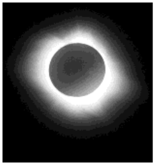
Thematic Information Extraction: Pattern Recognition

Dr. John R. Jensen
Department of Geography
University of South Carolina
Columbia, SC 29208



Classification

Multispectral classification may be performed using a variety of methods, including:

- algorithms based on *parametric* and *nonparametric* statistics that use ratio- and interval-scaled data and *nonmetric* methods that can also incorporate nominal scale data;
- the use of *supervised* or *unsupervised* classification logic;,
- the use of *hard* or *soft* (*fuzzy) set classification* logic to create hard or fuzzy thematic output products;
- the use of *per-pixel* or *object-oriented classification* logic, and
- *hybrid* approaches.

Jensen, 2005

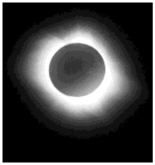
Classification

Parametric methods such as maximum likelihood classification and unsupervised clustering assume normally distributed remote sensor data and knowledge about the forms of the underlying class density functions.

Nonparametric methods such as nearest-neighbor classifiers, fuzzy classifiers, and neural networks may be applied to remote sensor data that are not normally distributed and without the assumption that the forms of the underlying densities are known.

Nonmetric methods such as rule-based decision tree classifiers can operate on both real-valued data (e.g., reflectance values from 0 to 100%) and nominal scaled data (e.g., class 1 = forest; class 2 = agriculture).

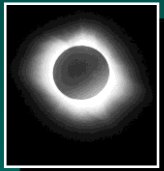
Jensen, 2005



Supervised Classification

In a *supervised classification*, the identity and location of some of the land-cover types (e.g., urban, agriculture, or wetland) are known a priori through a combination of fieldwork, interpretation of aerial photography, map analysis, and personal experience. The analyst attempts to locate specific sites in the remotely sensed data that represent homogeneous examples of these known land-cover types. These areas are commonly referred to as *training sites* because the spectral characteristics of these known areas are used to train the classification algorithm for eventual land-cover mapping of the remainder of the image. Multivariate statistical parameters (means, standard deviations, covariance matrices, correlation matrices, etc.) are calculated for each training site. Every pixel both within and outside the training sites is then evaluated and assigned to the class of which it has the highest likelihood of being a member.

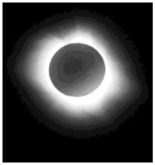
Jensen, 2005



Unsupervised Classification

In an *unsupervised classification*, the identities of land-cover types to be specified as classes within a scene are not generally known *a priori* because ground reference information is lacking or surface features within the scene are not well defined. The computer is required to group pixels with similar spectral characteristics into unique clusters according to some statistically determined criteria. The analyst then re-labels and combines the spectral clusters into information classes.

Jensen, 2005



Hard vs. Fuzzy Classification

Supervised and unsupervised classification algorithms typically use *hard classification* logic to produce a classification map that consists of hard, discrete categories (e.g., forest, agriculture).

Conversely, it is also possible to use *fuzzy set classification* logic, which takes into account the heterogeneous and imprecise nature of the real world.

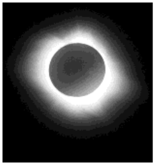
Jensen, 2005

Per-pixel vs. Object-oriented Classification

In the past, most digital image classification was based on processing the entire scene pixel by pixel. This is commonly referred to as *per-pixel classification.*

Object-oriented classification techniques allow the analyst to decompose the scene into many relatively homogenous image objects (referred to as patches or segments) using a multi-resolution image segmentation process. The various statistical characteristics of these homogeneous image objects in the scene are then subjected to traditional statistical or fuzzy logic classification. Object-oriented classification based on image segmentation is often used for the analysis of high-spatial-resolution imagery (e.g., 1 × 1 m Space Imaging IKONOS and 0.61 × 0.61 m Digital Globe QuickBird).

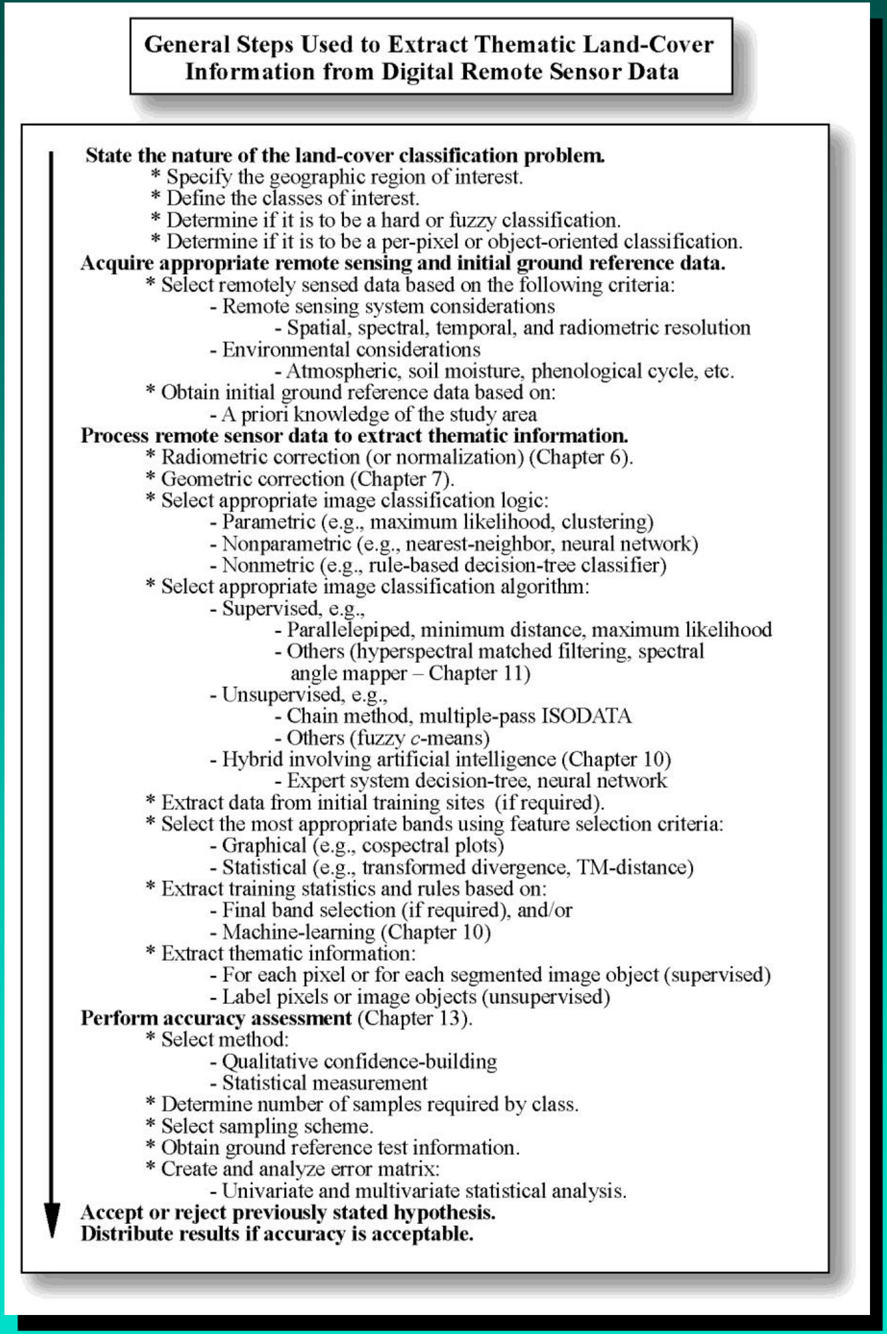
Jensen, 2005



Be Careful

No pattern classification method is inherently superior to any other. The nature of the classification problem, the biophysical characteristics of the study area, the distribution of the remotely sensed data (e.g., normally distributed), and *a priori* knowledge determine which classification algorithm will yield useful results. Duda et al. (2001) provide sound advice: *"We should have a healthy skepticism regarding studies that purport to demonstrate the overall superiority of a particular learning or recognition algorithm."*

Jensen, 2005



General Steps Used to Extract Thematic Land-Cover Information from Digital Remote Sensor Data

State the nature of the land-cover classification problem.

- * Specify the geographic region of interest.
- * Define the classes of interest.
- * Determine if it is to be a hard or fuzzy classification.
- * Determine if it is to be a per-pixel or object-oriented classification.

Acquire appropriate remote sensing and initial ground reference data.

- * Select remotely sensed data based on the following criteria:
    - Remote sensing system considerations
        - Spatial, spectral, temporal, and radiometric resolution
    - Environmental considerations
        - Atmospheric, soil moisture, phenological cycle, etc.
- * Obtain initial ground reference data based on:
    - A priori knowledge of the study area

Process remote sensor data to extract thematic information.

- * Radiometric correction (or normalization) (Chapter 6).
- * Geometric correction (Chapter 7).
- * Select appropriate image classification logic:
    - Parametric (e.g., maximum likelihood, clustering)
    - Nonparametric (e.g., nearest-neighbor, neural network)
    - Nonmetric (e.g., rule-based decision-tree classifier)
- * Select appropriate image classification algorithm:
    - Supervised, e.g.,
        - Parallelepiped, minimum distance, maximum likelihood
        - Others (hyperspectral matched filtering, spectral angle mapper – Chapter 11)
    - Unsupervised, e.g.,
        - Chain method, multiple-pass ISODATA
        - Others (fuzzy *c*-means)
    - Hybrid involving artificial intelligence (Chapter 10)
        - Expert system decision-tree, neural network
- * Extract data from initial training sites (if required).
- * Select the most appropriate bands using feature selection criteria:
    - Graphical (e.g., cospectral plots)
    - Statistical (e.g., transformed divergence, TM-distance)
- * Extract training statistics and rules based on:
    - Final band selection (if required), and/or
    - Machine-learning (Chapter 10)
- * Extract thematic information:
    - For each pixel or for each segmented image object (supervised)
    - Label pixels or image objects (unsupervised)

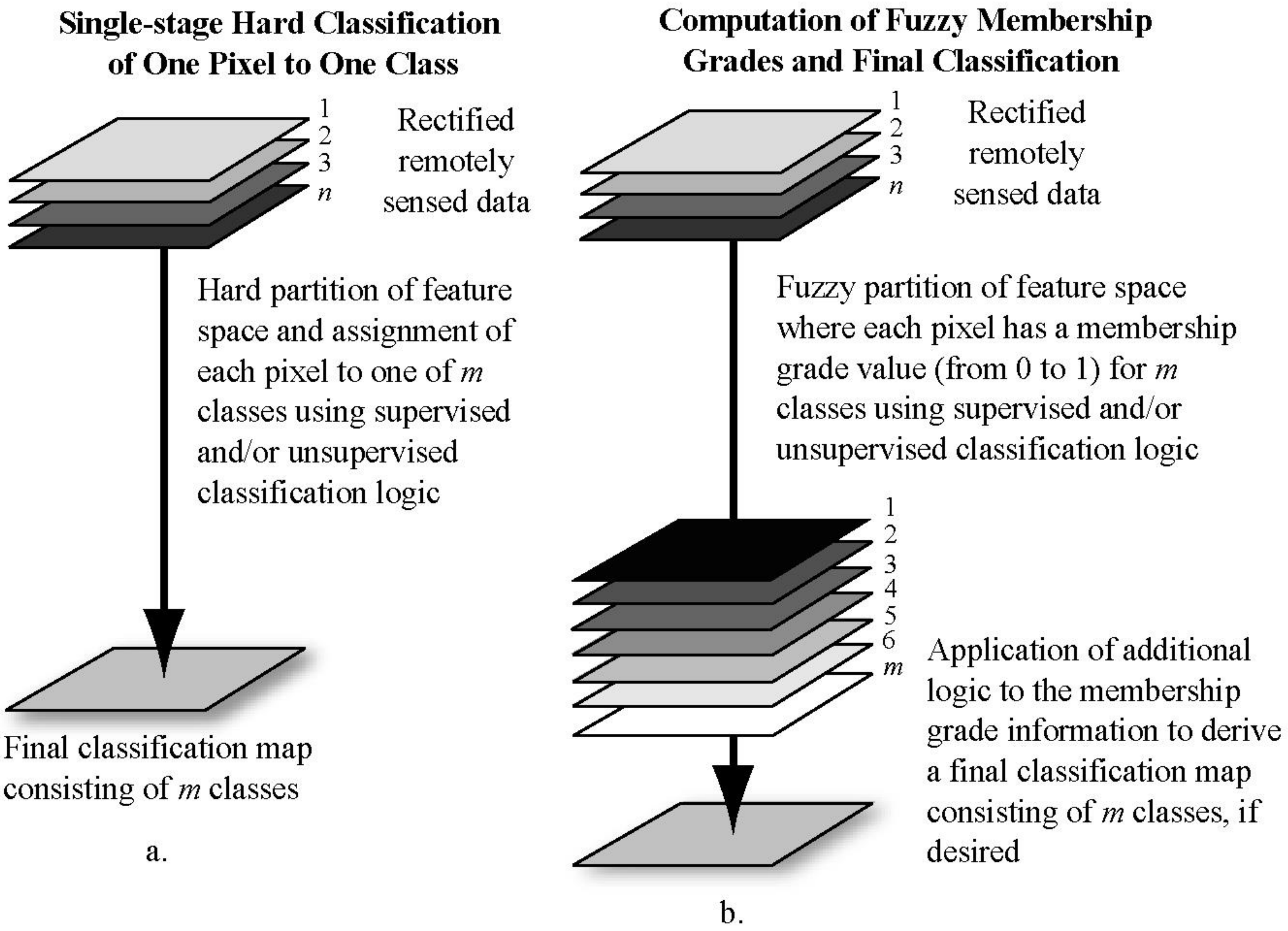
Perform accuracy assessment (Chapter 13).

- * Select method:
    - Qualitative confidence-building
    - Statistical measurement
- * Determine number of samples required by class.
- * Select sampling scheme.
- * Obtain ground reference test information.
- * Create and analyze error matrix:
    - Univariate and multivariate statistical analysis.

Accept or reject previously stated hypothesis.
Distribute results if accuracy is acceptable.

Jensen, 2005

Classification of Remotely Sensed Data Based on Hard versus Fuzzy Logic



Single-stage Hard Classification of One Pixel to One Class

Rectified remotely sensed data

Hard partition of feature space and assignment of each pixel to one of *m* classes using supervised and/or unsupervised classification logic

Final classification map consisting of *m* classes

a.

Computation of Fuzzy Membership Grades and Final Classification

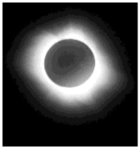
Rectified remotely sensed data

Fuzzy partition of feature space where each pixel has a membership grade value (from 0 to 1) for *m* classes using supervised and/or unsupervised classification logic

Application of additional logic to the membership grade information to derive a final classification map consisting of *m* classes, if desired

b.

Jensen, 2005



Land-use and Land-cover Classification Schemes

Land cover refers to the type of material present on the landscape (e.g., water, sand, crops, forest, wetland, human-made materials such as asphalt).

Land use refers to what people do on the land surface (e.g., agriculture, commerce, settlement).

The pace, magnitude, and scale of human alterations of the Earth's land surface are unprecedented in human history. Therefore, land-cover and land-use data are central to such United Nations' *Agenda 21* issues as combating deforestation, managing sustainable settlement growth, and protecting the quality and supply of water resources.

Jensen, 2005

Land-use and Land-cover Classification Schemes

All classes of interest must be selected and defined carefully to classify remotely sensed data successfully into land-use and/or land-cover information. This requires the use of a *classification scheme* containing *taxonomically* correct definitions of classes of information that are organized according to logical criteria. If a hard classification is to be performed, then the classes in the classification system should normally be:

- *mutually exclusive*,
- *exhaustive*, and
- *hierarchical*.

Jensen, 2005

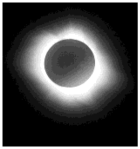
Land-use and Land-cover Classification Schemes

* *Mutually exclusive* means that there is no taxonomic overlap (or fuzziness) of any classes (i.e., deciduous forest and evergreen forest are distinct classes).

* *Exhaustive* means that all land-cover classes present in the landscape are accounted for and none have been omitted.

* *Hierarchical* means that sublevel classes (e.g., single-family residential, multiple-family residential) may be hierarchically combined into a higher- level category (e.g., residential) that makes sense. This allows simplified thematic maps to be produced when required.

Jensen, 2005



Land-use and Land-cover Classification Schemes

It is also important for the analyst to realize that there is a fundamental difference between *information* classes and *spectral* classes.

* *Information classes* are those that human beings define.

* *Spectral classes* are those that are inherent in the remote sensor data and must be identified and then labeled by the analyst.

Jensen, 2005

Land-use and Land-cover Classification Schemes

Certain *hard classification schemes* can readily incorporate land-use and/or land-cover data obtained by interpreting remotely sensed data, including the:

- American Planning Association *Land-Based Classification System* which is oriented toward detailed land-use classification;

- United States Geological Survey *Land-Use/Land-Cover Classification System for Use with Remote Sensor Data* and its adaptation for the U.S. National Land Cover Dataset and the NOAA Coastal Change Analysis Program (C-CAP);

- U.S. Department of the Interior Fish & Wildlife Service *Classification of Wetlands and Deepwater Habitats of the United States*;

- U.S. *National Vegetation and Classification System*;

- International Geosphere-Biosphere Program *IGBP Land Cover Classification System* modified for the creation of MODIS land-cover products

American Planning Association Land-Based Classification System

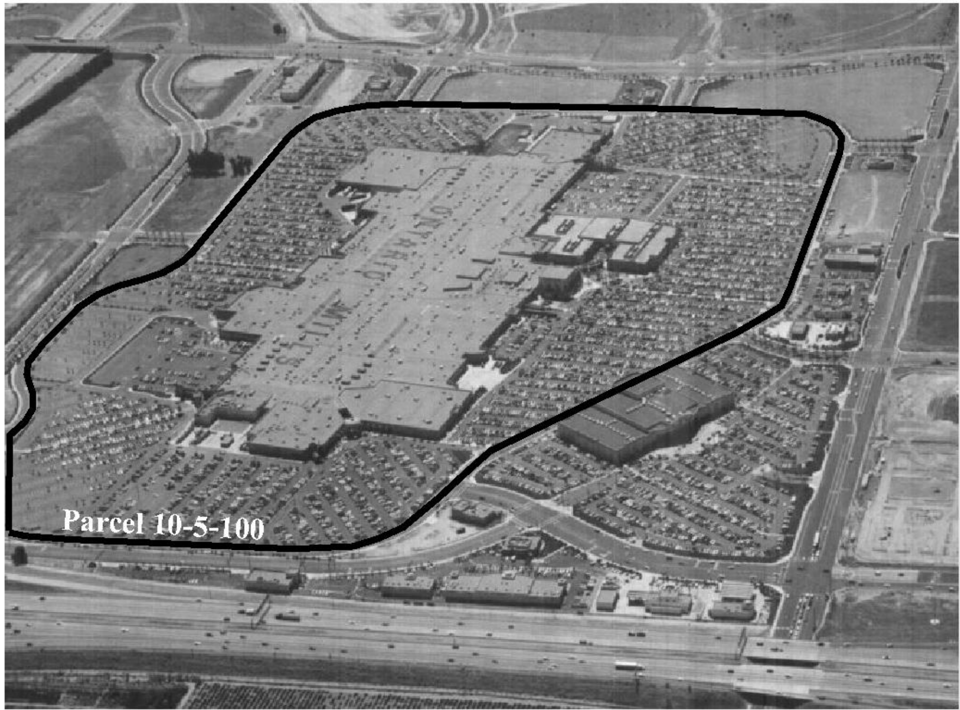


Table 1: Activity

Parcel ID	Activity	Description
10-5-100	2100	shopping
10-5-100	2200	restaurant
10-5-100	6600	social, religious assembly
10-5-100	2100	furniture
10-5-100	5210	vehicular parking
etc.	etc.	etc.

Table 2: Function

Function	Description
2110	retail sales and services
2510	full-service restaurant
6620	religious institutions
2121	furniture
5200	parking facilities
etc.	etc.

Table 3: Structure

Parcel ID	Structure	Description
10-5-100	2500	mall, shopping center

The *Land-Based Classification System* (LBCS) contains detailed definitions of urban/ suburban land use. The system incorporates information derived *in situ* and using remote sensing techniques. This is an oblique aerial photograph of a mall in Ontario, CA. Hypothetical activity and structure codes associated with this large parcel are identified. Site development and ownership information attribute tables are not shown (courtesy American Planning Association).

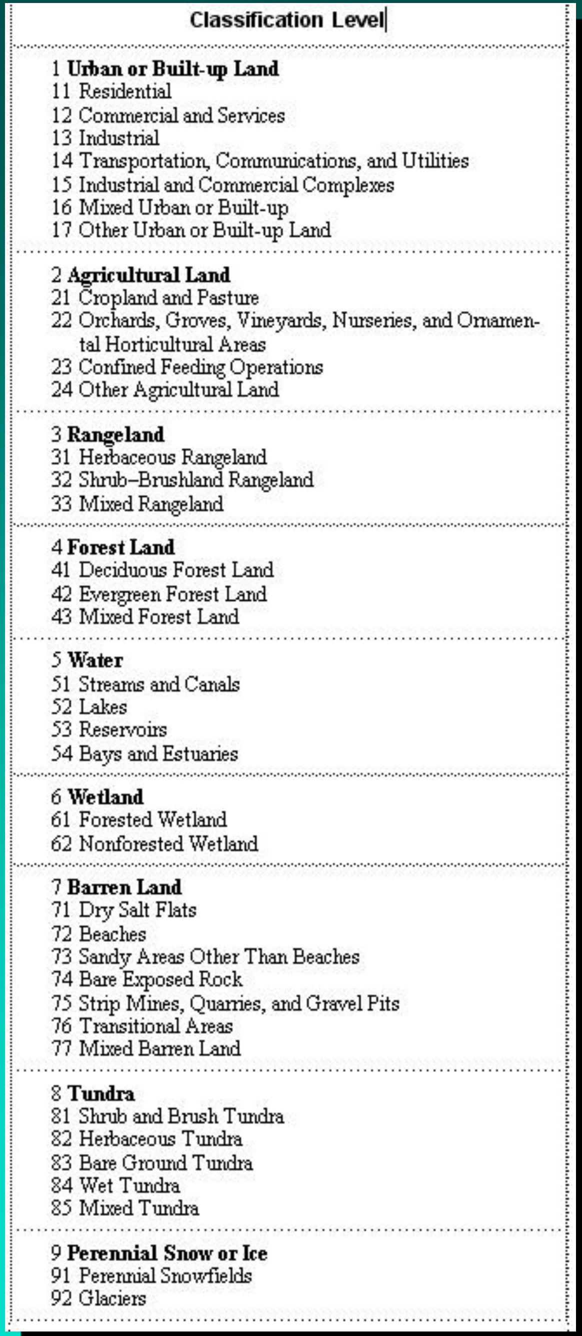
Jensen, 2005

U.S. Geological Survey's *Land-Use/Land-Cover Classification System for Use with Remote Sensor Data*

The U.S. Geological Survey's *Land-Use/Land-Cover Classification System for Use with Remote Sensor Data* is a resource-oriented land-cover classification system in contrast with people or activity land-use classification systems such as the APA's *Land-Based Classification System*. The USGS rationale is that *"although there is an obvious need for an urban-oriented land-use classification system, there is also a need for a resource-oriented classification system whose primary emphasis would be the remaining 95 percent of the United States land area."* The USGS system addresses this need with 8 of the 9 Level I categories that treat land area that is not in urban or built-up categories. The system is designed to be driven primarily by the interpretation of remote sensor data obtained at various scales and resolutions and not data collected *in situ*.

Jensen, 2005

Four Levels of the U.S. Geological Survey *Land-Use/Land-Cover Classification System for Use with Remote Sensor Data* and the type of remotely sensed data typically used to provide the information.



Classification Level

1 **Urban or Built-up Land**
11 Residential
12 Commercial and Services
13 Industrial
14 Transportation, Communications, and Utilities
15 Industrial and Commercial Complexes
16 Mixed Urban or Built-up
17 Other Urban or Built-up Land

2 **Agricultural Land**
21 Cropland and Pasture
22 Orchards, Groves, Vineyards, Nurseries, and Ornamental Horticultural Areas
23 Confined Feeding Operations
24 Other Agricultural Land

3 **Rangeland**
31 Herbaceous Rangeland
32 Shrub–Brushland Rangeland
33 Mixed Rangeland

4 **Forest Land**
41 Deciduous Forest Land
42 Evergreen Forest Land
43 Mixed Forest Land

5 **Water**
51 Streams and Canals
52 Lakes
53 Reservoirs
54 Bays and Estuaries

6 **Wetland**
61 Forested Wetland
62 Nonforested Wetland

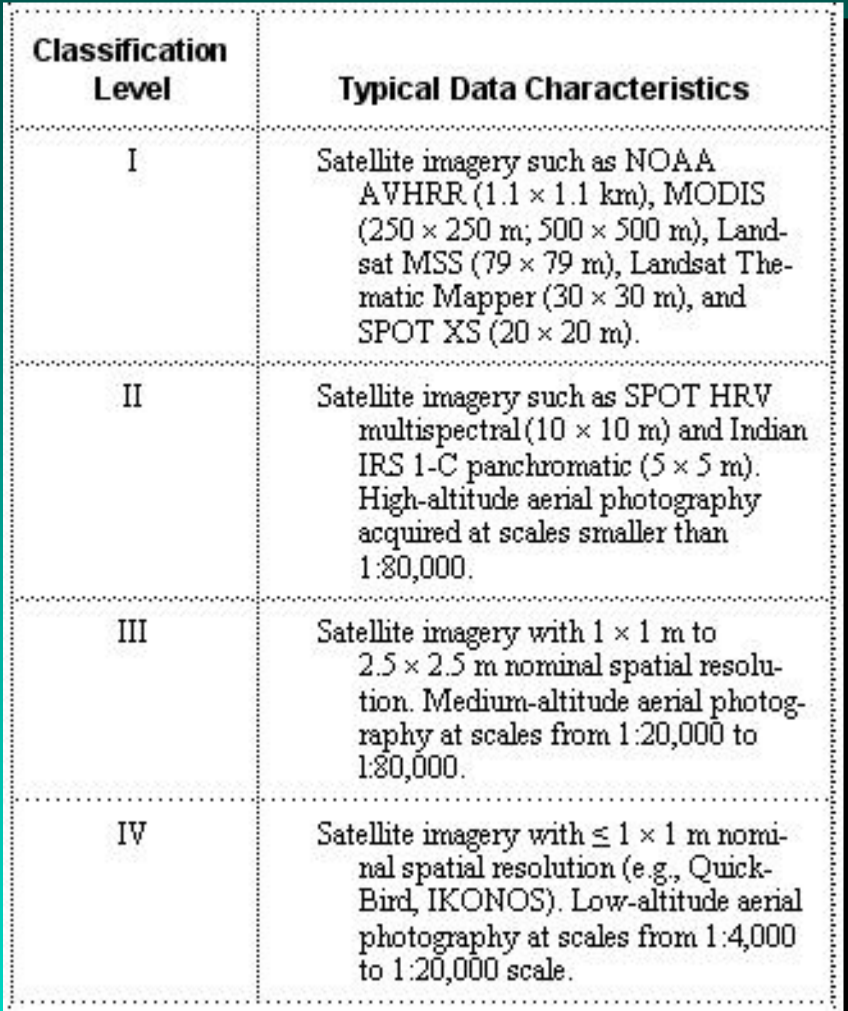
7 **Barren Land**
71 Dry Salt Flats
72 Beaches
73 Sandy Areas Other Than Beaches
74 Bare Exposed Rock
75 Strip Mines, Quarries, and Gravel Pits
76 Transitional Areas
77 Mixed Barren Land

8 **Tundra**
81 Shrub and Brush Tundra
82 Herbaceous Tundra
83 Bare Ground Tundra
84 Wet Tundra
85 Mixed Tundra

9 **Perennial Snow or Ice**
91 Perennial Snowfields
92 Glaciers

Jensen, 2005

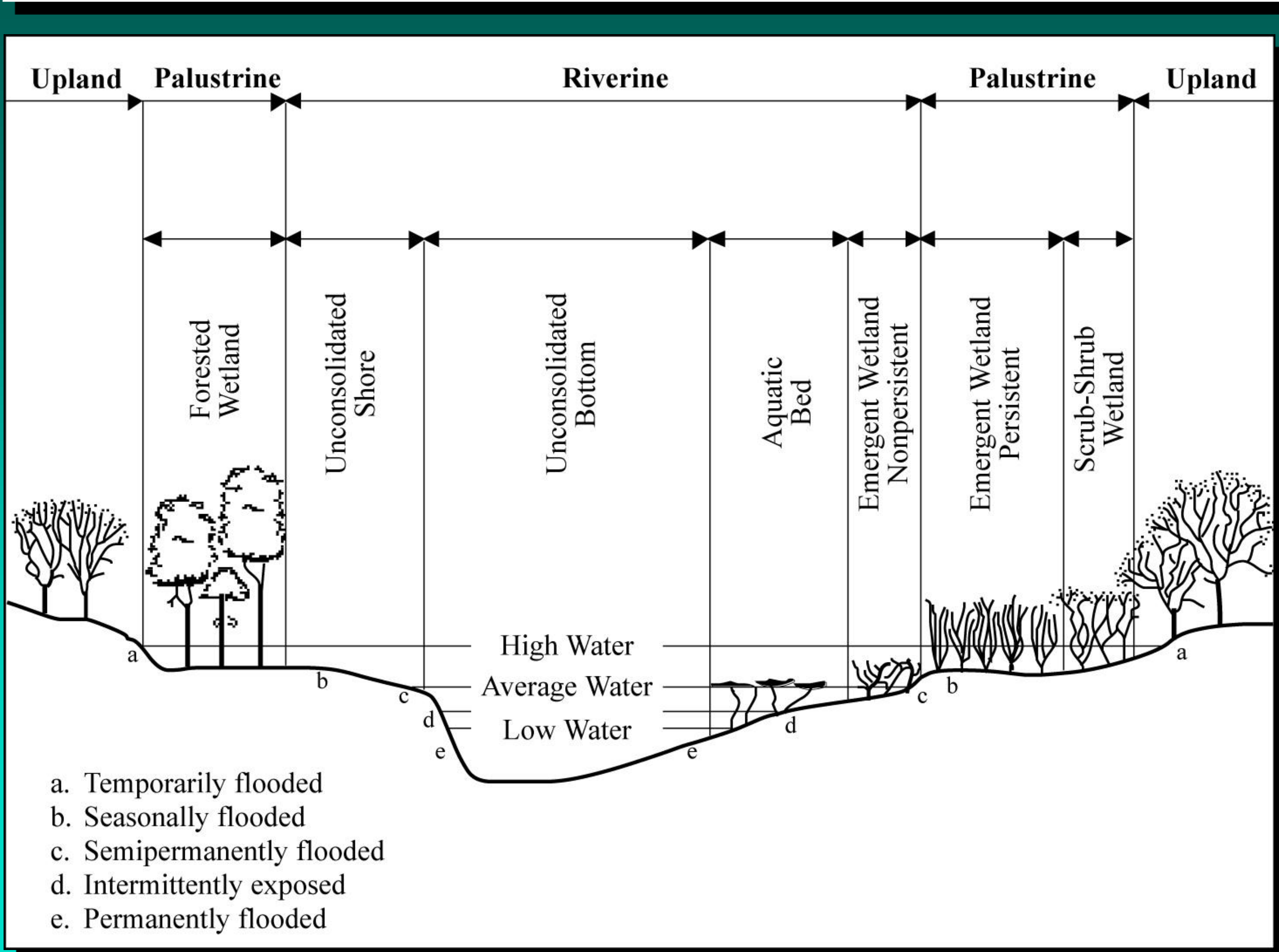
Four Levels of the U.S. Geological Survey *Land-Use/Land-Cover Classification System for Use with Remote Sensor Data* and the type of remotely sensed data typically used to provide the information.



Classification Level	Typical Data Characteristics
I	Satellite imagery such as NOAA AVHRR (1.1 × 1.1 km), MODIS (250 × 250 m; 500 × 500 m), Landsat MSS (79 × 79 m), Landsat Thematic Mapper (30 × 30 m), and SPOT XS (20 × 20 m).
II	Satellite imagery such as SPOT HRV multispectral (10 × 10 m) and Indian IRS 1-C panchromatic (5 × 5 m). High-altitude aerial photography acquired at scales smaller than 1:80,000.
III	Satellite imagery with 1 × 1 m to 2.5 × 2.5 m nominal spatial resolution. Medium-altitude aerial photography at scales from 1:20,000 to 1:80,000.
IV	Satellite imagery with ≤ 1 × 1 m nominal spatial resolution (e.g., QuickBird, IKONOS). Low-altitude aerial photography at scales from 1:4,000 to 1:20,000 scale.

Jensen, 2005

U.S. Department of the Interior Fish & Wildlife Service *Classification of Wetlands and Deepwater Habitats of the United States*



a. Temporarily flooded
b. Seasonally flooded
c. Semipermanently flooded
d. Intermittently exposed
e. Permanently flooded

Jensen, 2005

U.S. Department of the Interior Fish & Wildlife Service *Classification of Wetlands and Deepwater Habitats of the United States*

The U.S. Department of the Interior Fish & Wildlife Service is responsible for mapping and inventorying wetland in the United States. Therefore, they developed a wetland classification system that incorporates information extracted from remote sensor data and *in situ* measurement (Cowardin et al., 1979).

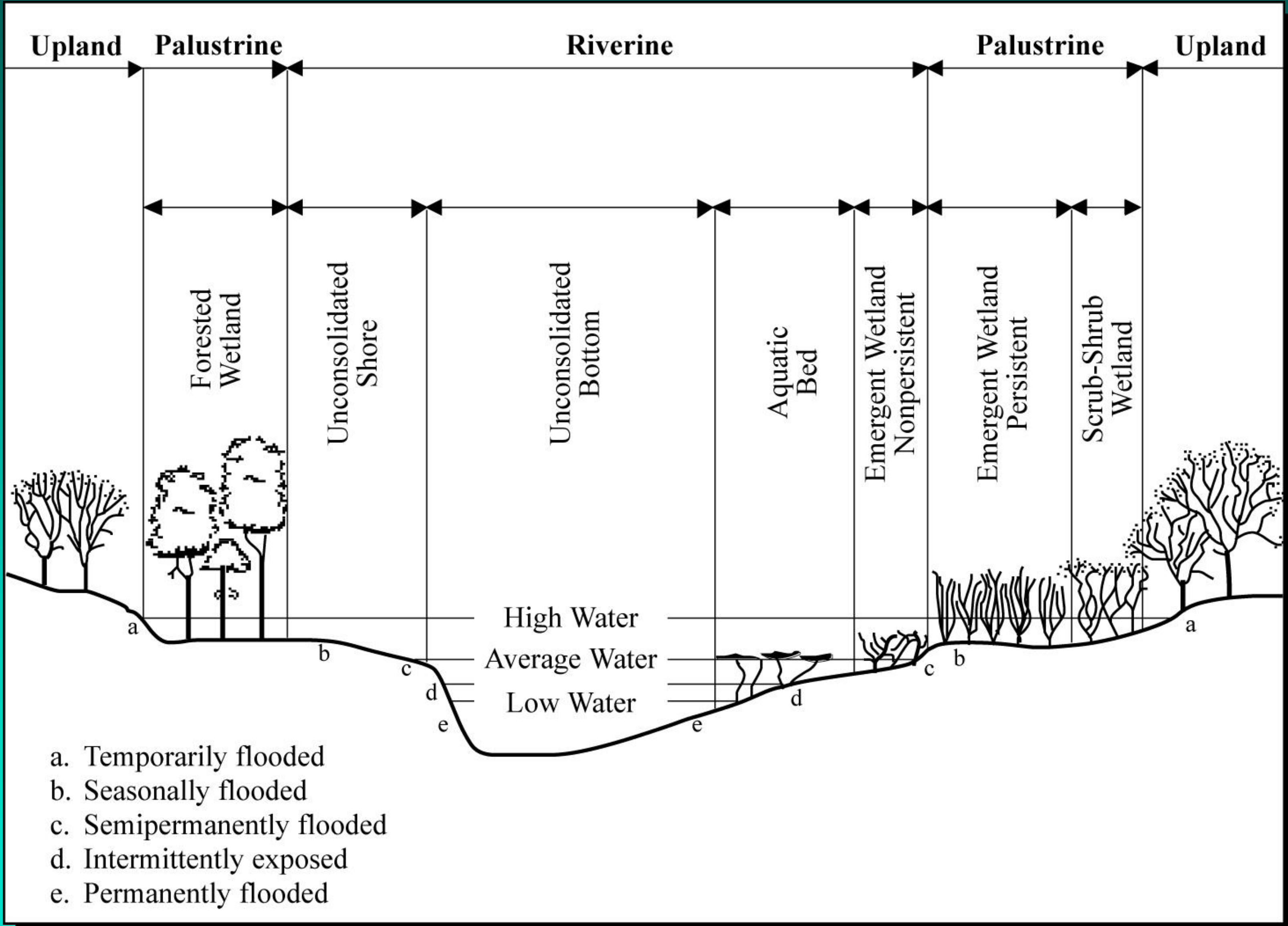
The *Cowardin system* describes ecological taxa, arranges them in a system useful to resource managers, and provides uniformity of concepts and terms. Wetlands are classified based on plant characteristics, soils, and frequency of flooding. Ecologically related areas of deep water, traditionally not considered wetlands, are included in the classification as deep-water habitats. Five systems form the highest level of the classification hierarchy: *marine, estuarine, riverine, lacustrine, and palustrine*. Marine and estuarine systems each have two subsystems: subtidal and intertidal. The riverine system has four subsystems: tidal, lower perennial, upper perennial, and intermittent. The lacustrine has two, littoral and limnetic, and the palustrine has no subsystem. Within the subsystems, classes are based on substrate material and flooding regime or on vegetative life form. The same classes may appear under one or more of the systems or subsystems.

U.S. Department of the Interior Fish & Wildlife Service *Classification of Wetlands and Deepwater Habitats of the United States*

The Cowardin system was adopted as the *National Vegetation Classification Standard for wetlands mapping and inventory* by the Wetlands Subcommittee of the Federal Geographic Data Committee (FGDC, 1996). The Cowardin wetland classification system is the most practical scheme to use if you are going to extract wetland information from remotely sensed data and share the information with others interested in wetland-related problems.

Jensen, 2005

U.S. Department of the Interior Fish & Wildlife Service *Classification of Wetlands and Deepwater Habitats of the United States*



a. Temporarily flooded
b. Seasonally flooded
c. Semipermanently flooded
d. Intermittently exposed
e. Permanently flooded

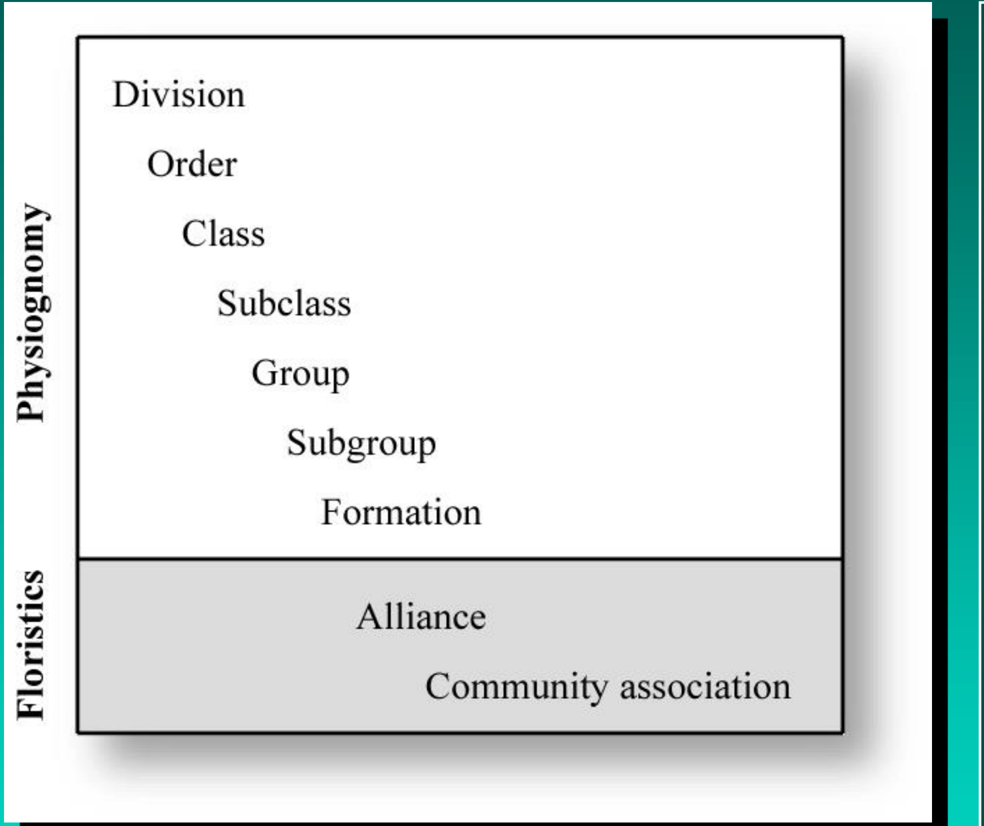
Jensen, 2005

U.S. *National Vegetation Classification System*

The Vegetation Subcommittee of the Federal Geographic Data Committee has endorsed the *National Vegetation Classification System* (NVCS) which produces uniform vegetation resource data at the national level. The NVCS uses a systematic approach to classifying a continuum of natural, existing vegetation. The combined physiognomic-floristic hierarchy uses both qualitative and quantitative data appropriate for conservation and mapping at various scales. Physiognomic characteristics include the more general and less precise levels of taxonomy, whereas the floristic characteristics are found in the more specific levels of taxonomy.

Jensen, 2005

U.S. *National Vegetation Classification System*



The Vegetation Subcommittee of the Federal Geographic Data Committee has endorsed the *National Vegetation Classification System* (NVCS) which produces uniform vegetation resource data at the national level. The NVCS uses a systematic approach to classifying a continuum of natural, existing vegetation. The combined physiognomic-floristic hierarchy uses both qualitative and quantitative data appropriate for conservation and mapping at various scales. Physiognomic characteristics include the more general and less precise levels of taxonomy, whereas the floristic characteristics are found in the more specific levels of taxonomy.

Jensen, 2005

International Geosphere-Biosphere Program *IGBP Land-Cover Classification System* Modified for the Creation of MODIS Land-Cover Products

If a scientist is interested in inventorying land cover at the regional, national, and global scale, then the modified International Geosphere-Biosphere Program *Land-Cover Classification System* may be appropriate. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS) of NASA's Earth Observing System (EOS) is providing global land-surface information at spatial resolutions of 250 to 1,000 m. There are approximately 44 standard MODIS-derived data products that scientists are using to study global change. The MODIS Land Science Team is producing a global land-cover change product at 1-km (0.6 mile) resolution to depict broad-scale land-cover changes.

The land-cover type and land-cover change parameters are produced at 1-km resolution on a quarterly basis. The land-cover parameter identifies *17 categories of land-cover* following the IGBP global vegetation database which defines nine classes of natural vegetation, three classes of developed lands, two classes of mosaic lands, and three classes of nonvegetated lands (snow/ice, bare soil/rocks, water). The first global land-cover map based on MODIS data was distributed in August, 2002.

Jensen, 2005

Observations about Classification Schemes

Geographical information (including remote sensor data) is often imprecise. For example, there is usually a gradual transition at the interface of forests and rangeland, yet many of the aforementioned classification schemes insist on a hard boundary between the classes at this transition zone. The schemes should contain fuzzy definitions because the thematic information they contain is fuzzy. Fuzzy classification schemes are not currently standardized. They are typically developed by individual researchers for site-specific projects. The fuzzy classification systems developed may not be transferable to other environments. *Therefore, we tend to see the use of existing hard classification schemes, which are rigid, based on a priori knowledge, and generally difficult to use.* They continue to be widely employed because they are scientifically based and different individuals using the same classification system can compare results

Jensen, 2005

Observations about Classification Schemes

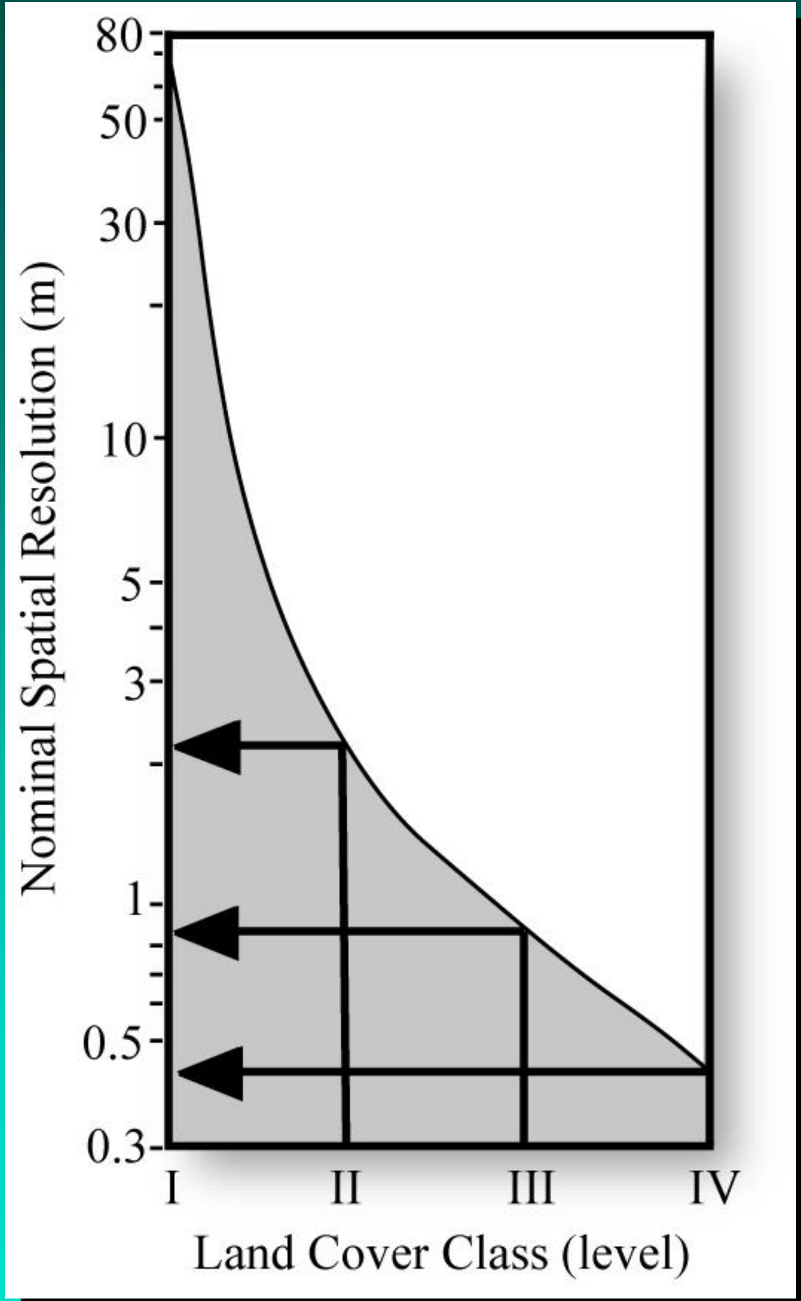
If a reputable classification system already exists, it is foolish to develop an entirely new system that will probably be used only by ourselves. *It is better to adopt or modify existing nationally or internationally recognized classification systems.* This allows us to interpret the significance of our classification results in light of other studies and makes it easier to share data.

Jensen, 2005

Observations about Classification Schemes

There is a relationship between the level of detail in a classification scheme and the spatial resolution of remote sensor systems used to provide information. Welch (1982) summarizes this relationship for mapping urban/suburban land use and land cover This suggests that the level of detail in the desired classification system dictates the spatial resolution of the remote sensor data that should be used. Of course, the spectral resolution of the remote sensing system is also an important consideration, especially when inventorying vegetation, water, ice, snow, soil, and rock.

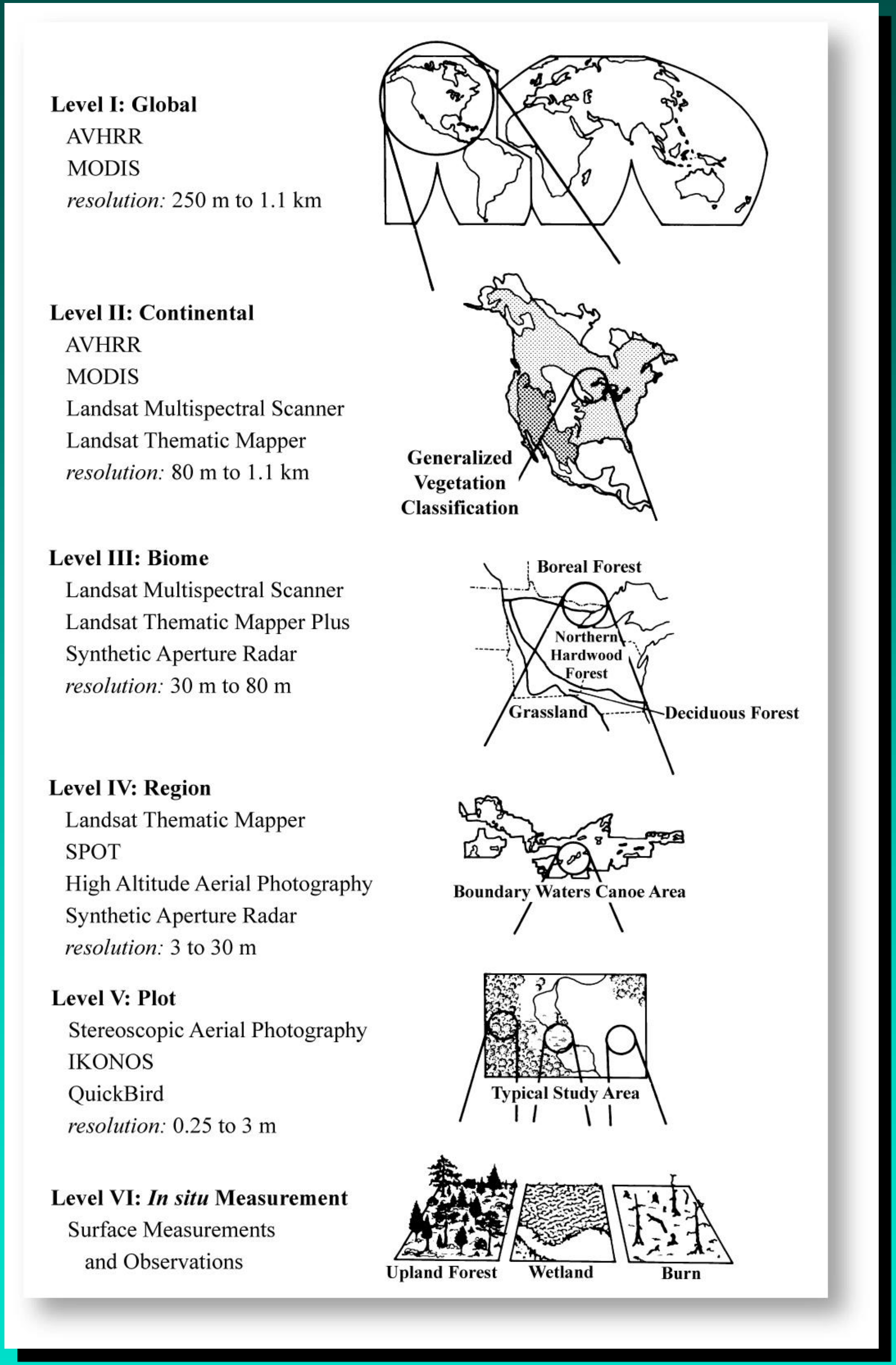
Jensen, 2005



Nominal spatial resolution requirements as a function of the mapping requirements for Levels I to IV land-cover classes in the United States (based on Anderson et al., 1976). Note the dramatic increase in spatial resolution required to map Level II classes.

Jensen, 2005

Relationship between the level of detail required and the spatial resolution of representative remote sensing systems for vegetation inventories.



Level I: Global
AVHRR
MODIS
resolution: 250 m to 1.1 km

Level II: Continental
AVHRR
MODIS
Landsat Multispectral Scanner
Landsat Thematic Mapper
resolution: 80 m to 1.1 km

Level III: Biome
Landsat Multispectral Scanner
Landsat Thematic Mapper Plus
Synthetic Aperture Radar
resolution: 30 m to 80 m

Level IV: Region
Landsat Thematic Mapper
SPOT
High Altitude Aerial Photography
Synthetic Aperture Radar
resolution: 3 to 30 m

Level V: Plot
Stereoscopic Aerial Photography
IKONOS
QuickBird
resolution: 0.25 to 3 m

Level VI: *In situ* Measurement
Surface Measurements
and Observations

Jensen, 2005

Training Site Selection and Statistics Extraction

An analyst may select *training sites* within the image that are representative of the land-cover or land-use classes of interest *after* the classification scheme is adopted. The training data should be of value if the environment from which they were obtained is relatively homogeneous. For example, if all the soils in a grassland region are composed of well-drained sandy-loam soil, then it is likely that grassland training data collected throughout the region would be representative. However, if the soil conditions change across the study area (e.g., one-half of the region has a perched water table with moist near-surface soil), it is likely that grassland training data acquired in the dry-soil part of the study area will *not* be representative of the spectral conditions for grassland found in the moist-soil portion of the study area. This is called a *geographic signature extension* problem, meaning that it may not be possible to extend the grassland remote sensing training data through x, y space.

Jensen, 2005

Training Site Selection and Statistics Extraction

The easiest way to remedy this situation is to apply *geographical stratification* during the early stages of a project. At this time all significant environmental factors that contribute to geographic signature extension problems should be identified, such as differences in soil type, water turbidity, crop species (e.g., two strains of wheat), unusual soil moisture conditions possibly caused by a thunderstorm that did not uniformly deposit its precipitation, scattered patches of atmospheric haze, and so on. Such environmental conditions should be carefully annotated on the imagery and the selection of training sites made based on the geographic stratification of these data. In such cases, it may be necessary to train the classifier over relatively short geographic distances. Each individual stratum may have to be classified separately. The final classification map of the entire region will then be a composite of the individual stratum classifications. However, if environmental conditions are homogeneous or can be held constant (e.g., through band ratioing or atmospheric correction), it may be possible to extend signatures vast distances in space, significantly reducing the training cost and effort. *Additional research is required before the concept of geographic and temporal (through time) signature extension is fully understood.*

Jensen, 2005

Training Site Selection and Statistics Extraction

Once spatial and temporal signature extension factors have been considered, the analyst selects representative *training sites* for each class and collects the spectral statistics for each pixel found within each training site.

Each site is usually composed of many pixels. The general rule is that if training data are being extracted from *n* bands then >10*n* pixels of training data are collected for each class. This is sufficient to compute the variance–covariance matrices required by some classification algorithms.

Jensen, 2005

Training Site Selection and Statistics Extraction

There are a number of ways to collect the *training site* data, including:

- collection of *in situ* information such as tree type, height, percent canopy closure, and diameter-at-breast-height (dbh) measurements,
- on-screen selection of polygonal training data, and/or
- on-screen seeding of training data.

Jensen, 2005

Training Site Selection and Statistics Extraction

The analyst may view the image on the color CRT screen and select *polygonal areas of interest (AOI)* (e.g., a stand of oak forest). Most image processing systems use a “rubber band” tool that allows the analyst to identify detailed AOIs. Conversely, the analyst may seed a specific location in the image space using the cursor. The *seed program* begins at a single *x, y* location and evaluates neighboring pixel values in all bands of interest. Using criteria specified by the analyst, the seed algorithm expands outward like an amoeba as long as it finds pixels with spectral characteristics similar to the original seed pixel. This is a very effective way of collecting homogeneous training information.

Jensen, 2005

Training Site Selection and Statistics Extraction

Each pixel in each training site associated with a particular class (c) is represented by a *measurement vector, X_c*:

$$X_c = \begin{bmatrix} BV_{i,j,1} \\ BV_{i,j,2} \\ BV_{i,j,3} \\ . \\ . \\ BV_{i,j,k} \end{bmatrix}$$

where $BV_{i,j,k}$ is the brightness value for the i,j^th pixel in band k.

Jensen, 2005

Training Site Selection and Statistics Extraction

The brightness values for each pixel in each band in each training class can then be analyzed statistically to yield a *mean measurement vector*, M_c, for each class:

$$M_c = \begin{bmatrix} \mu_{c1} \\ \mu_{c2} \\ \mu_{c3} \\ . \\ . \\ \mu_{ck} \end{bmatrix}$$

where μ_{ck} represents the mean value of the data obtained for *class c* in band *k*.

Jensen, 2005

Training Site Selection and Statistics Extraction

The raw measurement vector can also be analyzed to yield the covariance matrix for each *class c:*

$$V_c = V_{ckl} = \begin{bmatrix} \mathrm{cov}_{c11}\ \mathrm{cov}_{c12} \ldots \mathrm{cov}_{c1n} \\ \mathrm{cov}_{c21}\ \mathrm{cov}_{c22} \ldots \mathrm{cov}_{c2n} \\ . \\ . \\ \mathrm{cov}_{cn1}\ \mathrm{cov}_{cn2} \ldots \mathrm{cov}_{cnn} \end{bmatrix}$$

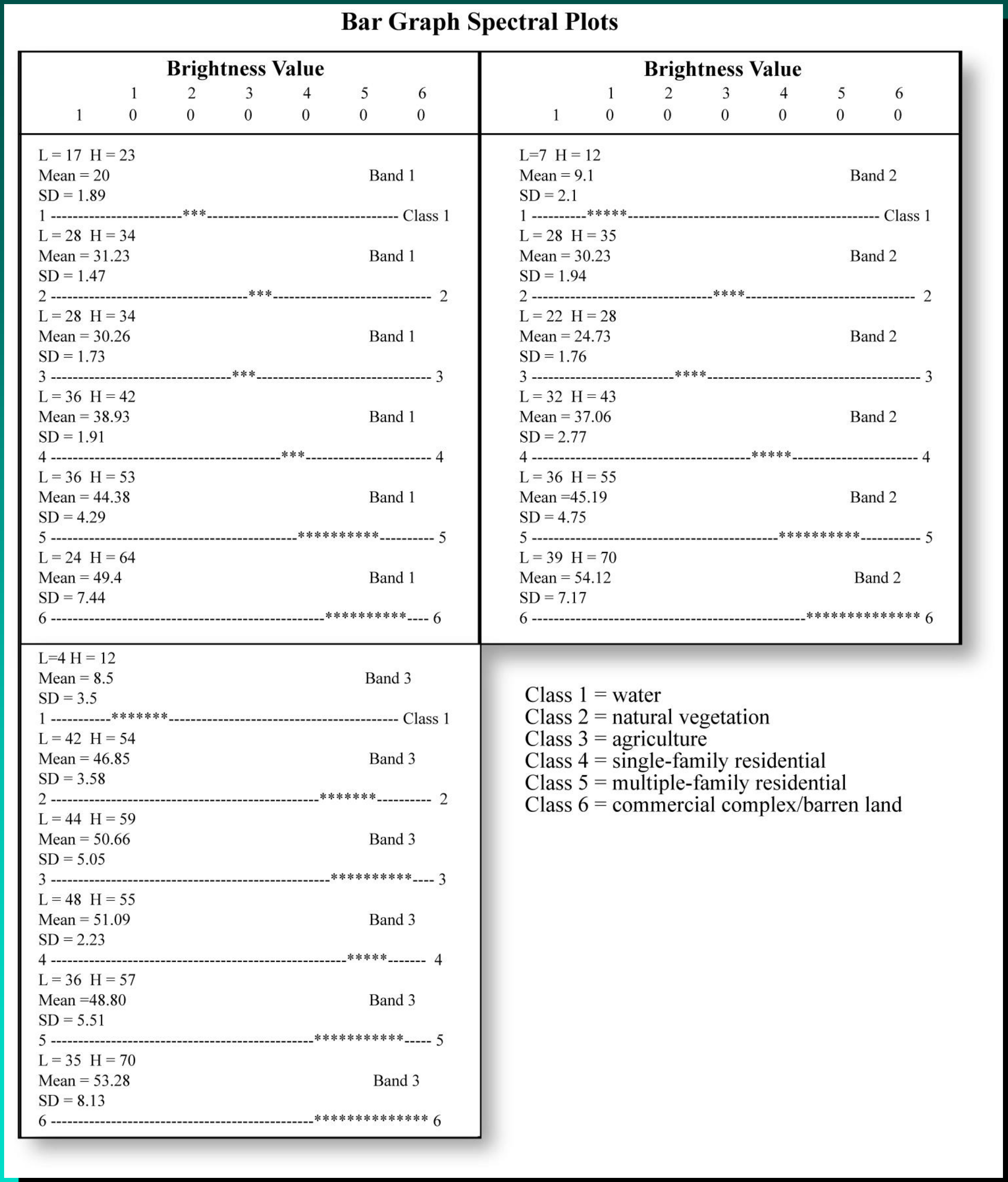
where Cov_{ckl} is the covariance of class *c* between bands *k* through *l*. For brevity, the notation for the covariance matrix for class *c* (i.e., V_{ckl}) will be shortened to just V_c. The same will be true for the covariance matrix of class *d* (i.e., $\mathrm{V}_{\mathrm{dkl}} = \mathrm{V}_{\mathrm{d}}$).

Jensen, 2005

Selecting the Optimum Bands for Image Classification: Feature Selection

Once the training statistics have been systematically collected from each band for each class of interest, a judgment must be made to determine the bands (channels) that are most effective in discriminating each class from all others. This process is commonly called *feature selection*. The goal is to delete from the analysis the bands that provide redundant spectral information. In this way the *dimensionality* (i.e., the number of bands to be processed) in the dataset may be reduced. This minimizes the cost of the digital image classification process (but should not affect the accuracy). Feature selection may involve both *statistical* and *graphical* analysis to determine the *degree of between-class separability* in the remote sensor training data. Using statistical methods, combinations of bands are normally ranked according to their potential ability to discriminate each class from all others using *n* bands at a time.

Jensen, 2005



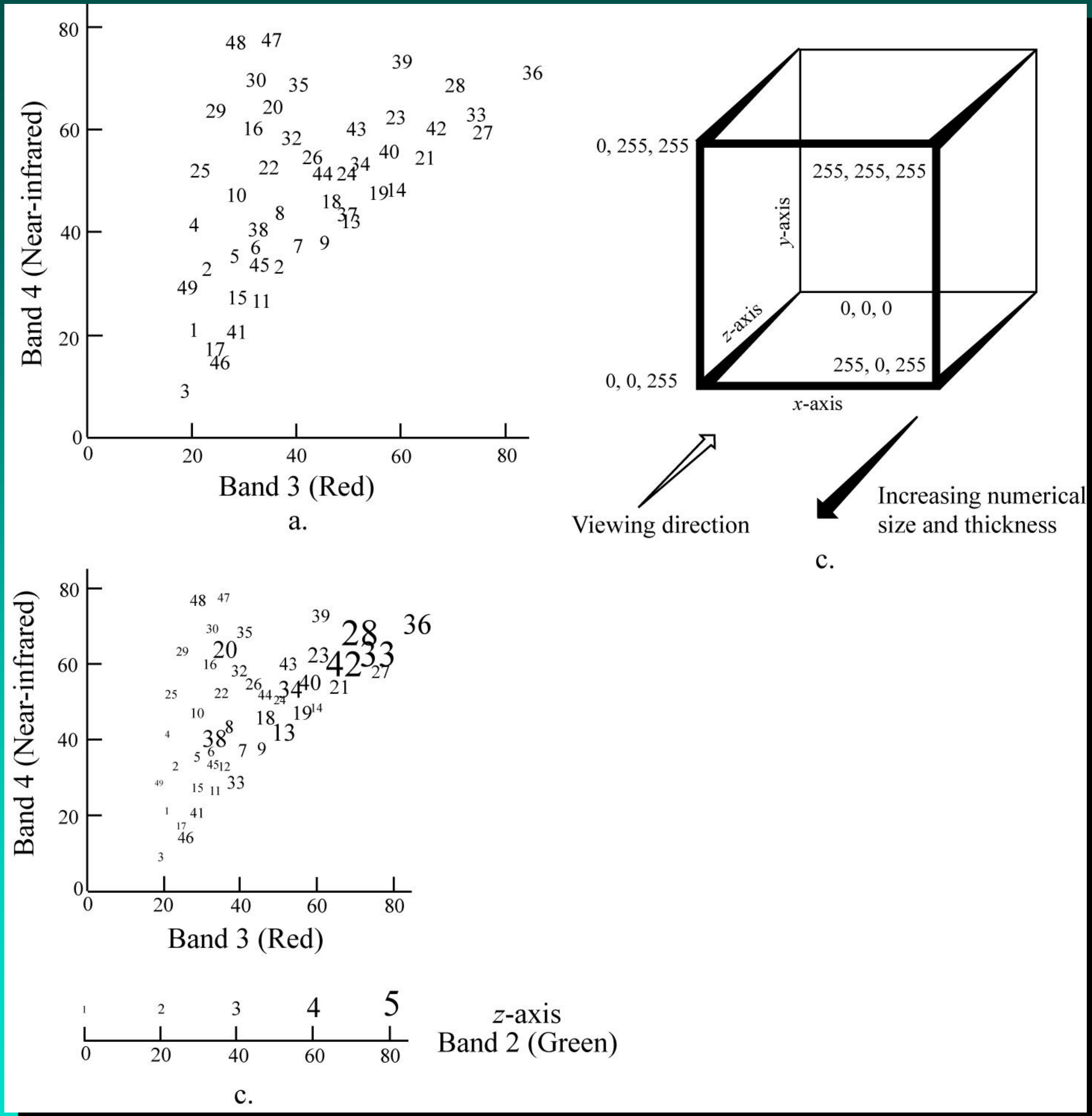
Bar Graph Spectral Plots

```
Brightness Value
        1    2    3    4    5    6
   1    0    0    0    0    0    0

L = 17  H = 23
Mean = 20                                   Band 1
SD = 1.89
1 ---------------------***--------------------------------- Class 1
L = 28  H = 34
Mean = 31.23                                Band 1
SD = 1.47
2 --------------------------------***---------------------------  2
L = 28  H = 34
Mean = 30.26                                Band 1
SD = 1.73
3 -----------------------------***------------------------------ 3
L = 36  H = 42
Mean = 38.93                                Band 1
SD = 1.91
4 -------------------------------------***---------------------- 4
L = 36  H = 53
Mean = 44.38                                Band 1
SD = 4.29
5 ----------------------------------------**********--------- 5
L = 24  H = 64
Mean = 49.4                                 Band 1
SD = 7.44
6 ---------------------------------------------**********---- 6

Brightness Value
        1    2    3    4    5    6
   1    0    0    0    0    0    0

L=7  H = 12
Mean = 9.1                                  Band 2
SD = 2.1
1 ---------*****------------------------------------------ Class 1
L = 28  H = 35
Mean = 30.23                                Band 2
SD = 1.94
2 -----------------------------****----------------------------  2
L = 22  H = 28
Mean = 24.73                                Band 2
SD = 1.76
3 -----------------------****------------------------------------ 3
L = 32  H = 43
Mean = 37.06                                Band 2
SD = 2.77
4 ------------------------------------*****---------------------- 4
L = 36  H = 55
Mean =45.19                                 Band 2
SD = 4.75
5 ----------------------------------------**********---------- 5
L = 39  H = 70
Mean = 54.12                                 Band 2
SD = 7.17
6 ---------------------------------------------************** 6

L=4 H = 12
Mean = 8.5                                  Band 3
SD = 3.5
1 ----------*******--------------------------------------- Class 1
L = 42  H = 54
Mean = 46.85                                Band 3
SD = 3.58
2 --------------------------------------------*******---------  2
L = 44  H = 59
Mean = 50.66                                Band 3
SD = 5.05
3 ----------------------------------------------**********---- 3
L = 48  H = 55
Mean = 51.09                                Band 3
SD = 2.23
4 -------------------------------------------------*****-------  4
L = 36  H = 57
Mean =48.80                                 Band 3
SD = 5.51
5 -------------------------------------------***********----- 5
L = 35  H = 70
Mean = 53.28                                 Band 3
SD = 8.13
6 -------------------------------------------************** 6
```

Class 1 = water
Class 2 = natural vegetation
Class 3 = agriculture
Class 4 = single-family residential
Class 5 = multiple-family residential
Class 6 = commercial complex/barren land

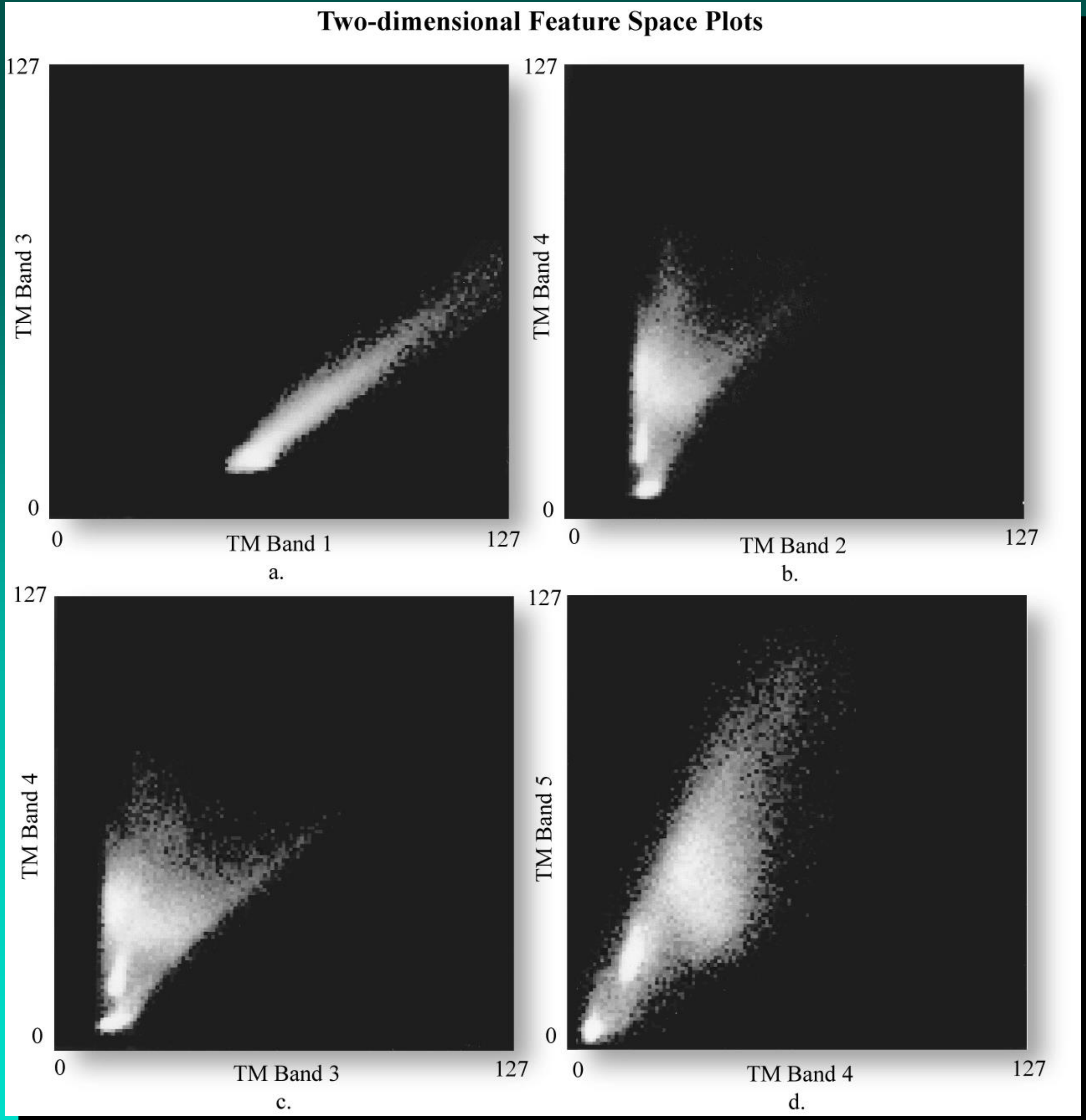
Bar graph spectral plots of data. Training statistics (the mean ±1σ) for six land-cover classes are displayed for three Landsat MSS bands. The simple display can be used to identify between-class separability for each class and single band.

Jensen, 2005



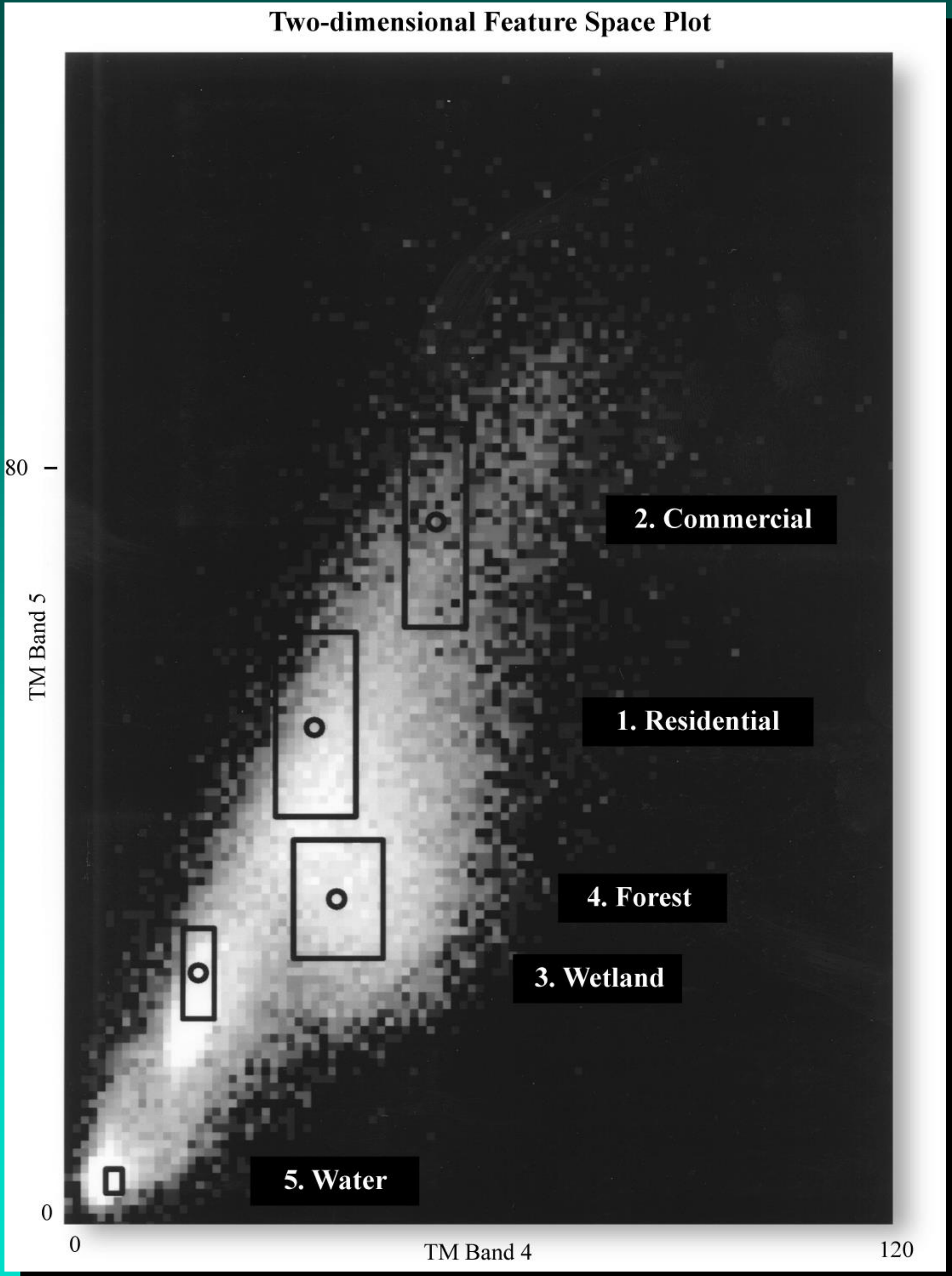
a) Cospectral mean vector plots of 49 clusters extracted from Charleston Landsat TM data bands 3 and 4. b) The logic for increasing numeral size and thickness along the *z*-axis. c) The introduction of band 2 information scaled according to size and thickness along the *z*-axis.

Jensen, 2005



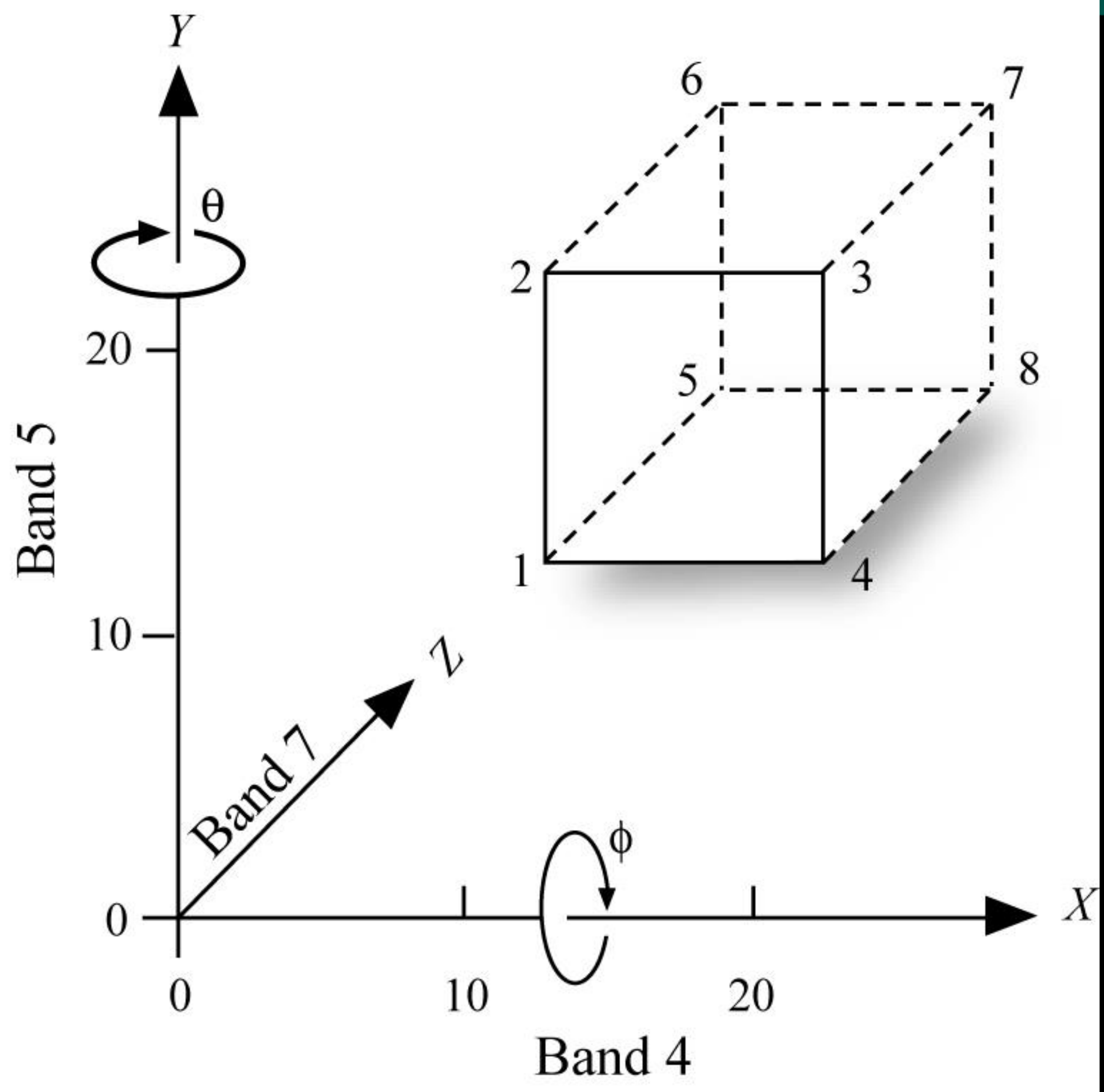
Two-dimensional feature space plots of four pairs of Landsat TM data of Charleston, SC. a) TM bands 1 and 3, b) TM bands 2 and 4, c) TM bands 3 and 4, and d) TM bands 4 and 5. The brighter a particular pixel is in the display, the more pixels within the scene having that unique combination of band values.

Jensen, 2005



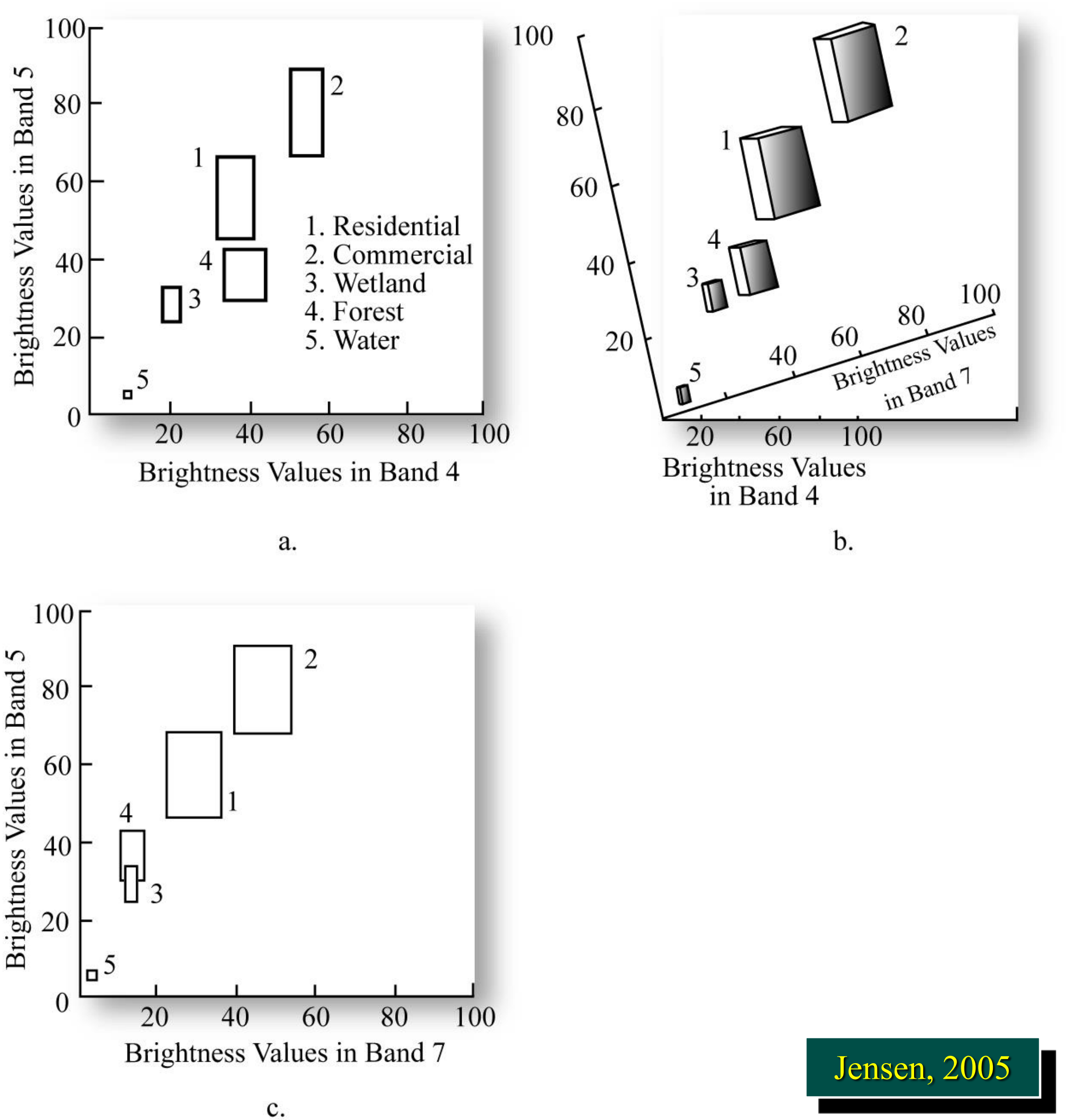
Plot of the Charleston, SC, Landsat TM training statistics for five classes measured in bands 4 and 5 displayed as cospectral parallelepipeds. The upper and lower limit of each parallelepiped is ±1σ. The parallelepipeds are superimposed on a feature space plot of bands 4 and 5.

Jensen, 2005



Simple parallelepiped displayed in pseudo three-dimensional space. Each of the eight corners represents a unique *x, y, z* coordinate corresponding to a lower or upper threshold value of the training data. For example, the original coordinates of point 4 are associated with 1) the upper threshold value of band 4, 2) the lower threshold value of band 5, and 3) the lower threshold value of band 7. The rotation matrix transformations cause the original coordinates to be rotated about the *y*-axis some θ radians, and the *x* axis some ϕ radians.

Jensen, 2005



a.

b.

c.

Jensen, 2005

Three-dimensional parallelepipeds of five Charleston, SC, training classes derived from Landsat TM data. Only bands 4, 5, and 7 are used in this investigation. The data are rotated about the *y*-axis, 0°, 45°, 90°. At 0° and 90° [parts (a) and (c), respectively]. We are actually looking at only two bands, analogous to the two-dimensional parallelepiped boxes shown previously. The third band lies perpendicular to the page we are viewing. Between such extremes, it is possible to obtain optimum viewing angles for visual analysis of training class statistics using three bands at once. Part (b) displays the five classes at a rotation of 45°, revealing that the classes are separable using this three-band combination. It probably is not necessary to use all three bands since bands 4 and 5 alone will discriminate satisfactorily between the five classes, as shown in part (a). There would be a substantial amount of overlap between classes if bands 5 and 7 were used.

Statistical Measures of Feature Selection

Statistical methods of *feature selection* are used to quantitatively select which subset of bands (also referred to as features) provides the greatest degree of statistical separability between any two classes *c* and *d*. The basic problem of spectral pattern recognition is that given a spectral distribution of data in *n* bands of remotely sensed data, we must find a discrimination technique that will allow separation of the major land-cover categories with a minimum of error and a minimum number of bands. This problem is demonstrated diagrammatically using just one band and two classes. Generally, the more bands we analyze in a classification, the greater the cost and perhaps the greater the amount of redundant spectral information being used. When there is overlap, any decision rule that one could use to separate or distinguish between two classes must be concerned with two types of error.

- A pixel may be assigned to a class to which it does not belong (an *error of commission*).
- A pixel is not assigned to its appropriate class (an *error of omission*).

Jensen, 2005

Statistical Measures of Feature Selection

Divergence was one of the first measures of statistical separability used in the machine processing of remote sensor data, and it is still widely used as a method of feature selection. It addresses the basic problem of deciding what is the best *q*-band subset of *n* bands for use in the supervised classification process. The number of combinations *c* of *n* bands taken *q* at a time is:

$$c\left(\frac{n}{q}\right) = \frac{n!}{q!\left(n-q\right)!}$$

Thus, if there are six TM bands and we are interested in the three best bands to use in the classification of the Charleston scene, this results in 20 combinations that must be evaluated:

$$c\left(\frac{6}{3}\right) = \frac{6!}{3!\left(6-3\right)!} = \frac{720}{6\left(6\right)} = 20\ combinations$$

Jensen, 2005

Statistical Measures of Feature Selection

Divergence is computed using the mean and covariance matrices of the class statistics collected in the training phase of the supervised classification. We will initiate the discussion by concerning ourselves with the statistical separability between just two classes, *c* and *d*. The degree of divergence or separability between *c* and *d*, $Diver_{cd}$, is computed according to the formula:

$$Diver_{cd} = \frac{1}{2} tr\left[\left(V_c - V_d\right)\left(V_d^{-1} - V_c^{-1}\right)\right]$$

$$+ \frac{1}{2} tr\left[\left(V_c^{-1} + V_d^{-1}\right)\left(M_c - M_d\right)\left(M_c - M_d\right)^T\right]$$

where *tr* [ ] is the trace of a matrix (i.e., the sum of the diagonal elements), V_c and V_d are the covariance matrices for the two classes under investigation, *c* and *d*, and M_c and M_d are the mean vectors for classes *c* and *d* (Konecny, 2003).

Jensen, 2005

Statistical Measures of Feature Selection

But what about the case where there are more than two classes? In this instance, the most common solution is to compute the *average divergence, Diver*$_{\text{avg}}$. This involves computing the average over all possible pairs of classes *c* and *d,* while holding the subset of bands *q* constant. Then, another subset of bands *q* is selected for the *m* classes and analyzed. The subset of features (bands) having the maximum average divergence may be the superior set of bands to use in the classification algorithm. This can be expressed:

$$Diver_{avg} = \frac{\sum_{c=1}^{m-1} \sum_{d=c+1}^{m} Diver_{cd}}{C}$$

Using this, the band subset *q* with the highest average divergence would be selected as the most appropriate set of bands for classifying the *m* classes.

Jensen, 2005

Statistical Measures of Feature Selection

The *Bhattacharyya distance* assumes that the two classes *c* and *d* are Gaussian and that the means M_c and M_d and covariance matrices V_c and V_d are available. It is computed (Duda et al., 2001):

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

To select the best *q* features (i.e., combination of bands) from the original *n* bands in an *m*-class problem, the Bhattacharyya distance is calculated between each *m*(*m* – 1)/2 pair of classes for each possible way of choosing *q* features from *n* dimensions. The best *q* features are those dimensions whose sum of the Bhattacharyya distance between the *m*(*m* – 1)/2 classes is highest.

Jensen, 2005

Select the Appropriate Classification Algorithm

Various supervised classification algorithms may be used to assign an unknown pixel to one of *m* possible classes. The choice of a particular classifier or decision rule depends on the nature of the input data and the desired output. *Parametric* classification algorithms assumes that the observed measurement vectors X_c obtained for each class in each spectral band during the training phase of the supervised classification are Gaussian; that is, they are normally distributed. *Nonparametric* classification algorithms make no such assumption.

Several widely adopted nonparametric classification algorithms include:

- one-dimensional *density slicing*
- parallepiped,
- minimum distance,
- nearest-neighbor, and
- neural network and expert system analysis.

The most widely adopted parametric classification algorithms is the:

- maximum likelihood.

Jensen, 2005

Parallelepiped Classification Algorithm

This is a widely used digital image classification decision rule based on simple *Boolean "and/or" logic*. Training data in *n* spectral bands are used to perform the classification. Brightness values from each pixel of the multispectral imagery are used to produce an *n*-dimensional mean vector, $M_{\mathrm{c}} = (\mu_{ck}, \mu_{c2}, \mu_{c3}, ..., \mu_{cn})$ with μ_{ck} being the mean value of the training data obtained for class *c* in band *k* out of *m* possible classes, as previously defined. σ_{ck} is the standard deviation of the training data class *c* of band *k* out of *m* possible classes. In this discussion we will evaluate all five Charleston classes using just bands 4 and 5 of the training data.

Jensen, 2005

Parallelepiped Classification Algorithm

Using a one-standard deviation threshold (as shown in Figure 9-15), a parallelepiped algorithm decides BV_{ijk} is in class c if, and only if:

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

where

$c = 1, 2, 3, \ldots, m,$ number of classes, and
$k = 1, 2, 3, \ldots, n,$ number of bands.

Therefore, if the low and high decision boundaries are defined as:

$$L_{ck} = \mu_{ck} - \sigma_{ck}$$

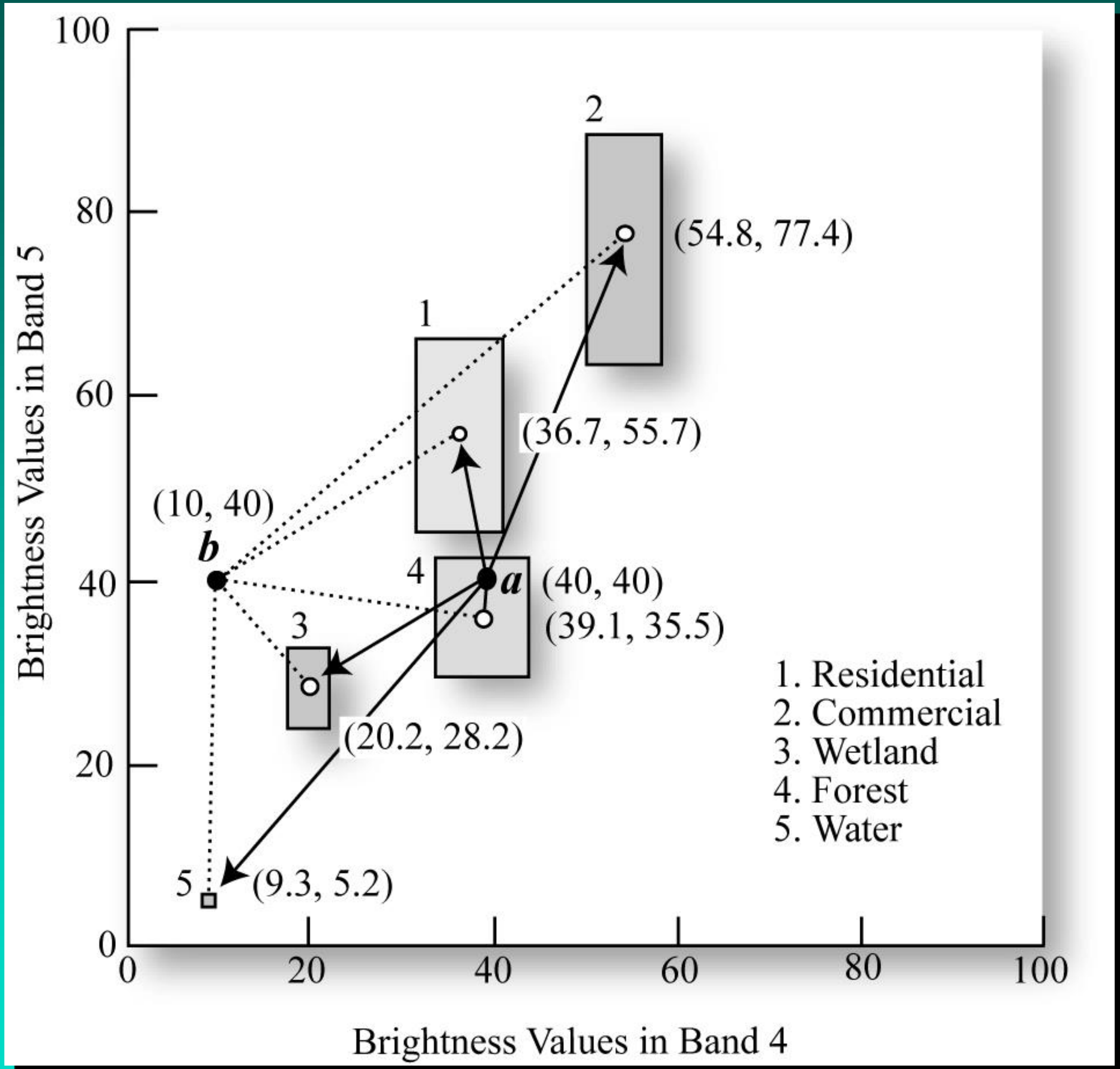
and

$$H_{ck} = \mu_{ck} + \sigma_{ck}$$

the parallelepiped algorithm becomes $L_{ck} \leq BV_{ijk} \leq H_{ck}$

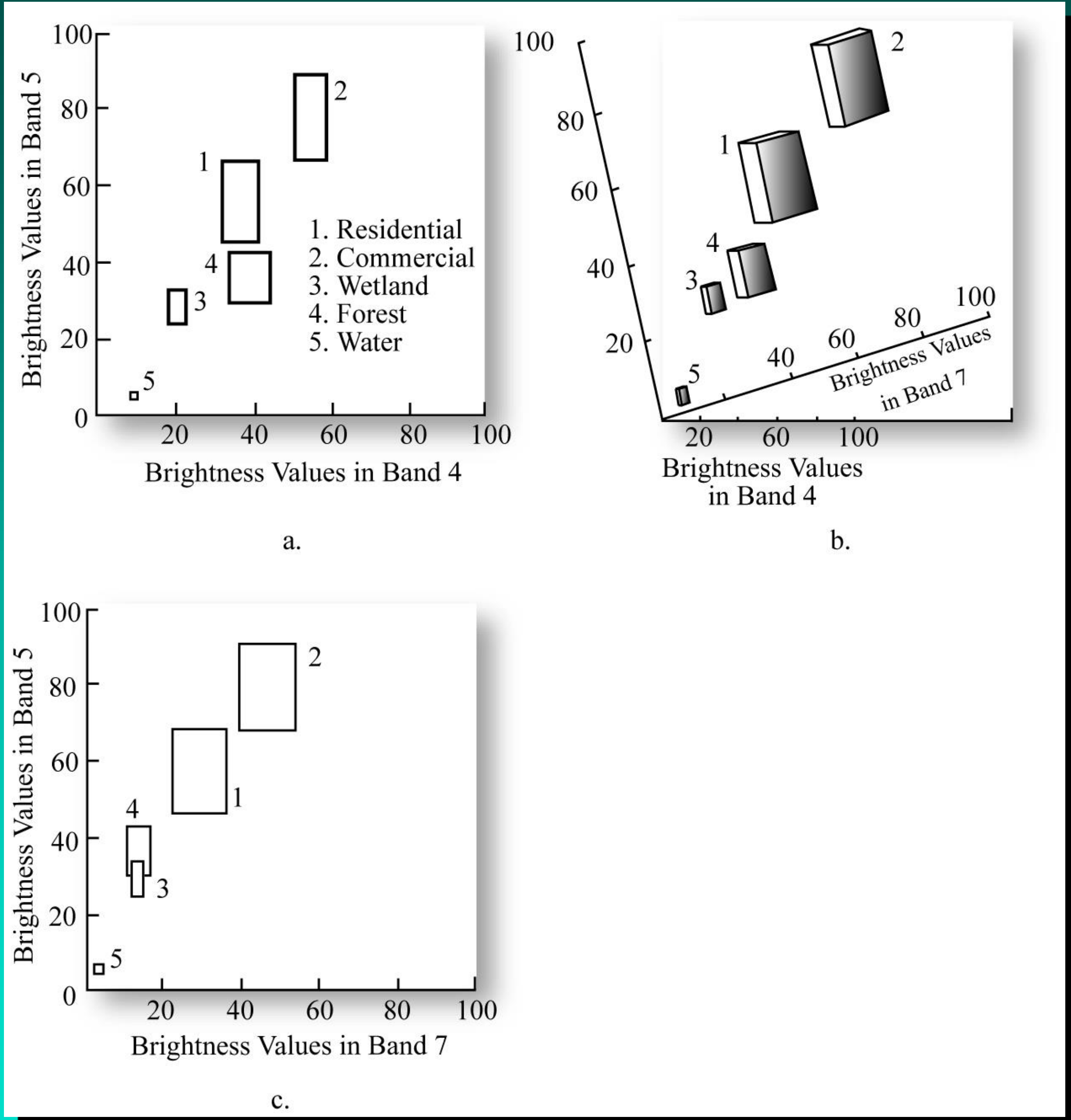
Jensen, 2005

Parallelepiped Classification Algorithm



Jensen, 2005

Points *a* and *b* are pixels in the image to be classified. Pixel *a* has a brightness value of 40 in band 4 and 40 in band 5. Pixel *b* has a brightness value of 10 in band 4 and 40 in band 5. The boxes represent the *parallelepiped* decision rule associated with a ±1σ classification. The vectors (*arrows*) represent the distance from *a* and *b* to the mean of all classes in a *minimum distance to means* classification algorithm. Refer to Tables 9-8 and 9-9 for the results of classifying points *a* and *b* using both classification techniques.



a.

b.

c.

Jensen, 2005

Minimum Distance to Means Classification Algorithm

The *minimum distance to means* decision rule is computationally simple and commonly used. When used properly it can result in classification accuracy comparable to other more computationally intensive algorithms such as the maximum likelihood algorithm. Like the parallelepiped algorithm, it requires that the user provide the mean vectors for each class in each band μck from the training data. To perform a minimum distance classification, a program must calculate the distance to each mean vector μck from each unknown pixel (BV_{ijk}). It is possible to calculate this distance using Euclidean distance based on the Pythagorean theorem or “round the block” distance measures. In this discussion we demonstrate the method of minimum distance classification using Euclidean distance measurements applied to the two unknown points (a and b).

Jensen, 2005

Minimum Distance to Means Classification Algorithm

The computation of the Euclidean distance from point *a* (40, 40) to the mean of class 1 (36.7, 55.7) measured in bands 4 and 5 relies on the equation:

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

where μ_{ck} and μ_{cl} represent the mean vectors for class *c* measured in bands *k* and *l*. In our example this would be:

$$Dist_{a\ \text{to class 1}} = \sqrt{\left(BV_{ij4} - \mu_{1,4}\right)^2 + \left(BV_{ij5} - \mu_{1,5}\right)^2}$$

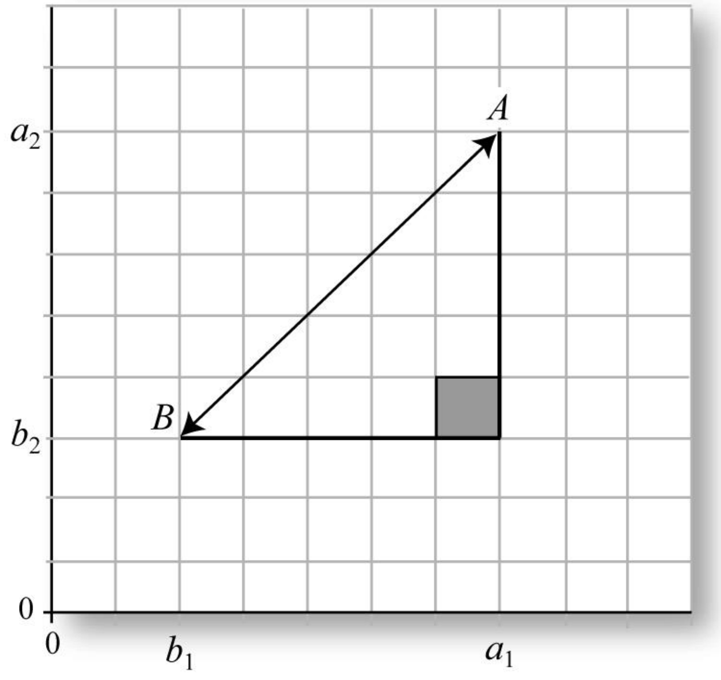
The distance from point *a* to the mean of class 2 in these same two bands would be:

$$Dist_{a\ \text{to class 2}} = \sqrt{\left(BV_{ij4} - \mu_{2,4}\right)^2 + \left(BV_{ij5} - \mu_{2,5}\right)^2}$$

Notice that the subscript that stands for class *c* is incremented from 1 to 2. By calculating the Euclidean distance from point *a* to the mean of all five classes, it is possible to determine which distance is shortest.

Jensen, 2005

Minimum Distance to Means Classification Algorithm



Euclidean distance

$$D_{AB} = \sqrt{\sum_{i=1}^{2} (a_i - b_i)^2}$$

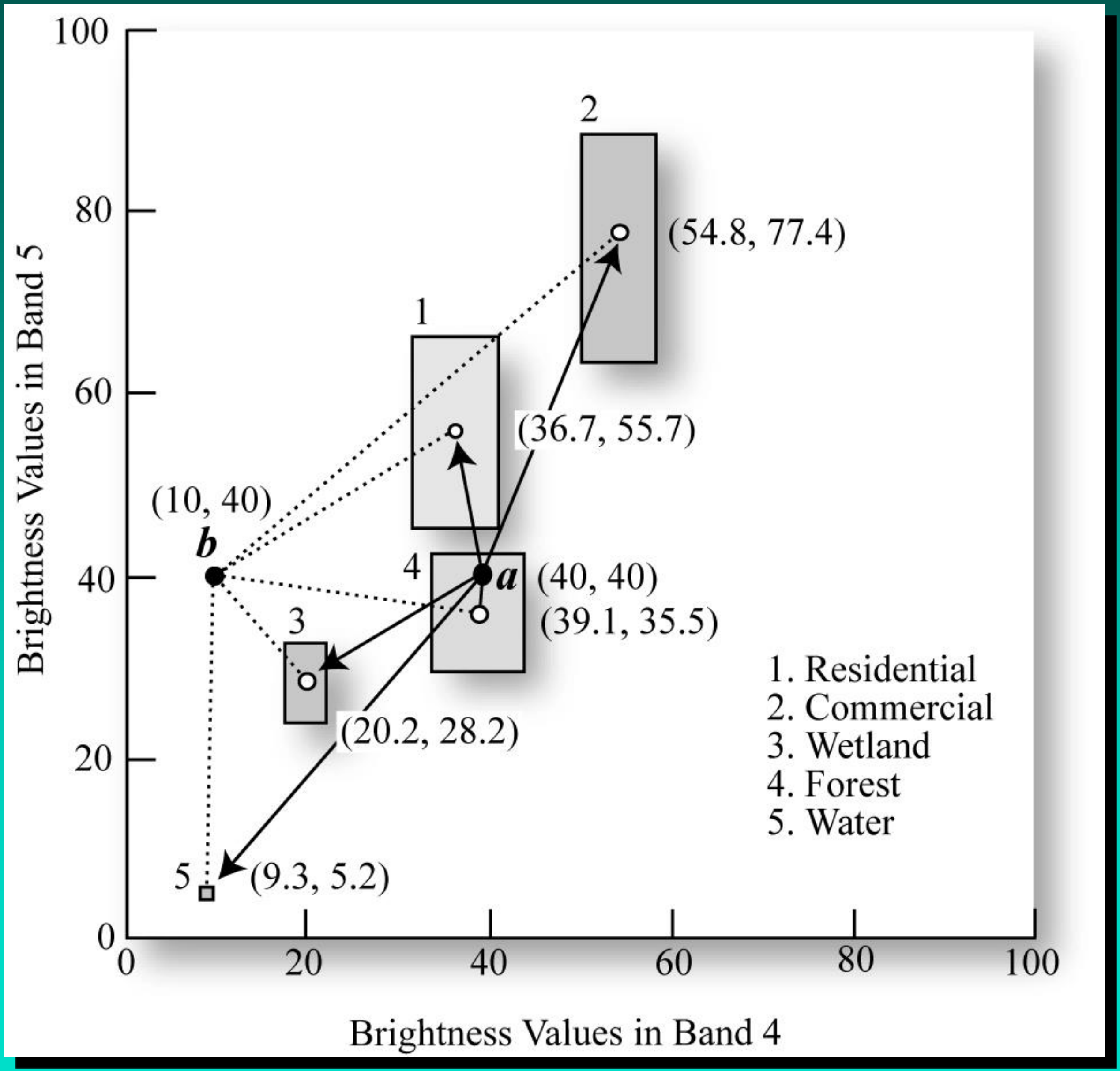
Round the block distance

$$D_{AB} = \sum_{i=1}^{2} |(a_i - b_i)|$$

The distance used in a *minimum distance to means* classification algorithm can take two forms: the Euclidean distance based on the Pythagorean theorem and the “round the block” distance. The Euclidean distance is more computationally intensive.

Jensen, 2005

Minimum Distance to Means Classification Algorithm



Jensen, 2005

Maximum Likelihood Classification Algorithm

The aforementioned classifiers were based primarily on identifying decision boundaries in feature space based on training class multispectral *distance* measurements. The *maximum likelihood decision rule* is based on *probability*.

- It assigns each pixel having pattern measurements or features *X* to the class *i* whose units are most probable or likely to have given rise to feature vector *X*.

- In other words, the probability of a pixel belonging to each of a predefined set of *m* classes is calculated, and the pixel is then assigned to the class for which the probability is the highest.

- The *maximum likelihood decision rule* is still one of the most widely used supervised classification algorithms.

Jensen, 2005

Maximum Likelihood Classification Algorithm

The maximum likelihood procedure assumes that the training data statistics for each 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. This should then produce unimodal, *Gaussian training class statistics* that fulfill the *normal distribution requirement*.

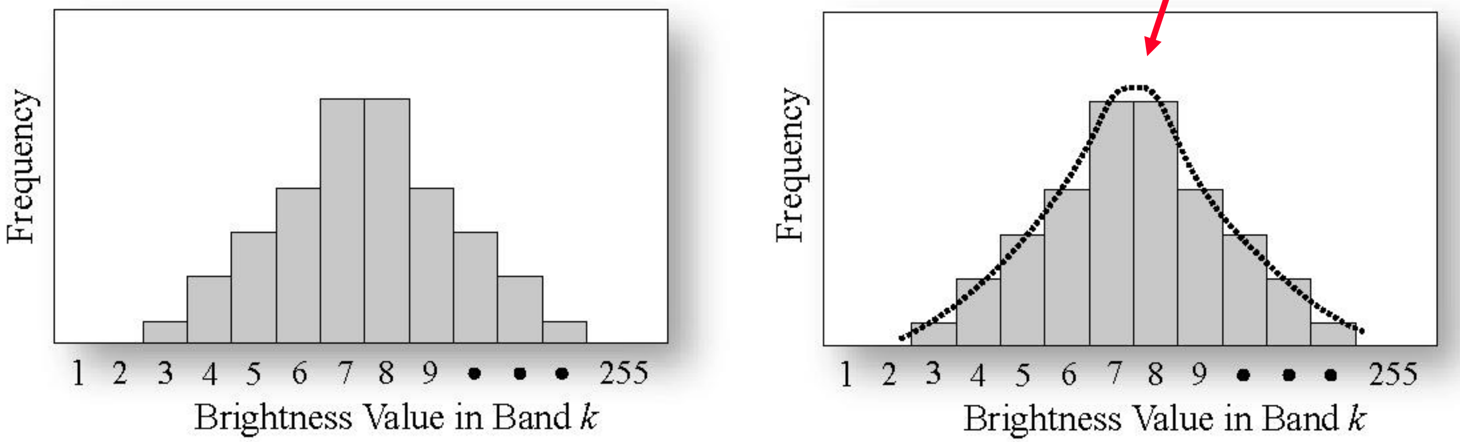
Jensen, 2005

Maximum Likelihood Classification Algorithm

But how do we obtain the probability information we will need from the remote sensing training data we have collected? The answer lies first in the computation of *probability density functions*. We will demonstrate using a single class of training data based on a single band of imagery.

Jensen, 2005

For example, consider the hypothetical histogram (data frequency distribution) of forest training data obtained in band *k*. We could choose to store the values contained in this histogram in the computer, but a more elegant solution is to approximate the distribution by a normal *probability density function (curve),* as shown superimposed on the histogram.



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

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

Jensen, 2005

Maximum Likelihood Classification Algorithm

The estimated *probability density function* for class w_i (e.g., forest) is computed using the equation:

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

where exp [ ] is *e* (the base of the natural logarithms) raised to the computed power, *x* is one of the brightness values on the *x*-axis, $\hat{\mu}_i$ is the estimated mean of all the values in the forest training class, and $\hat{\sigma}_i^{\,2}$ is the estimated variance of all the measurements in this class. *Therefore, we need to store only the mean and variance of each training class (e.g., forest) to compute the probability function associated with any of the individual brightness values in it.*

Jensen, 2005

Maximum Likelihood Classification Algorithm

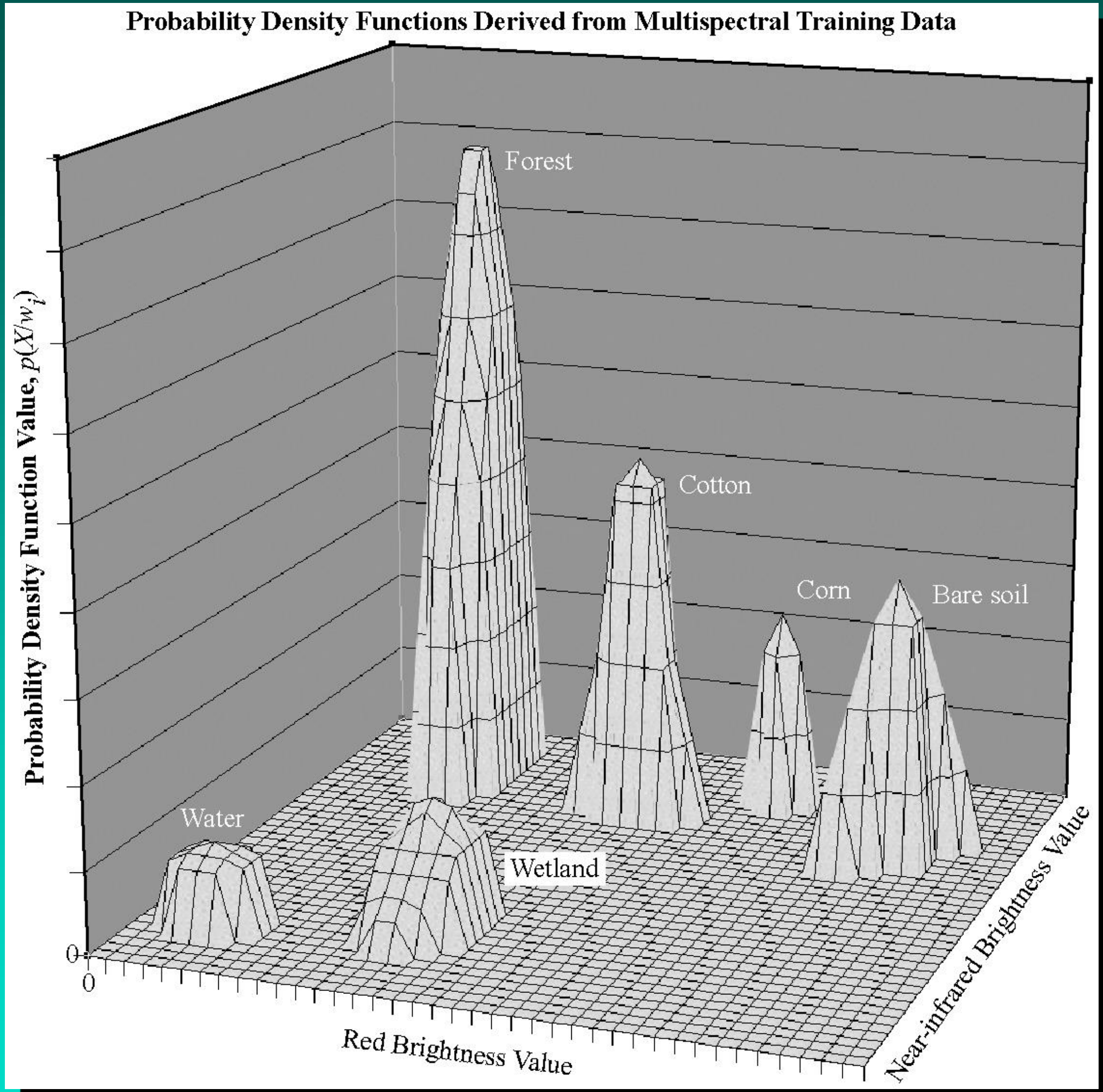
But what if our training data consists of multiple bands of remote sensor data for the classes of interest? In this case we compute an *n-dimensional multivariate normal density function* using:

$$p(X \mid w_i) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid V_i \mid^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (X - M_i)^T V_i^{-1} (X - M_i) \right]$$

where $|V_i|$ is the determinant of the covariance matrix, V_i^{-1} is the inverse of the covariance matrix, and $(X - M_i)^T$ is the transpose of the vector $(X - M_i)$. The mean vectors (M_i) and covariance matrix (V_i) for each class are estimated from the training data.

Jensen, 2005

Maximum Likelihood Classification Algorithm



For example, consider this illustration where the *bi-variate probability density functions* of six hypothetical classes are arrayed in red and near-infrared feature space. It is bi-variate because two bands are used. Note how the probability density function values appear to be normally distributed (i.e., bell-shaped). The vertical axis is associated with the probability of an unknown pixel measurement vector *X* being a member of one of the classes. In other words, if an unknown measurement vector has brightness values such that it lies within the wetland region, it has a high probability of being wetland.

data.

Jensen, 2005

Maximum Likelihood Classification Algorithm

If we assume that there are *m* classes, then $p(X/w_i)$ is the probability density function associated with the unknown measurement vector *X*, given that *X* is from a pattern in class w_i. In this case the *maximum likelihood decision rule* becomes:

Decide $X \in w_i$ if, and only if,

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

for all *i* and *j* out of 1, 2, ... *m* possible classes.

Therefore, to classify a pixel in the multispectral remote sensing dataset with an unknown measurement vector *X*, a maximum likelihood decision rule computes the product for each class and assigns the pattern to the class having the largest product. This assumes that we have some useful information about the prior probabilities of each class *i* (i.e., $p(w_i)$).

Jensen, 2005

Maximum Likelihood Classification Algorithm

Maximum Likelihood Classification Without Prior Probability Information:

In practice, we rarely have prior information about whether one class (e.g., forest) is expected to occur more frequently in a scene than any other class (e.g., 60% of the scene should be forest). This is called class *a priori* probability information (i.e., $p(w_i)$). Therefore, most applications of the maximum likelihood decision rule assume that each class has an equal probability of occurring in the landscape. This makes it possible to remove the prior probability term ($p(w_i)$) in Equation 9-23 and develop a simplified decision rule that can be applied to the unknown measurement vector X for each pixel in the scene:

Jensen, 2005

Maximum Likelihood Classification Algorithm

Maximum Likelihood Classification Without Prior Probability Information:

Decide unknown measurement vector *X* is in class *i* if, and only if,

$p_i \geq p_j$

for all *i* and *j* out of 1, 2, ... *m* possible classes

and

$$p_i = \frac{1}{2}\log_e |V_i| - \left[\frac{1}{2}\left(X - M_i\right)^T V_i^{-1}\left(X - M_i\right)\right]$$

where M_i is the mean measurement vector for class *i* and V_i is the covariance matrix of class *i* for bands *k* through *l*. Therefore, to assign the measurement vector *X* of an unknown pixel to a class, the maximum likelihood decision rule computes the value p_i for each class. Then it assigns the pixel to the class that has the largest (or maximum) value.

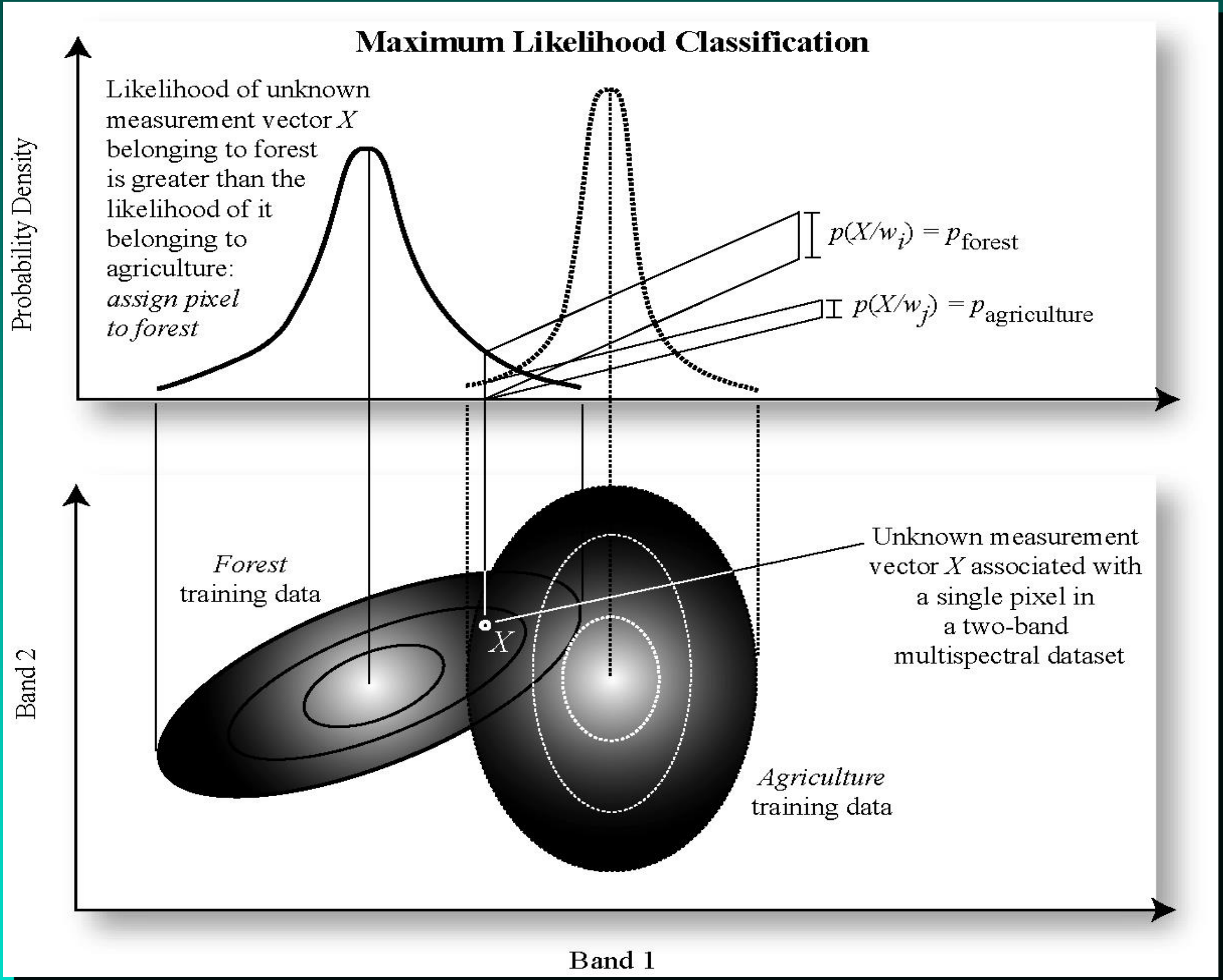
Jensen, 2005

Maximum Likelihood Classification Algorithm

Now let us consider the computations required. In the first pass, p_1 is computed with V_1 and M_1 being the covariance matrix and mean vectors for class 1. Next, p_2 is computed with V_2 and M_2 being the covariance matrix and mean vectors for class 2. This continues for all *m* classes. The pixel or measurement vector *X* is assigned to the class that produces the largest or maximum p_i. The measurement vector *X* used in each step of the calculation consists of *n* elements (the number of bands being analyzed). For example, if six Landsat TM bands (i.e., no thermal band) were being analyzed, each unknown pixel would have a measurement vector *X* of:

$$X = \begin{bmatrix} BV_{i,j,1} \\ BV_{i,j,2} \\ BV_{i,j,3} \\ BV_{i,j,4} \\ BV_{i,j,5} \\ BV_{i,j,7} \end{bmatrix}$$

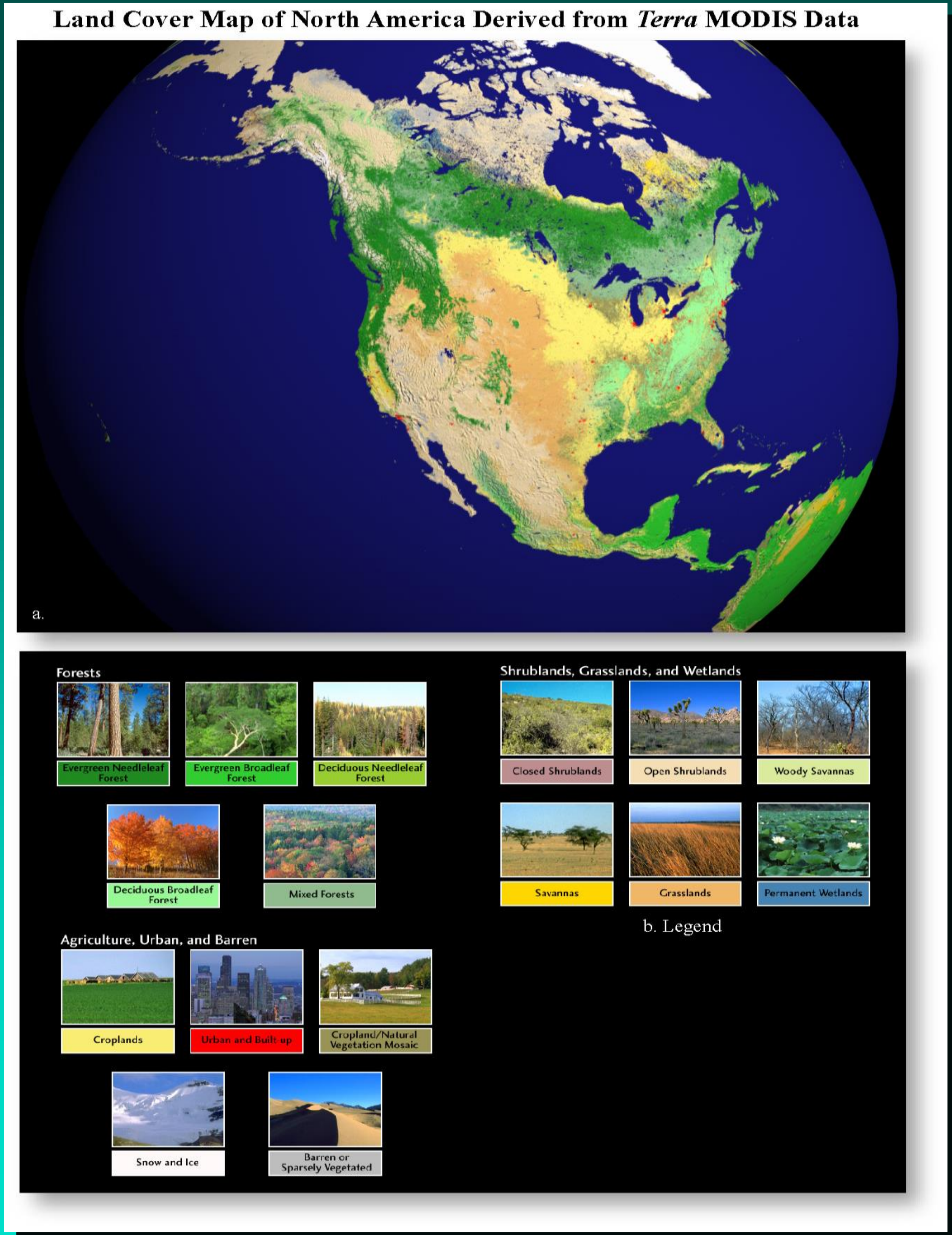
Jensen, 2005


Maximum Likelihood Classification
Probability Density
Likelihood of unknown measurement vector X belonging to forest is greater than the likelihood of it belonging to agriculture: *assign pixel to forest*
$p(X/w_i) = p_{\text{forest}}$
$p(X/w_j) = p_{\text{agriculture}}$
Band 2
Forest training data
X
Unknown measurement vector X associated with a single pixel in a two-band multispectral dataset
Agriculture training data
Band 1

What happens when the probability density functions of two or more training classes overlap in feature space? For example, consider two hypothetical normally distributed probability density functions associated with forest and agriculture training data measured in bands 1 and 2. In this case, pixel *X* would be assigned to forest because the probability density of unknown measurement vector *X* is greater for forest than for agriculture.

Jensen, 2005

MODIS Land Cover Mapping



Land Cover Map of North America Derived from *Terra* MODIS Data

a.
Forests
Evergreen Needleleaf Forest
Evergreen Broadleaf Forest
Deciduous Needleleaf Forest
Deciduous Broadleaf Forest
Mixed Forests
Shrublands, Grasslands, and Wetlands
Closed Shrublands
Open Shrublands
Woody Savannas
Savannas
Grasslands
Permanent Wetlands
Agriculture, Urban, and Barren
Croplands
Urban and Built-up
Cropland/Natural Vegetation Mosaic
Snow and Ice
Barren or Sparsely Vegetated
b. Legend

Jensen, 2005