keep splitting groups until purity can no longer be improved by splitting.

Non-parametric Classifiers (cont'd)

- A decision tree for wetland classification produced from Worldview and elevation data

- In classification of an unknown pixel (with a value in each of the variables), the rules are followed down the tree to an end node
  - The pixel is assigned to the class of that end node.



Non-parametric Classifiers (cont'd)

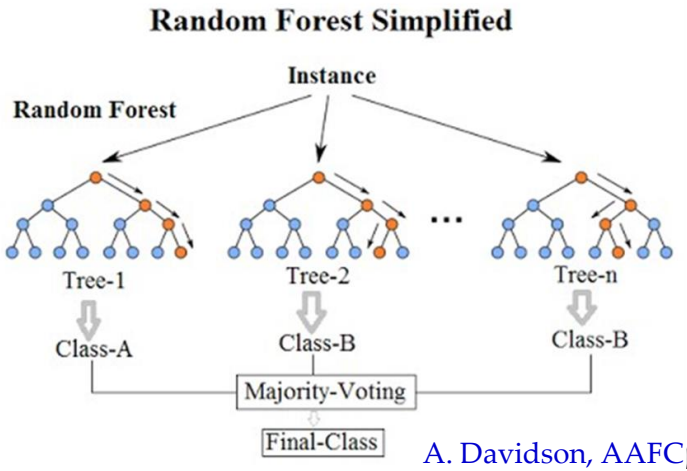
- Advantages of DTs:
  - Can handle non-parametric data types with different measurement scales and non-nonlinear relations between input data and target classes.
  - Computationally efficient. Does not need extensive design and training. Are easy to apply because fewer numbers of parameters need to be estimated.
  - Provides a hierarchical structure that is transparent. Rules are easy to interpret and used to derive a physical understanding of the classification process.
  - Can adapt when new learning data are provided.
- Disadvantages of DTs:
  - Sensitivity to too many variables, noisy data and over-fitting.
  - End nodes may not represent the classes of interest or may represent too detailed sub-classes; the tree may need to be pruned back.

A. Davidson, AAFC

Non-parametric Classifiers (cont'd)

Random Forest Classifier

- A set of many decision trees implemented in sequence.
  - An "ensemble" classifier; often 100 to 1000 trees.
- Each tree is derived using a random subset (e.g. 70%) of the training (reference) data
  - The remaining data are used to estimate classification error
  - These "Out-of-Bag" (OOB) data are returned to the training set; a new random selection is done for the next tree.
  - Overall OOB error is determined from aggregating OOB errors over all trees
    - OOB error can be optimistic, so generally it is best to set accuracy assessment data aside from the start
- If many input variables (sometimes tens or > 100), since all can't be processed in each of many trees, a subset of variables is randomly selected
  - m_{try} parameter specifies proportion to be selected each time
- In the end, after all 'n' trees are produced, each unknown pixel is passed through the trees and classified 'n' times. The class which represents the majority is assigned to the pixel.



A. Davidson, AAFC

Non-parametric Classifiers (cont'd)

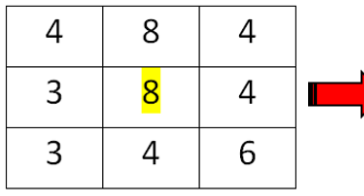
- Random forests also has capability to determine the importance of each input variable to the overall classification
  - Metrics of entropy reduction or information gain are aggregated over all trees for each variable.
    - Variables are then ranked based on these metric values.
  - It has become a common means to determine the optimal variable subset from among many possible variables.

- Other Non-parametric classifiers: Neural networks, Support Vector Machines, Fuzzy, etc.
  - Can learn about them in advanced remote sensing classes

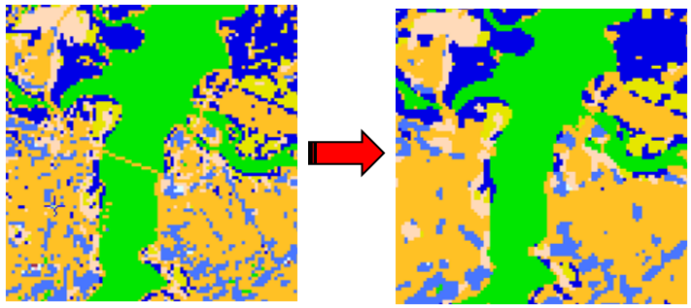
Post-Classification Smoothing

- Classification maps generated for individual pixels (i.e., not objects) often have a speckled appearance due to spectral variability between pixels
- Can smooth the output map to show the dominant classes more prominently.
  - Can't use an average filter because the cluster map pixel values correspond to class labels and are arbitrary
    - They are not continuous variables such as image brightness.
  - A mode (most frequently occurring class) filter is often used.
- This produces a nicer looking map, but always ask: Is important detail being lost?
  - E.g., In agricultural classification, may want one class per unit area (field)
  - But if, for example, you are classifying trees as bird habitat in farmlands, you may not want to smooth the map too much.
  - Relates to the selection of imagery (spatial and spectral resolution), the specification of the MMU, the attributes (classes) to be mapped and their spatial variability.

4	8	4
3	8	4
3	4	6



4	8	4
3	4	4
3	4	6



Classification Assessment

- Thematic map produced should be assessed for quality
  - For visual representation of classes
    - Are they where they should be?
    - Are there a lot of mixed class assignments for any class – produces speckled map.
    - Are there too many ‘unclassified’ pixels (if had a ‘null’ class)?
  - For accuracy: test pixels of known classes are cross-referenced with the thematic map pixel classes
    - Frequency analysis of %correct in each class
- If map is not adequate, training data can be edited, classes can be aggregated, other information (variables) can be added, etc. and the classification process repeated.

Thematic Map Error (Confusion) Matrix

- Generated from independent and random sample points of reference vs. map data.

Reference Class

Map Class	1	2	3	4	Sum	UA (%)
1						
2						
3						
4						
Sum						
PA (%)						

PA = 'Producer's' accuracy; UA = 'User's' (or 'Consumer's) accuracy

Note: rows and columns may be inversed or the table may have a different format in some software (e.g., PCI) or reports/papers.

Thematic Map Error (Confusion) Matrix

Reference Class

Map Class	1	2	3	4	Sum	UA (%)
1	20	3	5	0	28	71.4
2	10	45	3	8	66	68.2
3	0	3	72	23	98	73.4
4	2	5	21	46	74	62.1
Sum	32	56	101	77	266	
PA (%)	62.5	80.4	71.3	59.7		

Overall accuracy = # correct/total = Sum (diagonal)/total = 183/266 = 68.8%

PA = 100% - % errors of omission (i.e. errors in reference data)

UA = 100% - % errors of commission (i.e. errors in map data)

Other Accuracy Assessment Metrics: e.g. Kappa Coefficient

- The proportion of classification accuracy (0.0 – 1.0) greater than that which could be achieved by each reference pixel to randomly to one of the classes.

Reference Class

Map Class	1	2	3	4	Sum
1	20	3	5	0	28
2	10	45	3	8	66
3	0	3	72	23	98
4	2	5	21	46	74
Sum	32	56	101	77	266

$e = \Sigma \{Nrow \times Ncol / Ntot\}$

$\therefore \quad e_1 = 28(32)/266 = 3.37$

$e_2 = 66(56)/266 = 13.89$

$e_3 = 98(101)/266 = 37.21$

$e_4 = 74(77)/266 = 21.42$

$\therefore \quad e_{tot} = 3.37 + 13.89 + 37.21 + 21.42 = 75.89$

$$K = \frac{\text{\# correct - \# expected correct by chance}}{\text{\# samples - \# expected correct by chance}}$$

$K = (d\text{-}e) / (Ntot\text{-}e)$

$= (183 - 75.9) / (266 - 75.9)$

$= 0.56$

Accuracy Assessment

- Several other ways to use the Error Matrix to assess accuracy. E.g.,
  - 95% confidence intervals can be calculated for Overall Accuracy and for UA and PA of each class
    - Gives range within which we can state the class or overall accuracy with 95% confidence
  - Other accuracy metrics have been developed because Kappa has some limitations
  - Use errors in classes to determine range of area coverage for each class
- Much literature on accuracy assessment and using error estimates
  - E.g., Olofsson, et al. 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42-57.

Improving Map Accuracy

In iterative testing, can do any of the following:

- Edit training data to remove outliers
  - or use only a portion of the training data; e.g., +/- 2 std. dev.
- Merge similar classes to an aggregated class
- Eliminate a consistently poorly classified class
- Add more information to classifier
  - New spectral data, other remote sensing data such as radar
  - Image spatial information (e.g. texture metrics)
  - DEM information
  - Other GIS types of variables
- Change *a priori* probabilities if maximum likelihood classifier is used and have the information

Improving Map Accuracy

- Use a classifier that:
  - is more robust if classes non-normal or if data non-ratio
    - E.g, decision tree, neural network, expert classifier
  - gives greater precision of sub-pixel composition
    - E.g, fuzzy, spectral unmixing
  - classifies segmented objects and not individual pixels
- Process data to improve quality
  - Do atmospheric and/or topographic brightness correction, or try to improve on them if already done
  - Reduce spatial non-uniformity in image brightness/reflectance (optical, BRDF)
  - Transform data (PCA, vegetation indices, etc.)
    - can often reducethe above effects)
  - Reduce noise in image and other input data using filtering