Introductory Digital Image Processing

A Remote Sensing Perspective

Second Edition

John R. Jensen

Prentice Hall
Upper Saddle River, New Jersey 07458

Thematic Information Extraction: Image Classification

8

Introduction

Remotely sensed data of the Earth may be analyzed to extract useful thematic information. Notice that *data* are transformed into *information*. *Multispectral classification* is one of the most often used methods of information extraction. This procedure assumes that imagery of a specific geographic area is collected in multiple regions of the electromagnetic spectrum and that the images are in good geometric registration. The general steps required to extract land-cover information from digital remote sensor data are summarized in Figure 8-1. The actual multispectral classification may be performed using a variety of algorithms (Figures 8-1 and 8-2), including (1) hard classification using supervised or unsupervised approaches, (2) classification using fuzzy logic, and/or (3) hybrid approaches often involving the use of ancillary (collateral) information.

In a *supervised classification,* the identity and location of some of the land cover types, such as urban, agriculture, or wetland, are known *a priori* (before the fact) through a combination of fieldwork, analysis of aerial photography, maps, and personal experience (Mausel et al., 1990). 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 these training sites is then evaluated and assigned to the class of which it has the highest likelihood of being a member. This is often referred to as a hard classification (Figure 8-2a) because a pixel is assigned to only one class (e.g., forest), even though the sensor system records radiant flux from a mixture of biophysical materials within the IFOV, for example, 10% bare soil, 20% scrub shrub, and 70% forest (Foody et al., 1992).

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 (Jahne, 1991). The analyst then combines and relabels the spectral clusters into hard information classes (Figure 8-2a).

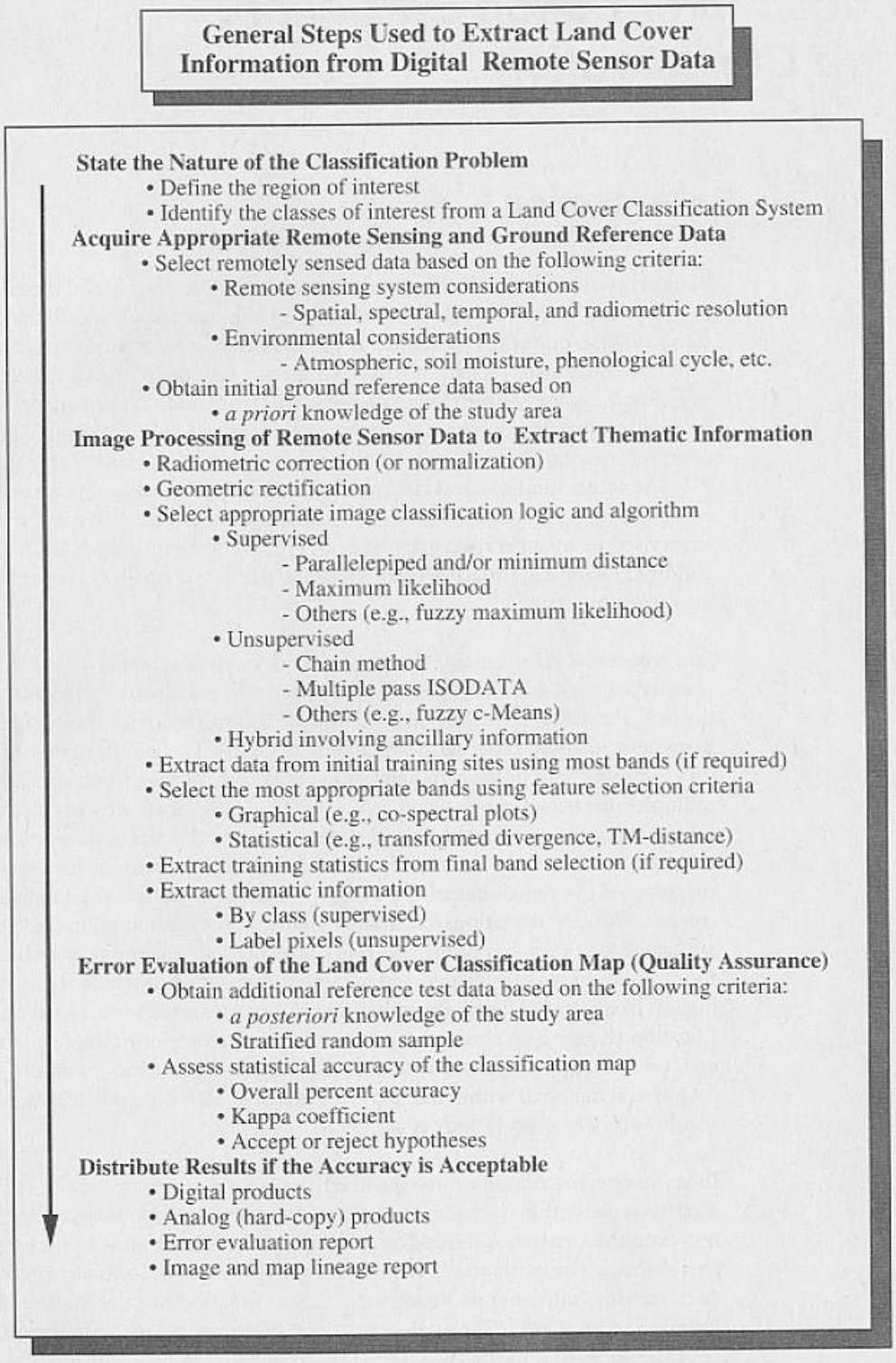


Figure 8-1 General steps required to extract land-cover information from digital remote sensor data.

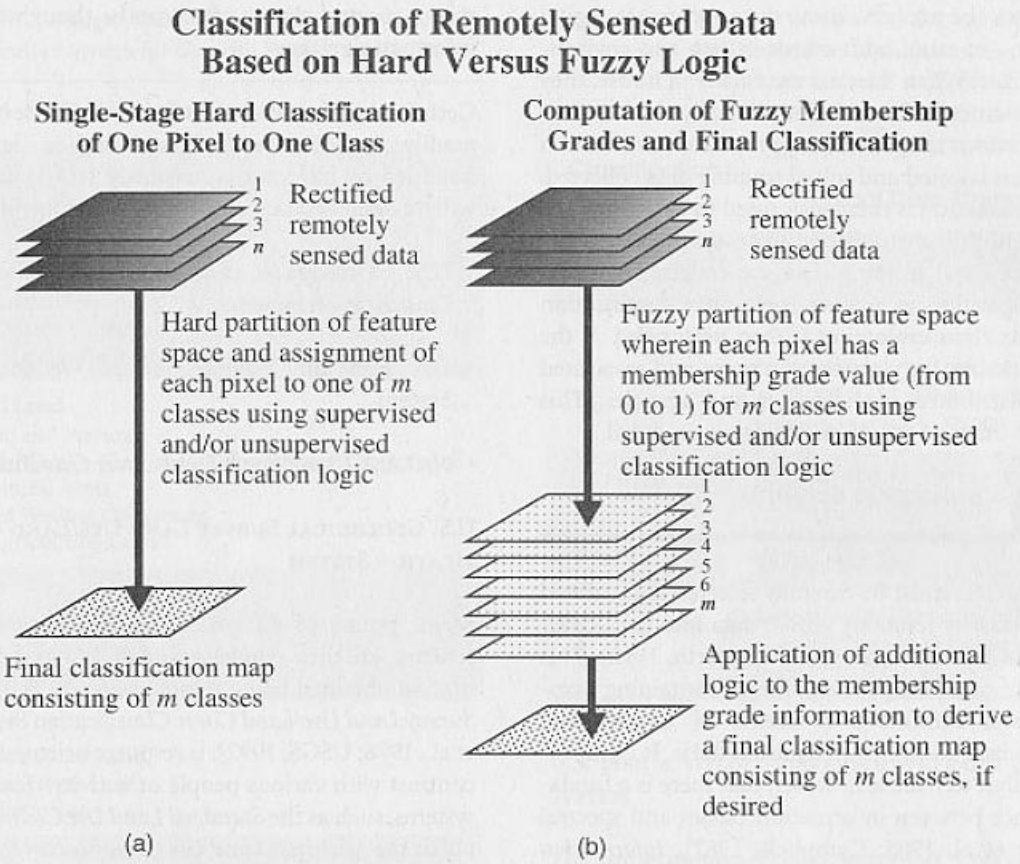


Figure 8-2 Difference between a traditional single-stage hard classification using supervised or unsupervised classification logic and classification using fuzzy logic.

Fuzzy set classification logic, which takes into account the heterogeneous and imprecise nature of the real world, may be used in conjunction with supervised and unsupervised classification algorithms. The IFOV of a sensor system normally records the reflected or emitted radiant flux from heterogeneous mixtures of biophysical materials such as soil, water, and vegetation. Also, the land-cover classes usually grade into one another without sharp, hard boundaries. Thus, reality is actually very imprecise and heterogeneous (Wang, 1990a and b; Lam, 1993). Unfortunately, we usually use very precise classical set theory to classify remotely sensed data into discrete, homogeneous information classes, ignoring the imprecision found in the real world. Instead of being assigned to a single class out of m possible classes, each pixel in a fuzzy classification has m membership grade values (to be discussed), each associated with how probable (or correlated) it is with each of the classes of interest (Figure 8-2b). This information may be used by the analyst to extract more precise land cover information, especially concerning the makeup of mixed pixels (Fisher and Pathirana, 1990; Foody and Trodd, 1993).

Sometimes it is necessary to include nonspectral *ancillary data* when performing a supervised, unsupervised, and/or fuzzy classification to extract the desired information. A variety of methods exists, including the use of geographic stratification, layered classification logic, and expert systems.

In this chapter, each major information extraction methodology is discussed in terms of (1) when it is appropriate, (2) important considerations that must be addressed, and (3) the nature of the expected results.

Supervised Classification

Useful supervised and unsupervised classification of remote sensor data may be obtained if the general steps summarized in Figure 8-1 are understood and carefully followed. The analyst first selects an appropriate region of interest on which to test hypotheses. The classes of interest to be tested in the hypothesis will dictate the nature of the classification system

to be used. Next, the analyst selects the appropriate digital imagery, keeping in mind both sensor system and environmental constraints. When the data are finally in house, they are usually radiometrically and geometrically corrected as discussed in previous chapters. An appropriate classification algorithm is then selected and initial training data collected. Feature (band) selection is then performed to determine the bands that are most likely to discriminate among the classes of interest. Additional training data are collected and the classification algorithm is applied, yielding a classification map. A rigorous error evaluation is then performed. If the results are acceptable, the classification maps and associated statistics are distributed to colleagues and agencies. This chapter reviews many of these considerations in detail.

Land-cover Classification Scheme

All classes of interest must be carefully selected and defined to successfully classify remotely sensed data into land-cover (or land-use) information (Gong and Howarth, 1992). This requires the use of a *classification scheme* containing taxonomically correct definitions of classes of information, which are organized according to logical criteria. It is important for the analyst to realize, however, that there is a fundamental difference between information classes and spectral classes (Jensen et al, 1983; Campbell, 1987). *Information classes* are those that human beings define. Conversely, *spectral classes* are those that are inherent in the remote sensor data and must be identified and then labeled by the analyst. For example, in a remotely sensed image of an urban area there is likely to be single-family residential housing. A relatively high spatial resolution (20×20 m) remote sensor such as SPOT might be able to record a few pure pixels of vegetation and a few pure pixels of asphalt road or shingles. However, it is more likely that in this residential area the pixel brightness values will be a function of the reflectance from mixtures of vegetation and concrete. Few planners or administrators want to see a map labeled with classes like (1) concrete, (2) vegetation, and (3) mixture of vegetation and concrete . Rather, they prefer the analyst to rename the mixture class as single-family residential (Westmoreland and Stow, 1992). The analyst should only do this if in fact there is a good association between the mixture class and single-family residential housing. Thus, we see that an analyst must often translate spectral classes into information classes to satisfy bureaucratic requirements. An analyst should understand well the spatial and spectral characteristics of the sensor system and be able to relate these system parameters to the types and proportions of materials found within the scene and within pixel IFOVs. If these parameters are understood, spectral classes often can be thoughtfully relabled as information classes.

Certain classification schemes have been developed that can readily incorporate land-use and/or land-cover data obtained by interpreting remotely sensed data. Only a few will be discussed here, including the following:

- U.S. Geological Survey Land Use/Land Cover Classification System
- U.S. Fish and Wildlife Service Wetland Classification System
- N.O.A.A. CoastWatch Land Cover Classification System

U.S. Geological Survey Land Use/Land Cover Classification System

Major points of difference between various classification schemes are their emphasis and ability to incorporate information obtained using remote sensing. The *U.S. Geological Survey Land Use/Land Cover Classification System* (Anderson et al., 1976; USGS, 1992), is resource oriented (land cover) in contrast with various people or activity (land use) oriented systems, such as the *Standard Land Use Coding (SLUC) Manual* or the *Michigan Land Use Classification System* (Jensen et al., 1983). 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 U.S.G.S. system addresses this need with eight of the nine level I categories treating land area that is not in urban or built-up categories (Table 8-1). The system is designed to be driven primarily by the interpretation of remote sensor data obtained at various scales and resolutions (Table 8-2) and not data collected *in situ*. It was initially developed to include land-use data that was visually photointerpreted, although it has been widely used for digital multispectral classification studies as well.

The *SLUC*, on the other hand, is land-use activity oriented and is primarily dependent on *in situ* observation to obtain remarkably specific land-use information, even to the contents of buildings (Rhind and Hudson, 1980). Obviously, there exists the need to merge the two approaches to produce a hybrid classification system that incorporates both land use interpreted from remote sensor data and very precise (and expensive) land-use information obtained *in situ* when necessary.

Table 8-1. U.S. Geological Survey Land Use/Land Cover Classification System for Use with Remote Sensor Data[a]

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
61 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

[a] Source: Anderson et al., 1976; USGS, 1992

Table 8-2. The Four Levels of the U.S. Geological Survey Land Use/Land Cover Classification System and the Type of Remotely Sensed Data Typically Used to Provide the Information

Classification Level	Typical Data Characteristics
I	Landsat MSS (79 × 79 m), Thematic Mapper (30 × 30 m), and SPOT XS (20 × 20 m)
II	SPOT Panchromatic (10 × 10 m) data or high-altitude aerial photography acquired at 40,000 ft (12,400 m) or above; results in imagery that is ≤1 : 80,000 scale
III	Medium-altitude data acquired between 10,000 and 40,000 ft (3100 and 12,400 m); results in imagery that is between 1 : 20,000 to 1 : 80,000 scale
IV	Low-altitude data acquired below 10,000 ft (3100 m); results in imagery that is larger than 1 : 20,000 scale

U.S. Fish & Wildlife Service Wetland Classification System

The conterminous United States lost 53% of its wetland to agricultural, residential, and/or commercial land use from 1780 to 1980 (Dahl, 1990). The U.S. Fish and Wildlife Service is responsible for mapping all 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 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 (Figure 8-3). 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. The distinguishing features of the riverine system are shown in Figure 8-4. This was the first nationally recognized wetland classification scheme.

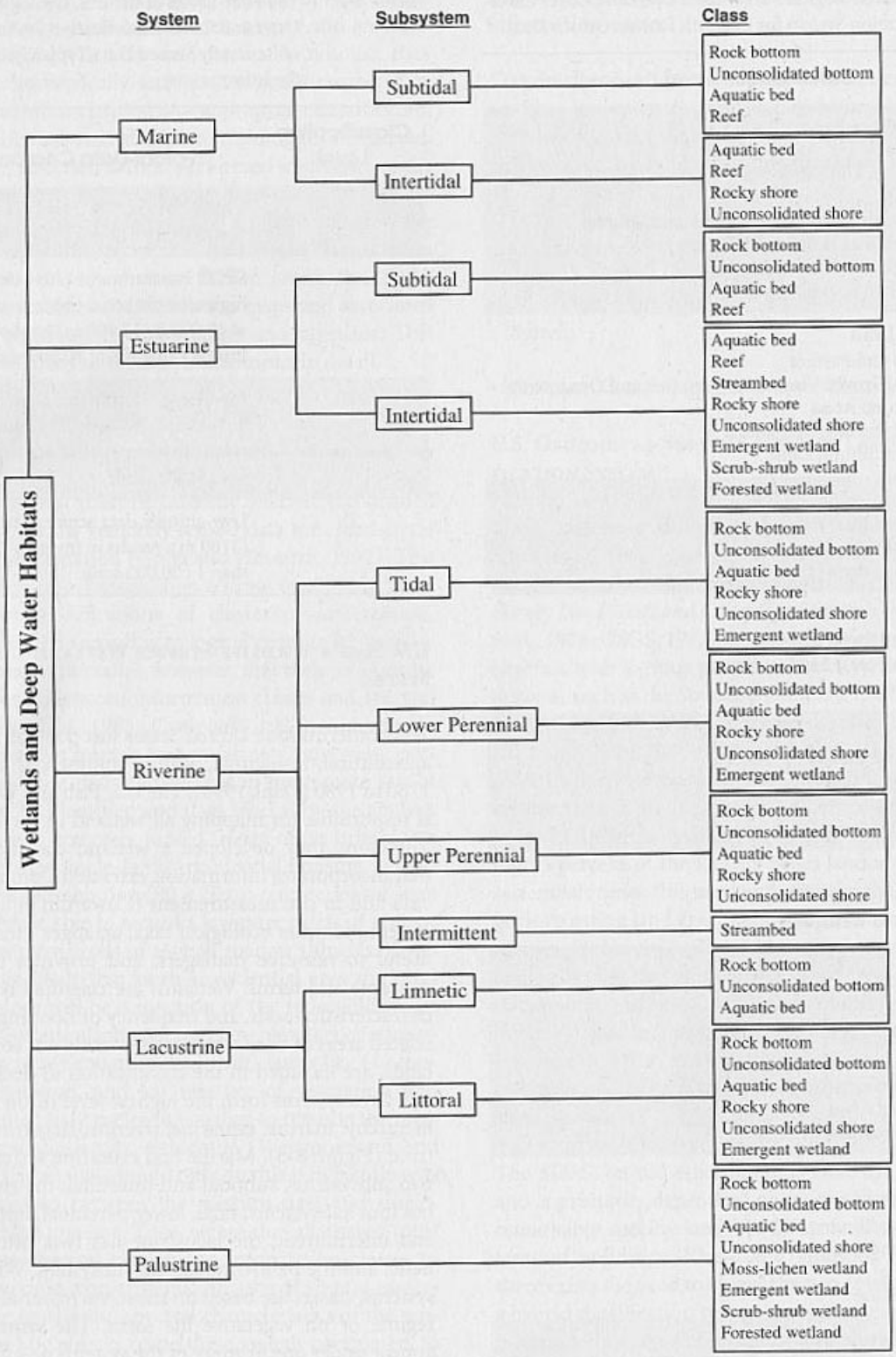


Figure 8-3 The U.S. Fish and Wildlife Service classification hierarchy of wetlands and deepwater habitats showing systems, subsystems, and classes (Cowardin et al., 1979). The palustrine system does not include deepwater habitats.

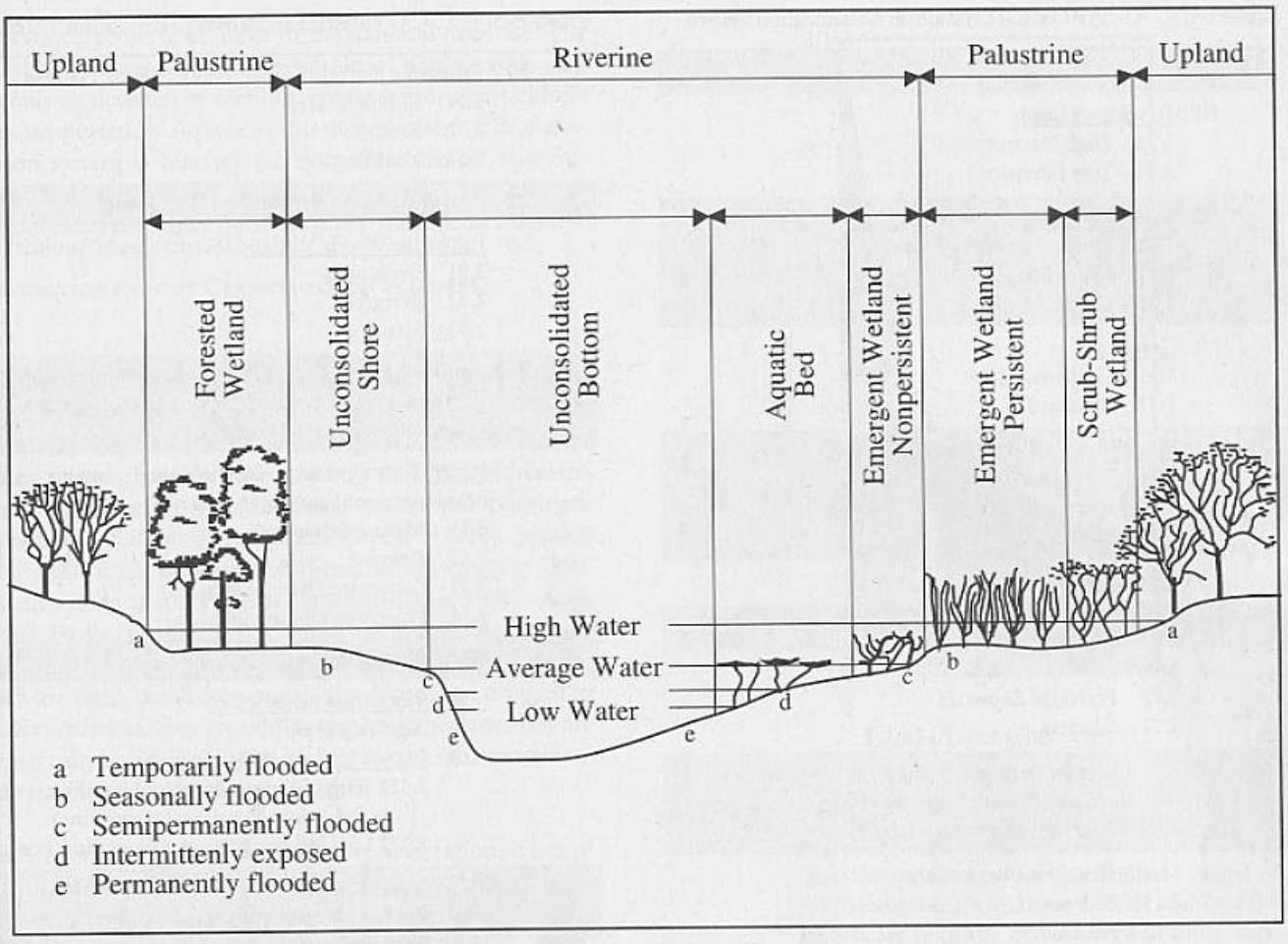


Figure 8-4 Distinguishing features and examples of habitats in the riverine system. (Cowardin et al., 1979).

NOAA CoastWatch Land Cover Classification System

Oil spills occurring throughout the world continue to devastate coastal wetland (Jensen et al., 1990). More abundant greenhouse gases in the atmosphere appear to be increasing Earth's average temperature and may produce a significant rise in global sea level, eventually inundating much of today's coastal wetlands (Cross and Thomas, 1992; Lee et al., 1992). Current projections for U.S. population growth in the coastal zone suggest accelerating losses of wetlands and adjacent habitats, as waste loads and competition for limited space and resources increase (U.S. Congress, 1989). Documentation of the loss or gain of coastal wetlands is needed now for their conservation and to effectively manage marine fisheries (Kiraly et al. 1990). Changes in wetlands are occurring too fast and too pervasively to be monitored as seldom as once a decade, or not at all in many instances.

For these reasons, the National Oceanic and Atmospheric Administration (NOAA) Coastal Ocean Program initiated the CoastWatch Change Analysis Project. The project utilizes digital remote sensor data, *in situ* measurement in conjunction with global positioning systems, and GIS technology to monitor changes in coastal wetland habitats and adjacent uplands (Dobson et al., 1995). Although the program stresses the use of satellite imagery (TM or SPOT), aerial photography may be used for mapping some upland and submerged habitats (Jensen et al., 1993a; Ferguson et al., 1993). The coastal regions of the United States are to be monitored every 1 to 5 years depending on the anticipated rate and magnitude of change in each region and the availability of suitable remote sensing and *in situ* measurements.

The CoastWatch database is taxonomically correct and in harmony with coastal wetland information derived from

Table 8-3. C-CAP Coastal Land Cover Classification System

- **1.0 Upland**
  - 1.1 Developed Land
    - 1.11 High Intensity
    - 1.12 Low Intensity
  - 1.2 Cultivated Land
    - 1.21 Orchards/Groves/Nurseries
    - 1.22 Vines/Bushes
    - 1.23 Cropland
  - 1.3 Grassland
    - 1.31 Unmanaged
    - 1.32 Managed
  - 1.4 Woody Land (Scrub-Shrub/Forested)
    - 1.41 Deciduous
    - 1.42 Evergreen
    - 1.43 Mixed
  - 1.5 Bare Land
  - 1.6 Tundra
  - 1.7 Snow/Ice
    - 1.71 Perennial Snow/Ice
    - 1.72 Glaciers
- **2.0 Wetland** (Excludes Bottoms, Reefs, Nonpersistent Emergent Wetlands, and Aquatic Beds, all of which are covered under 3.0, Water and Submerged Land)
  - 2.1 Marine/Estuarine Rocky Shore
    - 2.11 Bedrock
    - 2.12 Rubble
  - 2.2 Marine/Estuarine Unconsolidated Shore (Beach, Flat, Bar)
    - 2.21 Cobble-gravel
    - 2.22 Sand
    - 2.23 Mud/Organic
  - 2.3 Estuarine Emergent Wetland
    - 2.31 Haline (Salt Marsh)
    - 2.32 Mixohaline (Brackish Marsh)
  - 2.4 Estuarine Woody Wetland (Scrub-Shrub/Forested)
    - 2.41 Deciduous
    - 2.42 Evergreen
    - 2.43 Mixed
  - 2.5 Riverine Unconsolidated Shore (Beach, Flat, Bar)
    - 2.51 Cobble-gravel
    - 2.52 Sand
    - 2.53 Mud/Organic
  - 2.6 Lacustrine Unconsolidated Shore (Beach, Flat, Bar)
    - 2.61 Cobble-gravel
    - 2.62 Sand
    - 2.63 Mud/Organic
  - 2.7 Palustrine Unconsolidated Shore (Beach, Flat, Bar)
    - 2.71 Cobble-gravel
    - 2.72 Sand
    - 2.73 Mud/Organic
  - 2.8 Palustrine Emergent Wetland (Persistent)
  - 2.9 Palustrine Woody Wetland (Scrub-Shrub/Forested)
    - 2.91 Deciduous
    - 2.92 Evergreen
    - 2.93 Mixed
- **3.0 Water and Submerged Land** (Includes deepwater habitats and those wetlands with surface water but *lacking* trees, shrubs, and persistent emergents)
  - 3.1 Water (Bottoms and undetectable reefs, aquatic beds or nonpersistent emergent wetlands)
    - 3.11 Marine/Estuarine
    - 3.12 Riverine
    - 3.13 Lacustrine (Basin ≥20 acres)
    - 3.14 Palustrine (Basin <20 acres)
  - 3.2 Marine/Estuarine Reef
  - 3.3 Marine/Estuarine Aquatic Bed
    - 3.31 Algal (e.g., kelp)
    - 3.32 Rooted Vascular (e.g., seagrass)
      - 3.321 (High Salinity (≥5 ppt; Mesosaline, Polysaline, Eusaline, Hypersaline)
      - 3.322 Low Salinity (<5 ppt; Oligosaline, Fresh)
  - 3.4 Riverine Aquatic Bed
    - 3.41 Rooted Vascular/Algal/Aquatic Moss
    - 3.42 Floating Vascular
  - 3.5 Lacustrine Aquatic Bed (Basin ≥20 acres)
    - 3.51 Rooted Vascular/Algal/Aquatic Moss
    - 3.52 Floating Vascular
  - 3.6 Palustrine Aquatic Bed (Basin <20 acres)
    - 3.61 Rooted Vascular/Algal/Aquatic Moss
    - 3.62 Floating Vascular

other U.S. agencies (USGS, USF&WS, EPA). The CoastWatch Coastal Land Cover Classification System (Table 8-3) includes three Level I superclasses (Klemas et al., 1993):

1.0 Upland
2.0 Wetland
3.0 Water and Submerged Land

These are subdivided into classes and subclasses at levels II and III, respectively. While the latter two categories are the primary areas of interest, uplands are also included because they influence adjacent wetlands and water bodies. The underlined classes in Table 8-3 must be provided in regional

CoastWatch projects as input to the national database. The underlined classes, with the exception of aquatic beds, can generally be detected by satellite remote sensors, particularly when supported by surface *in situ* measurement. The classification system is hierarchical, reflects ecological relationships, and focuses on land-cover classes that can be discriminated primarily from satellite remote sensor data.

Observations about Classification Schemes

Geographical information (including remote sensor data) is often imprecise. For example, there is usually a gradual interface at the edge of forests and rangeland (where remote sensing mixed pixels are encountered), yet all the aforementioned classification schemes insist on a hard boundary between the classes. The schemes should actually contain fuzzy definitions because the thematic information contained in them is fuzzy (Fisher and Pathirana, 1990; Wang, 1990a). Fuzzy classification schemes are not currently available. Therefore, we must use existing classification schemes, which are rigid, based on *a priori* knowledge, and difficult to use. Nevertheless, they are widely employed because they are scientifically based, and individuals using the same classification system can compare their results.

This brings us to another important consideration. If a reputable classification system already exists, it is foolish to develop an entirely new system that will probably only be used by ourselves. It is better to adopt or modify existing nationally 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 (Rhind and Hudson, 1980).

Finally, it should be noted that 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) summarized this relationship for the mapping of urban/suburban land use and land cover in the United States (Figure 8-5). A similar relationship exists when mapping vegetation (Botkin et al., 1984). For example, the sensor systems and spatial resolutions useful for discriminating vegetation from a global to an *in situ* perspective are summarized in Figure 8-6. 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. Spectral resolution is also an important consideration. However, it is not as critical a parameter as spatial resolution since most of the sensor systems (e.g., Landsat MSS or SPOT HRV) record energy in approximately the same green, red, and near-infrared regions of the electromagnetic spectrum

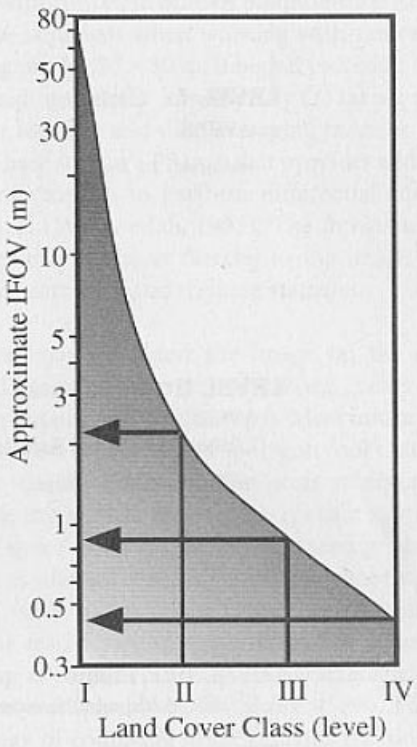


Figure 8-5 Spatial resolution (IFOV) requirements as a function of the mapping requirements for levels I to IV land-cover classes in the United States (based on Anderson et al., 1976). Levels I, II, III, and IV information are normally derived from satellite, high-, medium-, and low-altitude image data, respectively. Note the dramatic increase in resolution required to map level II classes (from Welch, 1982; Jensen et al., 1983).

(except for the Landsat TM, which has blue, middle-infrared, and thermal-infrared bands).

Training Site Selection and Statistics Extraction

An analyst may select *training sites* within the image that are representative of the land-cover 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 applicable. However, if the soil conditions should change dramatically across the study area (e.g., one-half of the region has a perched water table

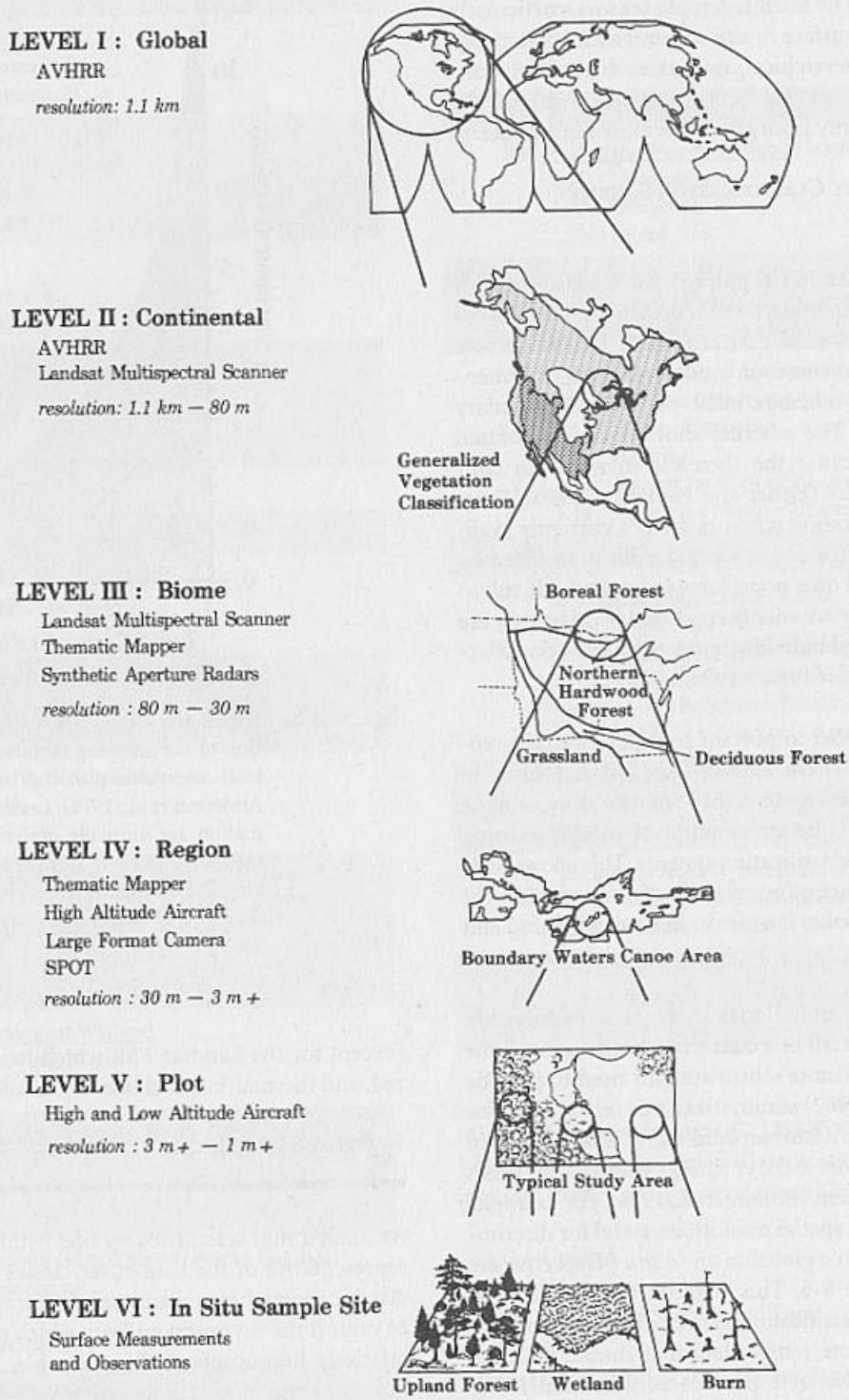


Figure 8-6 Relationship between the level of detail required and the spatial resolution of various remote sensing systems for vegetation inventories (Botkin, et al., 1984).

with very moist near-surface soil), it is likely that grassland training data acquired in the dry soil part of the study area would *not* be representative of the spectral conditions for grassland found in the moist soil portion of the study area. Thus, we have a *signature extension* problem meaning that it may not be possible to extend our grassland training data through x, y space.

The easiest way to remedy this situation is by using *geographical stratification* during the preliminary stages of a project. At this time all significant environmental factors that contribute to 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 thundershower 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 will probably have to be classified separately. The final classification map of the entire region will 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 cost and effort involved with retraining. Additional research is required before the concept of spatial and temporal (through time) signature extension is fully understood.

Once 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 $>10n$ pixels of training data are collected for each class. This is sufficient to compute the variance–covariance matrices required by some classification algorithms.

There are a number of ways to actually collect the *training site* data, including (1) collection of *in situ* information, such as tree height, percent canopy closure, and diameter-at-breast-height (dbh), (2) on-screen selection of polygonal training data, and/or (3) on-screen seeding of training data. Ideally, the sites are visited in the field and their perimeter and/or centroid coordinates obtained from a planimetric map or measured directly using a global positioning system (GPS). When U.S. government "selective availability" is "on" the GPS x, y coordinates from a single hand-held receiver should be within ±100 m of their planimetric position which may not be sufficient when working with remotely sensed data having pixels ≤30 × 30 m. If higher precision is required, the GPS readings may be improved by (1) taking more readings at one location and then averaging them or (2) having access to a base station GPS unit that provides additional calibration information to perform differential correction of the GPS data (Welch et al., 1992). The *in situ* x, y training coordinates may be input directly to the image processing system to extract per band training statistics.

The analyst may also view the image on the color CRT screen and select polygons of interest (e.g., fields containing different types of agricultural crops). Most image processing systems utilize a "rubber band" polygon tool that allows the analyst to identify fairly specific areas of interest (AOI). Conversely, the analyst may seed a specific x, y 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 amoebae as long as it finds pixels with characteristics similar to the original seed pixel (e.g., Skidmore, 1989). This is a very effective way of collecting homogeneous training information.

If the analyst trains on six bands of Landsat thematic mapper data, then each pixel in each training site is represented by a *measurement vector*, X_c, such that

$$X_c = \begin{matrix} BV_{ij1} \\ BV_{ij2} \\ BV_{ij3} \\ BV_{ij4} \\ \vdots \\ BV_{ijk} \end{matrix} \quad (8\text{-}1)$$

where BV_{ijk} is the brightness value for the i, jth pixel in band k. 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:

$$M_c = \begin{matrix} \mu_{c1} \\ \mu_{c2} \\ \mu_{c3} \\ \mu_{c4} \\ \vdots \\ \mu_{ck} \end{matrix} \quad (8\text{-}2)$$

where μ_{ck} represents the mean value of the data obtained for class c in band k. The raw measurement vector can also be analyzed to yield the covariance matrix for each class c:

$$V_c = V_{ckl} = \begin{bmatrix} \text{Cov}_{c11} & \text{Cov}_{c12} & \cdots & \text{Cov}_{c1n} \\ \text{Cov}_{c21} & \text{Cov}_{c22} & \cdots & \text{Cov}_{c2n} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}_{cn1} & \text{Cov}_{cn2} & \cdots & \text{Cov}_{cnn} \end{bmatrix} \quad (8\text{-}3)$$

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., $V_{dkl} = V_d$).

The mean, standard deviation, variance, minimum value, maximum value, variance–covariance matrix, and correlation matrix for the training statistics of five Charleston, S.C., land-cover classes (residential, commercial, wetland, forest, and water) are listed in Table 8-4. These represent fundamental information on the spectral characteristics of the five classes.

Observations About Training Class Selection

Sometimes the manual selection of polygons results in the collection of training data with multiple modes in a training class histogram. This suggests that there are at least two different types of land cover within the training area. This condition is not good when we are attempting to discriminate between individual classes. Therefore, it is a good practice to discard multimodal training data and retrain on specific parts of the polygon of interest until unimodal histograms are derived per class.

Positive spatial *autocorrelation* exists among pixels that are contiguous or close together (Griffith, 1987; Gong and Howarth, 1992). This means that adjacent pixels have a high probability of having similar brightness values. Training data collected from autocorrelated data tend to have reduced variance which may be caused more from the way the sensor is collecting the data than from actual field conditions (e.g., most detectors dwell on an individual pixel for a very short time and may smear spectral information from one pixel to an adjacent pixel). The ideal situation is to collect training data within a region using every nth pixel or some other sampling criteria (Labovitz and Masuoka, 1984). The goal is to get nonautocorrelated training data. Unfortunately, most digital image processing systems do not provide this option in training data collection modules.

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 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 process minimizes the cost of the digital image classification process (but should not affect the accuracy). Feature selection may involve both statistical and/or 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. Statistical measures such as divergence will be discussed shortly.

Why use graphical methods of feature selection if statistical techniques provide all the information necessary to select the most appropriate bands for classification? The reason is simple. An analyst may base a decision solely on the statistic, yet never obtain a fundamental understanding of the spectral nature of the data being analyzed. In effect, without ever visualizing where the spectral measurements cluster in n-dimensional feature space, each new supervised classification finds the analyst beginning anew, relying totally on the abstract statistical analysis. Many of the practitioners of remote sensing are by necessity very graphically literate; that is, they can readily interpret maps and graphs (Dent, 1993). Therefore, a graphic display of the statistical data is useful and often necessary for a thorough analysis of multispectral training data and feature selection. Several graphic feature selection methods have been developed for this purpose.

Graphic Methods of Feature Selection

Bar graph spectral plots were one of the first simple feature selection aids where the mean $\pm 1\sigma$ are displayed in a bar graph format for each band (Figure 8-7). This provides an effective visual presentation of the degree of between-class separability for one band at a time. In the example, band 3 is only useful for discriminating between water (class 1) and all other classes. Bands 1 and 2 appear to provide good separability between most of the classes (with the possible exception of classes 5 and 6). The display provides no information on how well any two bands would perform.

Table 8-4. Univariate and Multivariate Training Statistics for the Five Land-cover Classes Using Six Bands of Landsat Thematic Mapper Data Obtained over Charleston, South Carolina

a. Statistics for Residential

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Univariate statistics						
Mean	70.6	28.8	29.8	36.7	55.7	28.2
Std. dev.	6.90	3.96	5.65	4.53	10.72	6.70
Variance	47.6	15.7	31.9	20.6	114.9	44.9
Minimum	59	22	19	26	32	16
Maximum	91	41	45	52	84	48
Variance-covariance matrix						
1	47.65					
2	24.76	15.70				
3	35.71	20.34	31.91			
4	12.45	8.27	12.01	20.56		
5	34.71	23.79	38.81	22.30	114.89	
7	30.46	18.70	30.86	12.99	60.63	44.92
Correlation matrix						
1	1.00					
2	0.91	1.00				
3	0.92	0.91	1.00			
4	0.40	0.46	0.47	1.00		
5	0.47	0.56	0.64	0.46	1.00	
7	0.66	0.70	0.82	0.43	0.84	1.00

b. Statistics for Commercial

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Univariate statistics						
Mean	112.4	53.3	63.5	54.8	77.4	45.6
Std. dev.	5.77	4.55	3.95	3.88	11.16	7.56
Variance	33.3	20.7	15.6	15.0	124.6	57.2
Minimum	103	43	56	47	57	32
Maximum	124	59	72	62	98	57

b. Statistics for Commercial (Continued)

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Variance-covariance matrix						
1	33.29					
2	11.76	20.71				
3	19.13	11.42	15.61			
4	19.60	12.77	14.26	15.03		
5	–16.62	15.84	2.39	0.94	124.63	
7	–4.58	17.15	6.94	5.76	68.81	57.16
Correlation matrix						
1	1.00					
2	0.45	1.00				
3	0.84	0.64	1.00			
4	0.88	0.72	0.93	1.00		
5	–0.26	0.31	0.05	0.02	1.00	
7	–0.10	0.50	0.23	0.20	0.82	1.00

c. Statistics for Wetland

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Univariate statistics						
Mean	59.0	21.6	19.7	20.2	28.2	12.2
Std. dev.	1.61	0.71	0.80	1.88	4.31	1.60
Variance	2.6	0.5	0.6	3.5	18.6	2.6
Minimum	54	20	18	17	20	9
Maximum	63	25	21	25	35	16
Variance-covariance matrix						
1	2.59					
2	0.14	0.50				
3	0.22	0.15	0.63			
4	–0.64	0.17	0.60	3.54		
5	–1.20	0.28	0.93	5.93	18.61	
7	–0.32	0.17	0.40	1.72	4.53	2.55

c. Statistics for Wetland (Continued)

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Correlation matrix						
1	1.00					
2	0.12	1.00				
3	0.17	0.26	1.00			
4	–0.21	0.12	0.40	1.00		
5	–0.17	0.09	0.27	0.73	1.00	
7	–0.13	0.15	0.32	0.57	0.66	1.00

d. Statistics for Forest

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Univariate statistics						
Mean	57.5	21.7	19.0	39.1	35.5	12.5
Std. dev.	2.21	1.39	1.40	5.11	6.41	2.97
Variance	4.9	1.9	1.9	26.1	41.1	8.8
Minimum	53	20	17	25	22	8
Maximum	63	28	24	48	54	22
Variance-covariance matrix						
1	4.89					
2	1.91	1.93				
3	2.05	1.54	1.95			
4	5.29	3.95	4.06	26.08		
5	9.89	5.30	5.66	13.80	41.13	
7	4.63	2.34	2.22	3.22	16.59	8.84
Correlation matrix						
1	1.00					
2	0.62	1.00				
3	0.66	0.80	1.00			
4	0.47	0.56	0.57	1.00		
5	0.70	0.59	0.63	0.42	1.00	
7	0.70	0.57	0.53	0.21	0.87	1.00

e. Statistics for Water

:	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Univariate statistics						
Mean	61.5	23.2	18.3	9.3	5.2	2.7
Std. dev.	1.31	0.66	0.72	0.56	0.71	1.01
Variance	1.7	0.4	0.5	0.3	0.5	1.0
Minimum	58	22	17	8	4	0
Maximum	65	25	20	10	7	5
Variance-covariance matrix						
1	1.72					
2	0.06	0.43				
3	0.12	0.19	0.51			
4	0.09	0.05	0.05	0.32		
5	–0.26	–0.05	–0.11	–0.07	0.51	
7	–0.21	–0.05	–0.03	–0.07	0.05	1.03
Correlation matrix						
1	1.00					
2	0.07	1.00				
3	0.13	0.40	1.00			
4	0.12	0.14	0.11	1.00		
5	–0.28	–0.10	–0.21	–0.17	1.00	
7	–0.16	–0.08	–0.04	–0.11	0.07	1.00

Cospectral mean vector plots may be used to present statistical information about at least two bands at one time. Hodgson and Plews (1989) provided several methods for displaying the mean vectors for each class in two- and three-dimensional feature space. For example, in Figure 8-8a we see 49 mean vectors derived from Charleston, S.C., TM data arrayed in two-dimensional feature space (bands 3 and 4). Theoretically, the greater the distance is between numbers in the feature space distribution, the greater the potential for accurate between-class discrimination. Using this method, only two bands of data may be analyzed at one time. Therefore, they devised an alternative method whereby the size of the numeral depicts the location of information in a third dimension of feature space (Figure 8-8b). For example, Figure 8-8c depicts the same 49 mean vectors in simulated three-dimensional feature space (bands 2, 3, and 4). Normal viewing of the *trispectral mean vector plot* looks down the z axis; thus the z axis is not seen. Scaling of the numeral size is performed by linear scaling:

$$\text{Size} = \frac{BV_{ck}}{\text{quant}_k} * \text{MaxSize} \qquad (8\text{-}4)$$

where

Size = the numeral size in feature space
BV_{ck} = brightness value in class c for band k depicted by the z axis
quant_k = quantization level of band k (e.g., 0 to 255)
MaxSize = maximum numeral size

Size and *MaxSize* are in the units of the output device (e.g., inches for a pen plotter or pixels for a raster display). By

Spectral Plots

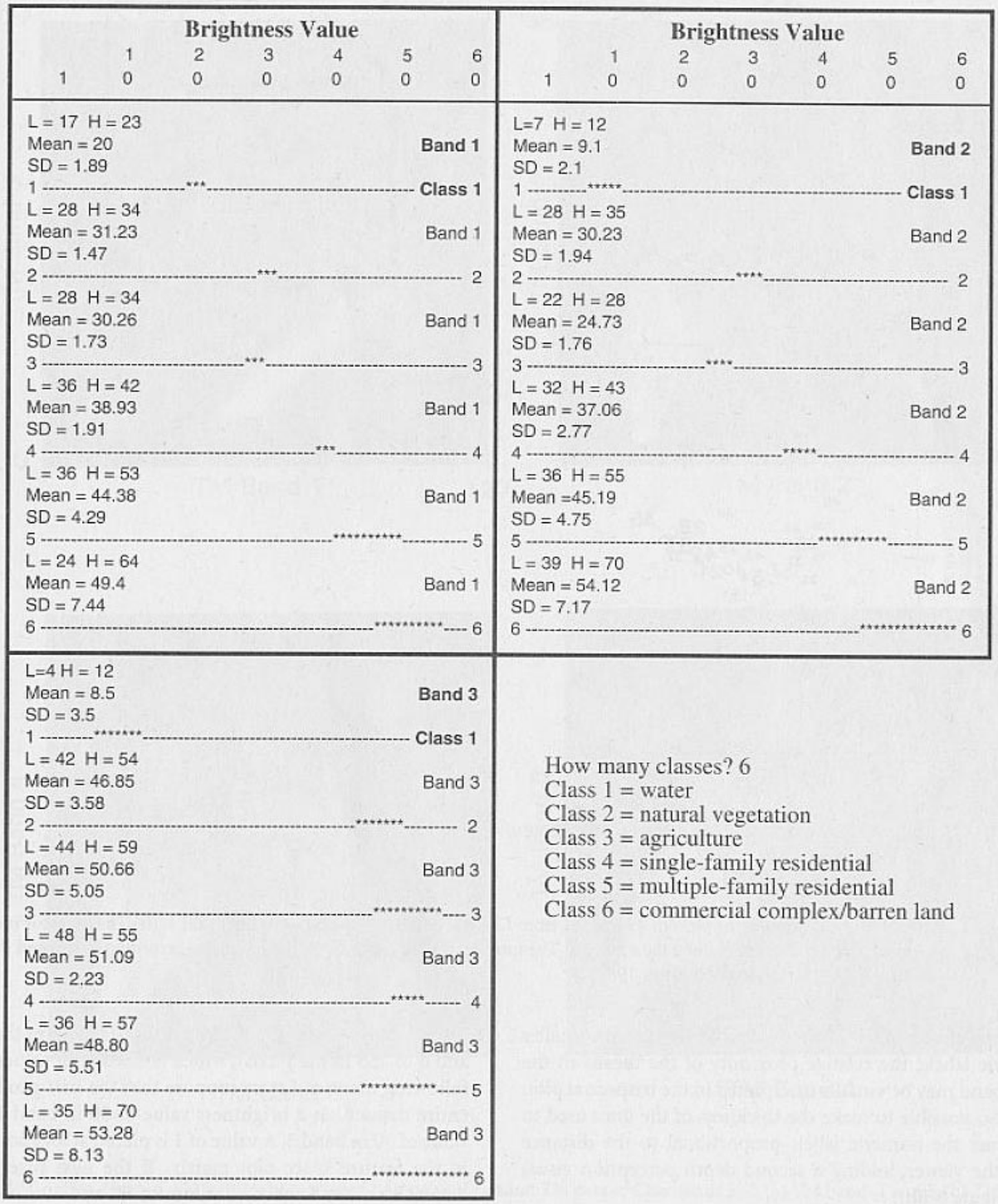


Figure 8-7 Bar graph spectral plots of data analyzed by Jensen (1979). Training statistics (the mean ±1 standard deviation) for six land-cover classes are displayed for three Landsat MSS bands. The display can be used to identify between-class separability for each class and single band.

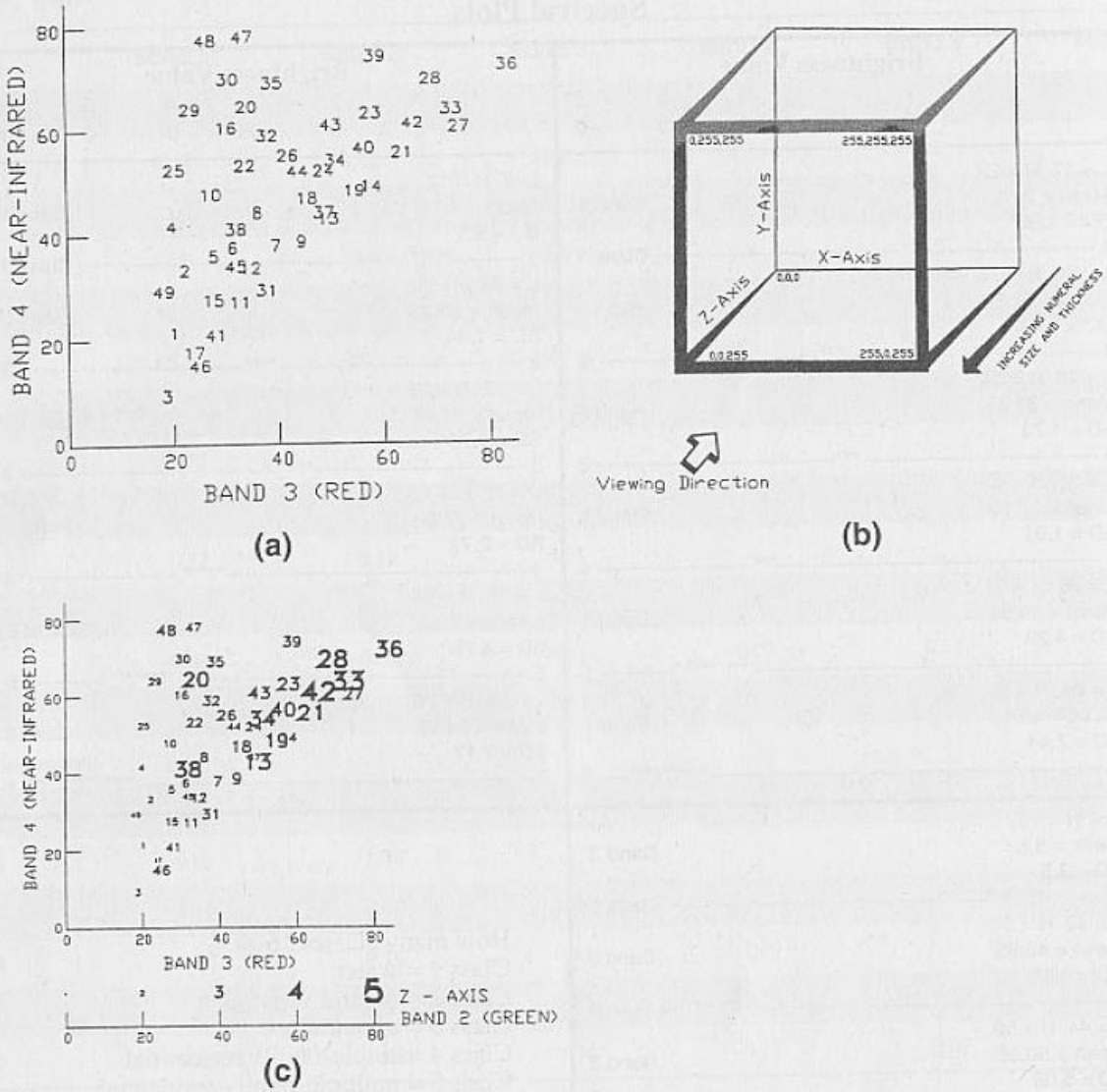


Figure 8-8 (a) Cospectral mean vector plots of 49 clusters from Charleston, S.C., TM data in bands 3 and 4. (b) The logic for increasing numeral size and thickness along the z axis. (c) The introduction of band 2 information scaled according to size and thickness along the z axis (Hodgson and Plews, 1989).

depicting cluster labels farther from the viewer with smaller numeric labels, the relative proximity of the means in the third band may be visually interpreted in the trispectral plot. It is also possible to make the thickness of the lines used to construct the numeric labels proportional to the distance from the viewer, adding a second depth perception visual cue (Figure 8-8b).

Feature space plots in two dimensions depict the distribution of all the pixels in the scene using two bands at a time (Figure 8-9). Such plots are often used as a backdrop for the display of various graphic feature selection methods. A typical plot usually consists of a 256×256 matrix (0 to 255 in the x axis and 0 to 255 in the y axis), which is filled with values in the following manner. Let us suppose that the first pixel in the entire dataset has a brightness value of 50 in band 1 and a value of 30 in band 3. A value of 1 is placed at location 50, 30 in the feature space plot matrix. If the next pixel in the dataset also has brightness values of 50 and 30 in bands 1 and 3, the value of this cell in the feature space matrix is incremented by 1, becoming 2. This logic is applied to each pixel in the scene. The brighter the pixel is in the feature space plot display, the greater the number of pixels having the same values in the two bands of interest. Feature space

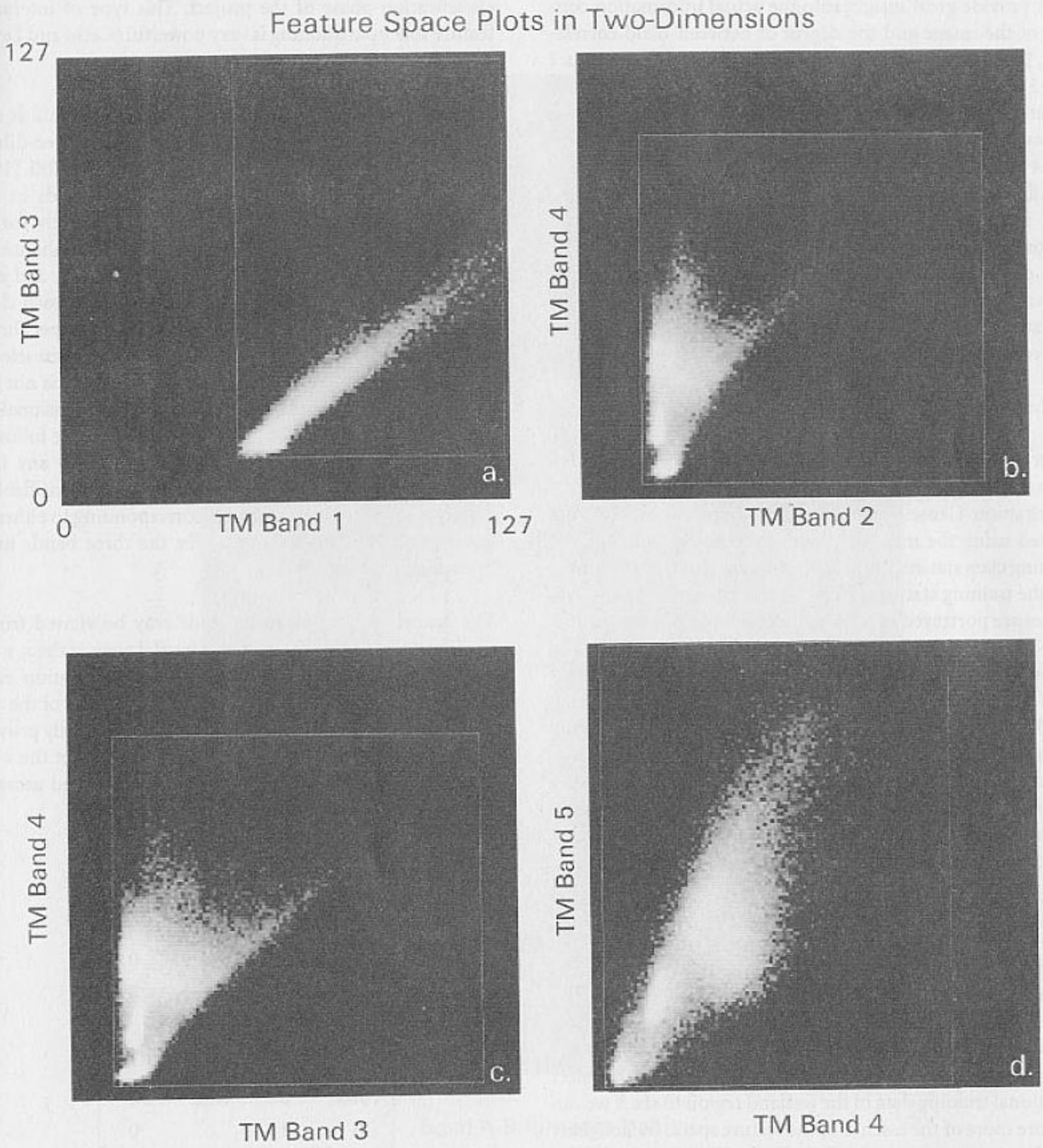


Figure 8-9 Two-dimensional feature space plots of four pairs of Landsat TM data of Charleston, S.C. (a) TM Bands 1 and 3, (b) TM bands 2 and 4, (c) TM bands 3 and 4, and (d) TM bands 4 and 5. The brighter a particular pixel is in the display, the more pixels within the scene having that unique combination of band values.

plots provide great insight into the actual information content of the image and the degree of between-band correlation. For example, in Figure 8-9a it is obvious that bands 1 and 3 are highly correlated and that atmospheric scattering in band 1 (blue) results in a significant shift of the brightness values down the x axis. Conversely, plots of bands 2 (green) and 4 (near-infrared) and 3 (red) and 4 have a much greater distribution of pixels within the spectral space and some very interesting bright locations, which correspond with important land cover types (Figures 8-9b and c). Finally, the plot of bands 4 (near-infrared) and 5 (middle infrared) shows exceptional dispersion throughout the spectral space and some very interesting bright locations (Figs 8-9d). For this reason, a spectral space plot of bands 4 and 5 will be used as a backdrop for the next graphic feature selection method.

Cospectral parallelepiped or *ellipse plots* in two-dimensional feature space provide useful visual between-class separability information (Jensen and Toll, 1982; Jain, 1989). They are created using the mean, μ_{ck}, and standard deviation, s_{ck}, of training class statistics for each class c and band k. For example, the training statistics for five Charleston, S.C. land-cover classes are portrayed in this manner and draped over the feature space plot of TM bands 4 and 5 in Figure 8-10. The lower and upper limits of the two-dimensional parallelepipeds (rectangles) were obtained using the mean $\pm 1\sigma$ of each band for each class. If only band 4 data were used to classify the scene, there would be confusion between classes 1 and 4, and if only band 5 data were used there would be confusion between classes 3 and 4. However, when band 4 and 5 data are used at the same time to classify the scene there appears to be good between-class separability among the five classes (at least a $\pm 1\sigma$). An evaluation of Figure 8-10 reveals that there are numerous water pixels in the scene found near the origin in bands 4 and 5. The water training class is located in this region. Similarly, the wetland training class is situated within the bright wetland region of band 4 and 5 spectral space. However, it appears that training data were not collected in the heart of the wetland region of spectral space. Such information is valuable because we may want to collect additional training data in the wetland region to see if we can capture more of the essence of the feature space. In fact, there may be two or more wetland classes residing in this portion of spectral space. Sophisticated image processing systems allow the analyst to select training data directly from this type of display, which contains (1) the training class parallelepipeds and (2) the feature space plot. The analyst uses the cursor to interactively select training locations (they may be polygonal areas, not just parallelepipeds) within the feature space (Baker et al., 1991). If desired, these feature space partitions can be used as the actual decision logic during the classification phase of the project. This type of interactive feature space partitioning is very powerful (Cetin and Levandowski, 1991).

It is possible to display three bands of training data at once using *trispectral parallelepipeds* or *ellipses* in three-dimensional feature space (Figure 8-11). Jensen and Toll (1982) presented a method of displaying parallelepipeds in synthetic three-dimensional space and of interactively varying the viewpoint azimuth and elevation angles to enhance feature analysis and selection. Again, the mean, μ_{ck}, and standard deviation, s_{ck}, of training class statistics for each class c and band k are used to identify the lower and upper threshold values for each class and band. The analyst then selects a combination of three bands to portray because it is not possible to use all six bands at once in a three-dimensional display. Landsat TM bands 4, 5, and 7 are used in the following example; however, the method is applicable to any three band subset. Each corner of a parallelepiped is identifiable by a unique set of x, y, z coordinates corresponding to either the lower or upper threshold value for the three bands under investigation (Figure 8-11).

The corners of the parallelepipeds may be viewed from a vantage point other than a simple frontal view of the x, y axes using three-dimensional coordinate transformation equations. The feature space may be rotated about any of the axes, although rotation around the x and y axes normally provides a sufficient number of viewpoints. Rotation about the x-axis ϕ radians and the y-axis θ radians is implemented using the following equations (Hodgson and Plews, 1989):

$$[X,Y,Z,1]^{P^{T'}} = [BVx,BVy,BVz,1]^{P^{T}} * \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi & 0 \\ 0 & \sin\phi & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8\text{-}5)$$

Negative signs of ϕ or θ are used for counterclockwise rotation and positive signs for clockwise rotation. This transformation causes the original brightness value coordinates, p^T, to be shifted about and contain depth information as vector $P^{T'}$. Display devices are two dimensional (e.g., plotter surfaces or cathode-ray-tube screens); only the x and y elements

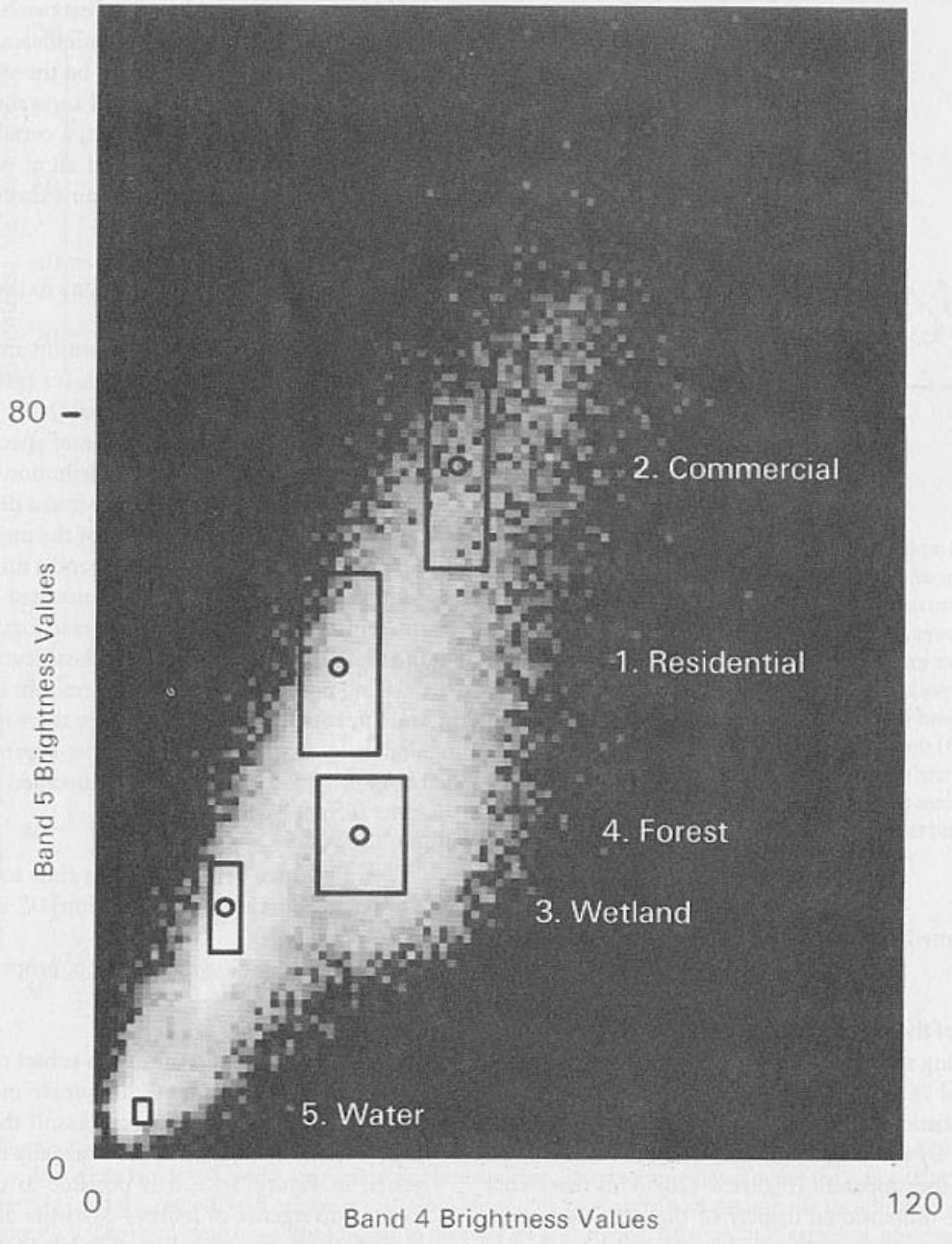


Figure 8-10 Plot of the Charleston, S.C., TM training statistics for five classes measured in bands 4 and 5 displayed as co-spectral parallelepipeds. The upper and lower limit of each parallelepiped is ±1 standard deviation. The parallelepipeds are superimposed on a feature space plot of bands 4 and 5.

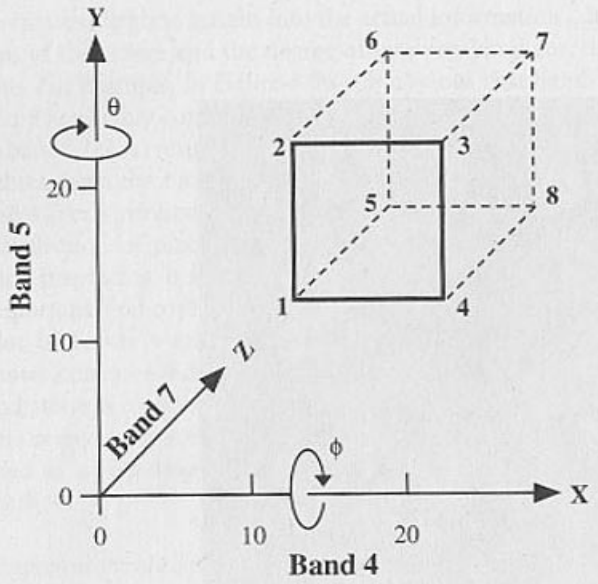


Figure 8-11 Simple parallelepiped displayed in pseudo three-dimensional space. Each of the eight corners represents a unique x, y, z coordinate corresponding to a lower or upper threshold value of the training data. For example, the original coordinates of point 4 are associated with (1) the upper threshold 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.

of the transformed matrix $P^{\overline{T}}$ are used to draw the parallelepipeds.

Manipulation of the transformed coordinates of the Charleston, S.C., training statistics is shown in Figure 8-12. All three bands (4, 5, and 7) are displayed in Figure 8-12a, except that the band 7 statistics are perpendicular (orthogonal) to the sheet of paper. By rotating the display 45°, the contribution of band 7 becomes apparent (Figure 8-12b). This represents a pseudo three-dimensional display of the parallelepipeds. As the display is rotated another 45° to 90°, band 7 data collapse onto what was the band 4 axis (Figure 8-12c). The band 4 axis is now perpendicular to the page, just as band 7 was originally. The band 7, band 5 plot (Figure 8-12c) displays some overlap between wetland (3) and forest (4). By systematically specifying various azimuth and elevation angles, it is possible to display the parallelepipeds for optimum visual examination. This allows the analyst to obtain insight as to the consistent location of the training data in three-dimensional feature space.

In this example it is evident that just two bands, 4 and 5, provide as good if not better separation than all three bands used together. However, this may not be the very best set of two bands to use. It might be useful to evaluate other two- or three-band combinations. In fact, a certain combination of perhaps four or five bands used all at one time might be superior. The only way to determine this is through statistical feature selection.

Statistical Methods of Feature Selection

Statistical methods of feature selection are used to quantitatively select which subset of bands (or 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 in Figure 8-13. 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 (Figure 8-13):

1. A pixel may be assigned to a class to which it does not belong (an error of commission).

2. A pixel is not assigned to its appropriate class (an error of omission).

The goal is to select an optimum subset of bands and apply appropriate classification techniques to minimize both types of error in the classification process. If the training data for each class from each band are normally distributed, as suggested in Figure 8-13, it is possible to use either a transformed divergence or Jeffreys–Matusita distance equation to identify the optimum subset of bands to use in the classification procedure.

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 (Swain and Davis, 1978; Mausel et al., 1990). It addresses the basic problem of deciding what is the best q-band subset of n

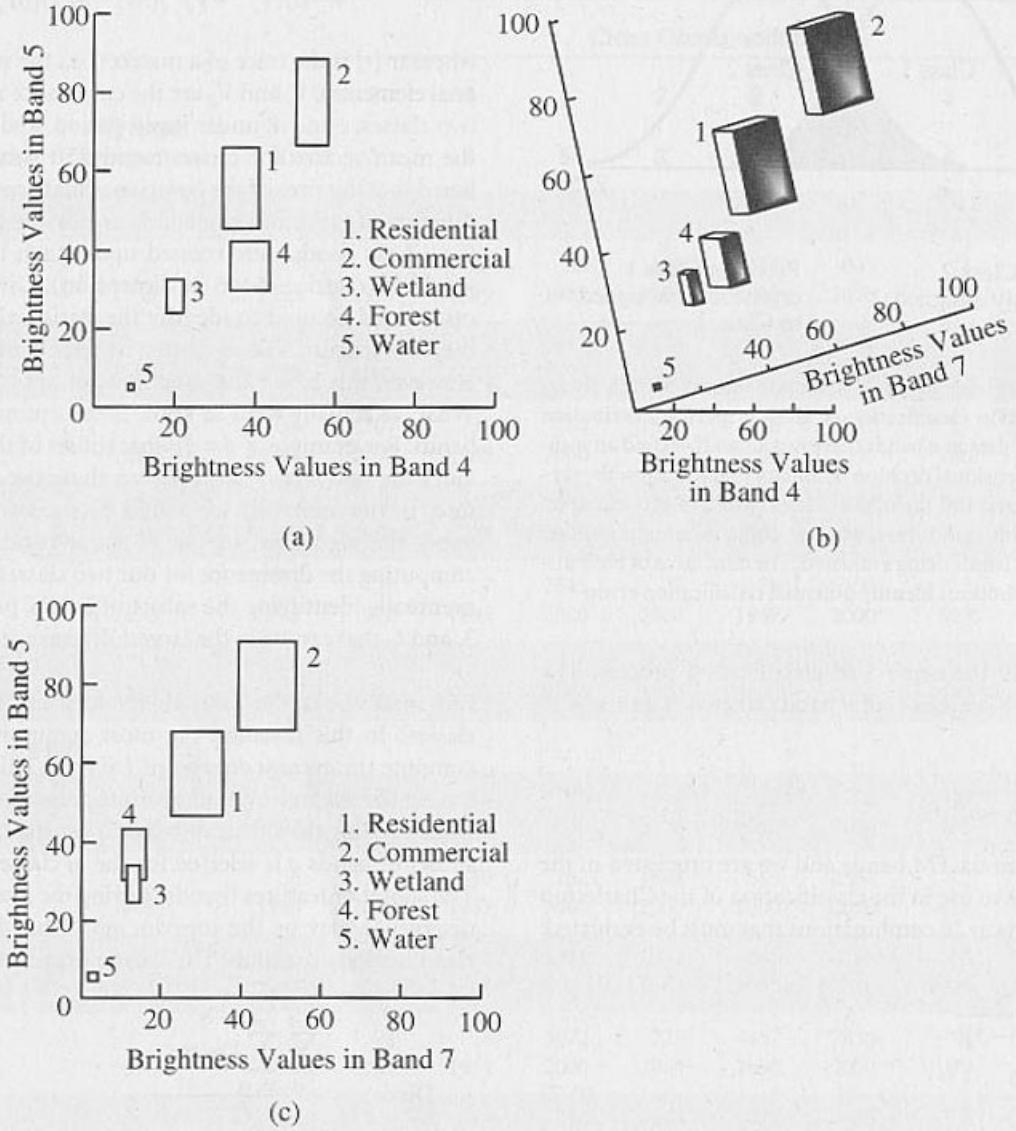


Figure 8-12 Development of the three-dimensional parallelepipeds of the five Charleston, S.C., training classes derived from the Thematic Mapper data. Only bands 4, 5, and 7 are used in this investigation. The data are rotated about the *y* axis, 0°, 45°, 90°. At 0° and 90° [parts (a) and (c), respectively], we are actually looking at only two bands, analogous to the two-dimensional boxes shown in Figure 8-10. The third band lies perpendicular to the page we are viewing. Between such extremes, however, 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°, demonstrating that the classes are entirely separable using this three band combination. However, 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.

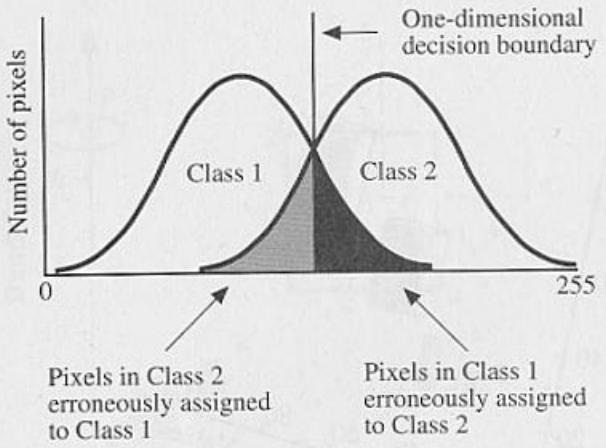


Figure 8-13 The basic problem in remote sensing pattern recognition classification is, given a spectral distribution of data in n bands (here just 1 band), to find an n-dimensional decision boundary that will allow the separation of the major classes (just 2 in this example) with a minimum of error and a minimum number of bands being evaluated. The dark areas of both distributions identify potential classification error.

bands for use in the supervised classification process. The number of combinations C of n bands taken q at a time is

$$C\left(\frac{n}{q}\right)=\frac{n!}{q!(n-q)!} \tag{8-6}$$

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:

$$\begin{aligned} C\left(\frac{6}{3}\right) &= \frac{6!}{3!(6-3)!} \\ &= \frac{720}{6(6)} \\ &= 20 \text{ combinations} \end{aligned} \tag{8-7}$$

If the best two band combinations were desired, it would be necessary to evaluate 15 possible combinations.

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

$$\begin{aligned} \text{Diver}_{cd} = &\frac{1}{2}\text{tr}\left[(V_c - V_d)(V_d^{-1} - V_c^{-1})\right] \\ &+\frac{1}{2}\text{tr}\left[(V_c^{-1} + V_d^{-1})(M_c - M_d)(M_c - M_d)^T\right] \end{aligned} \tag{8-8}$$

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, c and d, under investigation, and M_c and M_d are the mean vectors for classes c and d. It should be remembered that the sizes of the covariance matrices V_c and V_d are a function of the number of bands used in the training process (i.e., if six bands were trained upon, then both V_c and V_d would be matrices 6×6 in dimension). Divergence in this case would be used to identify the statistical separability of the two training classes using six bands of training data. However, this is not the usual goal of applying divergence. What we actually want to know is the optimum subset of q bands. For example, if $q = 3$, what subset of three bands provides the best separation between these two classes? Therefore, in our example, we would proceed to systematically apply the algorithm to the 20 three-band combinations, computing the divergence for our two classes of interest and eventually identifying the subset of bands, perhaps bands 2, 3, and 6, that results in the largest divergence value.

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_{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 as

$$\text{Diver}_{avg} = \frac{\sum_{c=1}^{m-1}\sum_{d=c+1}^{m}\text{Diver}_{cd}}{C} \tag{8-9}$$

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.

Unfortunately, outlying easily separable classes will weight average divergence upward in a misleading fashion to the extent that suboptimal reduced feature subsets might be indicated as best (Richards, 1986). Therefore, it is necessary to compute *transformed divergence*, TDiver_{cd}, expressed as

$$\text{TDiver}_{cd} = 2000\left[1 - \exp\left(\frac{-\text{Diver}_{cd}}{8}\right)\right] \tag{8-10}$$

Table 8-5. Divergence Statistics for the Five Charleston, South Carolina, Land-cover Classes Evaluated Using 1, 2, 3, 4, and 5 Thematic Mapper Band Combinations at One Time

Band Combinations	Average Divergence	Divergence (upper number) and Transformed Divergence (lower number) Class Combinations[a] 1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
a. One band at a time											
1		45	36	23	38	600	356	803	1	3	7
	1583	1993	1977	1889	1982	2000	2000	2000	198	651	1145
2		34	67	15	54	1036	286	1090	1	5	5
	1588	1970	2000	1786	1998	2000	2000	2000	246	988	890
3		54	107	39	160	1591	576	2071	1	3	1
	1525	1998	2000	1985	2000	2000	2000	2000	286	642	339
4		19	47	0	1238	209	13	3357	60	210	1466
	1748	1809	1994	70	2000	2000	1603	2000	1999	2000	2000
5		4	26	7	2645	77	29	5300	2	556	961
	1636	779	1920	1194	2000	2000	1947	2000	523	2000	2000
7		6	61	18	345	238	74	940	1	63	56
	1707	1061	1999	1795	2000	2000	2000	2000	213	1999	1998
b. Two bands at a time											
1 2		51	92	26	85	1460	410	1752	2	8	10
	1709	1997	2000	1919	2000	2000	2000	2000	463	1256	1457
1 3		56	125	40	182	1888	589	2564	2	7	11
	1709	1998	2000	1987	2000	2000	2000	2000	418	1196	1490
1 4		55	100	32	1251	941	446	3799	66	219	1525
	1996	1998	2000	1962	2000	2000	2000	2000	1999	2000	2000
1 5		54	71	28	3072	778	497	7838	6	585	1038
	1896	1998	2000	1939	2000	2000	2000	2000	1029	2000	2000
1 7		52	107	28	426	944	421	2065	3	63	76
	1852	1997	2000	1939	2000	2000	2000	2000	586	1999	2000
2 3		57	140	42	170	2099	593	2345	2	13	9
	1749	1998	2000	1990	2000	2000	2000	2000	524	1599	1382
2 4		35	103	28	1256	1136	356	3985	65	228	1529
	1992	1976	2000	1941	2000	2000	2000	2000	1999	2000	2000
2 5		35	86	20	2795	1068	328	6932	4	560	979
	1856	1976	2000	1826	2000	2000	2000	2000	760	2000	2000
2 7		37	111	24	423	1148	292	2192	2	69	66
	1829	1980	2000	1902	2000	2000	2000	2000	405	2000	1999

Table 8-5. Divergence Statistics for the Five Charleston, South Carolina, Land-cover Classes Evaluated Using 1, 2, 3, 4, and 5 Thematic Mapper Band Combinations at One Time (Continued)

Band Combinations	Average Divergence	Divergence (upper number) and Transformed Divergence (lower number) Class Combinations[a] 1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
3 4	2000	101 2000	124 2000	61 1999	1321 2000	1606 2000	905 2000	4837 2000	80 2000	210 2000	1487 2000
3 5	1895	59 1999	114 2000	45 1992	3206 2000	1609 2000	740 2000	9142 2000	5 964	597 2000	1024 2000
3 7	1845	63 1999	131 2000	41 1989	525 2000	1610 2000	606 2000	3122 2000	2 469	65 1999	59 1999
4 5	1930	21 1851	52 1997	11 1468	4616 2000	231 2000	37 1981	10376 2000	98 2000	889 2000	2902 2000
4 7	1970	20 1844	76 2000	21 1857	1742 2000	309 2000	79 2000	4740 2000	86 2000	285 2000	1599 2000
5 7	1795	6 1074	62 1999	24 1900	2870 2000	246 2000	97 2000	5956 2000	5 978	598 2000	989 2000
c. Three bands at a time											
1 2 3	1815	59 1999	154 2000	44 1992	191 2000	2340 2000	613 2000	2821 2000	3 643	16 1745	17 1774
1 2 4	1999	95 2000	142 2000	40 1986	1266 2000	1662 2000	675 2000	4381 2000	68 2000	236 2000	1573 2000
1 2 5	1909	58 1999	118 2000	32 1964	3201 2000	1564 2000	604 2000	9281 2000	7 1129	589 2000	1045 2000
1 2 7	1868	57 1998	146 2000	30 1953	493 2000	1653 2000	494 2000	3176 2000	4 732	69 2000	80 2000
1 3 4	2000	117 2000	150 2000	64 1999	1329 2000	1905 2000	985 2000	5120 2000	86 2000	219 2000	1534 2000
1 3 5	1920	60 1999	137 2000	51 1997	3569 2000	1902 2000	863 2000	11221 2000	7 1202	622 2000	1088 2000
1 3 7	1872	63 1999	157 2000	45 1993	580 2000	1935 2000	669 2000	3879 2000	4 731	66 1999	79 2000
1 4 5	1998	82 2000	105 2000	36 1979	4923 2000	978 2000	635 2000	12361 2000	104 2000	906 2000	2955 2000
1 4 7	1998	82 2000	129 2000	37 1980	1777 2000	1055 2000	610 2000	5452 2000	93 2000	288 2000	1669 2000
1 5 7	1924	56 1998	109 2000	37 1982	3405 2000	956 2000	508 2000	8948 2000	8 1261	627 2000	1077 2000

Table 8-5. Divergence Statistics for the Five Charleston, South Carolina, Land-cover Classes Evaluated Using 1, 2, 3, 4, and 5 Thematic Mapper Band Combinations at One Time (Continued)

Band Combinations	Average Divergence	Divergence (upper number) and Transformed Divergence (lower number) Class Combinations[a] 1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
2 3 4		117	156	63	1331	2119	956	4971	81	229	1530
	2000	2000	2000	1999	2000	2000	2000	2000	2000	2000	2000
2 3 5		62	147	47	3221	2120	749	9480	6	605	1034
	1908	1999	2000	1994	2000	2000	2000	2000	1082	2000	2000
2 3 7		66	160	46	541	2113	617	3480	3	74	69
	1865	1999	2000	1994	2000	2000	2000	2000	661	2000	2000
2 4 5		38	108	31	4674	1158	385	11402	103	896	2946
	1994	1984	2000	1956	2000	2000	2000	2000	2000	2000	2000
2 4 7		40	125	34	1771	1191	367	5511	90	300	1668
	1996	1986	2000	1970	2000	2000	2000	2000	2000	2000	2000
2 5 7		38	113	33	3050	1157	365	7757	7	594	1006
	1906	1982	2000	1968	2000	2000	2000	2000	1113	2000	2000
3 4 5		106	129	65	5031	1622	1037	13505	120	914	2935
	2000	2000	2000	1999	2000	2000	2000	2000	2000	2000	2000
3 4 7		111	144	63	1841	1644	955	6309	102	285	1626
	2000	2000	2000	1999	2000	2000	2000	2000	2000	2000	2000
3 5 7		66	134	63	3453	1648	823	9900	8	631	1054
	1927	1999	2000	1999	2000	2000	2000	2000	1268	2000	2000
4 5 7		22	83	26	5003	362	114	11477	105	944	2994
	1979	1870	2000	1923	2000	2000	2000	2000	2000	2000	2000
d. Four bands at a time											
1 2 3 4		167	177	65	1339	2361	1151	5259	87	238	1575
	2000	2000	2000	1999	2000	2000	2000	2000	2000	2000	2000
1 2 3 5		63	165	54	3582	2355	876	11525	8	630	1095
	1929	1999	2000	1998	2000	2000	2000	2000	1294	2000	2000
1 2 3 7		67	182	49	595	2369	683	4222	5	75	87
	1888	2000	2000	1996	2000	2000	2000	2000	885	2000	2000
1 2 4 5		115	147	46	4971	1696	901	13287	108	913	2987
	1999	2000	2000	1994	2000	2000	2000	2000	2000	2000	2000
1 2 4 7		110	165	45	1801	1731	868	6161	96	303	1725
	1999	2000	2000	1993	2000	2000	2000	2000	2000	2000	2000
1 2 5 7		61	148	41	3564	1665	614	10579	9	633	1085
	1932	1999	2000	1989	2000	2000	2000	2000	1331	2000	2000

Table 8-5. Divergence Statistics for the Five Charleston, South Carolina, Land-cover Classes Evaluated Using 1, 2, 3, 4, and 5 Thematic Mapper Band Combinations at One Time (Continued)

Band Combinations	Average Divergence	Divergence (upper number) and Transformed Divergence (lower number) Class Combinations[a] 1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
1 3 4 5		133	156	74	5293	1931	1283	15187	127	928	2976
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
1 3 4 7		134	172	69	1863	1955	1184	6814	110	289	1682
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
1 3 5 7		66	159	66	3919	1954	901	12411	10	665	1129
	1940	2000	2000	2000	2000	2000	2000	2000	1397	2000	2000
1 4 5 7		88	135	42	5422	1105	659	13950	112	970	3068
	1999	2000	2000	1990	2000	2000	2000	2000	2000	2000	2000
2 3 4 5		122	161	67	5040	2133	1093	13663	121	933	2981
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
2 3 4 7		132	173	65	1848	2143	1023	6509	103	302	1670
	2000	2000	2000	1999	2000	2000	2000	2000	2000	2000	2000
2 3 5 7		69	163	68	3476	2144	837	10308	9	639	1062
	1937	2000	2000	2000	2000	2000	2000	2000	1370	2000	2000
2 4 5 7		41	131	38	5079	1229	397	12641	110	951	3037
	1997	1987	2000	1983	2000	2000	2000	2000	2000	2000	2000
3 4 5 7		112	148	74	5436	1665	1066	14688	125	971	3030
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
e. Five bands at a time											
1 2 3 4 5		176	183	75	5302	2384	1422	15334	128	947	3019
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
1 2 3 4 7		176	196	71	1871	2393	1316	7015	111	305	1726
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
1 2 3 5 7		70	184	72	3940	2386	919	12798	11	673	1135
	1948	2000	2000	2000	2000	2000	2000	2000	1479	2000	2000
1 2 4 5 7		117	171	50	5487	1770	920	15021	115	977	3101
	2000	2000	2000	1996	2000	2000	2000	2000	2000	2000	2000
1 3 4 5 7		138	176	80	5803	1979	1294	16829	132	994	3089
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
2 3 4 5 7		134	177	77	5443	2161	1130	14893	126	987	3072
	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000

[a] Class numbers: 1, residential; 2, commercial; 3, wetland; 4, forest; 5, water.

This statistic gives an exponentially decreasing weight to increasing distances between the classes. It also scales the divergence values to lie between 0 and 2000. For example, Table 8-5 demonstrates which bands are most useful when taken 1, 2, 3, 4, or 5 at a time. There is no need to compute the divergence using all six bands since this represents the totality of the data set. It is useful, however, to calculate divergence with individual channels ($q = 1$), since a single channel might adequately discriminate among all classes of interest.

A transformed divergence value of 2000 suggests excellent between-class separation. Above 1900 provides good separation, while below 1700 is poor. It can be seen that for the Charleston study, using any single band (Table 8-5a) would not produce as acceptable results as using bands 3 and 4 together (Table 8-5b). Several three-band combinations should yield good between-class separation for all classes. Most of them understandably include bands 3 and 4. But why should we use three, four, five, or six bands in the classification when divergence statistics suggest that very good between-class separation is possible using just two bands? We probably should not if the dimensionality of the dataset can be reduced by a factor of 3 (from 6 to 2) and classification results appear promising using just the two bands.

There are other methods of feature selection also based on determining the separability between two classes at a time. For example, the *Bhattacharyya distance* assumes that the two classes c and d are Gaussian in nature and that the means and covariance matrices M_c and M_d and covariance matrices V_c and V_d are available. It is computed as

$$\text{Bhat}_{cd} = \frac{1}{8}\left(M_c - M_d\right)' \frac{(V_c + V_d)}{2}\left(M_c - M_d\right) + \frac{1}{2}\log_e \frac{\det \frac{V_c + V_d}{2}}{\sqrt{\det(V_c)}\sqrt{\det(V_d)}} \quad (8\text{-}11)$$

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 of the $m(m-1)/2$ pairs of classes for each of the possible ways 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 (Haralick and Fu, 1983).

A saturating transform applied to the Bhattacharyya distance measure yields the *Jeffreys–Matusita Distance* (often referred to as the JM distance):

$$\text{JM}_{cd} = \sqrt{2\left(1 - e^{-\text{Bhat}_{cd}}\right)} \quad (8\text{-}12)$$

The JM distance has a saturating behavior with increasing class separation like transformed divergence. However, it is not as computationally efficient as transformed divergence.

Mausel et al. (1990) evaluated four statistical separability measures to determine which would most accurately identify the best subset of four channels from an eight-channel (two date) set of multispectral video data for a computer classification of six agricultural features. Supervised maximum likelihood classification (to be discussed) was applied to all 70 possible four-band combinations. Transformed divergence and the Jeffreys–Matusita distance both selected the four-channel subset (bands 3, 4, 7, and 8 in their example), which yielded the highest overall classification accuracy of all the band combinations tested. In fact, the transformed divergence and JM-distance measures were highly correlated (0.96 and 0.97, respectively) with classification accuracy when all 70 classifications were considered. The Bhattacharyya distance and simple divergence selected the eleventh and twenty-sixth ranked four-channel subsets, respectively. A general rule of thumb is to use transformed divergence or JM-distance feature selection measures whenever possible.

Select the Appropriate Classification Algorithm

Various supervised classification algorithms may be used to assign an unknown pixel to one of a number of 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 assume 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 in nature; that is, they are normally distributed. *Nonparametric* classification algorithms make no such assumption. It is instructive to review the logic of several of the classifiers. Among the most frequently used classification algorithms are the parallelepiped, minimum distance, and maximum likelihood decision rules.

Parallelepiped Classification Algorithm

This is a widely used decision rule based on simple Boolean "and/or" logic. Training data in n spectral bands are used in performing the classification. Brightness values from each pixel of the multispectral imagery are used to produce an n-dimensional mean vector, $M_c = (\mu_{ck}, \mu_{c2}, \mu_{c3}, \ldots, \mu_{cn})$ 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. S_{ck}

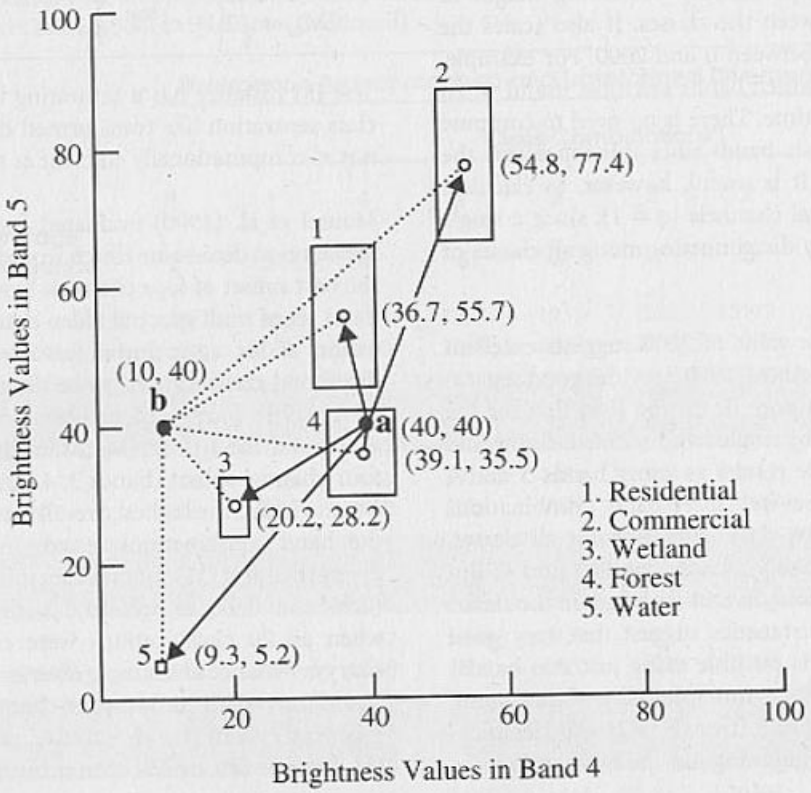


Figure 8-14 Points *a* and *b* are pixels in the image to be classified. Pixel *a* has a brightness value of 40 in band 4 and 40 in band 5. Pixel *b* has a brightness value of 10 in band 4 and 40 in band 5. The boxes represent the *parallelepiped* decision rule associated with a ±1 standard deviation 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 8-8 and 8-9 for the results of classifying points *a* and *b* using both classification techniques.

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.

Using a one-standard deviation threshold (as shown in Figure 8-14), a parallelepiped algorithm decides BV_{ijk} is in class c if , and only if,

$$\mu_{ck} - s_{ck} \leq BV_{ijk} \leq \mu_{ck} + s_{ck} \quad (8\text{-}13)$$

where

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

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

$$L_{ck} = \mu_{ck} - s_{ck} \quad (8\text{-}14)$$

and

$$H_{ck} = \mu_{ck} + s_{ck} \quad (8\text{-}15)$$

the parallelepiped algorithm becomes

$$L_{ck} \leq BV_{ijk} \leq H_{ck} \quad (8\text{-}16)$$

These decision boundaries form an n-dimensional parallelepiped in feature space. If the pixel value lies above the lower threshold and below the high threshold for all n bands evaluated, it is assigned to that class (see point *a* in Figure 8- 14). When an unknown pixel does not satisfy any of the Boolean logic criteria (point *b* in Figure 8-14), it is assigned to an unclassified category. Although it is only possible to analyze visually up to three dimensions, as described in the section on computer graphic feature analysis, it is possible to create an n-dimensional parallelepiped for classification purposes.

We will review how unknown pixels *a* and *b* are assigned to the forest and unclassified categories in Figure 8-14. The computations are summarized in Table 8-6. First, the stan-

Table 8-6. Example of Parallelepiped Classification Logic for Pixels *a* and *b* in Figure 8-14.

	Class	Lower Threshold, L_{ck}	Upper Threshold, H_{ck}	Does pixel *a* (40, 40) satisfy criteria for this class in this band? $L_{ck} \le a \le H_{ck}$	Does pixel *b* (10, 40) satisfy criteria for this class in this band? $L_{ck} \le b \le H_{ck}$
1.	Residential				
	Band 4	36.7 – 4.53 = 31.27	36.7 + 4.53 = 41.23	Yes	No
	Band 5	55.7 – 10.72 = 44.98	55.7 + 10.72 = 66.42	No	No
2.	Commercial				
	Band 4	54.8 – 3.88 = 50.92	54.8 + 3.88 = 58.68	No	No
	Band 5	77.4 – 11.16 = 66.24	77.4 + 11.16 = 88.56	No	No
3.	Wetland				
	Band 4	20.2 – 1.88 = 18.32	20.2 + 1.88 = 22.08	No	No
	Band 5	28.2 – 4.31 = 23.89	28.2 + 4.31 = 32.51	No	No
4.	Forest				
	Band 4	39.1 – 5.11 = 33.99	39.1 + 5.11 = 44.21	Yes	No
	Band 5	35.5 – 6.41 = 29.09	35.5 + 6.41 = 41.91	Yes, assign pixel to class 4, forest. STOP.	No
5.	Water				
	Band 4	9.3 – 0.56 = 8.74	9.3 + 0.56 = 9.86	—	No
	Band 5	5.2 – 0.71 = 4.49	5.2 + 0.71 = 5.91	—	No, assign pixel to unclassified category. STOP.

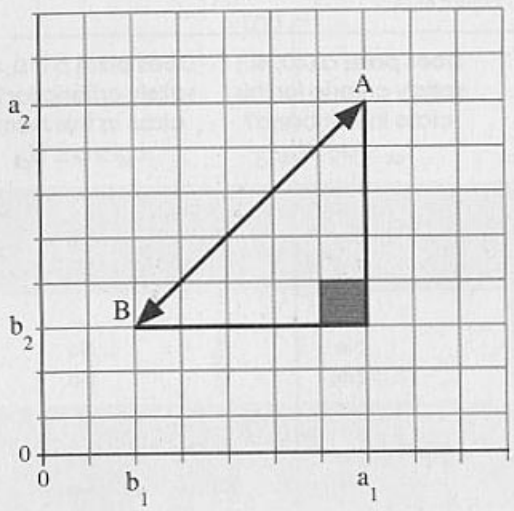
dard deviation is subtracted and added to the mean of each class and for each band to identify the lower (L_{ck}) and upper (H_{ck}) edge of the parallelepiped. In this case only two bands are used, 4 and 5, resulting in a two-dimensional box. This could be extended to n dimensions or bands. With the lower and upper thresholds for each box identified it is possible to determine if the brightness value of an input pixel in each band, k, satisfies the criteria of any of the five parallelepipeds. For example, pixel *a* has a value of 40 in both bands 4 and 5. It satisfies the band 4 criteria of class 1 (i.e., $31.27 \le 40 \le 41.23$), but does not satisfy the band 5 criteria. Therefore, the process continues by evaluating the parallelepiped criteria of classes 2 and 3, which are also not satisfied. However, when the brightness values of *a* are compared with class 4 thresholds, we find it satisfies the criteria for band 4 (i.e., $33.99 \le 40 \le 44.21$) and band 5 ($29.09 \le 40 \le 41.91$). Thus, the pixel is assigned to class 4, forest.

This same logic is applied to classify unknown pixel *b*. Unfortunately, its brightness values of 10 in band 4 and 40 in band 5 never fall within the thresholds of any of the parallelepipeds. Therefore, it is assigned to an unclassified category. Increasing the size of the thresholds to ±2 or 3 standard deviations would increase the size of the parallelepipeds. This might result in point *b* being assigned to one of the classes. However, this same action might also introduce a significant amount of overlap among many of the parallelepipeds resulting in classification error. Perhaps point *b* really belongs to a class that was not trained upon (e.g., dredge spoil).

The parallelepiped algorithm is a computationally efficient method of classifying remote sensor data. Unfortunately, because some parallelepipeds overlap, it is possible that an unknown candidate pixel might satisfy the criteria of more than one class. In such cases it is usually assigned to the first class for which it meets all criteria. A more elegant solution is to take this pixel that can be assigned to more than one class and use a minimum distance to means decision rule to assign it to just one class.

Minimum Distance to Means Classification Algorithm

This 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



Euclidean distance

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

"Round the block" distance

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

Figure 8-15 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.

the mean vectors for each class in each band μ_{ck}, from the training data. To perform a minimum distance classification, a program must calculate the distance to each mean vector, μ_{ck} from each unknown pixel (BV_{ijk}) (Jahne, 1991). It is possible to calculate this distance using Euclidean distance based on the Pythagorean theorem or "round the block" distance measures (Figure 8-15). In this discussion we demonstrate the method of minimum distance classification using Euclidean distance measurements applied to the two unknown points (*a* and *b*) shown in Figure 8-14.

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

$$\text{Dist} = \sqrt{\left(BV_{ijk} - \mu_{ck}\right)^2 + \left(BV_{ijl} - \mu_{cl}\right)^2} \qquad (8\text{-}17)$$

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

$$\text{Dist}_{a \text{ to class1}} = \sqrt{\left(BV_{ij4} - \mu_{1,4}\right)^2 + \left(BV_{ij5} - \mu_{1,5}\right)^2} \qquad (8\text{-}18)$$

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

$$\text{Dist}_{a \text{ to class 2}} = \sqrt{\left(BV_{ij4} - \mu_{2,4}\right)^2 + \left(BV_{ij5} - \mu_{2,5}\right)^2} \qquad (8\text{-}19)$$

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. Table 8-7 is a listing of the mathematics associated with the computation of distances for the five land-cover classes. It reveals that pixel *a* should be assigned to class 4 (forest) because it obtained the minimum distance of 4.59. The same logic can be applied to evaluating the unknown pixel *b*. It is assigned to class 3 (wetland) because it obtained the minimum distance of 15.75. It should be obvious that any unknown pixel will definitely be assigned to one of the five training classes using this algorithm. There will be no unclassified pixels.

Many minimum-distance algorithms let the analyst specify a distance or threshold from the class means beyond which a pixel will not be assigned to a category even though it is nearest to the mean of that category. For example, if a threshold of 10.0 was specified, point *a* would still be classified as class 4 (forest) because it had a minimum distance of 4.59, which was below the threshold. Conversely, point *b* would not be assigned to class 3 (wetland) because its minimum distance of 15.75 was greater than the 10.0 threshold. Instead, point *b* would be assigned to an unclassified category.

When more than two bands are evaluated in a classification, it is possible to extend the logic of computing the distance between just two points in *n* space using the equation (Schalkoff, 1992)

$$D_{AB} = \sqrt{\sum_{i=1}^{n} \left(a_i - b_i\right)^2} \qquad (8\text{-}20)$$

Figure 8-15 demonstrates how this algorithm is implemented.

Hodgson (1988) identified six additional Euclidean-based minimum distance algorithms that decreased computation time by exploiting two areas: (1) the computation of the distance estimate from the unclassified pixel to each candidate class and (2) the criteria for eliminating classes from the search process, thus avoiding unnecessary distance computations. Algorithms implementing these improvements were tested using up to 2, 4, and 6 bands of TM data and 5, 20, 50, and 100 classes. All algorithms were more efficient than the

Table 8-7. Example of Minimum Distance to Means Classification Logic for Pixels *a* and *b* in Figure 8-14.

	Class	Distance from pixel *a* (40, 40) to the mean of each class	Distance from pixel *b* (10, 40) to the mean of each class
1.	Residential	$\sqrt{(40-36.7)^2+(40-55.7)^2} = 16.04$	$\sqrt{(10-36.7)^2+(40-55.7)^2} = 30.97$
2.	Commercial	$\sqrt{(40-54.8)^2+(40-77.4)^2} = 40.22$	$\sqrt{(10-54.8)^2+(40-77.4)^2} = 58.35$
3.	Wetland	$\sqrt{(40-20.2)^2+(40-28.2)^2} = 23.04$	$\sqrt{(10-20.2)^2+(40-28.2)^2} = 15.75$ Assign pixel *b* to this class; it has the minimum distance
4.	Forest	$\sqrt{(40-39.1)^2+(40-35.5)^2} = 4.59$ Assign pixel *a* to this class; it has the minimum distance	$\sqrt{(10-39.1)^2+(40-35.5)^2} = 29.45$
5.	Water	$\sqrt{(40-9.3)^2+(40-5.2)^2} = 46.4$	$\sqrt{(10-9.3)^2+(40-5.2)^2} = 34.8$

traditional Euclidean minimum distance algorithm. Classification times for the six improved algorithms using a four band dataset are summarized in Figure 8-16. The simplest and slowest new algorithm (D2) does not compute the square root of the sum of squared partial distances (i.e., accumulated distance). The most computationally efficient algorithm incorporated three new ideas: (1) the accumulation of partial distances (ACCUM), (2) adding a check for one-half the nearest-neighbor distance (NND), and (3) first performing a sort of the classes in a single band (SORT). All algorithms result in the assignment of pixels to the same n classes, so any increase in efficiency is very important.

A traditional minimum distance to means classification algorithm was run on the Charleston, S.C., Thematic Mapper dataset using the training data previously described. The results are displayed as a color-coded thematic map in Figure 8-17 (color section). The total numbers of pixels in each class are summarized in Table 8-8. Error associated with the classification is discussed later in the accuracy assessment section of this chapter.

Table 8-8. Total Number of Pixels Classified into Each of the Five Charleston Land-cover Classes Shown in Figure 8-17

	Class	Total Number of Pixels
1.	Residential	14,398
2.	Commercial	4,088
3.	Wetland	10,772
4.	Forest	11,673
5.	Water	20,509

Maximum Likelihood Classification Algorithm

The *maximum likelihood* decision rule assigns each pixel having pattern measurements or features X to the class c whose units are most probable or likely to have given rise to feature vector X (Swain and Davis, 1978; Foody et al., 1992). It assumes that the training data statistics for each class in each band are normally distributed, that is, Gaussian (Blaisdell, 1993). In other words, training data with bi- or trimodal histograms in a single band are not ideal. In such cases the individual modes probably represent individual classes that should be trained upon individually and labeled as separate classes. This would then produce unimodal, Gaussian training class statistics that would fulfill the normal distribution requirement.

Maximum likelihood classification makes use of the statistics already computed and discussed in previous sections, including the mean measurement vector M_c for each class and the covariance matrix of class c for bands k through l, V_c. The decision rule applied to the unknown measurement vector X is (Swain and Davis, 1978; Schalkoff, 1992)

Decide X is in class c if, and only if,

$$p_c \geq p_i, \quad \text{where } i = 1, 2, 3, \ldots, m \text{ possible classes} \qquad (8\text{-}21)$$

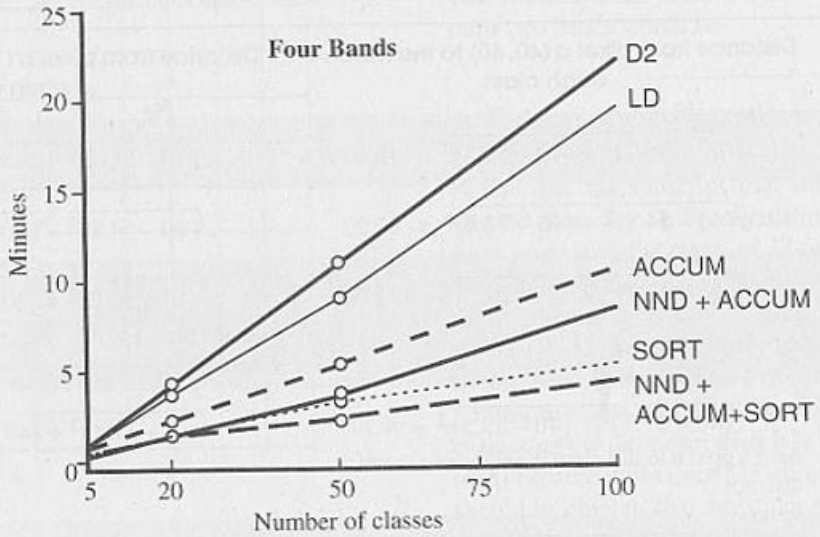


Figure 8-16 Results of applying six improved minimum distance to means classification algorithms to Charleston, S.C., TM data (from Hodgson, 1988).

and

$$p_c = \left\{-0.5\log_e\left[\det(V_c)\right]\right\} - \left[0.5(X-M_c)^T V_c^{-1}(X-M_c)\right] \quad (8\text{-}22)$$

and det (V_c) is the determinant of the covariance matrix V_c. Therefore, to classify the measurement vector X of an unknown pixel into a class, the maximum likelihood decision rule computes the value p_c for each class. Then it assigns the pixel to the class that has the largest (or maximum) value.

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 using V_2 and M_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_c. The measurement vector X used in each step of the calculation consists of n elements (the number of bands being analyzed). For example, if all six bands were being analyzed, each unknown pixel would have a measurement vector X of

$$X = \begin{bmatrix} BV_{i,j,1} \\ BV_{i,j,2} \\ BV_{i,j,3} \\ BV_{i,j,4} \\ BV_{i,j,5} \\ BV_{i,j,6} \end{bmatrix} \quad (8\text{-}23)$$

Equation 8-22 assumes that each class has an equal probability of occurring in the terrain. Common sense reminds us that in most remote sensing applications there is a high probability of encountering some classes more often than others. For example, in the Charleston scene the probability of encountering residential land use is approximately 20% (or 0.2); commercial, (0.1); wetland, (0.3); forest, (0.1); and water, (0.3). Thus, we would expect more pixels to be classified as water simply because it is more prevalent in the terrain. It is possible to include this valuable *a priori* (prior knowledge) information in the classification decision. We can do this by weighting each class c by its appropriate *a priori* probability, a_c. The equation then becomes

Decide X is in class c, if and only if,

$$p_c(a_c) \geq p_i(a_i), \quad (8\text{-}24)$$

where

$$i = 1, 2, 3, \ldots, m \text{ possible classes}$$

and

$$p_c(a_c) = \log_e(a_c) - \left\{0.5\log_e\left[\det(V_c)\right]\right\} - \left[0.5(X-M_c)^T (V_c^{-1})(X-M_c)\right] \quad (8\text{-}25)$$

This Bayes's decision rule is identical to the maximum likelihood decision rule except that it does not assume that each class has equal probabilities (Hord, 1982). *A priori* probabilities have been used successfully as a way of incorporating the

effects of relief and other terrain characteristics in improving classification accuracy (Strahler, 1980). Haralick and Fu (1983) provide an in-depth discussion of the probabilities and mathematics of the maximum likelihood and Bayes's decision rules. The maximum likelihood and Bayes's classifications require many more computations per pixel than either the parallelepiped or minimum-distance classification algorithms. They do not always produce superior results.

The maximum likelihood classification of remotely sensed data involves considerable computational effort because it calculates a large amount of information on the class membership characteristics of each pixel. Unfortunately, little of this information is made available in the conventional output which consists simply of the most likely class of membership per pixel. Foody et al. (1992) suggest that more of the information generated in the classification can be output, specifically, the *a posteriori* probabilities of class membership can be computed. For example, the *a posteriori* probability of a pixel X belonging to class c is

$$L(c \mid X) = \frac{a_c p(X \mid c)}{\sum_{r=1}^{m} a_r p(X \mid r)} \qquad (8\text{-}26)$$

where $p(X \mid c)$ is the probability density function for a pixel X as a member of class c, a_c is the *a priori* probability of membership of class c, and m is the total number of classes. The *a posteriori* probabilities sum to 1.0 for each pixel. The *a posteriori* information may be used to assess how much confidence should be placed on the classification of each pixel. For example, the analyst may decide to only keep pixels that had an *a posteriori* probability >0.85. Additional fieldwork and perhaps retraining may be required for those pixels or regions in the image that do not meet these criteria. Foody et al. (1992) suggest several other ways to improve the accuracy of maximum likelihood classification algorithm using probabilistic measures of class membership.

Unsupervised Classification

In contrast to supervised classification, *unsupervised classification* requires only a minimal amount of initial input from the analyst. It is a process whereby numerical operations are performed that search for natural groupings of the spectral properties of pixels, as examined in multispectral feature space. The user allows the computer to select the class means and covariance matrices to be used in the classification. Once the data are classified, the analyst then attempts *a posteriori* (after the fact) to assign these natural or *spectral* classes to the *information* classes of interest. This may not be easy. Some clusters may be meaningless because they represent mixed classes of Earth's surface materials. It takes careful thinking by the analyst to unravel such mysteries. The analyst should understands the spectral characteristics of the terrain well enough to label certain clusters as representing information classes.

Hundreds of methods of clustering have been developed for a wide variety of purposes apart from pattern recognition in remote sensing. Clustering algorithms used for the unsupervised classification of remotely sensed data generally vary according to the efficiency with which the clustering takes place. Different criteria of efficiency lead to different approaches (Haralick and Fu, 1983). Two examples of conceptually simple but not necessarily efficient clustering algorithms will be used to demonstrate the fundamental logic of unsupervised classification.

Unsupervised Classification Using the Chain Method

The clustering algorithm discussed here operates in a two-pass mode (i.e., it passes through the registered multispectral dataset two times). In the first pass, the program reads through the dataset and sequentially builds clusters (groups of points in spectral space). A mean vector is associated with each cluster (Jain, 1989). In the second pass, a minimum distance to means classification algorithm similar to the one previously described is applied to the whole dataset on a pixel-by-pixel basis whereby each pixel is assigned to one of the mean vectors created in pass 1. The first pass, therefore, automatically creates the cluster signatures to be used by the supervised classifier.

Pass 1: Cluster Building

During the first pass, the analyst is required to supply four types of information:

1. R, a radius distance in spectral space used to determine when a new cluster should be formed (e.g., when raw remote sensor data are used, it might be set at 15 brightness value units)

2. C, a spectral space distance parameter used when merging clusters (e.g., 30 units) when N is reached

3. N, the number of pixels to be evaluated between each major merging of the clusters (e.g., 2000 pixels)

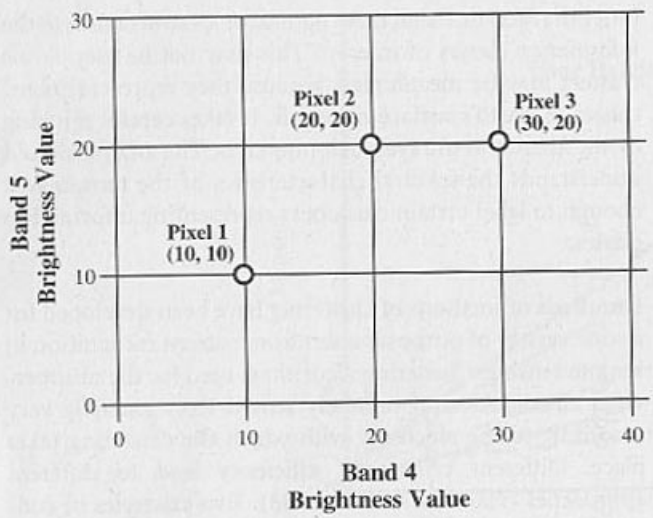


Figure 8-18 Original values of pixels 1, 2, and 3 as measured in bands 4 and 5 of the hypothetical remotely sensed data.

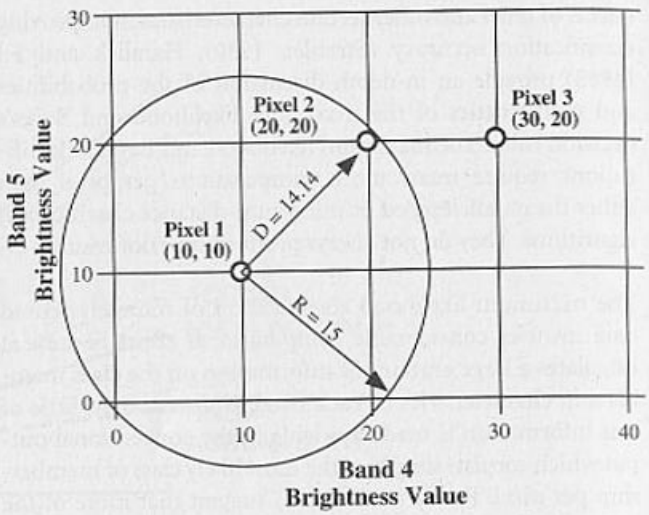


Figure 8-19 The distance (D) in two-dimensional spectral space between pixel 1 (cluster 1) and pixel 2 (cluster 2) in the first iteration is computed and tested against the value of R, the minimum acceptable radius. In this case, D does not exceed R; therefore, we merge clusters 1 and 2 as shown in the next illustration.

4. C_{max}, the maximum number of clusters to be identified by the algorithm (e.g., 20 clusters)

These can be set to default values if no initial human interaction is desired.

Starting at the origin of the multispectral dataset (i.e., line 1, column 1), pixels are evaluated sequentially from left to right as if in a chain. After one line is processed, the next line of data is considered. We will only analyze the clustering of the first three pixels in a hypothetical image and label them pixels 1, 2, and 3. The pixels have brightness values in just two bands, 4 and 5. Their spatial relationships in two-dimensional spectral space are shown in Figure 8-18.

First, we let the brightness values associated with pixel 1 in the image represent the mean data vector of cluster 1 (i.e., $M_1 = \{10, 10\}$). Remember, it is an n-dimensional mean data vector with n being the number of bands used in the unsupervised classification. In our example, just two bands are being used, so $n = 2$. Because we have not identified all 20 spectral clusters (C_{max}) as yet, pixel 2 will be considered as the mean data vector of cluster 2 (i.e., $M_2 = \{20, 20\}$). If the spectral distance D between cluster 2 to cluster 1 is greater than R, then cluster 2 will remain as cluster 2. However, if the spectral distance D is less than R, then the mean data vector of cluster 1 becomes the average of the first and second pixel brightness values and the weight (or count) of cluster 1 becomes 2 (Figure 8-19). In our example, the distance D between cluster 1 (actually pixel 1) and pixel 2 is 14.14. Because the radius R was initially set at 15.0, pixel 2 does not satisfy the criteria for being cluster 2 because its distance from cluster 1 is <15. Therefore, the mean data vectors of cluster 1 and pixel 2 are averaged, yielding the new location of cluster 1 at $M_1 = \{15, 15\}$ as shown in Figure 8-20. The spectral distance D is computed using the Pythagorean theorem as discussed previously.

Next, pixel 3 is considered as the mean data vector of cluster 2 (i.e., $M_2 = \{30, 20\}$). The distance from pixel 3 to the revised location of cluster 1, $M_1 = \{15, 15\}$, is 15.81 (Figure 8-20). Because it is >15, the mean data vector of pixel 3 becomes the mean data vector of cluster 2.

This cluster accumulation continues until the number of pixels evaluated is greater than N. At that point, the program stops evaluating individual pixels and looks closely at the nature of the clusters obtained thus far. It calculates the distance between each cluster and every other cluster. Any two clusters separated by a spectral distance less than C are merged. Such a new cluster mean vector is the weighted average of the two original clusters, and the weight is the sum of the two individual weights. This proceeds until there are no clusters with a separation distance less than C. Then the next pixel is considered. This process continues to iterate until the entire multispectral dataset is examined.

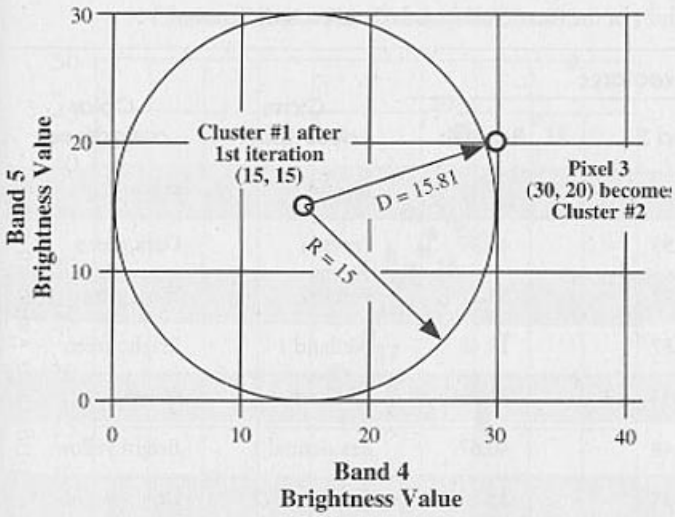


Figure 8-20 Pixels 1 and 2 now represent cluster 1. Note that the location of cluster 1 has migrated from 10, 10 to 15, 15 after the first iteration. Now, pixel 3 distance (D) is computed to see if it is greater than the minimum threshold, R. It is, so pixel location 3 becomes cluster 2. This process continues until all 20 clusters are identified. Then the 20 clusters are evaluated using a distance measure, C (not shown), to merge the clusters that are closest to one another.

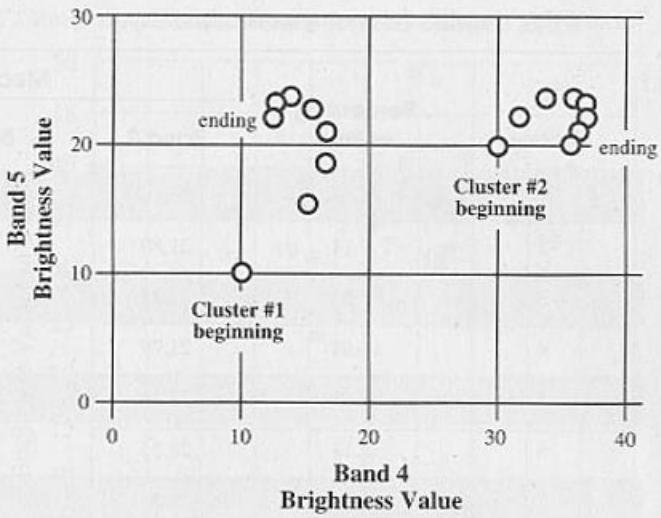


Figure 8-21 How clusters migrate during the several iterations of a clustering algorithm. The final ending point represents the mean vector that would be used in phase 2 of the clustering process when the minimum distance classification is performed.

Schowengerdt (1983) suggests that virtually all the commonly used clustering algorithms use iterative calculations to find an optimum set of decision boundaries for the data set. It should be noted that some clustering algorithms allow the analyst to initially seed the mean vector for several of the important classes. The seed data are usually obtained in a supervised fashion, as discussed previously. Others allow the analyst to utilize *a priori* information to direct the clustering process (Wharton and Turner, 1981).

Some programs do not evaluate every line and every column of the data when computing the mean vectors for the clusters. Instead, they may sample every ith row and jth column to identify the C_{max} clusters. If computer resources are abundant, then every pixel may be sampled. If resources are scarce, then acceptable results may usually be obtained by sampling the data. Obviously, a tremendous number of computations are performed during this initial pass through the dataset.

A hypothetical diagram showing the cluster migration for our two-band dataset is shown in Figure 8-21. Notice that as more points are added to a cluster the mean shifts less dramatically since the new mean computed is weighted by the number of pixels currently in a cluster. The ending point is the spectral location of the final mean vector that is used as a signature in the minimum distance classifier applied in pass 2.

Pass 2: Assignment of Pixels to One of the C_{max} Clusters Using Minimum Distance Classification Logic

The final cluster mean data vectors are used in a minimum distance to means classification algorithm to classify all the pixels in the image into one of the C_{max} clusters. The analyst usually produces a cospectral plot display to document where the clusters reside in three-dimensional feature space (Baker et al., 1991). It is then necessary to evaluate the location of the clusters in the image, label them if possible, and see if any should be combined. It is usually necessary to combine some clusters. This is where an intimate knowledge of the terrain is critical.

An unsupervised classification of the Charleston, S.C., TM scene is displayed in Figure 8-22 (color section). It was created using TM bands 2, 3, and 4. The analyst stipulated that a total of 20 clusters (C_{max}) should be extracted from the data. The mean data vectors for each of the final 20 clusters are summarized in Table 8-9. These mean vectors represented the data used in the minimum-distance classification of every pixel in the scene into one of the 20 cluster categories.

Cospectral plots of the mean data vectors for each of the 20 clusters using bands 2 and 3 and bands 3 and 4 are displayed in Figures 8-23 and 8-24, respectively. The 20 clusters lie on a diagonal extending from the origin in the band 2 versus

Table 8-9. Results of Clustering on Thematic Mapper Bands 2, 3, and 4 of the Charleston, South Carolina TM Scene

Cluster	Percent of scene	Mean vector			Class description	Color assignment
		Band 2	Band 3	Band 4		
1	24.15	23.14	18.75	9.35	Water	Dark blue
2	7.14	21.89	18.99	44.85	Forest 1	Dark green
3	7.00	22.13	19.72	38.17	Forest 2	Dark green
4	11.61	21.79	19.87	19.46	Wetland 1	Bright green
5	5.83	22.16	20.51	23.90	Wetland 2	Green
6	2.18	28.35	28.48	40.67	Residential 1	Bright yellow
7	3.34	36.30	25.58	35.00	Residential 2	Bright yellow
8	2.60	29.44	29.87	49.49	Parks, golf	Gray
9	1.72	32.69	34.70	41.38	Residential 3	Yellow
10	1.85	26.92	26.31	28.18	Commercial 1	Dark red
11	1.27	36.62	39.83	41.76	Commercial 2	Bright red
12	0.53	44.20	49.68	46.28	Commercial 3	Bright red
13	1.03	33.00	34.55	28.21	Commercial 4	Red
14	1.92	30.42	31.36	36.81	Residential 4	Yellow
15	1.00	40.55	44.30	39.99	Commercial 5	Bright red
16	2.13	35.84	38.80	35.09	Commercial 6	Red
17	4.83	25.54	24.14	43.25	Residential 5	Bright yellow
18	1.86	31.03	32.57	32.62	Residential 6	Yellow
19	3.26	22.36	20.22	31.21	Commercial 7	Dark red
20	0.02	34.00	43.00	48.00	Commercial 8	Bright red

band 3 plot. Compare this distribution of cluster means with the feature space plot using the same bands in Figure 8-9a. Unfortunately, the water cluster was located in the same spectral space as forest and wetland when viewed using just bands 2 and 3. Therefore, this scatterplot was not used to *label* or assign the clusters to *information* classes. Conversely, a cospectral plot of bands 3 and 4 mean data vectors is relatively easy to interpret and looks very much like the perpendicular vegetation index distribution shown earlier in Figure 7-41. This is not surprising since this is a red (band 3) versus near-infrared (band 4) plot.

Cluster labeling is usually performed by interactively displaying all the pixels assigned to an individual cluster on the screen with a color composite of the study area in the background. In this manner it is possible to identify the location and spatial association among clusters. This interactive visual analysis in conjunction with the information provided in the co-spectral plot, allows the analyst to group the clusters into information classes as shown in Figure 8-25 and Table 8-9. It is instructive to review some of the logic that resulted in the final unsupervised classification (Figure 8-22) (color section).

Cluster 1 occupied a distinct region of spectral space (Figure 8-25). It was not difficult to assign it to the information class water. Clusters 2 and 3 had high reflectance in the near-infrared (band 4) with low reflectance in the red (band 3) due to

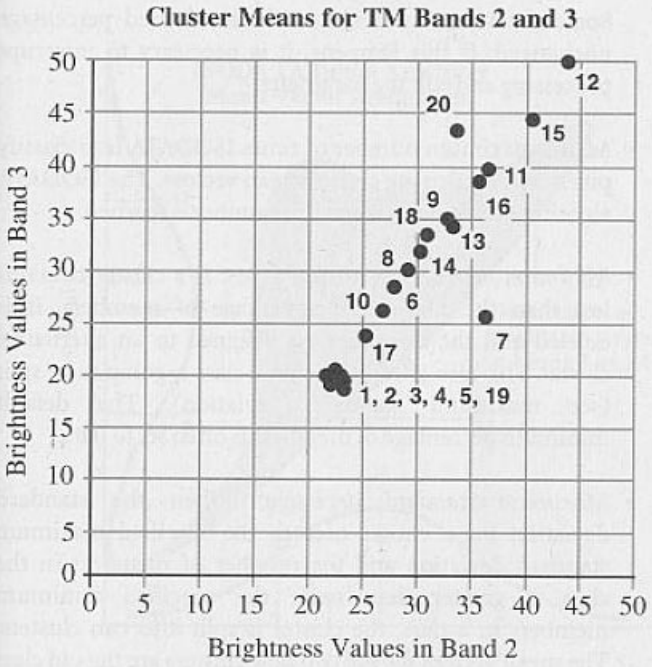


Figure 8-23 The mean vectors of the 20 clusters displayed in Figure 8-22 are shown here using only bands 2 and 3. The mean vector values are summarized in Table 8-9. Notice the substantial amount of overlap among clusters 1 through 5 and 19.

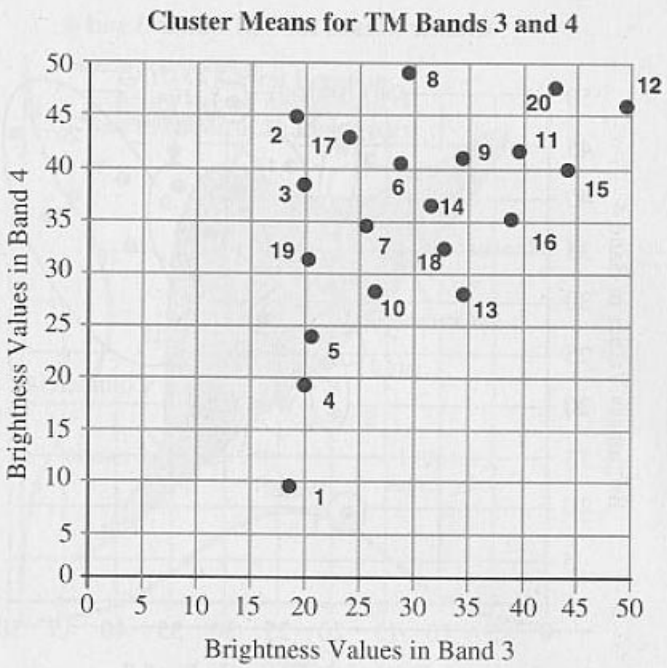


Figure 8-24 The mean vectors of the 20 clusters displayed in Figure 8-22 are shown here using only band 3 and 4 data. The mean vectors values are summarized in Table 8-9. Compare the spatial distribution of these 20 clusters in the red and near-infrared feature space with what is expected in a typical perpendicular vegetation index as discussed in Chapter 7 and Figure 7-41.

chlorophyll absorption. These two clusters were both assigned to the forest class and color coded dark green (refer to Table 8-9). Clusters 4 and 5 were situated alone in spectral space between the forest (2 and 3) and water (1) and were comprised of a mixture of moist soil and abundant vegetation. Therefore, it was not difficult to assign both these clusters to a wetland class. They were given different color codes to demonstrate that, indeed, two separate classes of wetland were identified.

Six clusters were associated with residential housing. These clusters were situated between the forest and commercial clusters (to be discussed). This is not unusual since residential housing is composed of a mixture of vegetated and non-vegetated (asphalt and concrete) surfaces, especially at TM spatial resolutions of 30×30 meters. Based on where they were located in feature space, the six clusters were collapsed into just two: bright yellow (6, 7, 17) or yellow (9,14,18).

Eight clusters were associated with commercial land use. Four of the clusters (11, 12, 15, 20) reflected high amounts of both red and near-infrared energy as commercial land use composed of concrete and bare soil often does. Two other clusters (13 and 16) were associated with commercial strip areas, particularly the downtown areas. Finally, there were two clusters (10 and 19) that were definitely commercial in character but that had a substantial amount of associated vegetation. They were mainly found along major thoroughfares in the residential areas where vegetation is more plentiful. These three subgroups of commercial land use were assigned bright red, red, and dark red, respectively (Table 8-11).

Cluster 8 did not fall nicely into any group. It experienced very high near-infrared reflectance and chlorophyll absorption often associated with very well kept lawns or parks. In fact, this is precisely what it was labeled, "parks and golf."

The 20 clusters and their color assignments are shown graphically in Figure 8-25. There is more information present in this unsupervised classification than in the supervised classification. Except for water, there are at least two classes in each land-use category that could be successfully identified using the unsupervised technique. The supervised classification simply did not sample many of these classes during the training process.

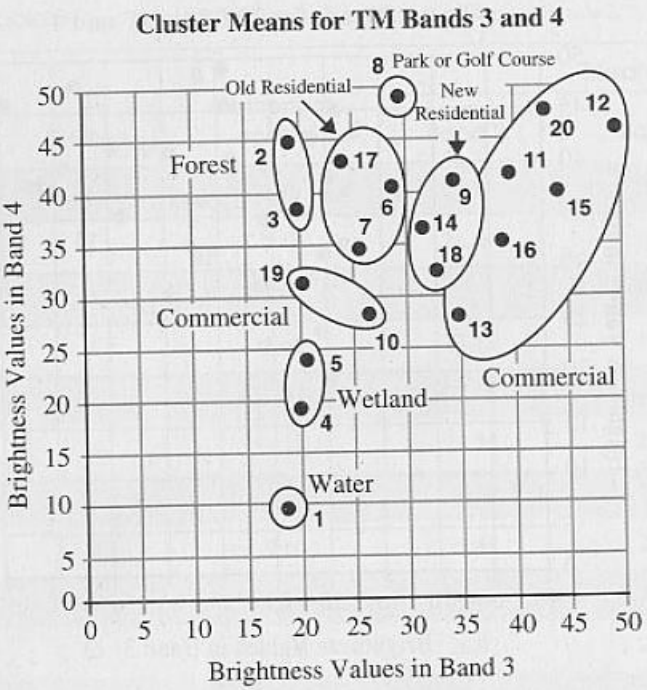


Figure 8-25 Grouping (relabeling) of the original 20 spectral clusters into information classes. The relabeling was performed by analyzing the mean vector locations in bands 3 and 4.

Unsupervised Classification Using the ISODATA Method

Another widely used clustering algorithm is the Iterative Self-Organizing Data Analysis Technique (ISODATA) (Tou and Gonzalez, 1977; Sabins, 1987; Jain, 1989). ISODATA represents a fairly comprehensive set of heuristic (rule-of-thumb) procedures that have been incorporated into an iterative classification algorithm (ERDAS, 1994; USGS, 1990; Hayward, 1993). Many of the steps incorporated into the algorithm are a result of experience gained through experimentation.

ISODATA is self-organizing because it requires relatively little human input. A sophisticated ISODATA algorithm normally requires the analyst to specify the following criteria:

- C_{max}: the maximum number of clusters to be identified by the algorithm (e.g., 20 clusters). However, it is not uncommon for less to be found in the final classification map after splitting and merging take place.

- *T:* the maximum percentage of pixels whose class values are allowed to be *unchanged* between iterations. When this number is reached, the ISODATA algorithm terminates. Some datasets may never reach the desired percentage unchanged. If this happens, it is necessary to interrupt processing and edit the parameter.

- *M:* the maximum number of times ISODATA is to classify pixels and recalculate cluster mean vectors. The ISODATA algorithm terminates when this number is reached.

- *Minimum members in a cluster (%):* If a cluster contains less than the minimum percentage of members, it is deleted and the members are assigned to an alternative cluster. This also affects whether a class is going to be split (see maximum standard deviation). The default minimum percentage of members is often set to 0.01.

- *Maximum standard deviation:* When the standard deviation for a cluster exceeds the specified maximum standard deviation and the number of members in the class is greater than twice the specified minimum members in a class, the cluster is split into two clusters. The mean vectors for the two new clusters are the old class centers ±1 standard deviation. Maximum standard deviation values between 4.5 and 7 are typical.

- *Split Separation Value:* If this value is changed from 0.0, it takes the place of the standard deviation in determining the locations of the new mean vectors plus and minus the split separation value.

- *Minimum distance between cluster means:* Clusters with a weighted distance less than this value are merged. A default of 3.0 is often used.

ISODATA Initial Arbitrary Cluster Allocation

ISODATA is iterative because it makes a large number of passes through the remote sensing dataset until specified results are obtained, instead of just two passes. Also, ISODATA does not allocate its initial mean vectors based on the analysis of pixels in the first line of data like the two-pass algorithm. Rather, an initial arbitrary assignment of all C_{max} clusters takes place along an n-dimensional vector that runs between very specific points in feature space. The region in feature space is defined using the mean, μ_k, and standard deviation, σ_k, of each band in the analysis. A hypothetical two-dimensional example using bands 3 and 4 is presented in Figure 8-26a, in which five mean vectors are distributed along the vector beginning at location $\mu_3 - \sigma_3$, $\mu_4 - \sigma_4$ and ending at $\mu_3 + \sigma_3$, $\mu_4 + \sigma_4$. This method of automatically seeding the original C_{max} vectors makes sure that the first few lines of data do not bias the creation of clusters. Note that the two-dimensional parallelepiped (box) does not capture all

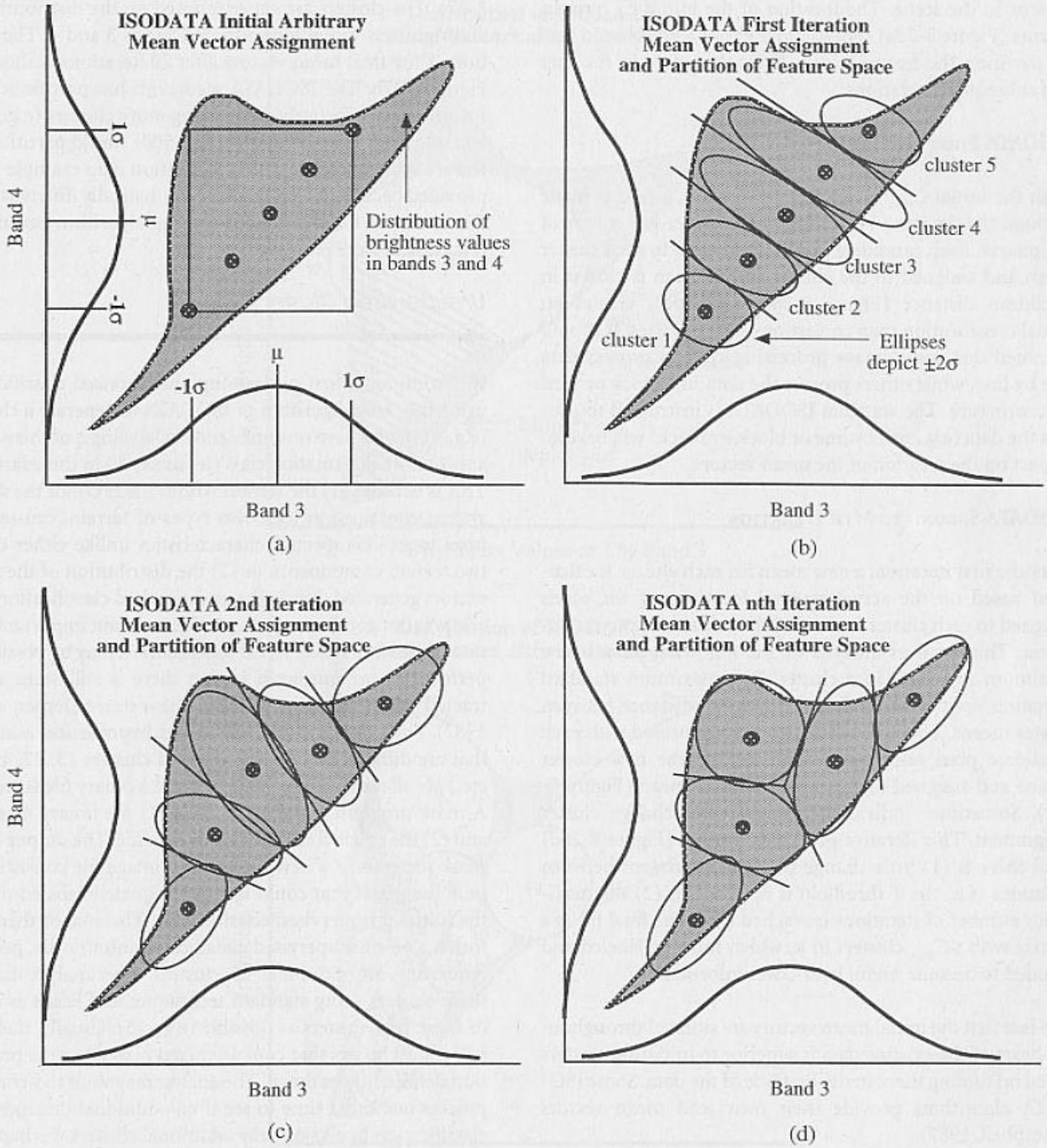


Figure 8-26 (a) ISODATA initial distribution of five hypothetical mean vectors using $\pm 1\sigma$ standard deviations in both bands as beginning and ending points. (b) In the first iteration, each candidate pixel is compared to each cluster mean and assigned to the cluster whose mean is closest in Euclidean distance. (c) During the second iteration, a new mean is calculated for each cluster based on the actual spectral locations of the pixels assigned to each cluster, instead of the initial arbitrary calculation. This involves analysis of several parameters to merge or split clusters. After the new cluster mean vectors are selected, every pixel in the scene is once again assigned to one of the new clusters. (d) This split–merge–assign process continues until there is little change in class assignment between iterations (the T threshold is reached) or the maximum number of iterations is reached (M).

the possible band 3 and 4 brightness value combinations present in the scene. The location of the initial C_{max} mean vectors (Figure 8-26a) should move about somewhat to better partition the feature space. This takes place in the first and subsequent iterations.

ISODATA First Iteration

With the initial C_{max} mean vectors in place, a pass is made through the database beginning in the upper-left corner of the matrix. Each candidate pixel is compared to each cluster mean and assigned to the cluster whose mean is closest in Euclidean distance (Figure 8-26b). This pass creates an actual classification map consisting of C_{max} classes. It should be noted that some image processing systems process data line by line, while others process the data in a block or tiled data structure. The way that ISODATA is instructed to process the data (e.g., line by line or block by block) will have an impact on the creation of the mean vectors.

ISODATA Second to *M*th Iteration

After the first iteration, a new mean for each cluster is calculated based on the actual spectral locations of the pixels assigned to each cluster, instead of the initial arbitrary calculation. This involves analysis of the following parameters: minimum members in a cluster (%), maximum standard deviation, split separation, and minimum distance between cluster means. Then the entire process is repeated with each candidate pixel once again compared to the new cluster means and assigned to the nearest cluster mean (Figure 8-26c). Sometimes individual pixels do not change cluster assignment. This iterative process continues (Figure 8-26d) until there is (1) little change in class assignment between iterations (i.e., the T threshold is reached) or (2) the maximum number of iterations is reached (M). The final file is a matrix with $\leq C_{max}$ clusters in it, which must be labeled and recoded to become useful land-cover information.

The fact that the initial mean vectors are situated throughout the heart of the existing data is superior to initiating clusters based on finding them in the first line of the data. Some ISODATA algorithms provide their own seed mean vectors (Campbell, 1987).

The iterative ISODATA algorithm is relatively slow, and image analysts are notoriously impatient. Analysts must allow the ISODATA algorithm to iterate enough times to generate meaningful mean vectors.

An ISODATA classification was performed using the Charleston TM bands 3 and 4 data. The locations of the clusters (mean $\pm 2\sigma$) after just one iteration are shown in Figure 8-27a. The clusters are superimposed on the distribution of all brightness values found in TM bands 3 and 4. The location of the final mean vectors after 20 iterations is shown in Figure 8-27b. The ISODATA algorithm has partitioned the feature space effectively. Requesting more clusters (e.g., 100) and allowing more iterations (e.g., 500) would partition the feature space even better. A classification map example is not provided because it would not be dramatically different from the results of the two-pass clustering algorithm because so few clusters were requested.

Unsupervised Cluster Busting

It is common when performing unsupervised classification using the chain algorithm or ISODATA to generate n clusters (e.g., 100) and have no confidence in labeling q of them to an appropriate information class (let us say 30 in this example). This is because (1) the terrain within the IFOV of the sensor system contained at least two types of terrain, causing the pixel to exhibit spectral characteristics unlike either of the two terrain components, or (2) the distribution of the mean vectors generated during the unsupervised classification process was not good enough to partition certain important portions of feature space. When this occurs, it may be possible to perform *cluster busting* if in fact there is still some unextracted information of value in the dataset (Jensen et al., 1987). First, the q clusters (30 in our hypothetical example) that are difficult to label (e.g., mixed clusters 13, 22, 45, 92, etc.) are all recoded to a value of 1 and a binary file is created. A mask program is then run using (1) the binary mask file and (2) the original remote sensor data file. The output of the mask program is a new multi-band image file consisting of only the pixels that could not be adequately labeled during the initial unsupervised classification. The analyst then performs a new unsupervised classification on this file, perhaps requesting an additional 25 clusters. The analyst displays these clusters using standard techniques and keeps as many of these new clusters as possible (e.g., 15). Usually, there are still some clusters that contain mixed pixels, but the proportion definitely goes down. The analyst may want to iterate the process one more time to see if an additional unsupervised classification breaks out any additional clusters. Perhaps five good clusters are extracted during the final iteration.

In this hypothetical example, the final cluster map would be composed of the 70 good clusters from the initial classification, 15 good clusters from the first cluster-busting pass (recoded as values 71 to 85), and 5 from the second pass (recoded as values 86 to 90). The final cluster map file may be put together using a simple GIS maximum dominate func-

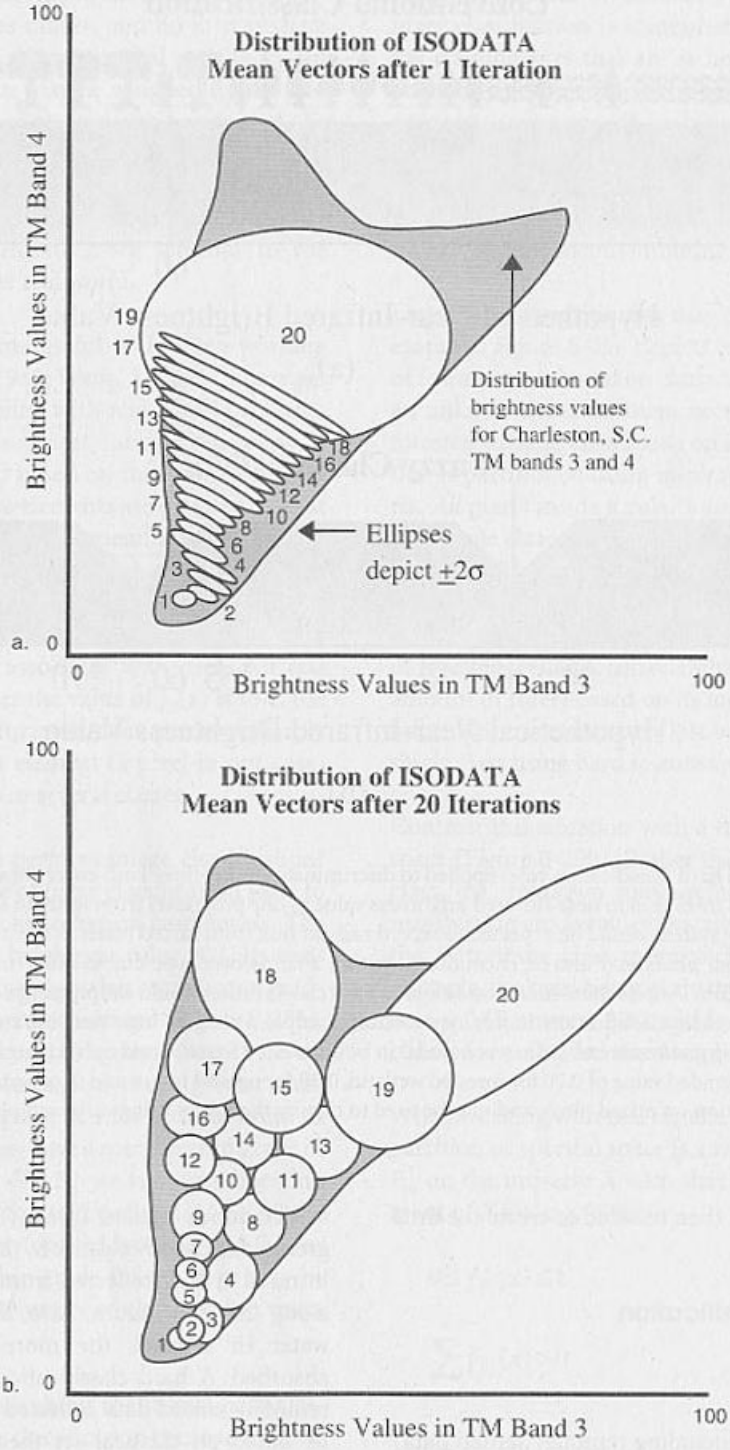


Figure 8-27 (a) Distribution of 20 ISODATA mean vectors after just one iteration using Landsat TM band 3 and 4 data of Charleston, S.C. Notice how the initial mean vectors are distributed along a diagonal in two-dimensional feature space according to the $\pm 2\sigma$ standard deviation logic discussed. (b) Distribution of 20 ISODATA mean vectors after 20 iterations. The bulk of the important feature space (the gray background) is partitioned rather well after just 20 iterations.

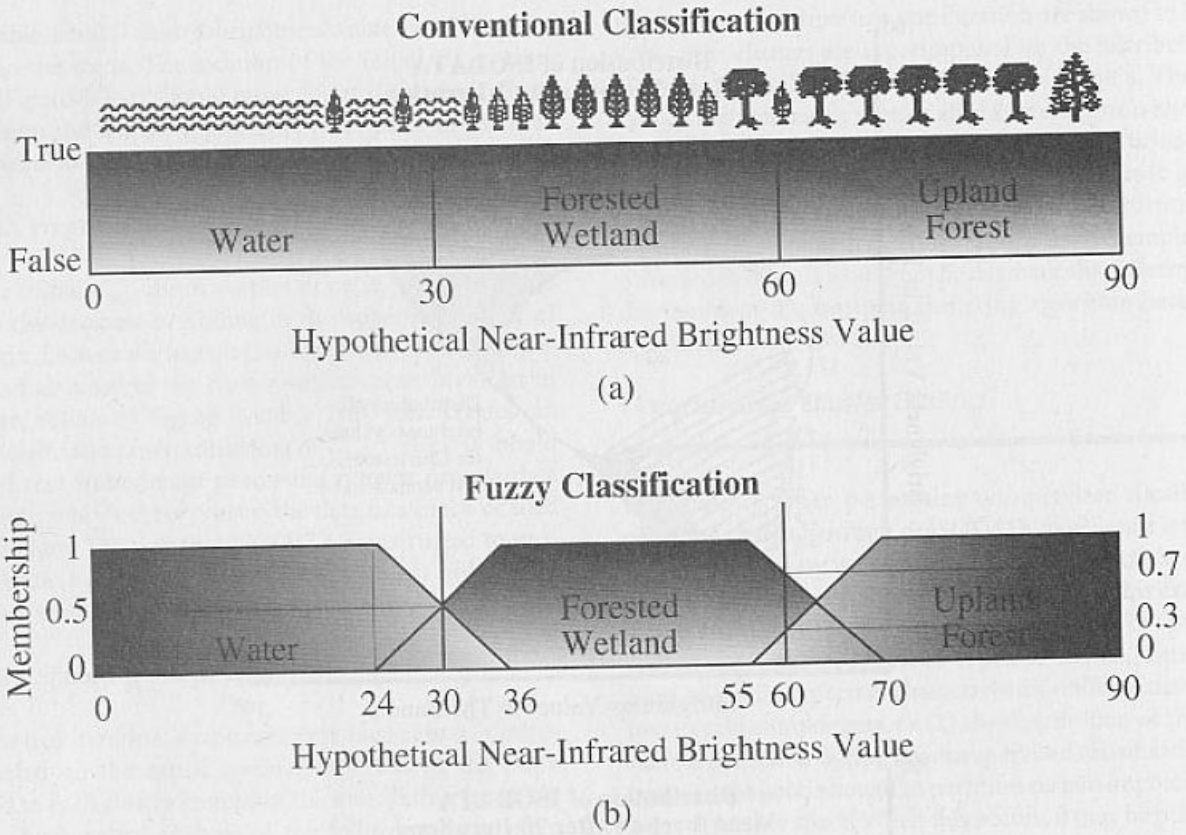


Figure 8-28 (a) Conventional hard classification rules applied to discriminate among three land-cover classes. The terrain icons suggest that there is a gradual transition in near-infrared brightness value as one progresses from water to forested wetland to upland forest. A remote sensing system would be expected to record radiant flux from mixed pixels at the interface between the major land-cover types. Mixed pixels may also be encountered within a land-cover type due to differences in species, age, or functional health of vegetation. Despite these fuzzy conditions, a hard classification would simply assign a pixel to one and only one class. (b) The logic of a fuzzy classification. In this hypothetical example, a pixel having a near-infrared brightness value of <24 would have a *membership grade* value of 1.0 in water and 0 in both forested wetland and upland forest. Similarly, a brightness value of 60 would have a graded value of 0.70 for forested wetland, 0.30 for upland forest, and 0 for water. The membership grade values provide information on mixed pixels and may be used to classify the image using various types of logic.

tion. The final cluster map is then recoded to create the final classification map.

Fuzzy Classification

Geographical information (including remotely sensed data) is imprecise, meaning that the boundaries between different phenomena are fuzzy, and/or there is heterogeneity within a class, perhaps due to differences in species, health, age, etc.

For example, terrain in the southeastern United States often exhibits a gradual transition from water, to forested wetland, to deciduous upland forest (Figure 8-28a.) Normally, the greater the canopy closure is, the greater the amount of near-infrared energy reflected from within the IFOV of a pixel along this continuum. Also, the greater the proportion of water in a pixel, the more near-infrared radiant flux absorbed. A hard classification algorithm applied to these remotely sensed data collected along this continuum would be based on classical set theory, which requires precisely defined set boundaries for which an element (e.g., a pixel) is either a member (true = 1) or not a member (false = 0) of a given set. For example, if we made a classification map using just a single near-infrared band (i.e., one-dimensional density slicing), the decision rules might be as shown in Figure 8-28a: 0 to 30 = water, 31 to 60 = forested wetland, and 61 to

90 = upland forest. The classic approach creates three discrete classes with specific class ranges, and no intermediate situations are allowed. Thus, using classical set theory, an unknown measurement vector may be assigned to one and only one class (Figure 8-28a). But everyone knows that the phenomena grade into one another and that mixed pixels are present, especially around the values of 24 to 36 and 55 to 70, as shown in the figure. Clearly, there needs to be a way to make the classification algorithms more sensitive to the imprecise (fuzzy) nature of the real world.

Fuzzy set theory provides some useful tools when working with imprecise data (Zadeh, 1965; Wang, 1990ab). Fuzzy set theory is better suited for dealing with real-world problems than traditional logic because most human reasoning is imprecise (ACM, 1984) and is based on the following logic. First, let X be a universe whose elements are denoted x. That is, $X = \{x\}$. As previously mentioned, membership in a classical set A of X is often viewed as a binary characteristic function x_A from X {0 or 1} such that $x_A(x) = 1$ if and only if $x \in A$. Conversely, a *fuzzy set* B in X is characterized by a *membership function* f_B that associates with each x a real number from 0 to 1. The closer the value of $f_B(x)$ is to 1, the more x belongs to B. Thus, a fuzzy set does not have sharply defined boundaries, and a set element (a pixel in our case) may have partial membership in several classes.

So how is fuzzy logic used to perform image classification? Figure 8-28b illustrates the use of fuzzy classification logic to discriminate among the three hypothetical land covers. The vertical boundary for water at brightness value 30 (Figure 8-28a) is replaced by a graded boundary that represents a gradual transition from water to forested wetland (Figure 8-28b). In the language of fuzzy set theory, *BV*s of less than 24 have a *membership grade* of 1 for water, and those greater than about 70 have a membership grade of 1 for upland forest. At several other locations a *BV* may have a membership grade in two classes. For example, at *BV* 30 we have membership grades of 0.5 water and 0.5 of forested wetland. At *BV* 60 the membership grades are 0.7 for forested wetland and 0.3 for upland forest. This membership grade information may be used by the analyst to create a variety of classification maps to be described.

All that has been learned before about traditional hard classification is pertinent for fuzzy classification because training still takes place, feature space is partitioned, and it is possible to assign a pixel to a single class, if desired. However, the major difference is that it is also possible *to obtain information on the various constituent classes found in a mixed pixel*, if desired. It is instructive to review how this is done.

First, the process of collecting training data as input to a fuzzy classification is somewhat different. Instead of selecting training sites that are as homogeneous as possible, the analyst may desire to select training areas that contain heterogeneous mixtures of biophysical materials in order to understand them better and to hopefully create a more accurate representation of the real world in the final classification map. Thus, a combination of pure (homogeneous) and mixed (heterogeneous) training sites may be selected.

Feature space partitioning may be dramatically different. For example, Figure 8-29a depicts a hypothetical hard partition of feature space based on classical set theory. In this example, an unknown measurement vector (pixel) is assigned to the forested wetland class based on its location in a feature space that is partitioned using minimum distance to means criteria. *All* pixels inside a partitioned feature space are assigned to a single class, no matter how close they may be to a partition line. The assignment implies full membership in that class and no membership in other classes. In this example, it is likely that the pixel under investigation probably has a lot of forested wetland, considerable water, and perhaps a small amount of forest based on its location in feature space. Such information is completely lost when the pixel is assigned to a single class using hard feature space partitioning.

Contrast this situation with a fuzzy partition of the feature space (Figure 8-29b). Rather than being assigned to a single class, the unknown measurement vector (pixel) now has *membership grade* values describing how close the pixel is to the m training class mean vectors (Wang, 1990a). In this example, the pixel being evaluated has values of forested wetland = 0.65, water = 0.30, and forest = 0.05. The values for all classes for each pixel must total 1.0.

When working with real remote sensor data, the actual fuzzy partition of spectral space is a family of fuzzy sets, $F_1, F_2, \ldots, F_m$ on the universe X such that for every x which is an element of X (Wang, 1990b):

$$0 \le f_{F_i}(x) \le 1 \tag{8-27}$$

$$\sum_{x \in X} f_{F_i}(x) > 0 \tag{8-28}$$

$$\sum_{i=1}^{m} f_{F_i}(x) = 1 \tag{8-29}$$

where $F_1, F_2, \ldots, F_m$ represent the spectral classes, X represents all pixels in the dataset, m is the number of classes

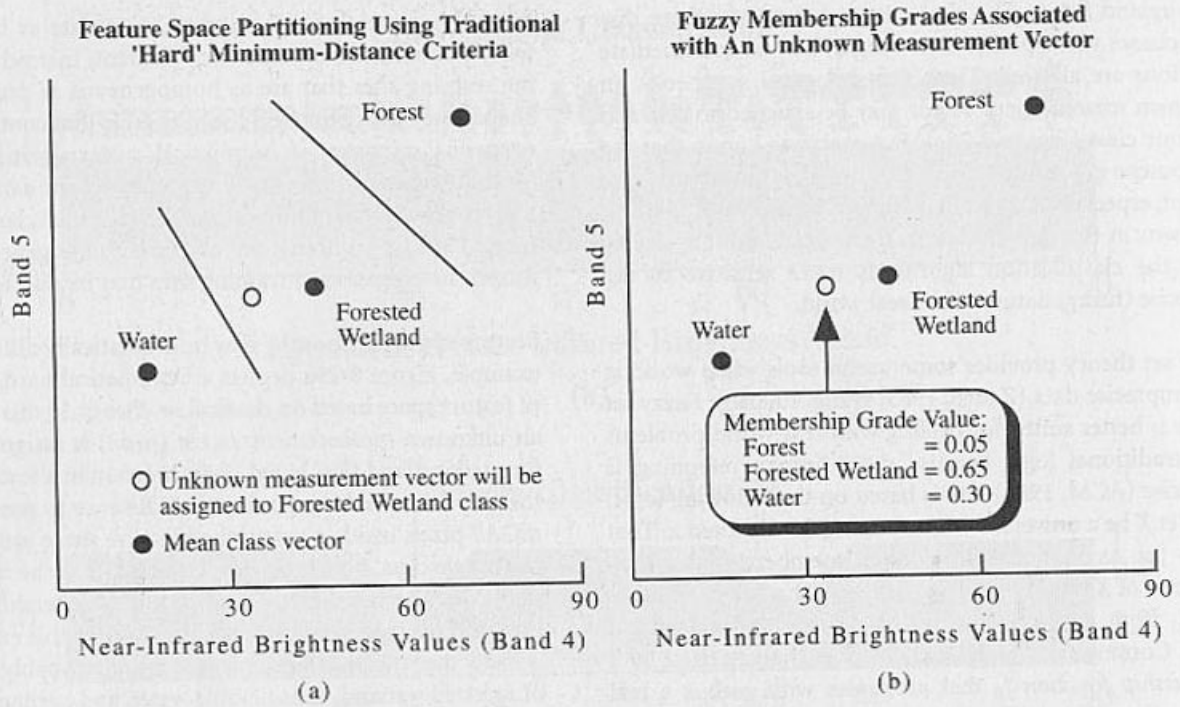


Figure 8-29 (a) Hypothetical hard partition of feature space based on classical set theory and a minimum-distance decision rule. (b) Each unknown measurement vector (pixel) has membership grade values describing how close the pixel is to the m training class mean vectors. In this example, we are most likely working with a mixed pixel predominantly composed of forested wetland (0.65) and water (0.30) based on the membership grade values. This membership grade information may be used to classify the image using various types of logic.

trained upon, x is a pixel measurement vector, and f_{F_i} is the membership function of the fuzzy set F_i $(1 \le i \le m)$. The fuzzy partition may be recorded in a fuzzy partition matrix:

$$\begin{bmatrix} f_{F_1}(x_1) & f_{F_1}(x_2) & \cdots & f_{F_1}(x_n) \\ f_{F_2}(x_1) & f_{F_2}(x_2) & \cdots & f_{F_2}(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ f_{F_m}(x_1) & f_{F_m}(x_2) & \cdots & f_{F_m}(x_n) \end{bmatrix}$$

where n is the number of pixels, and x_i is the ith pixel's measurement vector $(1 \le i \le n)$.

Fuzzy logic may be used to compute fuzzy mean and covariance matrices. For example, the fuzzy mean may be expressed as (Wang, 1990a)

$$\mu_c^* = \frac{\sum_{i=1}^{n} f_c(x_i) x_i}{\sum_{i=1}^{n} f_c(x_i)} \qquad (8\text{-}30)$$

where n is the total number of sample pixel measurement vectors, f_c is the membership function of class c, and x_i is a sample pixel measurement vector $(1 \le i \le n)$. The fuzzy covariance matrix V_c^* is computed as

$$V_c^* = \frac{\sum_{i=1}^{n} f_c(x_i)(x_i - \mu_c^*)(x_i - \mu_c^*)^T}{\sum_{i=1}^{n} f_c(x_i)} \qquad (8\text{-}31)$$

When calculating a fuzzy mean for class c, a sample pixel measurement vector x is multiplied by its membership grade in c, $f_c(x)$ before being added to the sum. Similarly, in calculating a fuzzy covariance matrix for class c, $(x_i - \mu_c^*)(x_i - \mu_c^*)^T$ is multiplied by $f_c(x)$ before being added.

To perform a fuzzy feature space partition, a membership function must be defined for each class. The following example is based on the maximum likelihood classification algorithm with fuzzy mean μ^* and fuzzy covariance matrix V^* replacing the conventional mean and covariance matrix

Table 8-10. Fuzzy Classification Membership Grades for Eight Selected Pixels (A to H) from Landsat MSS Data of Hamilton City, Ontario, Canada[a]

Pixel	A	B	C	D	E	F	G	H
Water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Industrial	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Residential	0.00	0.00	0.00	0.00	0.99	0.64	0.48	0.24
Forest	0.99	0.77	0.54	0.00	0.00	0.13	0.00	0.00
Grass	0.00	0.33	0.45	0.87	0.00	0.22	0.17	0.00
Pasture	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Bare soil	0.00	0.00	0.00	0.12	0.00	0.00	0.35	0.62

[a] Source: Wang, 1990a

(Wang, 1990b). The following is the definition of the membership function for class c:

$$f_c(x) = \frac{P_c^*(x)}{\sum_{i=1}^{m} P_i^*(x)} \tag{8-32}$$

where

$$P_i^*(x) = \frac{1}{(2\pi)^{N/2}\left|V_i^*\right|^{1/2}} \times \exp\left[-0.5\left(x-\mu_i^*\right)^T V_i^{*-1}\left(x-\mu_i^*\right)\right] \tag{8-33}$$

and N is the dimension of the pixel vectors, m is the number of classes, and $1 \le i \le m$.

The membership grades of a pixel vector x depend on x's position in the spectral space (Wang, 1990a). $f_c(x)$ increases exponentially with the decrease of $(x-\mu_c^*)^T V_i^{*-1}(x-\mu_c^*)$, that is, the Mahalanobis distance between x and class c. The value

$$\sum_{i=1}^{m} P_i^*(x) \tag{8-34}$$

is a normalizing factor.

Applying this type of fuzzy logic creates a membership grade matrix for each pixel. An example based on the work of Wang (1990a) using Landsat MSS data of Hamilton City, Ontario, Canada is shown in Table 8-10. Eight pixels (labeled A to H) in the scene are arrayed according to their membership grades. Homogeneous and mixed pixels may be easily differentiated by analyzing the membership grades, for example, pixels C, F, G, and H are mixed pixels, while A, B, D, and E are relatively homogeneous. Proportions of component cover classes in a pixel can be estimated from the membership grades. For example, it can be estimated that pixel A in the dataset contained 99% forest, B contained 77% forest and 33% grass, C contained 54% forest and 45% grass, and D contained 87% grass and 12% bare soil (Table 8-10). This is very useful information. It may be used to produce one map or a series of maps that contain(s) robust ecological information because the map(s) may more closely resemble the real-world situation. For example, an analyst could apply simple Boolean logic to the membership-grade dataset to make a map showing only the pixels that had a forest grade value of >70% and a grass value of >20% (pixel B in the example would meet this criteria). Conversely, a hard partition can be derived from the fuzzy partition matrix by changing the maximum value in each column into a 1 and others into 0. A hardened classification map can then be generated by assigning the label of the row with the value 1 of each column to the corresponding pixel.

Scientists have also applied fuzzy logic to perform unsupervised classification and to perform change detection (Bezdek et al., 1984; Fisher and Pathirana, 1990, 1993; Foody and Trodd, 1993; Rignot, 1994). Fuzzy set theory is not a panacea (ACM, 1984), but it does offer significant potential for extracting information on the makeup of the biophysical materials within a mixed pixel, a problem that will always be with us.

Incorporating Ancillary Data in the Classification Process

An analyst photointerpreting a color aerial photograph of the terrain has at his or her disposal (1) systematic knowledge about the soils, geology, vegetation, hydrology, and geography of the area, (2) the ability to visualize and comprehend the landscape's color, texture, height, and shadows, (3) the ability to place much of this diverse information in context to understand site conditions and associations among phenomena, and (4) historical knowledge about the area (Mason et al., 1988). Conversely, 95% of all remote sensing digital image classifications attempt to accomplish the same task using a single variable, an object's color and/or black-and-white tone. Therefore, it is not surprising that there is error in remote-sensing-derived classification maps (Meyer and Wirth, 1990). Why should we expect the maps to be extremely accurate when the information provided to the classification algorithm is so rudimentary?

Numerous scientists recognize this condition and have attempted to improve the accuracy and quality of remote-sensing-derived land-cover classification by incorporating ancillary data in the classification process (e.g., Strahler et al., 1978; Hutchinson, 1982; Trotter, 1991). *Ancillary data* are any type of spatial or nonspatial information that may be of value in the image classification process, including elevation, slope, aspect, geology, soils, hydrology, transportation networks, political boundaries, and vegetation maps. Ancillary data are not without error. Analysts who desire to incorporate ancillary data into the remote sensing classification process must be aware of several shortcomings.

Problems Associated Ancillary Data

Ancillary data were produced for a specific purpose and it was not to improve remote sensing classification accuracy. Second, the nominal, ordinal, or interval thematic attributes on the collateral maps may be inaccurate or incomplete (Mason et al., 1988). For example, Kuchler (1967) pointed out that polygonal boundaries between vegetation types on his respected regional maps may or may not actually exist on the ground! Great care must be exercised when generalizing the classes found on the ancillary map source materials as we try to make them compatible with the remote sensing investigation classes of interest.

Third, most ancillary information is stored in analog map format. The maps must be coordinate digitized, translated, rotated, rescaled, and often resampled to bring the dataset into congruence with the remote sensing map projection. During this process the locational attributes of the phenomena may be moved from their true planimetric position (Lunetta et al., 1991). This assumes that the ancillary data were planimetrically accurate to begin with. Unfortunately, considerable ancillary data were never recorded in their proper planimetric position. For example, old soil surveys published by the U.S. Soil Conservation Service were compiled onto uncontrolled photomosaics. Analysts trying to use such data must be careful that they do not introduce more error into the classification process than they are attempting to remove.

Approaches to Incorporating Ancillary Data to Improve Remote Sensing Classification Maps

Several approaches may be used to incorporate ancillary data in the image classification process that should improve results. These include incorporating the data before, during or after classification through geographical stratification, classifier operations, and/or postclassification sorting (Hutchinson, 1982). Combinations of these methods may be implemented using layered classification logic or rule-based expert systems (McKeown et al., 1985; Mason et al., 1988). It is instructive to review each of these alternatives.

Geographical Stratification

Ancillary data may be used *prior* to classification to subdivide the regional image into strata, which may then be processed independently. The goal is to increase the homogeneity of the individual stratified image datasets to be classified. For example, what if we wanted to locate spruce–fir in the Colorado Rockies but often encountered misclassification up and down the mountainside. One approach would be to stratify the scene into just two files: one with elevations from 0 to 2600 ft above sea level (dataset 1) and another with elevation > 2600 ft ASL (dataset 2). We would then classify the two datasets independently. Spruce–fir do not grow below 2600 ft ASL, therefore, during the classification process we would not label *any* of the pixels in dataset 1 as spruce–fir. This would keep spruce–fir pixels from being assigned to forested areas that cannot ecologically support them. Errors of commission for spruce–fir should be reduced when datasets 1 and 2 are put back together to compile the final map and compared to a traditional classification. If specific ecological principles are known, the analyst could stratify the area further using slope and aspect criteria to refine the classification (Franklin and Wilson, 1992).

Stratification is a conceptually simple tool and, carefully used, can be effective in improving classification accuracy. Illogical stratification can have severe implications. For example, differences in training set selection for individual strata and/or the vagaries of clustering algorithms, if used, may produce different spectral classes on either side of strata boundaries (Hutchinson, 1982). Edge-matching problems become apparent when the final classification map is put together from the maps derived from the individual strata.

Classifier Operations

Several methods may be used to incorporate ancillary data during the image classification process. A per-pixel logical channel classification includes ancillary data as one of the channels (features) used by the classification algorithm. For example, a dataset might consist of three SPOT bands of spectral data plus two additional bands (percent slope and aspect) derived from a digital elevation model. The entire five-band dataset is acted on by the classification algorithm in a per-pixel classification. Such methods have met with mixed results (Jones et al., 1988; Franklin and Wilson, 1992).

The context of a pixel refers to its spatial relationship with any other pixel or group of pixels throughout the entire scene (Gurney and Townshend, 1983). *Contextual logical channel* classification occurs when information about the neighboring (surrounding) pixels is used as one of the features in the classification. *Texture* is one simple contextual measure that may be extracted from an $n \times n$ window (Chapter 7) and then added to the original image dataset prior to classification (Jensen and Toll, 1982; Franklin and Peddle, 1989; Pedley and Curran, 1991). There are other contextual measurements besides texture that can be computed. For example, Gong and Howarth (1992) synthesized a frequency-based contextual measure from special principal component SPOT images of Toronto, Canada. Urban-land use information derived from their contextual data was significantly more accurate than that obtained using conventional per-pixel minimum distance classification. Gong and Howarth (1992) summarized several other contextual classification alternatives that yielded mixed results. It is important to remember that contextual information may also be derived from *non*-image ancillary sources, such as maps showing proximity to roads, streams, and so on.

A second approach involves the use of *a priori* probabilities in the classification algorithm (Strahler, 1980). The analyst gets the *a priori* probabilities by evaluating historical summaries of the region (e.g., last year cotton accounted for 80% of the acreage, hay 15%, and barley 5%). These statistics are incorporated directly into the maximum likelihood classification algorithm as weights to the classes (Equation 8-25). This has proved to be a useful way of separating classes with similar spectral responses (Mather, 1985) or for decreasing the chance of misclassifying the spatially more extensive classes (Kenk et al., 1988; Pedley and Curran, 1991). Of course, the maximum likelihood algorithm assumes that the ancillary *a priori* data are normally distributed and this is rarely the case (Watson et al., 1992).

The use of ancillary data directly in the classification process generally improves accuracy, but also increases costs (Pedley and Curran, 1991). The results are unpredictable.

Post-Classification Sorting

This method involves the application of very specific rules to (1) initial remote sensing classification results and (2) spatially distributed ancillary information. For example, Hutchinson (1982) classified Landsat MSS data of a desert area in California into nine initial classes. He then registered slope and aspect maps derived from a digital elevation model with the classification map and applied 20 if–then rules to the datasets (e.g., if the pixel was initially classified as an active sand dune and if the slope <1%, then the pixel is a dry lake bed). This eliminated confusion between several of the more prominent classes in this region [e.g., between the bright surfaces of a dry lake bed (playa) and the steep sunny slopes of large sand dunes]. Similarly, Cibula and Nyquist (1987) used postclassification sorting to improve the classification of Landsat MSS data for Olympic National Park. Topographic (elevation, slope, and aspect) and watershed boundary data (precipitation and temperature) were analyzed in conjunction with the initial land-cover classification using Boolean logic. The result was a 21-class forest map that was just as accurate as the initial map but contained much more information. Janssen et. al (1990) used an initial per pixel classification of TM data and digital terrain information to improve classification accuracy for areas in the Netherlands by 12% to 20%.

Postclassification sorting is only as good as the quality of the rules and ancillary data used. For example, some have found that digital elevation information extracted from Army Map Service 1 : 250,000 topographic maps is unsuitable for stratification or postclassification sorting (Stow and Estes, 1981). Most prefer to use U.S. Geological Survey digital elevation models derived from the 7.5-minute 1 : 24,000 map series.

Layered Classification to Incorporate Ancillary Information

Single-step classification algorithms use all available information in a single decision rule to classify all pixels to their most detailed class. *Layered classification* is a hierarchical process whereby two or more decision rules are used in the classification process (Jensen, 1978). Layered classification uses remote sensing data and/or ancillary data in a series of separate decisions (Campbell, 1987). For example, Franklin and Wilson (1992) performed a three-stage layered classification to analyze mountainous terrain using Landsat MSS data. Pixels were classified at the earliest stage possible to reduce unnecessary computation:

- *Stage 1:* The image was segmented into homogeneous regions (clusters) using a quad-tree image segmentation approach. This stage was unsupervised and knowledge about the classes or distribution of features in the image was not required. When a homogeneous quadrant was discovered, the mean and variance were compared to the mean and variance of known training class data. Each pixel in a quadrant was tested using a variance test based on the F statistic. The second test was a Student's t test that compared means of the sample (the homogeneous quadrant) and the population (the training data for each class). If no significant differences were found, each pixel in the quadrant was assigned to that particular class and eliminated from further processing. If several classes passed the test, the quadrant was assigned to the class with the lowest cumulative t test value in all the available bands.

- *Stage 2:* A minimum-distance to means calculation with stringent acceptance criteria was applied to those pixels not assigned a class in stage 1.

- *Stage 3:* Those pixels not labeled in the previous two stages were evaluated in light of digital elevation information and an examination of spectral reflectance curves of the training data to determine if some pixels were in shadow, or the like.

The overall accuracy of their final map was 87% versus 71% achieved with the maximum likelihood classification alone. The three-stage layered classification was also more efficient than the traditional maximum likelihood classification.

Expert Systems that Incorporate Ancillary Information

An expert system is the embodiment within a computer of a knowledge base that can be acted on to offer intelligent advice or take intelligent action (Forsyth, 1984). The system must also have the capability to justify its line of reasoning. Most expert systems consist of four components: (1) a knowledge base that consists of facts and rules, (2) an inference engine that evaluates the knowledge base using *forward chaining* reasoning (from data to hypotheses) or *backward chaining* (starting with a hypothesis and ending with data), (3) a knowledge acquisition module that automates the knowledge acquisition process, and (4) an explanatory interface that can be used to justify the reasoning process. A true expert system will have all four components while a knowledge-base or rule-based system may lack one or two of them (Forsyth, 1984). Expert systems attempt to present human knowledge and mimic the human reasoning process, both of which are often fuzzy or imprecise in nature.

Rule-based image classification systems that incorporate ancillary information have been available for some time (e.g., McKeown et al., 1985; Goodenough et al., 1987). McKeown et al. (1985) used ancillary map data in a rule-based system to identify airport infrastructure in aerial photography. Map information was used to decide where in the image to look and what to look for. Their approach extracted segments characterized as islands of reliability for particular classes of objects. These local regions were then analyzed by modules that brought to bear more object-specific knowledge to confirm or refute the initial hypothesis. New regions were created by merging two regions that shared a weak edge. Each time a new region was created it was scored against a specified set of area, intensity, and shape criteria to determine if it was more similar to the prototype region than were the two original regions; if so, it was retained. The idea underlying the scheme was that if a feature existed with the required characteristics the feature would eventually be merged into a single region.

Mason et al. (1988) built on this logic and developed a rule-based system based on digitized topographic map information that was segmented and then used in conjunction with principal component images of an agricultural area in England. Their methods improved both image segmentation and final classification. Their regions were constructed by applying very specific rules involving the following:

- The texture of each region was computed using the mean of the absolute differences in intensity between the pixel and each of its four axial neighbors, as described in Chapter 7.

- *Concavity:* the area of the convex hull of the region minus the area of the region, divided by the area of the convex hull.

- *Compactness:* the region area divided by the square of the region perimeter. Agricultural fields generally have high compactness.
- *Boundary straightness:* the percentage of the boundary pixels of a region that belongs to straight segments. Man-made boundaries tend to be straight; thus agricultural regions tend to have high boundary straightness.
- Spectral characteristics of water based on the simple red/near-infrared ratios discussed in Chapter 7.
- Size of the region.

Classification errors were "substantially lower (by a factor of about 3) than those of the per-pixel classifier" (Mason et al, 1988). Similarly, Bolstad and Lillesand (1992) developed a rule-based classification model based on Landsat TM data, soil texture data, and topographic position data. The rule-based approach resulted in statistically significant improvements in classification accuracy (>15%). Westmoreland and Stow (1992) used a rule-based integrated image processing/GIS system to update urban land use polygons in San Diego, California. Their approach was based on analysis of remotely sensed data (1988 Landsat TM), map ancillary data (San Diego 1987 land use forecast and 1989 general land use plan), and a series of Boolean logic decision rules. About 75% of the change in land use was correctly labeled into 19 categories using their method.

The incorporation of ancillary data in the remote sensing classification process is an important alternative to studies based solely on the analysis of spectral information analyzed on a per-pixel basis. However, the choice of variables to be included is critical. Common sense suggests that the analyst should thoughtfully select only variables with conceptual and practical significance to the classification problem at hand. Incorporating illogical or suspect ancillary information can rapidly consume limited data analysis resources and lead to inaccurate results.

Land-use Classification Map Accuracy Assessment

There must be a method for quantitatively assessing *classification accuracy* if remote-sensing-derived land-use or land-cover maps and associated statistics are to be useful (Meyer and Werth, 1990). Classification accuracy assessment was an afterthought rather than an integral part of many remote sensing studies in 1970s and 1980s. Unfortunately, many studies still simply report a single number (e.g., 85%) to express classification accuracy. Such nonsite-specific accuracy assessments completely ignore locational accuracy. In other words, only the total amount of a category is considered without regard for its location. A nonsite-specific accuracy assessment yields very high accuracy but misleading results when all the errors balance out in a region.

To correctly perform classification accuracy assessment, it is necessary to compare two sources of information: (1) the *remote-sensing-derived classification map* and (2) what we will call *reference test information* (which may in fact contain error). The relationship between these two sets of information is commonly summarized in an *error matrix* (Table 8-11). An error matrix is a square array of numbers laid out in rows and columns that expresses the number of sample units (i.e., pixels, clusters of pixels, or polygons) assigned to a particular category relative to the actual category as verified in the field. The columns normally represent the reference data, while the rows indicate the classification generated from the remotely sensed data. An error matrix is a very effective way to represent accuracy because the accuracy of each category is clearly described, along with both the errors of inclusion (commission errors) and errors of exclusion (omission errors).

But how do we obtain unbiased ground reference information to compare with the remote sensing classification map and fill the error matrix with values? Basically, the following issues must be addressed:

- Use of training versus test reference information
- Total number of samples to be collected by category
- Sampling scheme
- Appropriate descriptive and multivariate statistics to be applied

Training versus Test Reference Information

Some analysts continue to perform error evaluation based only on the *training pixels* used to train or seed the classification algorithm. Unfortunately, the locations of these training sites are usually not random. They are biased by the analyst's *a priori* knowledge of where certain land-cover types existed in the scene. Because of this bias, the classification accuracies for pixels found within the training sites are generally higher than for the remainder of the map. Therefore, this biased procedure is born of expediency and can have little use in any serious attempt at accuracy assessment (Campbell, 1987).

Table 8-11. Error Matrix of the Classification Map Derived from Landsat TM Data of Charleston, South Carolina

	Reference Data					
Classification	**Residential**	**Commercial**	**Wetland**	**Forest**	**Water**	**Row Total**
Residential	70	5	0	13	0	88
Commercial	3	55	0	0	0	58
Wetland	0	0	99	0	0	99
Forest	0	0	4	37	0	41
Water	0	0	0	0	121	121
Column Total	73	60	103	50	121	407
Overall Accuracy = 382/407 = 93.86%						

Producer's Accuracy (measure of omission error)

Residential = 70/73 =	96%	4% omission error
Commercial = 55/60 =	92%	8% omission error
Wetland = 99/103 =	96%	4% omission error
Forest = 37/50 =	74%	26% omission error
Water = 121/121 =	100%	0% omission error

User's Accuracy (measure of commission error)

Residential = 70/88 =	80%	20% commission error
Commercial = 55/58 =	95%	5% commission error
Wetland = 99/99 =	100%	0% commission error
Forest = 37/41 =	90%	10% commision error
Water = 121/121 =	100%	0% commission error

Computation of K_{hat} Coefficient

$$K_{\text{hat}} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}$$

where $N = 407$

$$\sum_{i=1}^{r} x_{ii} = (70 + 55 + 99 + 37 + 121) = 382$$

$$\sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right) = (88 \times 73) + (58 \times 60) + (99 \times 103) + (41 \times 50) + (121 \times 121) = 36{,}792$$

$$\text{therefore } K_{\text{hat}} = \frac{407(382) - 36{,}792}{407^2 - 36{,}792} = \frac{155{,}474 - 36{,}792}{165{,}649 - 36{,}792} = \frac{118{,}682}{128{,}857} = 92.1\%$$

The ideal situation is to locate *reference test pixels* in the study area. These sites are *not* used in the training of the classification algorithm and therefore represent unbiased reference information. It is possible to collect some test reference information prior to the classification, perhaps at the same time as the training data. But the majority of test reference information is collected after the classification has been performed, so some sort of stratified random sample can be utilized to collect the appropriate number of samples per category. Landscapes often change rapidly. Therefore, it is desirable to collect both the training and reference information as close to the date of data acquisition as possible.

Sample Size

The actual number of pixels to be referenced on the ground and used to assess the accuracy of individual categories in the

remote sensing classification map is often difficult to determine. Some analysts use an equation based on the binomial distribution or the normal approximation to the binomial distribution to compute the required sample size. These techniques are statistically sound for computing the sample size needed to compute the overall accuracy of a classification. The equations are based on the proportion of correctly classified samples (e.g., pixels, clusters, or polygons) and on some allowable error. For example, Fitzpatrick-Lins (1981) suggests that the sample size N to be used to assess the accuracy of a land-use classification map should be determined from the formula for the binomial probability theory:

$$N = \frac{Z^2(p)(q)}{E^2} \tag{8-35}$$

where p is the expected percent accuracy, $q = 100 - p$, E is the allowable error, and $Z = 2$ from the standard normal deviate of 1.96 for the 95% two-sided confidence level. For a sample for which the expected accuracy is 85% at an allowable error of 5%, the number of points necessary for reliable results is

$$N = \frac{2^2(85)(15)}{5^2} = \text{a minimum of 204 points} \tag{8-36}$$

With expected map accuracies of 85% and an acceptable error of 10%, the sample size for a map would be 51. The greater the allowable error is, the fewer points that need to be collected to evaluate the classification accuracy.

While this method is acceptable for selecting the total number of pixels to be sampled, it was never designed to select a sample size for filling an error matrix (Congalton, 1991). Because of the large number of pixels in remotely sensed data, traditional thinking about sampling does not apply (Glantz, 1993). Even a 0.5% sample of a single TM scene can contain over 300,000 pixels. A balance between what is statistically sound and what is practicably attainable must be found. Congalton (1991) suggests that a good rule of thumb is to collect a minimum of 50 samples for each land-cover class in the error matrix. If the area is especially large (i.e., more than 1 million acres) or the classification has a large number of land use categories (i.e., more than 12 classes), the minimum number of samples should be increased to 75 or 100 samples per class. The number of samples can also be adjusted based on the relative importance of that category within the objectives of the project or by the inherent variability within each category. It may be useful to take fewer samples in categories that show little variability, such as water or forest plantations, and increase the sampling in the categories that are more variable, such as uneven aged forest or riparian areas. The goal is to balance the statistical recommendation in order to get an adequate sample to generate an appropriate error matrix with the time, cost, and practical limitations associated with the remote sensing project.

Using this logic, approximately 250 randomly selected reference sites were necessary to assess the accuracy of the Charleston, S.C., classification map (i.e., 5 classes at 50 pixels each). To be on the safe side, 407 test reference pixels were selected (Table 8-11).

Sampling Strategy

Most of the robust error evaluation statistical measures to be discussed assume that the test reference data are randomly sampled. Simple *random sampling* without replacement always provides adequate estimates of the population parameters, provided the sample size is sufficient (Congalton, 1988). However, random sampling may undersample small but possibly very important classes unless the sample size is significantly large. Systematic or stratified systematic unaligned sampling should be used with great caution as they tend to overestimate the population parameters (Congalton, 1988). For these reasons, most analysts prefer *stratified random sampling* by which a minimum number of samples are selected from each strata (i.e., land-use category). Some combination of random and stratified sampling provides the best balance between statistical validity and practical application (Dicks and Lo, 1990). Such a system may employ random sampling to collect some assessment data early in the project, while random sampling within strata should be used after the classification has been completed to assure that enough samples were collected for each category and to minimize any periodicity (spatial autocorrelation) in the data (Congalton, 1988). Ideally, the x, y location of the reference test sites is determined using global positioning system (GPS) instruments (Abler, 1993).

The 407 reference pixels for the Charleston, S.C. study were collected based on stratified random sampling after classification. The methodology involved making five separate files, each containing only the pixels having a specific land cover class (i.e., the classified land-cover pixels in each file were recoded to a value of 1 and the unclassified background to a value of 0). A random number generator was then used to identify random x, y coordinates within each of these stratified files until a sufficient number of points was collected (i.e., at least 50 for each class). The result was a stratified random sample of the five classes. All locations were then visited in the field or evaluated using large-scale orthophotography acquired during the same month as the Landsat TM overpass.

Evaluation of Error Matrices

After the test reference information has been collected from the randomly located sites, it is compared on a pixel-by-pixel basis with the information present in the remote-sensing-derived classification map. Agreement and disagreement are summarized in the cells of the error matrix. Information in the error matrix may be evaluated using (1) simple descriptive statistics and/or (2) discrete multivariate analytical statistical techniques.

Descriptive Evaluation of Error Matrices

Overall accuracy is computed by dividing the total correct (sum of the major diagonal) by the total number of pixels in the error matrix. Computing the accuracy of individual categories, however, is more complex because the analyst has the choice of dividing the number of correct pixels in the category by the total number of pixels in the corresponding row or column. Traditionally, the total number of correct pixels in a category is divided by the total number of pixels of that category as derived from the reference data (i.e. the column total). This statistic indicates the probability of a reference pixel being correctly classified and is a measure of omission error. This statistic is called the *producer's accuracy* because the producer (the analyst) of the classification is interested in how well a certain area can be classified. If the total number of correct pixels in a category is divided by the total number of pixels that were actually classified in that category, the result is a measure of commission error. This measure, called the *user's accuracy* or reliability, is the probability that a pixel classified on the map actually represents that category on the ground (Story and Congalton, 1986).

Sometimes we are producers of classification maps and sometimes we are users of them. Therefore, we should always report all three accuracy measures; overall accuracy, producer's accuracy, and user's accuracy, because we never know how the classification may be used (Felix and Binney, 1989). For example, the remote-sensing-derived error matrix in Table 8-11 has an overall classification accuracy of 93.86%. However, what if we were primarily interested in the ability to classify just residential land use using Landsat TM data of Charleston, S.C.? The producer's accuracy for this category was calculated by dividing the total number of correct pixels in the category (70) by the total number of residential pixels as indicated by the reference data (73), yielding 96%, which is quite good. We might conclude that because the overall accuracy of the entire classification was 93.86% and the producer's accuracy of the residential land use class was 96% the procedures and Landsat TM data used are quite adequate for identifying residential land use in this area. Such a conclusion could be a mistake. We should not forget the user's accuracy, which is computed by dividing the total number of correct pixels in the residential category (70) by the total number of pixels classified as residential (88), yielding 80%. In other words, although 96% of the residential pixels were correctly identified as residential, only 80% of the areas called residential are actually residential. A careful evaluation of the error matrix reveals that there was confusion when discriminating residential land use from commercial and forest land cover. Therefore, although the producer of this map can claim that 96% of the time an area that was residential was identified as such, a user of this map will find that only 80% of the time will an area she or he visits in the field using the map actually be residential. The user may feel that an 80% user's accuracy is unacceptable.

Discrete Multivariate Analytical Techniques Applied to the Error Matrix

Discrete multivariate techniques have been used to statistically evaluate the accuracy of remote-sensing-derived classification maps and error matrices since 1983 and are now widely adopted (Congalton and Mead, 1983; Hudson and Ramm, 1987; Campbell, 1987). The techniques are appropriate because remotely sensed data are discrete rather than continuous and are also binomially or multinomially distributed rather than normally distributed. Statistical techniques based on normal distributions simply do not apply.

It is instructive to review several multivariate error evaluation techniques using the error matrix found in Table 8-11. First, the raw error matrix may be *normalized* (standardized) by applying an iterative proportional fitting procedure that forces each row and column in the matrix to sum to 1 (not shown). In this way, differences in sample sizes used to generate the matrices are eliminated and individual cell values within the matrix are directly comparable. In addition, because as part of the iterative process the rows and columns are totaled (i.e., the marginals), the resulting normalized matrix is more indicative of the off-diagonal cell values (i.e. the errors of omission and commission). In other words, all the values in the matrix are iteratively balanced by row and column, thereby incorporating information from that row and column into each individual cell value. This process then changes the cell values along the major diagonal of the matrix (correct classification), and therefore a normalized overall accuracy can be computed for each matrix by summing the major diagonal and dividing by the total of the entire matrix. Therefore, it may be argued that the normalized overall accuracy is a better representation of accuracy than is the overall accuracy computed from the original

matrix because it contains information about the off-diagonal cell values (Congalton, 1991).

Standardized error matrices are of value for another reason. Consider a situation where analyst 1 uses classification algorithm A and analyst 2 uses classification algorithm B on the same study area to extract the same four classes of information. Analyst A evaluates 250 random locations to derive error matrix A and analyst B evaluates 300 random locations to derive error matrix B. After the two error matrices are standardized, it is possible to directly compare cell values between the two matrices to see which of the two algorithms was better. Therefore, the normalization process provides a convenient way of comparing individual cell values between error matrices regardless of the number of samples used to derive the matrix.

KAPPA analysis is a discrete multivariate technique of use in accuracy assessment (Congalton and Mead, 1983). KAPPA analysis yields a K_{hat} statistic (an estimate of KAPPA) that is a measure of agreement or accuracy (Rosenfield and Fitzpatrick-Lins, 1986; Congalton, 1991). The K_{hat} statistic is computed as

$$K_{hat} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)} \quad (8\text{-}37)$$

where r is the number of rows in the matrix, x_{ii} is the number of observations in row i and column i, and x_{i+} and x_{+i} are the marginal totals for row i and column i, respectively, and N is the total number of observations.

The computation of the K_{hat} statistic for the Charleston, S.C., dataset is summarized in Table 8-13. The overall classification accuracy was 93.86%, while the K_{hat} statistic is 92.1%. The results are different because the two measures incorporated different information. The overall accuracy only incorporated the major diagonal and excluded the omission and commission errors. Conversely, K_{hat} computation incorporated the off-diagonal elements as a product of the row and column marginals. Therefore, depending on the amount of error included in the matrix, these two measures may not agree. Congalton (1991) suggests that overall accuracy, normalized accuracy, and K_{hat} be computed for each matrix to "glean as much information from the error matrix as possible." Computation of the K_{hat} statistic may also be used (1) to determine whether the results presented in the error matrix are significantly better than a random result (i.e., a null hypothesis of $K_{hat} = 0$) or (2) to compare two similar matrices (consisting of identical categories) to determine if they are significantly different.

Finn (1993) proposed an alternative method for comparing maps based on classical information theory. His measure of shared information, *average mutual information* (AMI), is based on the use of *a posteriori* entropies for one map given that the class identity from the second map allows evaluation of individual class performance. Unlike the percentage correct and/or KAPPA, which measure correctness, the AMI measures consistency between two maps. It provides an alternative viewpoint because it can be used to assess the similarity of maps. For example, it can be used to compare the consistency between two maps of the same region that have entirely different themes (e.g., one could be a soils map with five classes and one could be a remote sensing classification map with ten classes). Further research will determine the significance of the technique for evaluating error in remote sensing classification maps.

Procedures such as those discussed allow land-use maps derived from remote sensing to be quantitatively evaluated to determine overall and individual category classification accuracy. Their proper use enhances the credibility of using remote-sensing-derived land-use information.

Lineage (Genealogy) of Maps and Databases Derived from Digital Image Processing

Lineage documentation records the history of all analytical operations performed on a dataset and its resultant products. Unfortunately, manual bookkeeping of the processes used to create a final product is cumbersome and rarely performed. Some digital image processing systems do provide history or audit files to keep track of the iterations and operations performed. However, none of these methods is capable of fulfilling the information requirements of a true lineage report that itemizes the characteristics of image and cartographic sources, the topological relationships among source, intermediate, and final product layers, and a history of the transformations applied to the sources to derive the output products (Lanter, 1990; 1991).

The National Committee for Digital Cartographic Data Standards proposed that lineage information be included in every "quality report" of a digital cartographic product (NCDCDS, 1988) and should contain the following:

- Source material from which the data were derived

- Methods of derivation, including transformations applied
- Reference to specific control used (e.g., National Geodetic Reference System) or if other points are used then sufficient detail must be provided to allow recovery
- Description of the mathematical transformations of coordinates used in each step from source material to final product

Lineage documentation should be an integral part of the annotation of remote sensing or GIS products. Software designed to document lineage must have the following components: (1) lineage tracing, (2) maintenance of data quality information, (3) automatic error detection, (4) rule building (i.e., users should be able to build their own rules into a knowledge base about how their GIS and image data should be handled), (5) graphical user interface, and 6) project management (such as keeping track of times, dates, and user names to show who did what to the database and when) (Lanter, 1990). Quality assurance is an important part of life today. Image analysts extracting thematic information from remotely sensed data add value and rigor to the product by documenting its lineage.

References

Abler, R. F., 1993, "Everything in Its Place: GPS, GIS, and Geography in 1990s," *Professional Geographer*, 45(2):131–139.

ACM, 1984, "Coping with the Imprecision of the Real World: An Interview with Lotfi A. Zadeh," *Communications of the Association of Computing Machinery*, 27:304–311.

Anderson, J. R., E. Hardy, J. Roach, and R. Witmer, 1976, *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*, Washington, DC: U.S. Geological Survey Profession Paper 964, 28 p.

Baker, J. R., S. A. Briggs, V. Gordon, A. R. Jones, J. J. Settle, J. R. Townshend, and B. K. Wyatt, 1991, "Advances in Classification for Land Cover Mapping Using SPOT HRV Imagery," *International Journal of Remote Sensing*, 12(5):1071–1085.

Bezdek, J. C., R. Ehrlich, and W. Full, 1984, "FCM: The Fuzzy c–Means Clustering Algorithm," *Computers & Geosciences*, 10(2-3):191–203.

Blaisdell, E. A., 1993, *Statistics in Practice*. New York: Harcourt Brace Javanovich, 653 p.

Bolstad, P. V. and T. M. Lillesand, 1992, "Rule-based Classification Models: Flexible Integration of Satellite Imagery and Thematic Spatial Data," *Photogrammetric Engineering & Remote Sensing*, 58(7):965–971.

Botkin, D. B., J. E. Estes, R. B. MacDonald, and M. V. Wilson, 1984, "Studying the Earth's Vegetation from Space," *Bioscience*, 34(8):508–514.

Campbell, J., 1987, *Introduction to Remote Sensing*, New York: Guilford Press, 551 p.

Cetin, H. and D. Levandowski, 1991, "Interactive Classification and Mapping of Multi-Dimensional Remotely Sensed Data Using *n*-Dimensional Probability Density Functions (nPDF)," *Photogrammetric Engineering & Remote Sensing*, 57(12):1579–1587.

Cibula, W. G. and M. O. Nyquist, 1987, "Use of Topographic and Climatological Models in a Geographical Data Base to Improve Landsat MSS Classification for Olympic National Park," *Photogrammetric Engineering & Remote Sensing*, 53:67–75.

Congalton, R. G., 1988, "Using Spatial Autocorrelation Analysis to Explore the Errors in Maps Generated from Remotely Sensed Data," *Photogrammetric Engineering & Remote Sensing*, 54(5):587–592.

Congalton, R. G., 1991, "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data," *Remote Sensing of Environment*, 37:35–46.

Congalton, R. G. and R. A. Mead, 1983, "A Quantitative Method to Test for Consistency and Correctness in Photointerpretation," *Photogrammetric Engineering & Remote Sensing*, 49(1):69–74.

Cowardin, L. M., V. Carter, F. C. Golet, and E. T. LaRoe, 1979, *Classification of Wetlands and Deepwater Habitats of the United States*, Washington, DC: U.S. Fish and Wildlife Service, FWS/OBS-79/31, 103 p.

Cross, A. M., D. C. Mason, and S. J. Dury, 1988, "Segmentation of Remotely Sensed Images by a Split-and-Merge Process," *International Journal of Remote Sensing*, 9:1329–1345.

Cross, F. A. and J. P. Thomas, 1992, "CoastWatch Change Analysis Program Chesapeake Bay Regional Project," *ASPRS/ACSM 92 Technical Papers*. Bethesda, MD: American Society for Photogrammetry & Remote Sensing, 1:57.

Dahl, T. W., 1990, *Wetlands Losses in the United States 1780s to 1980s*. Washington, DC: U.S. Department of Interior, U.S. Fish and Wildlife Service, 21 p.

Dent, B. D., 1993, *Cartography—Thematic Map Design*. Dubuque, Iowa: W.C. Brown, 1–23

Dobson, J. R., E. A. Bright, R. L. Ferguson, D. W. Field, L. L. Wood, K. D. Haddad, H. Iredale, J. R. Jensen, V. V. Klemas, R. J. Orth, and J. P. Thomas, 1995, *NOAA Coastal Change Analysis Program (C-CAP): Guidance for Regional Implementation*. Washington, DC: National Oceanic and Atmospheric Administration, NMFS 123, 92 p.

ERDAS, 1994, *ERDAS Field Guide*, Atlanta, GA: ERDAS, Inc., 628 p.

Felix, N. A. and D. L. Binney, 1989, "Accuracy Assessment of a Landsat-assisted Vegetation Map of the Coastal Plain of the Arctic National Wildlife Refuge," *Photogrammetric Engineering & Remote Sensing*, 55(4):475–478.

Ferguson, R. L., L. L. Wood, and D. B. Graham, 1993, "Monitoring Spatial Change in Seagrass Habitat with Aerial Photography," *Photogrammetric Engineering & Remote Sensing*, 59(6):1033–1038.

Finn, J. T., 1993, "Use of the Average Mutual Information Index in Evaluating Classification Error and Consistency," *International Journal of Geographical Information Systems*, 7(4):349–366.

Fitzpatrick-Lins, K., 1981, "Comparison of Sampling Procedures and Data Analysis for a Land-use and Land-cover Map," *Photogrammetric Engineering & Remote Sensing*, 47(3):343–351.

Fisher, P. F. and S. Pathirana, 1990, "The Evaluation of Fuzzy Membership of Land Cover Classes in the Suburban Zone," *Remote Sensing of Environment*, 34:121–132.

Fisher, P. F. and S. Pathirana, 1993, "The Ordering of Multitemporal Fuzzy Land-cover Information Derived from Landsat MSS Data," *Geocarto International*, 8(3):5–14.

Foody, G. M. and N. M. Trodd, 1993, "Non-Classificatory Analysis and Representation of Heathland Vegetation from Remotely Sensed Imagery," *GeoJournal*, 29(4):343–350.

Foody, G. M., N. A. Campbell, N. M. Trood, and T. F. Wood, 1992, "Derivation and Applications of Probabilistic Measures of Class Membership from the Maximum-likelihood Classification," *Photogrammetric Engineering & Remote Sensing*, 58(9):1335–1341.

Forsyth, R., 1984, *Expert Systems: Principles and Case Studies*. London: Chapman and Hall, 3–17.

Franklin, S. E. and B. A. Wilson, 1992, "A Three-stage Classifier for Remote Sensing of Mountain Environments," *Photogrammetric Engineering & Remote Sensing*, 58(4):449–454.

Franklin, S. W. and D. R. Peddle, 1989, "Spectral Texture for Improved Class Discrimination in Complex Terrain," *International Journal of Remote Sensing*, 10:1437–1443.

Glantz, S. A., 1993, *Bio-Statistics*, New York: McGraw-Hill, 440 p.

Gong, P. and P. Howarth, 1992, "Frequency-based Contextual Classification and Gray-level Vector Reduction for Land-use Identification," *Photogrammetric Engineering & Remote Sensing*, 58(4):423–437.

Goodenough, D. G., M. Goldberg, G. Plunkett, and J. Zelek, 1987, "An Expert System for Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, GE-25(3):349–359.

Griffith, D. A., 1987, *Spatial Autocorrelation*. Washington, DC: Association of American Geographers, 85 p.

Gurney, C. M. and J. R. G. Townshend, 1983, "The Use of Contextual Information in the Classification of Remotely Sensed Data," *Photogrammetric Engineering & Remote Sensing*, 49:55–64.

Haralick, R. M. and K. Fu, 1983. "Pattern Recognition and Classification," Chapter 18 in *Manual of Remote Sensing*, R. Colwell, ed. Falls Church, VA: American Society of Photogrammetry, 1:793–805.

Hayward, D., *Correspondence about Earth Resource Mapping ISODATA Algorithm*. San Diego, CA: ERM, Inc., 21 p.

Hodgson, M. E., 1988, "Reducing the Computational Requirements of the Minimum-distance Classifier," *Remote Sensing of Environment*, 25:117–128.

Hodgson, M. E. and R. W. Plews, 1989, "*N*-dimensional Display of Cluster Means in Feature Space," *Photogrammetric Engineering & Remote Sensing*, 55(5):613–619.

Hord, R. M., 1982, *Digital Image Processing of Remotely Sensed Data*. New York: Academic Press, 256 p.

Hudson, W. and C. Ramm, 1987, "Correct Formulation of the Kappa Coefficient of Agreement," *Photogrammetric Engineering & Remote Sensing*, 53(4):421–422.

Hutchinson, C. F., 1982, "Techniques for Combining Landsat and Ancillary Data for Digital Classification Improvement," *Photogrammetric Engineering & Remote Sensing*, 48(1):123–130.

Jahne, B., 1991, *Digital Image Processing*. New York: Springer-Verlag, pp. 219–230.

Jain, A. K., 1989, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, pp. 418-421.

Janssen, L. F., J. Jaarsma, and E. van der Linden, 1990, "Integrating Topographic Data with Remote Sensing for Land-cover Classification," *Photogrammetric Engineering & Remote Sensing*, 56(11):1503–1506.

Jensen, J. R., 1978, "Digital Land Cover Mapping Using Layered Classification Logic and Physical Composition Attributes," *American Cartographer*, 5:121–132.

Jensen, J. R. et al., 1983, "Urban–Suburban Land Use Analysis," Chapter 30 in *Manual of Remote Sensing*, R. Colwell, ed., Falls Church, VA: American Society of Photogrammetry, 2:1571–1666.

Jensen, J. R. and D. L. Toll, 1982, "Detecting Residential Land Use Development at the Urban Fringe," *Photogrammetric Engineering & Remote Sensing*, 48:629–643.

Jensen, J. R., D. J. Cowen, S. Narumalani, J. D. Althausen, and O. Weatherbee, 1993a, "An Evaluation of CoastWatch Change Detection Protocol in South Carolina," *Photogrammetric Engineering & Remote Sensing*, 59(6):1039–1046.

Jensen, J. R., S. Narumalani, O. Weatherbee, and H. E. Mackey, 1993b, "Measurement of Seasonal and Yearly Cattail and Waterlily Changes Using Multidate SPOT Panchromatic Data," *Photogrammetric Engineering & Remote Sensing*, 59(4):519–525.

Jensen, J. R., E. W. Ramsey, J. M. Holmes, J. E. Michel, B. Savitsky, and B. A. Davis, 1990, "Environmental Sensitivity Index (ESI) Mapping for Oil Spills Using Remote Sensing and Geographic Information System Technology," *International Journal of Geographical Information Systems*, 4(2):181–201.

Jensen, J. R., E. W. Ramsey, H. E. Mackey, E. Christensen, and R. Shartiz, 1987, "Inland Wetland Change Detection Using Aircraft MSS Data," *Photogrammetric Engineering & Remote Sensing*, 53(5):521–529.

Jones, A. R., J. J. Settle, and B. K. Wyatt, 1988, "Use of Digital Terrain Data in the Interpretation of SPOT-1 HRV Multispectral Imagery," *International Journal of Remote Sensing*, 9(4):669–682.

Kenk, E., M. Sondheim, and B. Yee, 1988, "Methods for Improving Accuracy of Thematic Mapper Ground Cover Classifications," *International Journal of Geographical Information Systems*, 14:17–31.

Kiraly S. J., F. A. Cross, and J. D. Buffington (eds.), 1990, "Federal Coastal Wetland Mapping Programs," *U.S. Fish & Wildlife Service Biological Report* , 90(18):1–7.

Klemas, V. V., J. E. Dobson, R. L. Ferguson, and K. D. Haddad, 1993, "A Coastal Land Cover Classification System for the NOAA CoastWatch Change Analysis Program," *Journal of Coastal Research*, 9(3):862–872.

Kuchler, A. W., 1967, *Vegetation Mapping*. New York: Ronald Press, 472 p.

Labovitz, M. L. and E. J. Masuoka, 1984, "The Influence of Autocorrelation in Signature Extraction—An example from a Geobotanical Investigation of Cotter Basin, Montana," *International Journal of Remote Sensing*, 5(2):315–332.

Lam, S., 1993, "Fuzzy Sets Advance Spatial Decision Analysis," *GIS World*, 6(12):58–59.

Lanter, D. P., 1990, *Lineage in GIS: The Problem and a Solution*. Santa Barbara, CA: National Center for Geographic Iinformation and Analysis, Technical Paper 90-6, 1–16.

Lanter, D. P., 1991, "Design of a Lineage-based Meta-database for GIS," *Cartography and Geographic Information Systems*, 18(4):255–261.

Lee, J. K., R. A. Park, and P. W. Mausel, 1992, "Application of Geoprocessing and Simulation Modeling to Estimate Impacts of Sea Level Rise on the Northeast Coast of Florida," *Photogrammetric Engineering & Remote Sensing*, 58:1579–1586.

Lunetta, R. S., R. G. Congalton, L. K. Fenstermaker, J. R. Jensen, K. C. McGwire, and L. R. Tinney, 1991, "Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues," *Photogrammetric Engineering & Remote Sensing*, 57(6):677–687.

Mason, D. C., D. G. Corr, A. Cross, D. C. Hoggs, D. Lawrence, M. Petrou, and A. M. Tailor, 1988, "The Use of Digital Map Data in the Segmentation and Classification of Remotely Sensed Data," *International Journal of Geographical Information Systems*, 2(3):195–215.

Mather, P. M., 1985, "A Computationally Efficient Maximum-likelihood Classifier Employing Prior Probabilities for Remotely Sensed Data," *International Journal of Geographical Information Systems*, 6:369–376.

Mausel, P. W., W. J. Kamber, and J. K. Lee, 1990, "Optimum Band Selection for Supervised Classification of Multispectral Data," *Photogrammetric Engineering & Remote Sensing*, 56(1):55–60.

McKeown, D. M., W. A. Harvey, and J. McDermott, 1985, "Rule-based Feature Extraction from Aerial Imagery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:570–585.

Meyer, M. and L. Werth, 1990, "Satellite Data: Management Panacea or Potential Problem?" *Journal of Forestry*, 88(9):10–13.

NCDCDS, 1988, "The Proposed Standard for Digital Cartographic Data," *American Cartographer*, 15(1):142 p.

Pedley, M. I. and P. J. Curran, 1991, "Per-field Classification: An Example Using SPOT HRV Imagery," *International Journal of Remote Sensing*, 12(11):2181–2192.

Rhind, D., and R. Hudson, 1980, *Land Use*. New York: Methuen, 272 p.

Richards, J. A., 1986, *Remote Sensing Digital Image Analysis*. New York: Springer-Verlag, 281 p.

Richards, J. A., D. A. Landgrebe, and P. H. Swain, 1982, "A Means for Utilizing Ancillary Information in Multispectral Classification," *Remote Sensing of Environment*, 12:463–477.

Rignot, E. J., 1994, "Unsupervised Segmentation of Polarimetric SAR Data," *NASA Tech Briefs*, 18(7):46–47.

Rosenfield, G. H. and K. Fitzpatrick-Lins, 1986, "A Coefficient of Agreement as a Measure of Thematic Classification Accuracy," *Photogrammetric Engineering & Remote Sensing*, 52(2):223–227.

Sabins, M. J., 1987, "Convergence and Consistency of Fuzzy c-Means/ISODATA Algorithms," *IEEE Transactions Pattern Analysis & Machine Intelligence*, 9:661–668.

Schalkoff, R., 1992, *Pattern Recognition: Statistical, Structural and Neural Approaches*. New York: John Wiley, 364 p.

Schowengerdt, R. A., 1983, *Techniques for Image Processing and Classification in Remote Sensing*. New York: Academic Press, 249 p.

Skidmore, A. K., 1989, "Unsupervised Training Area Selection in Forests Using a Nonparametric Distance Measure and Spatial Information," *International Journal of Remote Sensing*, 10(1):133–146.

Story, M. and R. Congalton, 1986, "Accuracy Assessment: A User's Perspective," *Photogrammetric Engineering & Remote Sensing*, 52(3):397–399.

Stowe, D. A. and J. E. Estes, 1981, "Landsat and Digital Terrain Data for County-Level Resource Management," *Photgrammetric Engineering and Remote Sensing*, 47(2): 215–222.

Strahler, A. H., 1980, "The Use of Prior Probabilities in Maximum Likelihood Classification of Remotely Sensed Data," *Remote Sensing of Environment*, 10:135–163.

Strahler, A. H., T. L. Logan, and N. A. Bryant, 1978, "Improving Forest Cover Classification Accuracy from Landsat by Incorporating Topographic Information," *Proceedings*, 12th International Symposium on Remote Sensing of the Environment, 927–942.

Swain, P. H. and S. M. Davis, 1978, *Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 166–174.

Tou, J. T. and R. C. Gonzalez, 1977, *Pattern Recognition Principles*. Reading, MA: Addison–Wesley, 377 p.

Trotter, C. M., 1991, "Remotely-sensed Data as an Information Source for Geographical Information Systems in Natural Resource Management: A Review," *International Journal of Geographical Information Systems*, 5(2):225–239.

U.S. Congress, 1989, "Coastal Waters in Jeopardy: Reversing the Decline and Protecting America's Coastal Resources," *Oversight Report of the Committee on Merchant Marine and Fisheries*, Serial 100-E. Washington, DC: U.S. Government. Printing Office, 47 p.

USGS, 1992, *Standards for Digital Line Graphs for Land Use and Land Cover Technical Instructions*, Referral STO-1-2. Washington, DC: US Government Printing Office, 60 p.

USGS, 1990, *Land Analysis System (LAS) V.5 User Guide*, Sioux Falls, SD: EROS Data Center, 330 p.

Wang, F., 1990a, "Improving Remote Sensing Image Analysis through Fuzzy Information Representation," *Photogrammetric Engineering & Remote Sensing*, 56(8):1163–1169.

Wang, F., 1990b, "Fuzzy Supervised Classification of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, 28(2):194–201.

Wang, F., 1991, "Integrating GIS's and Remote Sensing Image Analysis Systems by Unifying Knowledge Representation Schemes," *IEEE Transactions on Geoscience and Remote Sensing*, 29(4): 656–665.

Watson, A. I., R. A. Vaughn, and M. Powell, 1992, "Classification Using the Watershed Method," *International Journal of Remote Sensing*, 13(10):1881–1890.

Welch, R. A., 1982, "Spatial Resolution Requirements for Urban Studies," *International Journal of Remote Sensing*, 3:139–146.

Welch, R., M. Remillard, and J. Alberts, 1992, "Integration of GPS, Remote Sensing, and GIS Techniques for Coastal Resource Management," *Photogrammetric Engineering & Remote Sensing*, 58(11):1571–1578.

Westmoreland, S. and D. A. Stow, 1992, "Category Identification of Changed Land-use Polygons in an Integrated Image Processing/GIS," *Photogrammetric Engineering & Remote Sensing*, 58(11):1593–1599.

Wharton, S. W. and B. J. Turner, 1981, "ICAP: An Interactive Cluster Analysis Procedure for Analyzing Remotely Sensed Data," *Remote Sensing of Environment*, 11:279–293.

Zadeh, L. A., 1965, "Fuzzy Sets," *Information and Control*, 8:338–353.