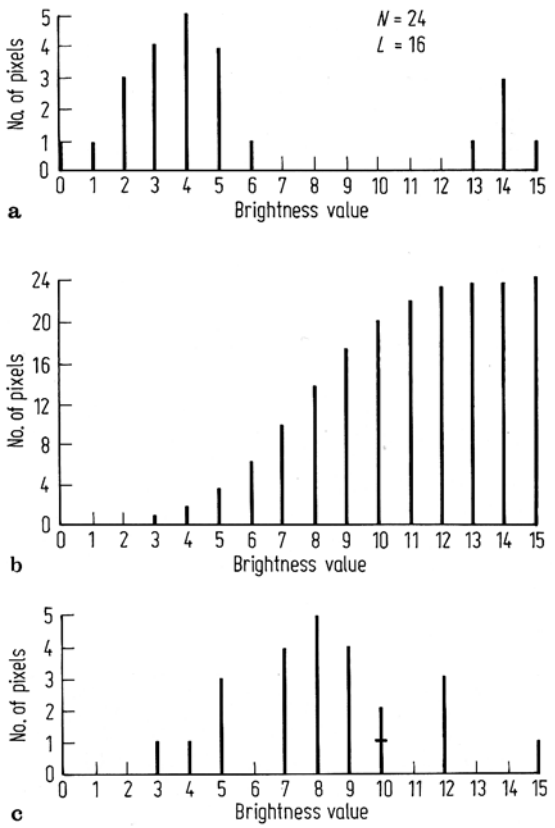


Fig. 4.15. Illustration of the modification of an image histogram to a pseudo-Gaussian shape. **a** Original histogram; **b** Cumulative normal histogram; **c** Histogram matched to Gaussian reference

4.6 Density Slicing

4.6.1 Black and White Density Slicing

A point operation often performed with remote sensing image data is to map *ranges* of brightness value to particular shades of grey. In this way the overall discrete number of brightness values used in the image is reduced and some detail is lost. However the effect of noise can also be reduced and the image becomes segmented,

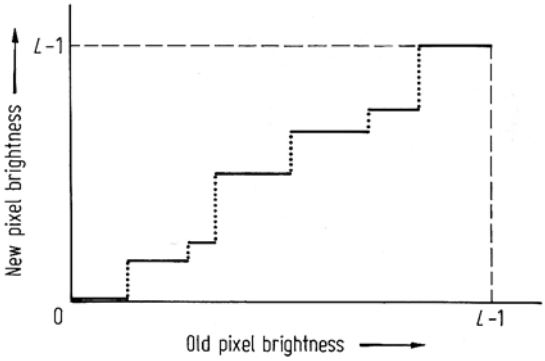


Fig. 4.16. The brightness value mapping function corresponding to black and white density slicing. The thresholds are user specified

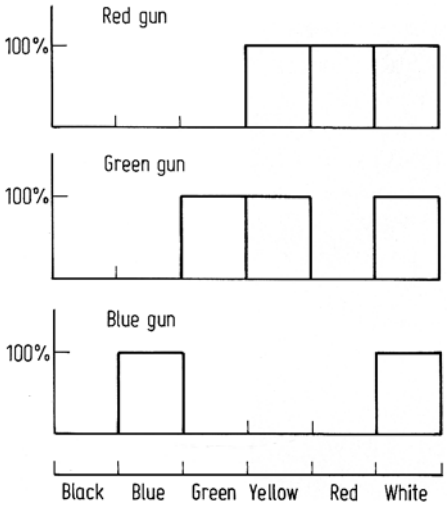


Fig. 4.17. Simple example of creating the look-up tables for a colour display device to implement colour density slicing. Here only six colours have been chosen for simplicity

or sometimes contoured, in sections of similar grey level, in which each segment is represented by a user specified brightness. The technique is known as density slicing and finds value, for example, in highlighting bathymetry in images of water regions when penetration is acceptable. When used generally to segment a scalar image into significant regions of interest it is acting as a simple one dimensional parallelepiped classifier (see Sect. 8.4). The brightness value mapping function for density slicing is as illustrated in Fig. 4.16. The thresholds in such a function are entered by the user. An image in which the technique has been used to highlight bathymetry is shown in Fig. 4.18. Here differences in Landsat multispectral scanner visible imagery, at brightnesses too low to be discriminated by eye, have been mapped to new grey levels to make the detail apparent.

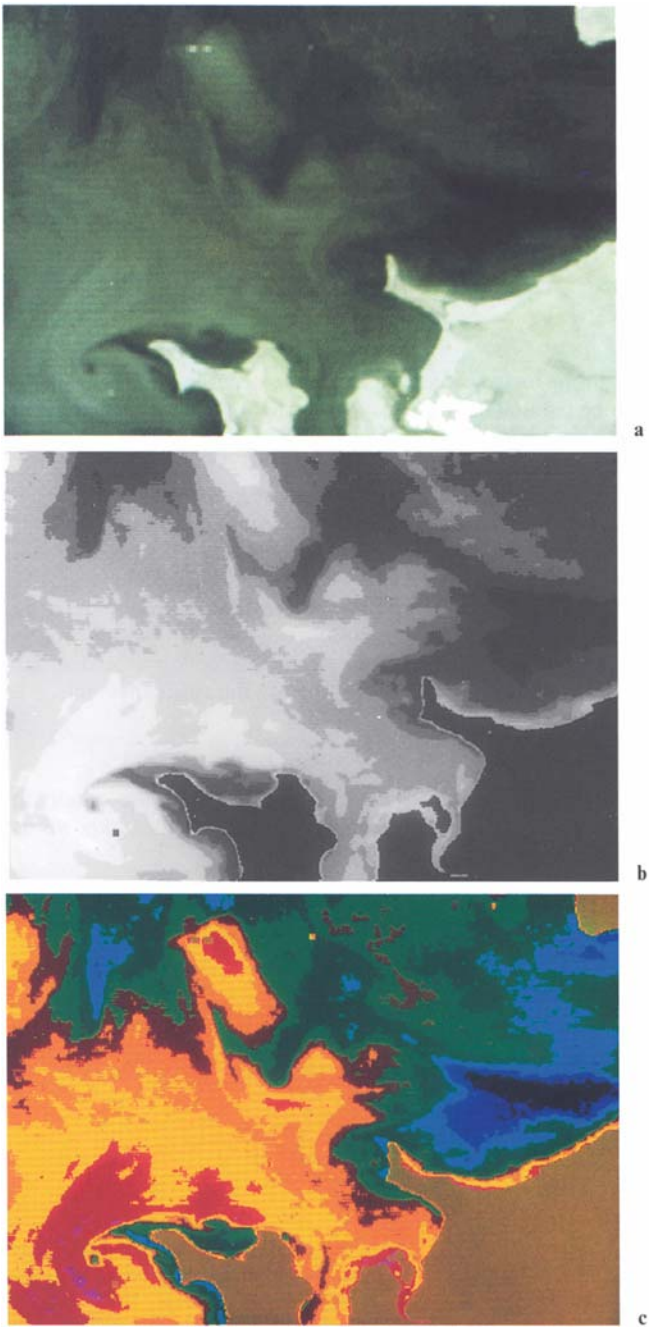


Fig. 4.18. Illustration of contouring in water detail using density slicing. **a** The image used is a band 5 + band 7 composite Landsat multispectral scanner image, smoothed to reduce line striping and then density sliced; **b** Black and white density slicing; **c** Colour density slicing

4.6.2 Colour Density Slicing and Pseudocolouring

A simple yet lucid extension of black and white density slicing is to use colours to highlight brightness value ranges, rather than simple grey levels. This is known as colour density slicing. Provided the colours are chosen suitably, it can allow fine detail to be made immediately apparent. It is a particularly simple operation to implement on a display system by establishing three brightness value mapping functions in the manner depicted in Fig. 4.17. Here one function is applied to each of the colour primaries used in the display device. An example of the use of colour density slicing, again for bathymetric purposes, is given in Fig. 4.18.

This technique is also used to give a colour rendition to black and white imagery. It is then usually called pseudocolouring. Where possible this uses as many distinct hues as there are brightness values in the image. In this way the contours introduced by density slicing are avoided. Moreover it is of value in perception if the hues used are graded continuously. For example, starting with black, moving from dark blue, mid blue, light blue, dark green, etc. through to oranges and reds will give a much more acceptable pseudocoloured product than one in which the hues are chosen arbitarily.

References for Chapter 4

Much of the material on contrast enhancement and contrast matching treated in this chapter will be found also in Castleman (1996) and Gonzalez and Woods (1992) but in more mathematical detail. Passing coverages are also given by Moik (1980) and Hord (1982). More comprehensive treatments will be found in Schowengerdt (1997), Jensen (1986), Mather (1987) and Harrison and Jupp (1990).

The papers by A. Schwartz (1976) and J.M. Soha et al. (1976) give examples of the effect of histogram equalization and of Gaussian contrast stretching. Chavez et al. (1979) have demonstrated the performance of multicycle contrast enhancement, in which the brightness value mapping function $y = f(x)$ is cyclic. Here, several sub-ranges of input brightness value x are each mapped to the full range of output brightness value y. While this destroys the radiometric calibration of an image it can be of value in enhancing structural detail.

K.R. Castleman, 1996: Digital Image Processing, 2e, N.J., Prentice-Hall.

P.S. Chavez, G.L. Berlin, and W.B. Mitchell, 1979: Computer Enhancement Techniques of Landsat MSS Digital Images for Land Use/Land Cover Assessment. Private Communication, US Geological Survey, Flagstaff, Arizona.

R.C. Gonzalez and R.E. Woods, 1992: Digital Image Processing, Mass., Addison-Wesley.

B.A. Harrison and D.L.B. Jupp, 1990: Introduction to Image Processing, Canberra, CSIRO.

A. Hogan, 1981: A Piecewise Linear Contrast Stretch Algorithm Suitable for Batch Landsat Image Processing. Proc. 2nd Australasian Conf. on Remote Sensing, Canberra, 6.4.1–6.4.4.

R.M. Hord, 1982: Digital Image Processing of Remotely Sensed Data, N.Y., Academic.

J.R. Jensen, 1986: Introductory Digital Image Processing — a Remote Sensing Perspective. N.J., Prentice-Hall.

The sensor collects some of the electromagnetic radiation (radiance[12]) that propagates upward from the earth and forms an image of the earth's surface on its focal plane. Each detector integrates the energy that strikes its surface (irradiance[13]) to form the measurement at each pixel. Due to several factors, the actual area integrated by each detector is somewhat larger than the *GIFOV*-squared (Chapter 3). The integrated irradiance at each pixel is converted to an electrical signal and quantized as a integer value, the *Digital Number (DN)*.[14] As with all digital data, a finite number of bits, Q, is used to code the continuous data measurements as binary numbers. The number of discrete *DN*s is given by,

$$N_{DN} = 2^{Q} \tag{1-4}$$

and the *DN* can be any integer in the range,

$$DN_{range} = [0, 2^{Q} - 1] \ . \tag{1-5}$$

The larger the value of Q, the more closely the quantized data approximates the original continuous signal generated by the detectors, and the higher the *radiometric resolution* of the sensor.

12. Radiance is a precise scientific term used to describe the power density of radiation; it has units of W-m^{-2}-sr^{-1}-μm^{-1}, i.e., watts per unit source area, per unit solid angle, and per unit wavelength. For a thorough discussion of the role of radiometry in optical remote sensing, see Slater (1980) and Schott (1996).

13. Irradiance has units of W-m^{-2}-μm^{-1}. The relationship between radiance and irradiance is discussed in Chapter 2.

14. Also known as *Digital Count*.

1.5 Image Display Systems

Computer image displays convert the digital image data to a continuous, analog image for viewing. They are usually preset to display 8 bits/pixel in greyscale, or 24 bits/pixel in additive color, achieved with red, green, and blue primary screen colors. Three bands of a multispectral image are processed by three hardware *Look-Up Tables (LUTs)* to convert the integer *DN*s of the digital image to integer *Grey Levels (GLs)* in each band,

$$GL = LUT_{DN}. \tag{1-7}$$

The *DN* serves as an integer index in the *LUT*, and the *GL* is an integer index in the video memory of the display (Fig. 1-24). The range in image *DN*s is given by Eq. (1-5), while GL typically has a range,

$$GL_{range} = [0, 255] \tag{1-8}$$

in each color. The hardware *LUT* can be used to apply a "stretch" transformation to the image *DN*s to improve the displayed image's contrast or, if the *DN* range of the original image is greater than the *GL* range, the *LUT* can be used to "compress" the range for display. The output of the *LUT* will always be limited according to Eq. (1-8).

Color images are formed from composites of the triplet of *GL*s corresponding to any three bands of a multispectral image. With a 24-bit display, each band is assigned to one of three 8-bit integers corresponding to the display colors: red (R), green (G), or blue (B). There are therefore 256 *GL*s available for each band. Every displayed pixel has a color defined by a triplet of *GL*s, which we may consider a three-dimensional column vector ***RGB***,[17]

$$\boldsymbol{RGB} \quad [GL_R, GL_G, GL_B]^T$$

There are 256^3 possible ***RGB*** vectors, but fewer distinguishable display colors because there are no monitors that can display all of the colors in the color cube. The exact color displayed for a given ***RGB*** data vector depends on the phosphor characteristics and control settings of the monitor.

Computer monitors create color in an *additive* way; that is, a pixel with equal *GL*s in red, green, and blue will appear as grey on the screen (if the monitor is properly adjusted); a pixel with equal amounts of red and green, and with no blue, appears as yellow, and so forth. Certain color combinations are widely used in remote sensing (Table 1-6). However, since many commonly-used bands are not even in the visible spectrum, color assignment in the display image is arbitrary. The "best"

17. The triplet is conveniently written as a row vector in Eq. (1-9). The superscript *T* converts it to a column vector by a transpose operation.

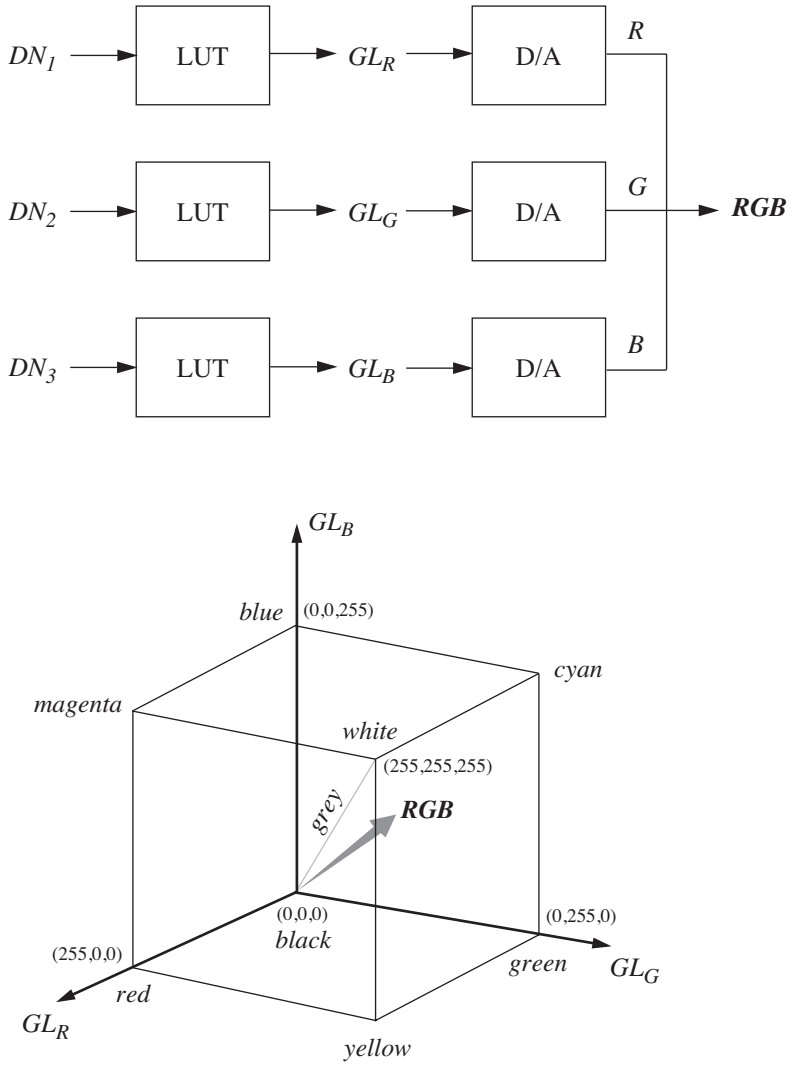


FIGURE 1-24. The conversion from DN to GL to color in a 24 bits/pixel digital video display. Three images are converted by individual LUTs to the three display GLs that determine the amplitude of the primary display colors, red, green and blue. The last step is a digital-to-analog (D/A) conversion and combination of the three channels to achieve the color seen on the monitor. At the bottom, the color cube for a 24 bits/pixel display is shown; the vector RGB specifies any triplet of GLs within the cube.

colors to use are those that enhance the data of interest. The popularity of the *Color IR (CIR)* type of display derives from its emulation of color IR photography, in which vegetation appears as red because of its relatively high reflectance in the NIR and low reflectance in the visible (Fig. 1-7). Anyone with photointerpretation experience is usually accustomed to interpreting such images. The natural color composite is sometimes called a "true" color composite, but that is misleading as there is no "true" color in remote sensing—natural color is more appropriate for the colors seen by the eye. The bands used to make TM CIR and natural color composites are shown in Plate 1-9.

Single bands of a multispectral image can be displayed as a greyscale image or *pseudo-colored* by converting each *DN* or range of *DN*s to a different color using the display's *LUT*s. Pseudo-coloring makes it easier to see small differences in *DN*.

TABLE 1-6. Sensor band mapping to RGB display color for standard color composites. A general false color composite is obtained by combining any three sensor bands.

sensor	**composite type**	
	natural color	*Color IR (CIR)*
generic	red:green:blue	NIR:red:green
ALI	4:3:2	5:4:2
ASTER	NA	3:2:1
AVIRIS	30:20:9	45:30:20
Hyperion	30:21:10	43:30:21
MODIS	13:12:9	16:13:12
MSS	NA	4:2:1
SPOT	NA	3:2:1
TM, ETM+	3:2:1	4:3:2

Ratio Images

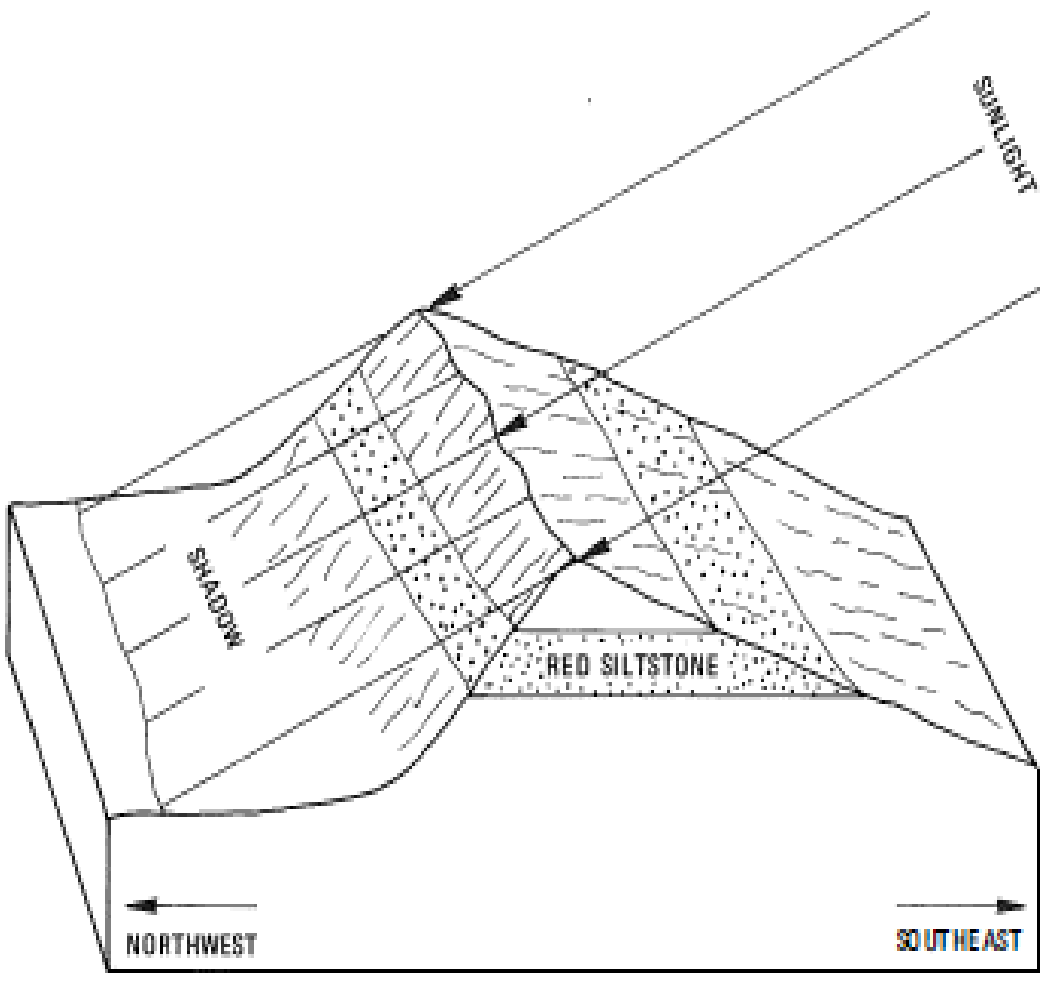
Ratio images are prepared by dividing the DN value in one band by the corresponding DN value in another band for each pixel. The resulting values are plotted as a ratio image. Figure 8-29 illustrates some ratio images prepared from TM bands of the Thermopolis subscene. In a ratio image the black and white extremes of the gray scale represent pixels having the greatest difference in reflectivity between the two spectral bands. The darkest signatures are areas where the denominator of the ratio is greater than the numerator. Conversely, the numerator is greater than the denominator for the brightest signatures. Where denominator and numerator are the same, there is no difference between the two bands.

For example, the spectral reflectance curve for vegetation (Figure 3-1) shows a maximum reflectance in TM band 4 (reflected IR) and a lower reflectance in band 2 (green). Figure 8-29C is the ratio image 4/2, which is produced by dividing DNs for band 4 by the DNs for band 2. The brightest signatures in this image correlate with the cultivated fields along the Wind River and Owl Creek (Figure 3-7H). Figure 8-29A is the ratio image 3/1 (red/blue), in which red beds of the Chugwater outcrops have very bright signatures.

Any three ratio images may be combined in red, green, and blue to produce a color image. In Plate 14C the ratio images 3/1, 5/7, and 3/5 are combined as red, green, and blue, respectively. Compare this image with the various Thermopolis color images (Plate 2) and the interpretation map (Figure 3-8). The signatures of the ratio color image express more geologic information and have greater contrast between units than do color images of individual TM bands. An advantage of ratio images is that they extract and emphasize differences in spectral reflectance of materials. A disadvantage of ratio images is that they suppress differences in albedo; materials that have different albedos but similar spectral properties may be indistinguishable in ratio images. Another disadvantage is that any noise is emphasized in ratio images.

Ratio images also minimize differences in illumination conditions, thus suppressing the expression of topography. In Figure 8-30 a red siltstone bed crops out on both the sunlit and shadowed sides of a ridge. In the individual Landsat TM bands 1 and 3, the DNs of the siltstone are lower in the shadowed area than in the sunlit outcrop, which makes it difficult to follow the siltstone bed around the hill. Values of the ratio image 3/1, however, are identical in the shadowed and sunlit areas, as shown by the chart in Figure 8-30; thus the siltstone has similar signatures throughout the ratio image. Highlights and shadows are notably lacking in the ratio images of Figure 8-29.

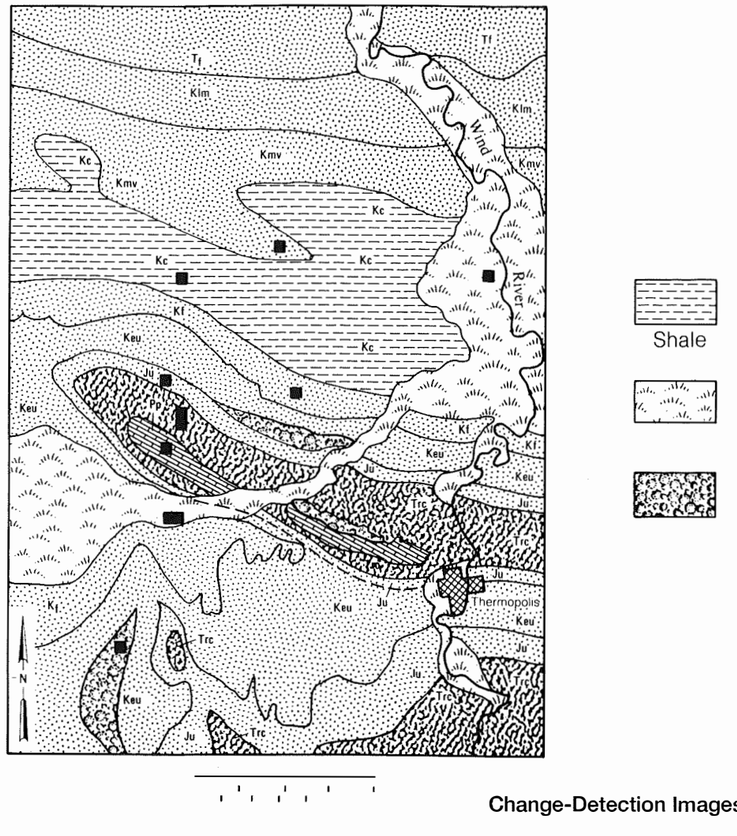
In addition to ratios of individual bands, a number of other ratios may be computed. An individual band may be divided by the average for all the bands, resulting in a normalized ratio image. Another ratio combination is produced by dividing the difference between two bands by their sum; for example, (band 4 – band 5)/(band 4 + band 5). Ratios of this type are used to process AVHRR data, as described in Chapter 9.



SILTSTONE REFLECTANCE

ILLUMINATION	TM BAND 3	TM BAND 1	RATIO 3/1
Sunlight	94	42	2.24
Shadow	76	34	2.23

Figure 8-30 Suppression of illumination differences on a ratio image.



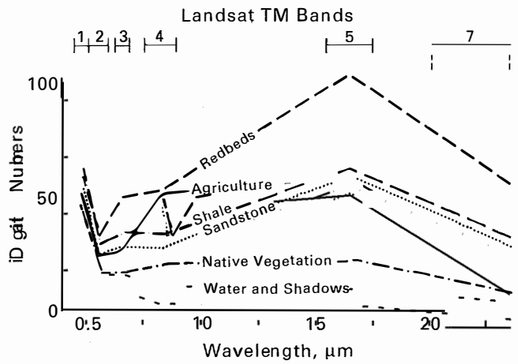


Figure 8-34 Reflectance spectra (from TM bands) of terrain categories shown in Figure 8-33

Change-Detection Images

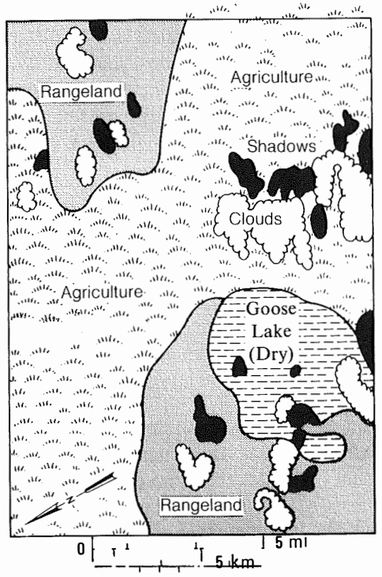
Change-detection images provide information about seasonal or other changes. The information is extracted by comparing two or more images of an area that were acquired at different times. The first step is to register the images using corresponding GCPs. Following registration, the digital numbers of one image are subtracted from those of an image acquired earlier or later. The resulting values for each pixel are positive, negative, or zero; the latter indicates no change. The next step is to plot these values as an image in which a neutral gray tone represents zero. Black and white tones represent the maximum negative and positive differences, respectively. Contrast stretching is employed to emphasize the differences.

The change-detection process is illustrated with Landsat MSS band 5 images of the Goose Lake area of Saskatchewan, Canada (Figure 8-35). The DN of each pixel in the June 27, 1973, image (Figure 8-35B) is subtracted from the DN of the corresponding registered pixel in the September 7, 1973, image (Figure 8-35A). The resulting values are linearly stretched and displayed as the change-detection, or difference, image (Figure 8-35C). The location map aids in understanding

A. September 7, 1973, image.

B. June 27, 1973, image.

C. Difference image (image A minus image B).



D. Terrain map.

Figure 8-35 Change-detection image computed from seasonal Landsat MSS images, Saskatchewan, Canada. From Rifman and others (1975, Figures 2-14, 2-15, 2-17).

signatures in the difference image. Neutral gray tones representing areas of little change are concentrated in the northwest and southeast and correspond to forest terrain. Forest terrain has a similar dark signature on both original images. Some patches within the ephemeral Goose Dry Lake have similar light signatures on images A and B, resulting in a neutral gray tone on the difference image. The clouds and shadows that are present only on image B produce dark and light tones, respectively, on the difference image. The agricultural practice of seasonally alternating between cultivated and fallow fields is clearly shown by the light and dark tones on the difference image. On the original images, the fields with light tones have crops or stubble and the fields with dark tones are bare earth.

Change-detection processing is also used to produce difference images for other remote sensing data, such as between nighttime and daytime thermal IR images (Chapter 5).

HARDWARE AND SOFTWARE FOR IMAGE PROCESSING

The image-processing routines are implemented on computer systems that consist of hardware and software, both of which are evolving at a rapid rate. For example, when the second edition of this book was written, in 1986, a typical state-of-the-art image-processing system cost several hundred thousand dollars and was supported by a supercomputer and peripheral hardware costing tens of million dollars. In 1996 a stand-alone desktop computer system costing $15,000 replaces the earlier system.

Hardware

Personal computers can be classified according to their operating system: Mac OS (Macintosh Operating System), Windows 95, or Unix. This book is not the forum to debate the relative merits of these systems. Most systems are acceptable; the choice is largely based on personal preference. Figure 8-36 shows the basic components of a typical image processing system, such as the one I use.

The data input/output components are used to read original data, such as TM, into the system. The common current formats are 8-mm tape and CD-ROM. The second function of the input/output components is to record the digitally processed data for later playback into image prints (hard copy). The output function is also used to make backup records that safeguard against loss of data.

The original TM data are loaded onto the hard drive (not shown in Figure 8-36) which stores the original data. Several different data sets (TM, SPOT, digital terrain data) may be stored and used concurrently to create a combination image. During a processing session, a number of intermediate data sets are generated and stored on the hard drive. Therefore the hard drive should have several gigabytes (10^9 bytes) of memory, which is relatively inexpensive. The data are processed by

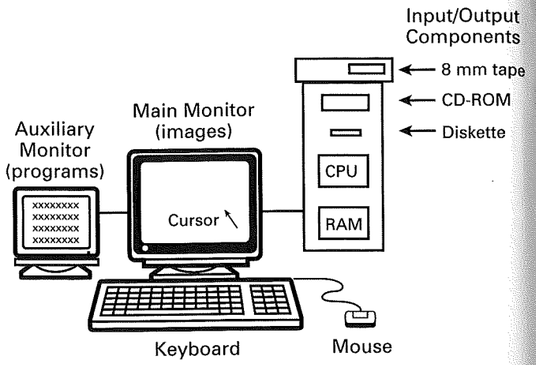


Figure 8-36 Components of a typical interactive image-processing system.

the central processing unit (CPU) which incorporates the random access memory (RAM), shown in Figure 8-36. The amount of RAM influences the speed at which the data are processed. RAM is relatively expensive, but several tens of megabytes are a minimum requirement. The system in Figure 8-36 has 96 megabytes of RAM and performs satisfactorily. New systems operate in a parallel-processing mode that accelerates processing speeds. Data are divided into subsets, typically four, which are processed simultaneously and recombined into output data.

Images and program information are shown on monitors, which are digitally controlled color display screens. Large-area, high-resolution monitors are desirable for most operations. The system in Figure 8-36 employs a second optional smaller monitor to display program information, which eliminates clutter on the large monitor. The keyboard and mouse enable the operator to interact with the system.

Software

Software is a set of instructions that commands the hardware system. A basic set of software (DOS, Apple, or Unix) is installed to operate the system. Additional application software is required for specific tasks, such as image processing. A variety of image-processing software is available for each of the three operating systems. Table 8-1 is a representative list of commercial vendors of image-processing software. All of the software packages include routines for the basic operations described in this chapter (restoration, enhancement, and information extraction), although the organization and nomenclature may be different. This book is not the forum to review the merits of different software packages, except to note that there are several ways to evaluate software:

1. Contact established users with applications similar to yours and get their opinions.
2. The Internet has bulletin boards that can be helpful.
3. Ask vendors for demonstration versions of their software that can be installed on your system for evaluation.
4. Many vendors have demonstration booths at the remote sensing conferences listed in Chapter 1, where you can try out the software.

In addition to commercial software, there are a variety of public-domain packages that are accessible via Internet. (*Note:* I am not suggesting "pirating," or copying copyrighted software. Not only is this practice illegal; it also denies compensation to the developers of the software. Without compensation, there is no incentive to develop new software that we will need in the future.)

Interactive Image-Processing Session

A typical Landsat image-processing session proceeds as follows:

1. After the CCT is loaded, the analyst selects three TM bands and assigns them to the blue, green, and red channels of the monitor. A typical combination is bands 2-4-7 shown in BGR (blue, green, and red). The TM image consists of 5667 lines, each with 6167 pixels, but the monitor in Figure 8-36 displays only 1152 lines by 870 pixels. The image is resampled to fit on the monitor; every eighth line and every eighth pixel are displayed. The mouse and on-screen cursor are used to select a representative subscene, which is displayed at full resolution with each line and pixel of original data shown on the monitor.
2. The display is examined for defects, such as banding and line dropouts, which are then restored.
3. The next step is to enhance the image. Each of the three spectral bands, together with its histogram, is viewed separately, and the contrast is enhanced. The three bands are then viewed as a color display, which typically is undersaturated. The image is transformed into its IHS components, and saturation is enhanced. At this stage it is generally useful to apply a nondirectional edge-enhancement filter to the intensity component. The image is transformed back into an enhanced BGR display. Coordinate grids (latitude and longitude; UTM) are added and the image is ready for plotting as hard copy.
4. Information extraction begins with the original data, not with the enhanced data from step 3. Images of principal components, ratios, and classifications are generated and interactively modified to suit the application.

Table 8-1 Image-processing software and vendors

Dimple
Cherwell Scientific Publishing, Inc.
744 San Antonio Road, Suite 27A
Palo Alto, CA 94303
Telephone: 415-852-0720
Fax: 415-852-0723

EASI/PACE
PCI Enterprises
50 West Wilmot Street
Richmond Hill, Ontario
Canada L4B 1M5
Telephone: 905-764-0614
Fax: 905-764-9604

ENVI
Research Systems, Inc.
2995 Wilderness Place
Boulder, CO 80301
Telephone: 303-786-9900
Fax: 303-786-9909

ER Mapper
Earth Resource Mapping
4370 La Jolla Drive, Suite 900
San Diego, CA 92122
Telephone: 619-558-4709
Fax: 619-558-2657

ImageStation
Intergraph Corp.
Huntsville, AL 35894
Telephone: 205-730-2000
Fax: 205-730-1263

IMAGINE
ERDAS
2801 Buford Highway NE, Suite 300
Atlanta, GA 30329
Telephone: 404-248-9000
Fax: 404-248-9400

TNTmips
MicroImages, Inc.
201 North 8th Street
Lincoln, NB 68508
Telephone: 402-477-9554
Fax: 402-477-9559

VI²STA
International Imaging Systems
1500 Buckeye Drive
Milpitas, CA 95035
Telephone: 408-432-3400
Fax: 408-433-0965

AN ON-LINE FLOOD DATABASE FOR GREECE SUPPORTED BY EARTH OBSERVATION DATA AND GIS

Nikolaidou M.[1], Mouratidis A.[2,3], Doxani G.[2], Oikonomidis D.[3], Tsakiri-Strati M.[1] and Sarti F.[2]

[1] *Department of Cadastre, Photogrammetry and Cartography, School of Rural and Surveying Engineering, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, melinikol@gmail.com, martsaki@topo.auth.gr*
[2] *European Space Agency (ESA/ESRIN), Via Galileo Galilei, 00044 Frascati, Italy, antonios.mouratidis.esa.int@gmail.com, georgia.doxani@esa.int, francesco.sarti@esa.int*
[3] *Department of Physical and Environmental Geography, School of Geology, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, amourati@geo.auth.gr, oikonomi@geo.auth.gr*

ABSTRACT

Flooding, a result of both natural and anthropogenic factors, poses serious risks for life, properties and infrastructure. On global scale, with respect to the direct or indirect impact of natural disasters on millions of people per year, floods are ranked as number one catastrophe. In Greece, despite the plethora of major and catastrophic events in the country during the past decades, floods had not been adequately studied. Furthermore, although Earth Observation (EO) data and Geographic Information Systems (GIS) provide a secure and economic way of delineating, monitoring and ultimately managing flood events, they have been hardly used in relevant studies in Greece. The main objectives of the present study were the production of an online flood database for Greece, the use of EO and GIS techniques for flood mapping and the provision of freely available geospatial data. EO data used mainly included Synthetic Aperture Radar (SAR) (e.g. ENVISAT/ASAR, ERS) and medium to high resolution optical satellite imagery (e.g. Landsat, SPOT) for delineating flooded areas and producing flood and flood-risk maps. The results of the investigation are publically available over the internet through the project's website, with potential for further updates and expansion. Overall, it is envisaged that the data provided through this project, shall serve as a basis for flood disaster management in the future, both during as well as in the pre- and post-crisis phases.

Keywords: Greece, floods, database, GIS, Earth observation, mapping

1. INTRODUCTION

Flooding, a result of both natural and anthropogenic factors, poses serious risks for life, properties and infrastructure. On global scale, with respect to the direct or indirect impact of natural disasters on millions of people per year, floods are ranked as number one catastrophe (Bell, 1999).

During the last decades, the escalating frequency of flood events around the world, together with the evidences and warnings about global climatic change, rendered this phenomenon a very serious issue. Efficient flood management is thereafter a fundamental necessity, in order to minimize the adverse consequences, in terms of human safety and damage to property.

To this end, European Union (EU) Member States are conducting a preliminary flood risk assessment and subsequently developing flood hazard (i.e. showing flood probability) and flood risk maps (i.e. related to the potential adverse consequences of a flood), whereas by 2015 flood risk management plans will be drawn for high risk zones (Directive 2007/60/EC). This three stage process applies to all kinds of floods (river, lakes, flash floods, urban floods, coastal floods, including storm surges and tsunamis) on all of the EU territory and will need to be reviewed every six years (European Union 2007).

According to the definitions given by the U.S. National Oceanic and Atmospheric Administration (NOAA), floods are overflows of water onto normally dry land that may last days or weeks. They are caused by rising water in an existing waterway, such as a river, stream, or drainage ditch, with the ponding of water occurring at or near the point where the rain fell. A flash flood is caused by heavy or excessive rainfall in a short period of time, generally less than six hours. Flash floods are usually characterized by raging torrents after heavy rains that rip through river beds, urban streets, or mountain

canyons sweeping everything before them. They can occur within minutes or a few hours of excessive rainfall. They can also occur even if no rain has fallen, for instance after a levee or dam has failed, or after a sudden release of water by a debris or ice jam. Depending also on the geological and geomorphological regime, the water can remain in the affected area for several days or, more commonly, run off within just a few hours.

Earth Observation (EO) data, along with Remote Sensing and Geographical Information Systems (GIS) techniques, provide safe and cost-effective tools for monitoring, mapping and assessing the evolution and damages caused by flood events. Initiatives, dedicated centres, institutions and services, such as (i) the International Charter (http://www.disasterscharter.org), (ii) the Centre for Satellite Based Crisis Information (ZKI, http://www.zki.dlr.de/), (iii) Services and Applications For Emergency Response (Safer, http://safer.emergencyresponse.eu) and (iv) SERTIT (Service Régional de Traitement d'Image et de Télédétection, http://sertit.u-strasbg.fr/), use satellite images for Earth monitoring, offering substantial support to major flood events and natural disasters in general, around the world.

In Greece, despite the plethora of major and catastrophic events in the country during the past decades, floods have not been adequately studied. In particular, although EO data and GIS provide a secure and economic way of delineating, monitoring and ultimately managing flood events, they have been hardly used in relevant studies in Greece. This has been mainly attributed to the fact that most floods occurring on Greek territory are relatively (with respect to the global average) small scale flash-floods, meaning that EO data of high spatial and temporal resolution are required, in order to "capture" the disaster. Regrettably, this kind of data is generally not available and hence it is considered as highly unlikely for flash-floods to be recorded with appropriate satellite acquisitions. However, in early 2011, a project focusing on the creation of a flood database for Greece and its combined use with EO data and GIS was carried out (Mouratidis 2011, Mouratidis et al. 2011, Mouratidis et al. 2012). One year later, Diakakis et al. (2012) published another historical flood catalogue together with a statistical and spatial analysis of flood events, while the Ministry of Environment Energy & Climate Change of Greece published a detailed report on the assessment and management of flood risks in Greece, in accordance to the European Union Directive (YPEKA, 2012).

The main objectives of the present study were the production of an online flood database for Greece, the use of EO techniques and GIS for flood mapping and the provision of freely available geospatial data. EO data used mainly included: a) Synthetic Aperture Radar (SAR) satellite imagery (ERS , ENVISAT/ASAR, ALOS/PALSAR) for delineating flooded areas by change detection techniques and b) high resolution optical satellite images (e.g. Landsat, SPOT, IKONOS, Quickbird) are related classification techniques, mainly for the creation of flood risk maps and, where feasible, also for the delineation of flooded areas.

2.4 Image processing

2.4.1 SAR image processing

Contrary to optical sensors, radar systems with their all-weather, day and night applicability make SAR data more appropriate for monitoring flood events (e.g. Badji and Dautrebande 1997; Yésou et al. 2000; Sarti 2004; Li et al. 2005), as the latter are normally associated with bad meteorological conditions and a high percentage of cloud coverage.

During the pre-processing of SAR images using NEST (Next ESA SAR Toolbox) and Envi™, the following steps took place:

1. Image calibration, by converting pixel values from digital number (DN) into backscattering coefficient (σ^0) following Rosich & Meadows (2004), so that the value of the same pixel in each image, as well as the different pixel values in the same image, would become comparable.
2. Co-registration of all SAR images in order to ensure geographical and geometrical overlap.
3. Optional orthorectification of the co-registered images using a Digital Elevation Model (DEM) (e.g. SRTM).

Subsequently, the main SAR image processing for the detection and mapping of flooded regions included:

1. The production of the average "dry" image, by calculating the mean values for each pixel using all the available "dry" images.
2. The implementation of a Change Detection Analysis (CDA) approach. CDA encompasses a broad range of methods used to identify, describe and quantify differences between images of the same scene at different times. In this study, the False Colour Composition (FCC) was adopted and applied. Typically, two images were used; one before (mean of the dry images-where more than one of them were available) and one during/after the flood. In each case study, the flood image was assigned to both the red (R) and the green (G) channels while the ''dry'' image was assigned to the blue (B) channel, in order to create an RGB false colour composite.

With respect to the interpretation of these RGB image products, unchanged or almost unchanged features appear in variations of grey, whereas any change in the scene (denoting change in backscattering) from one acquisition to the other appears in colour, so that:

1. Regions with significantly lower backscatter in the flood image appear in blue, indicating possible flooded areas.
2. Regions with significantly higher backscatter in the flood image appear in yellow (yellowish), indicating a possible increase in soil moisture.
3. Areas with little or no change are depicted as grey, the result thus being approximately equivalent with that of each image independently (Fig. 8-11).

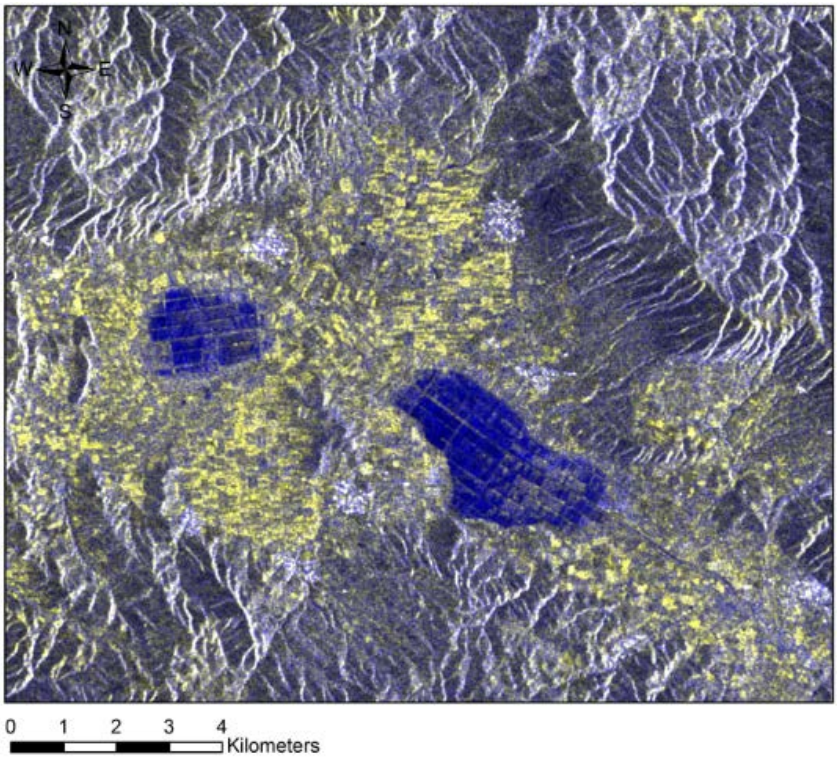


Figure 8. Delineation of flooded areas by change detection techniques, in the post-crisis period (two days after the flash-flood) in the prefecture of Thessaloniki (case study 1), using an ENVISAT/ASAR IMG mode false colour composite: R=G=10/10/2006 (flood image), B=(5/10/2004 + 25/10/2005)/2 (average of two images during the same season, but under dry conditions). Blue colour depicts flooded regions, while yellow colour depicts wet soil.

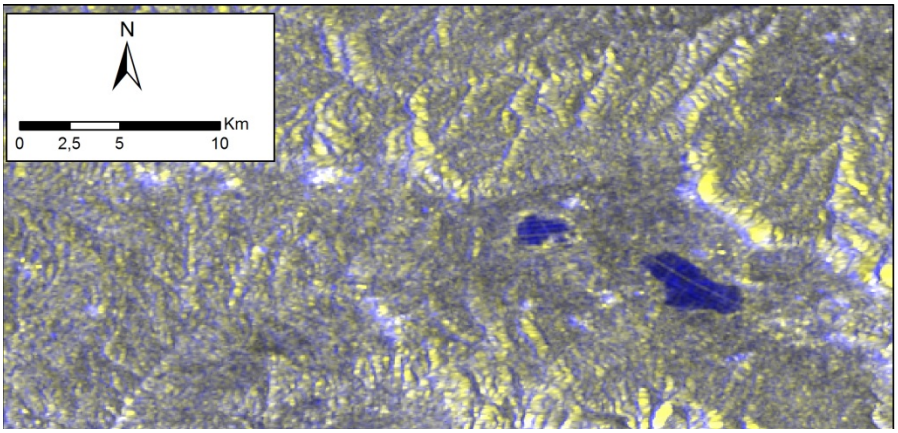


Figure 9. Delineation of flooded areas in the prefecture of Thessaloniki in 2011 (case study 2), during the post-crisis phase (four days after the flash-flood), using an RGB false colour composite of orthorectified medium spatial resolution (150m) ASAR Wide Swath Mode data. R=G=25/09/2011 (flood image), B=25/09/2011 (dry conditions). Blue colour depicts flooded regions, while yellow colour depicts wet soil.

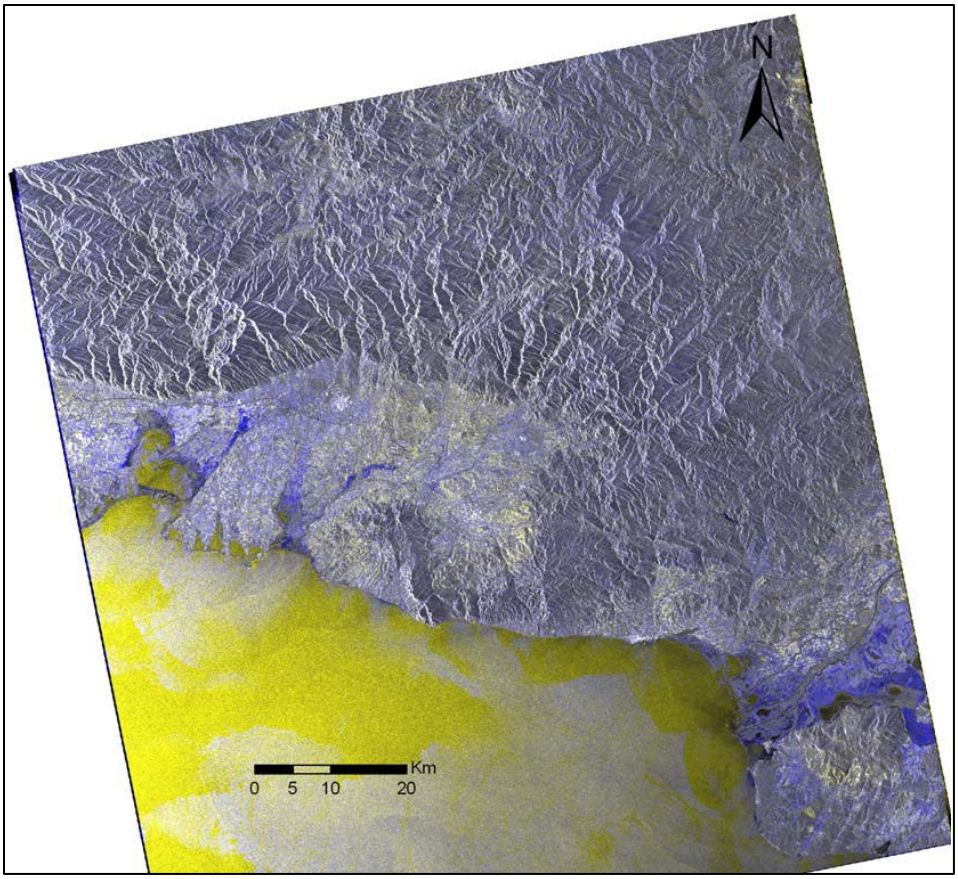


Figure 10: Overview of flooded areas in the prefecture of Thrace in 2007 (case study 3), during the crisis period, by change detection techniques, using an ENVISAT/ASAR false colour composite: R=G=18/11/2007 (flood image), B= average of seven images taken under dry conditions. Blue colour depicts flooded regions while yellow colour depicts wet soil.

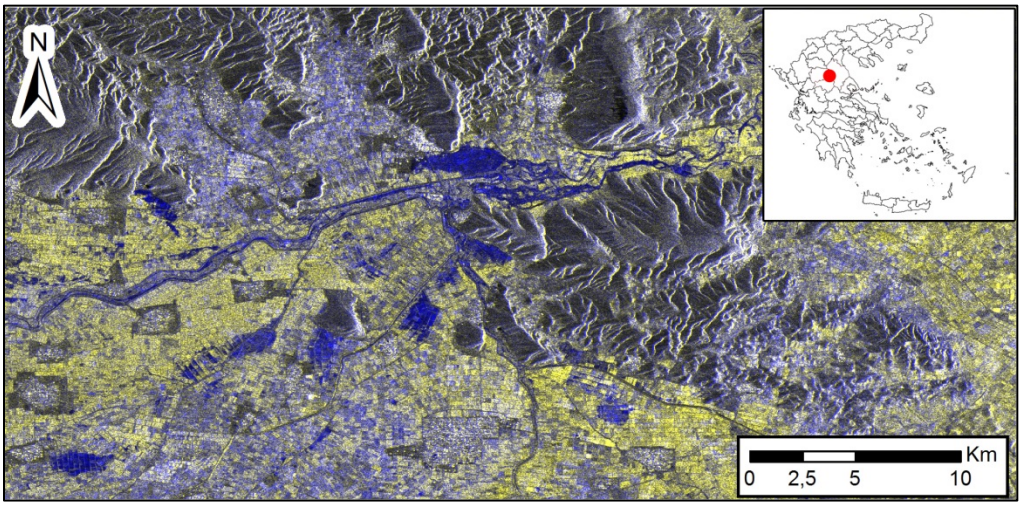


Figure 11: Floods along Pinios river, near Piniada, Farkadona and surrounding areas in Thessaly, in 2003 (case study 4), captured by ERS-2 during the crisis phase. SAR RGB false colour composite: R=G=02/02/2003 (flood image), B= 06/02/2005 (dry conditions). Blue colour depicts flooded regions while yellow colour depicts wet soil.

Badji, M., Dautrebande, S. (1997). Characterization of flood inundated areas and delineation of poor drainage soil using ERS-1 SAR imagery. Hydrol. Process. 11:1441–1450.

Bell, F.G., (1999). Geological Hazards. Their assessment, avoidance and mitigation. E and FN SPON, London.

Castilla, G., Hay, G., (2006). Uncertainties in land use data. Hydrology and Earth System Sciences, 3, 3439-3472.

Chavez, P.S., Sides, Jr., S.C., Anderson, J.A. (1991). Comparison of three different methods to merge multi-resolution and multi-sectoral data: Landsat TM and SPOT Panchromatic. Photogrammetric Engineering and Remote Sensing, 57(3), 295-303.

Diakakis M., Mavroulis S., Deligiannakis G. (2012). Floods in Greece, a statistical and spatial approach. Natural Hazards, 62 (2), 485-500.

European Union (2007). Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks – Official Journal of the European Union, L 288/27111.

Li, J., Yésou, H., Huang, S., Li, J., Li, X., Xin, J., Wang, X., Andreoli, R. (2005). ENVISAT ASAR medium and high resolution images for near real time flood monitoring in China during the 2005 flood season. In: Proceedings. 2005 dragon symposium, mid-term results, Santorini, Greece 27 June–1 July 2005 (ESA SP-611).

Mouratidis, A. (2011). Contribution of earth observation data and geographical information systems to mapping and managing flood events in Greece, Final Report, Project funded by the John S. Latsis Public Benefit Foundation, 37 p.

Mouratidis A., Sarti F. (2013). Flash-Flood Monitoring and Damage Assessment with SAR Data: Issues and Future Challenges for Earth Observation from Space Sustained by Case Studies from the Balkans and Eastern Europe. In: J. M. Krisp et al. (eds.), Earth Observation of Global Changes (EOGC), Lecture Notes in Geoinformation and Cartography, Springer, Berlin, pp. 125-136.

Mouratidis, A., Nikolaidou, M., Doxani, G., Sarti, F., Lampiri, M., Tsakiri-Strati, M. (2011). Flood studies in Greece using Earth Observation data and Geographical Information Systems. Annual General Meeting of the Geological Remote Sensing Group (GRSG), Geological Society of London/Remote Sensing and Photogrammetry Society, 7th-9th December 2011, ESA ESRIN, Frascati, Italy.

Mouratidis, A., Doxani, G., Nikolaidou, M., Lampiri, M., Sarti, F., Tsakiri-Strati, M. (2012). Contribution of Earth Observation Data and GIS to mapping and managing flood events in Greece. GRSS, IEEE International Geoscience and Remote Sensing Symposium, Remote Sensing for a Dynamic Earth, 22-27 July 2012, Munich, Germany.

Nikolaidou, M. (2009). Utilization and contribution of Remote Sensing and Geographical Information Systems (GIS) technology to the detection of flooded areas south of lake Volvi: An environmental approach, Master Thesis, School of Geology, Aristotle University of Thessaloniki (In Greek, with English abstract).

Nikolaidou, M., Mouratidis, A., Oikonomidis, D., Astaras, T. (2010). Mapping the catastrophic October 2006 flood events in Thessaloniki and Halkidiki with Envisat/ASAR data, Proceedings of the 9th Pan-Hellenic Geographical Conference, 4-6 November 2010, Athens, pp. 163-171.

Phol, C., Van Genderen, J.L (1998). Multisensor image fusion in remote sensing: concepts, methods and applications. International Journal of Remote Sensing, 19(5), 823-854.

Rosich, B., Meadows, P. (2004). Absolute calibration of ASAR Level 1 products, ESA/ESRIN, ENVI-CLVL-EOPG-TN-03-0010, Issue 1, Revision 5.

Sarti, F. (2004) Potentiels, limitations et évolutions de la télédétection optique-radar et de l'interférométrie radar pour le suivi des changements et des déformations de surface: applications scientifiques et applications pré-opérationnelles de gestion des risques naturels. Thèse, Univ. Paul Sabatier Toulouse.

Saaty, T.L., 1980: The Analytical Hierarchy Process, NY, McGraw Hill.

Tsakiri-Strati, M., Papadopoulou, M., Georgoula, O. (2002). Fusion of XS SPOT4 and PAN SPOT2 Images and Assessment of the Spectral Quality of the Products. Tech. Chron. Sci. J. TCG, I, 22(3), 9-22.

Yésou, H., Chastanet, P., Fellah, K., Jeanblanc, Y., De Fraipont, P., Bequignon, J. (2000). Contribution of ERS SAR images and ERS coherence data to a flood system on the Meuse basin-France. In: Proceedings of ERS-ENVISAT Symposium ‘‘looking at our Earth for the new millennium’’, 16–20 October 2000, Gothenburg (ESA SP-461).

YPEKA (2012). Report of the assessment and management of flood risks in Greece according to the European Union Member States directive, Ministry of Environment Energy & Climate Change of Greece (http://www.ypeka.gr/Default.aspx?tabid=252&language=en-US&SkinSrc=[G]Skins%2F_default%2FNo+Skin&ContainerSrc=[G]Containers%2F_default%2FNo+Container&dnnprintmode=true).

Internet links:
http://ceogis-floods.web.auth.gr
http://www.gdem.aster.ersdac.or.jp
http://geodata.gov.gr
http://gis.ktimanet.gr/wms/ktbasemap
http://glcf.umiacs.umd.edu
http://sertit.u-strasbg.fr/
http://srtm.csi.cgiar.org

Table 2-3 Terrain signatures on normal color film and IR color film

Subject	*Normal color film*	*IR color film*
Healthy vegetation:		
Broadleaf type	Green	Red to magenta
Needle-leaf type	Green	Reddish brown to purple
Stressed vegetation:		
Previsual stage	Green	Pink to blue
Visual stage	Yellowish green	Cyan
Autumn leaves	Red to yellow	Yellow to white
Clear water	Blue-green	Dark blue to black
Silty water	Light green	Light blue
Damp ground	Slightly darker than dry soil	Distinctly darker than dry soil
Shadows	Blue with details visible	Black with few details visible
Water penetration	Good	Moderate to poor
Contacts between land and water	Poor to fair discrimination	Excellent discrimination
Red bed outcrops	Red	Yellow

Figure 2-28 is a diagrammatic cross section of a leaf that explains these spectral signatures of vegetation. The transparent epidermis allows incident sunlight to penetrate into the mesophyll, which consists of two layers: (1) the palisade parenchyma of closely spaced cylindrical cells, and (2) the spongy parenchyma of irregular cells with abundant interstices filled with air. Both types of mesophyll cells contain chlorophyll, which reflects part of the incident green wavelengths and absorbs all of the blue and red energy for photosynthesis. The longer wavelengths of photographic IR energy penetrate into the spongy parenchyma, where the energy is strongly scattered and reflected by the boundaries between cell walls and air spaces. The high IR reflectance of leaves is caused not by chlorophyll but by the internal cell structure. Gausman (1985) gives details of optical properties of plant leaves in the visible and reflected IR regions. Buschmann and Nagel (1993) describe the roles of chlorophyll and cell structure in spectral reflectance of leaves.

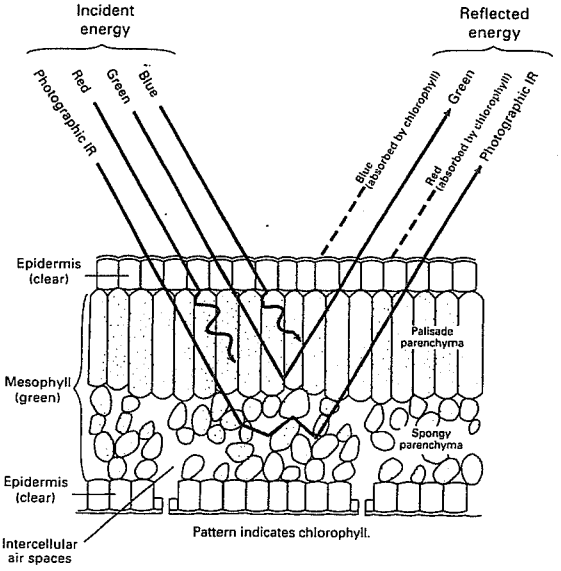
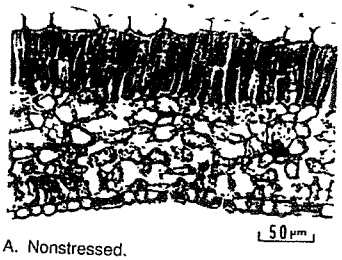


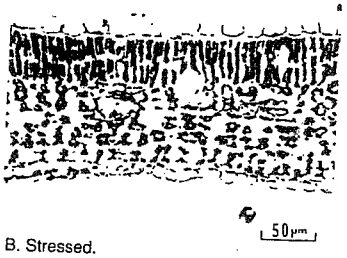
Figure 2-28 Diagrammatic cross section of a leaf, showing interaction with incident energy. Incident blue and red wavelengths are absorbed by chlorophyll in the process of photosynthesis. Incident green wavelengths are partially reflected by chlorophyll. Incident IR energy is strongly scattered and reflected by cell walls in the mesophyll. Modified from Buschmann and Nagel (1993, Figure 9).

Detection of Stressed Vegetation

Vegetation may be stressed because of drought, disease, insect infestation, or other factors that deprive the leaves of water. Figure 2-29 compares the internal structure of nonstressed and stressed leaves. The nonstressed leaf (Figure 2-29A) has a cell structure and reflectance characteristics comparable to those in Figure 2-28. In the stressed leaf (Figure 2-29B), the shortage of water causes the mesophyll cells to collapse, which strongly



A. Nonstressed.



B. Stressed.

Figure 2-29 Photomicrographs of cross sections of nonstressed and stressed leaves. Collapse of cells in the mesophyll layer strongly reduces reflectance of incident IR energy. From Everitt and Nixon (1986, Figure 1). Courtesy J. H. Everitt, U.S. Department of Agriculture.

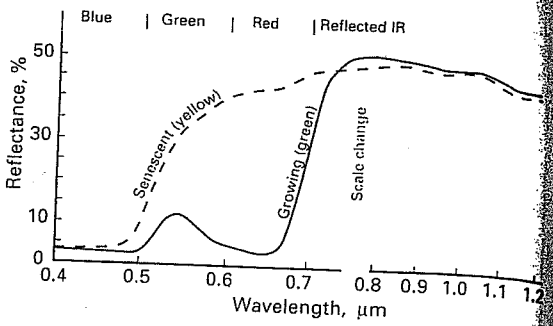


Figure 2-30 Reflectance spectra of green and senescent foliage. In the autumn, chlorophyll deteriorates, which reduces the absorption of incident red energy. The development of anthocyanin and tannin causes the yellow-red fall colors. From Schwaller and Tkach (1985, Figure 2).

reduces IR reflectance from the spongy parenchyma. This decreased reflectance diminishes the red signature in IR color photographs. Chlorophyll is still present, and the foliage may have a green signature in normal color photographs for some time after the onset of stress. In IR color photographs, however, stressed foliage has a distinctive blue signature. The loss of IR reflectance is a *previsual symptom* of plant stress because it often occurs days or even weeks before the visible green color begins to change. The previsual effect may be used for early detection of disease and insect damage in crops and forests. Evidence of plant stress is seen in the intramural playing field east of Drake Stadium (Figure 2-27 and Plate 1C,D), which is watered by a sprinkler system. In the normal color photograph the field is entirely green, but in the IR color photograph the red signature is interrupted by blue strips that indicate inadequately watered turf.

Autumn Senescence of Vegetation

In the autumn, leaves of deciduous trees undergo senescence and turn red, yellow, and brown. Figure 2-30 compares spectra of green and senescent foliage. The green chlorophyll has decayed, and red wavelengths are no longer absorbed. The organic compounds anthocyanin and tannin are formed, causing the familiar autumn colors (Boyer and others, 1988). The spectrum for senescent foliage (Figure 2-30) shows nearly equal reflectance values in the green, red, and photographic IR bands, which results in a white signature in IR color photographs. Boyer and others also describe the changes in leaf physiology and spectral reflectance during senescence.

Signatures of Other Terrain Features

The small lake north of UCLA has a dark green signature in the normal color photograph that blends with the vegetation. In the IR color photograph the lake has a dark blue signature that contrasts with the red signature of vegetation. This ability to enhance the difference between vegetation and water is especially valuable for mapping drainage patterns in heavily forested terrain. Silty water has a light blue signature in IR color photographs. One can recognize damp ground on IR color photographs by its relatively darker signature, caused by absorption of IR energy. Shadows are darker in IR color photographs than in normal color photographs because the yellow filter eliminates blue light.

The IR color photograph in Plate 1D has a better contrast ratio than the normal color photograph, for two reasons:

1. The yellow filter eliminates blue light, which is preferentially scattered by the atmosphere, as shown by the curve in Figure 2-26C. Eliminating much of the scattering improves the contrast ratio.
2. For vegetation, soils, and rocks, reflectance differences are commonly greater in the photographic IR region than in the visible region.

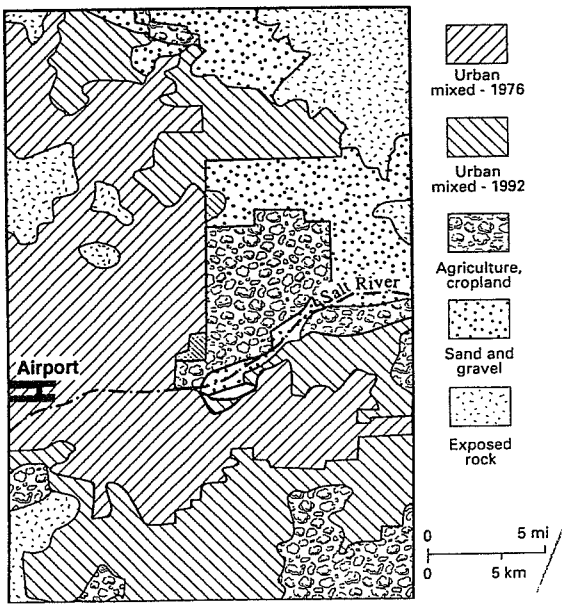


Figure 12-10 Changes in land use from 1976 to 1992 for the eastern portion of Phoenix, Arizona. Changes were interpreted from Landsat MSS images acquired in 1976 and 1992 (Plate 23C, D).

corner reflectors in urban environments. The reduced brightness is explained by the orientation of the street patterns and buildings relative to the Seasat look direction. Street patterns in this Arizona image are oriented north-south and east-west; therefore, the northeast Seasat look direction is oblique to most buildings, which reduces the intensity of radar backscatter. This orientation effect was described in Chapter 6. The retirement community of Sun City (Figure 12-9) northwest of Phoenix is an exception to the orthogonal street pattern, because the streets are laid out in curved and circular patterns. The radar-bright patches within Sun City (Figure 12-8A) are caused by groups of houses oriented with some walls normal to the northeast radar look direction.

The dark linear features in the urban areas are major highways and irrigation canals. Airport runways (Figure 12-9) also have dark signatures. The narrow irregular dark features in Sun City are golf courses and parks with smooth lawns.

The irrigated agricultural areas surrounding Phoenix belong to the major categories of "cropland and pasture" (210) and "orchards, groves, vineyards, nurseries, and ornamental horticulture" (220). In the Seasat image, the fields east of Phoenix have a distinctly darker tone than the fields west of Phoenix. When the Seasat and Landsat images were acquired in 1978, cotton, wheat, and alfalfa were the major crops in the eastern area. Vegetables, citrus fruits, and grapes along with some cotton and wheat were grown in the western area, which was more intensively irrigated. The relatively bright Seasat signatures of the western area may be caused by the different crops and soil moisture contents.

Changing Patterns of Land Use

Phoenix is one of the "sunbelt cities" in the southwestern United States that have experienced explosive urban growth in the past two decades. Repeated Landsat images are well suited to interpret the expansion of urban land use. Plate 23C,D shows MSS images acquired in 1976 and 1992 of the eastern portion of Phoenix and vicinity. The images coincide with a portion of the Seasat coverage as shown by the rectangle in Figure 12-9. Comparing the MSS images shows the extent and location of urban growth over a 13-year period. The map in Figure 12-10 shows level II land-use categories. The "mixed urban" category (180) is shown in contrasting patterns for 1976 and 1992. The "urban" category expanded by replacing adjacent agricultural land. The 1992 image clearly shows agricultural lands that are candidates for the next cycle of urban expansion. Repetitive images provide valuable information for regulating urban growth and planning the infrastructure of new schools, utilities, parks, and public services.

In this example, changes in land use were interpreted manually. Alternatively, each Phoenix image could have been digitally classified, such as in the example of Las Vegas. A change-detection algorithm applied to the registered 1976 and 1992 classifications would show graphically and quantitatively the changes in land use.

VEGETATION MAPPING WITH AVHRR IMAGES

Vegetation, both native and cultivated (agriculture), covers much of the earth and strongly influences the environment. Vegetation provides food, fiber, and building material. Until recently adequate data were lacking for mapping the composition, concentration, and dynamics of the world's vegetation. Now the advanced very high resolution radiometer (AVHRR) system on the NOAA environmental satellites provides worldwide coverage twice daily. The AVHRR was described in Chapter 4.

Normalized Difference Vegetation Index

Figure 12-11 shows spectral reflectance curves for soil and vegetation together with wavelength ranges of AVHRR bands 1 and 2, which were selected to record significant properties of vegetation. Band 1 (red, or R) records the absorption of red wavelengths by chlorophyll; lower values indicate higher chlorophyll content. Band 2 (reflected IR, or RIR) records the

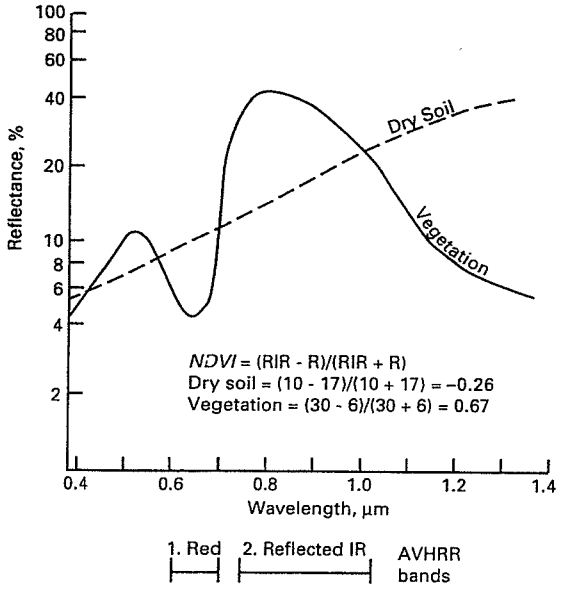


Figure 12-11 Calculation of *NDVI* for vegetation and soil from AVHRR bands 1 and 2.

reflection of IR wavelengths by the cell structure of leaves; higher values indicate more vigorous growth. These bands may be combined in various mathematical formulas to produce vegetation indexes. Richardson and Everitt (1992) describe the eight more commonly used indexes. By far the most widely employed version is the *normalized difference vegetation index* (*NDVI*), which is defined as

$$NDVI = \frac{(RIR - R)}{(RIR + R)} \qquad (12\text{-}1)$$

Values for *NDVI* range from 1.0 to −1.0. Higher values indicate higher concentrations of green vegetation. Lower values indicate nonvegetated features, such as water, barren land, ice, snow, or clouds. For the vegetation spectrum in Figure 12-11, the *NDVI* is calculated as 0.67. For the dry soil spectrum, the *NDVI* is only −0.26. The *NDVI* is also useful because it largely compensates for differences in solar illumination.

The global image in Plate 18B shows concentrations of vegetation for the continents based on *NDVI* values. Dark-green signatures represent the highest values and orange the lowest. High concentrations of vegetation are shown in the equatorial regions of South America, Africa, and Southeast Asia. Vegetation patterns in the United States are clearly shown.

Vegetation Maps Using AVHRR

Until recently, AVHRR data at the full spatial resolution of 1.1 km were not readily available. Earlier studies used data that were resampled as Global Area Coverage (GAC) with 4-km pixels or as Global Vegetation Index (GVI) data with 16-km pixels. Tucker, Townshend, and Goff (1985) derived the *NDVI* from GAC data to map major vegetation types and seasonal changes for Africa over a 19-month period in 1982 and 1983. Townshend, Justice, and Kalb (1987) used GAC and GVI data of South America to evaluate different approaches for mapping land cover. Goward, Tucker, and Dye (1985) derived the *NDVI* from GVI data of North America at 3-week intervals from April through November 1982. Seasonal *NDVI* patterns were associated with major land-cover regions, and multidate images portray patterns of vegetation growth and senescence. Lloyd (1990) mapped worldwide vegetation cover by a supervised classification of multidate GVI data.

A few early analyses employed 1.1-km AVHRR data. Tucker, Gatlin, and Schneider (1984) used 1.1-km data of the Nile Delta acquired from May to October 1981. They noted changes in greenness that corresponded to known vegetation cycles and agricultural practices. Gervin and others (1985) compared 1.1-km data for the Washington region with Landsat MSS data. Rather than calculating the *NDVI*, they performed an unsupervised classification of level I land use using AVHRR bands 1 through 4 for a single image acquired in July 1981. The results were compared with the MSS classification. Overall accuracy was 72 percent for AVHRR and 77 percent for MSS.

Eidenshink (1992A) of the U.S. Geological Survey published an AVHRR mosaic of North America for the period August 11 to 20, 1990. Each 1.1-km pixel is shown with a color code for the highest *NDVI* value recorded during the 10-day period. This maximum value represents the peak of vegetation "greenness," which is a measure of photosynthetic activity. The mosaic is a graphic portrayal of vegetation vigor for that period on a continent-wide scale. Zhu and Evans (1994) of the U.S. Department of Agriculture Forest Service used 1.1-km *NDVI* data to classify the forests of the United States (including Alaska and Hawaii) into 25 categories. They also used this information to estimate the percentage of forest cover for the conterminous United States.

Biweekly *NDVI* Maps

Eidenshink (1992B) and associates at the EROS Data Center (EDC) compiled 19 biweekly *NDVI* images of the conterminous United States for the 1990 growing season (March 16 to December 20). Plate 24 shows 15 of these images that were selected to illustrate the seasonal variation of vegetation cover. These images, plus images for the period 1991 to 1995 and

supporting information, are available on a CD-ROM from the EDC. The following section describes how Eidenshink (1992B) prepared the images.

Digital Processing The AVHRR data were resampled to 1.0-km pixels, which were processed at the EDC with the *Land Analysis System* (LAS) software described by Ailts and others (1990). Each biweekly *NDVI* image was produced by the following procedure.

1. **Scene selection** For each biweekly period all images without major cloud cover were selected, which typically amounted to 20 images. A single composite image was generated for each biweekly period, using the steps outlined below.
2. **Correcting for atmospheric scattering** The angular field of view for the AVHRR is 56° on either side of nadir. Toward either side of the image, the path length is much greater than at the nadir (directly beneath the satellite) and these longer paths are more severely affected by atmospheric scattering. In order to correct for atmospheric scattering, the relationship between solar illumination and satellite viewing geometry must be determined. These relationships are calculated from satellite orbital characteristics and used to correct for atmospheric scattering.
3. **Radiometric calibration** Performance of the AVHRR sensors is known to have degraded after launch because of exposure to the space environment. Data from bands 1 and 2 are calibrated using measurements of desert targets.
4. **Geometric registration** In order to produce a composite biweekly image the daily data sets (step 1) are registered to a common map projection; this ensures that each pixel is referenced to the correct ground location. A master reference image was prepared with a geographic error of less than 1.0 pixel. Automated correlation techniques are used to register all images to the reference image.
5. **Calculation of *NDVI*** For each daily mosaic, equation 12-1 was used to calculate the *NDVI* for each pixel.
6. **Image compositing** At this stage there are approximately 20 registered daily mosaics for each biweekly period. In other words, there are 20 *NDVI* values for each pixel in a biweekly period. For each pixel the maximum *NDVI* value is selected, which represents the peak greenness for the period.
7. **Products** For each biweekly period an image is produced that shows the maximum *NDVI* for each pixel. Plate 24 shows 15 of the images that are greatly reduced from the original scale (1:5,000,000). Maximum *NDVI* is shown in dark green; intermediate is yellow; minimum is brown and red. Water is light blue. Clouds are white. Statistical tables (not shown) are produced for each period and include a summary of the mean *NDVI* for each county in the United States.

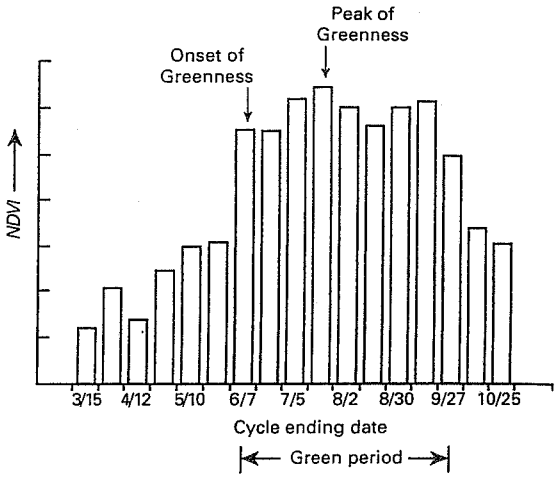


Figure 12-12 Derivation of greenness attributes from *NDVI* data of biweekly cycles. Total *NDVI* is the cumulative *NDVI* for the green period. From Loveland and others (1991, Figure 5).

The images in Plate 24 graphically show seasonal changes of vegetation patterns in the United States. In the spring of 1990 (Plate 24B) greening began in the Gulf and Pacific Coasts and expanded progressively inland. In mid-July and late August (Plate 24I,K) the northeastern United States was notably greener than the southeast portion, probably because of rainfall patterns. In the fall season, greenness diminished in a pattern that is essentially the reverse of the spring greening.

Wade and others (1994) used similar methods to prepare 12 biweekly *NDVI* maps of the USA for the growing seasons in both 1992 and 1993. In 1993 crops were affected by floods in the northeast USA and drought in the southeast. The year 1992 was normal. The two years were compared by computing difference images (Chapter 8). For each biweekly period the 1992 *NDVI* values were subtracted from the 1993 values on a pixel-by-pixel basis. The resulting difference maps graphically show the extent and severity of floods and droughts on crops.

Greenness Attributes from Biweekly NDVI Data Further processing of biweekly *NDVI* data derives additional vegetation information, called *greenness attributes*. Figure 12-12 is a histogram of the mean *NDVI* values for deciduous forests, plotted for each biweekly interval during the 1990 season. The following greenness attributes were derived from such plots and displayed as maps (U.S. Government Printing Office Document: 1993-556-415):

1. **Onset of greenness** The date when the *NDVI* first exceeds a threshold value, which occurs in early June for this exam-

and recalling that $R = L_\uparrow/E$, we see that for a perfect Lambertian surface

$$R = 1/\pi \tag{3.5}$$

We have said that the specular and Lambertian scatterers represent idealised, extreme forms of behaviour which are seldom realised in practice. However, it may often happen that the scattering from a real surface bears enough similarity to the ideal case for it to be described as *quasi-specular* or *quasi-Lambertian*.

The behaviour of real scattering surfaces is often specified, not by using the BRDF, but instead by measuring the *bidirectional reflectance factor* (BRF). This is defined as the ratio of the flux scattered into a given direction by a surface under given conditions of illumination to the flux scattered in the same direction by a perfect Lambertian scatterer under identical conditions. The utility of this function is that surfaces can be manufactured which have a BRF very close to unity for a fairly wide range of wavelengths and of incidence and scattering angles. The most common materials are barium sulphate which, as a pressed powder and for θ less than 45°, has a BRF greater than 0.99 for wavelengths between 0.37 μm and 1.15 μm, and magnesium oxide, which has BRF greater than 0.98 over roughly the same range of conditions.

3.2.2 The Rayleigh criterion

We have distinguished between the behaviour of a perfectly smooth surface, and a Lambertian surface which is in one sense perfectly rough. It is clear that in order to assess which of these forms of behaviour provides the better model of a real surface, some measure of roughness must be adopted. That which is usually adopted is the *Rayleigh criterion.*

Consider the diagram of fig. 3.3, in which radiation is incident on and reflected from a surface irregularity of height Δh, at an angle θ_0. It is clear that the path difference between the scattered ray and a ray which is reflected at the same angle from a height $\Delta h = 0$ is $2\Delta h \cos\theta_0$, and thus the phase difference $\Delta\phi$ is given by

$$\Delta\phi = 4\pi\Delta h \cos\theta_0/\lambda,$$

where λ is the wavelength. A surface can be defined as smooth enough for scattering to be specular if $\Delta\phi$ is less than some arbitrarily defined value of the order of 1 radian. The conventional value is $\pi/2$; this is called the Rayleigh criterion. Thus for a surface to be smooth according to this criterion,

$$\Delta h \cos\theta_0/\lambda < 1/8 \tag{3.6}$$

Note that other criteria, such as $\pi/8$, have also been adopted for the value of $\Delta\phi$ at which a surface becomes effectively smooth. A common definition which provides for the possibility of some intermediate condition between rough and smooth is that if $\Delta\phi$ is greater than $\pi/2$ the surface is rough, and if $\Delta\phi$ is less than $4\pi/25$, it is smooth.

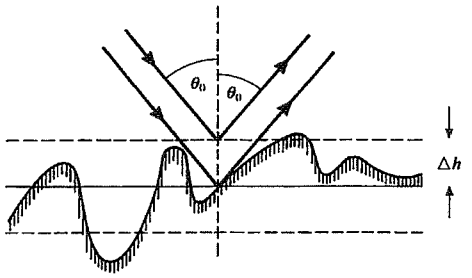
Equation (3.6) evidently dictates that for a surface to be effectively smooth at normal incidence, any irregularities must be less than about $\lambda/8$ in height. Thus for a surface to give specular reflexion at optical wavelengths ($\lambda = 0.5\ \mu$m, say), Δh must be less than about 60 nanometres. This is a condition of smoothness likely to be met only in certain man-made surfaces such as sheets of glass and metal. On the other hand, if the surface is to be examined using VHF radio waves (say $\lambda = 3$ m), Δh need only be less than about 0.4 metres, a condition which could be met by a number of naturally occurring surfaces.

A further aspect of (3.6) is that the restriction on Δh for a surface to reflect effectively specularly becomes less severe as the incidence angle θ_0 is increased. Thus at glancing incidence (large θ_0) a surface may appear quite smooth, whereas at $\theta_0 = 0$ it appears rough. This fact is familiar to anyone who has had to endure the glare of reflected sunlight from a low sun over an ordinary road surface. Although the scattering is by no means specular, the component of the BRDF in the specular direction is greatly enhanced.

3.2.3 Intermediate cases

We have derived the BRDF for the simple limiting cases of perfectly smooth and perfectly rough surfaces but, as we remarked before, these are

Fig. 3.3. The Rayleigh criterion. Radiation is specularly reflected at an angle θ_0 from a surface whose r.m.s. height deviation is Δh. The difference in the lengths of the two rays is $2\Delta h \cos\theta_0$.



Wind Speed, Direction, and Backscatter

So how is this roughness related to the wind speed? Unfortunately, there is no theory relating the wind speed to the size of these capillary waves and the backscatter, so the

wind speed dependence of sigma naught is determined empirically. There are various algorithms depending on the scatterometer frequency.

It should be noted that the wind speed that gives rise to the small scale roughness is not the same as the wind which is usually measured at say ten meters above the surface. While these two speeds are connected, the stability of the atmosphere determines how they are coupled. The stability depends on the vertical distribution of atmospheric properties (namely temperature). Therefore, some assumptions about the relationship between the 10-meter wind speed and the surface wind (called the friction velocity) are generally made so that the wind speed can be determined.

Since the capillary waves align themselves roughly perpendicular to the wind, the backscatter due to Bragg scattering will depend on the direction of the radiation (look direction) relative to the wind speed (figure 3)

$$\lambda_s = \frac{\lambda_R \sin\phi}{2 \sin\theta}. \qquad (10.4)$$

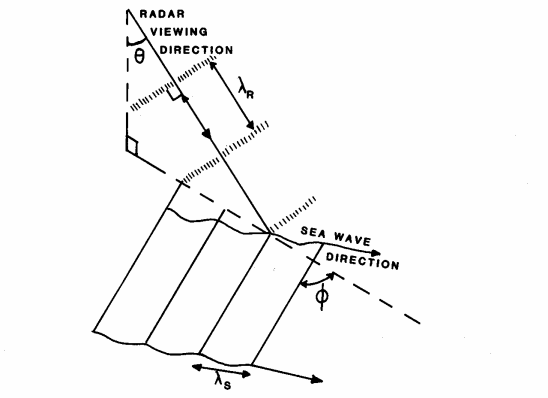


Fig 10.3 – Bragg scattering geometry: surface-wave direction at angle ϕ to plane of radar waves

Figure 3. Dependence of scattering on wind direction. For small angles, the wave crests align to promote Bragg scattering. If the angle is large, the crests are mostly parallel to the look direction, and so there is no series of crests to cause Bragg scattering (from Robinson, 1985)

The backscatter will be highest parallel to the wind direction and lowest perpendicular to the wind direction (since there would be no series of wave crests for resonant scattering in this direction). This means that if we look at a spot from all angles relative to the wind direction, as the azimuth changes from 0 degrees to 360 degrees the backscatter will peak initially (looking upwind, wave crests are all aligned with look direction), reduce to a minimum at 90 degrees (wave crests run parallel to look direction), and then increase to a maximum again at 180 degrees (looking downwind), to another minimum at 270 degrees, then maximum again at 360 (back to upwind). This pattern is depicted by the curve W1 in the figure 4.

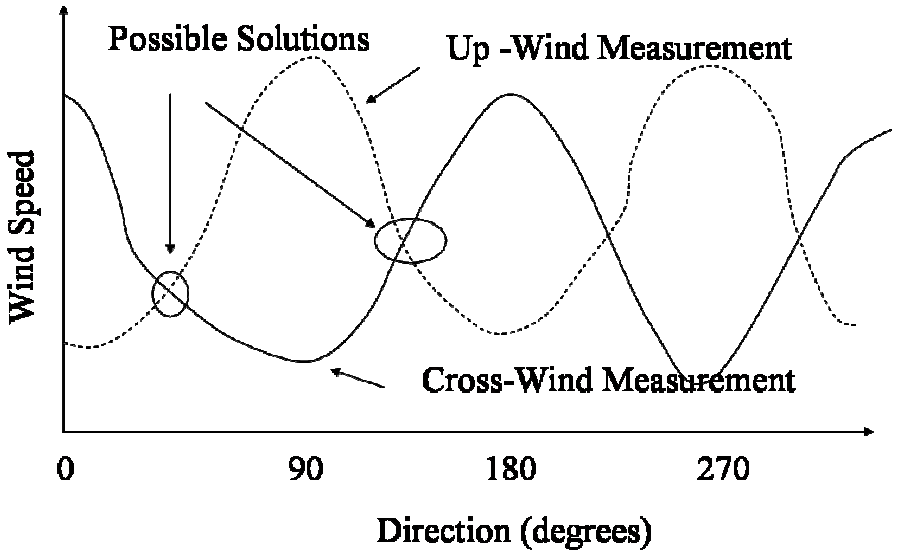


Figure 4. Solution curves for wind speed and direction for a measured backscatter. The two curves show two measurements made of the same point in two different directions. Since the backscatter varies with direction, the measured backscatter will be different. As the solution must satisfy both measurements, the only possible solutions are where the curves intersect.

From this behavior, we see there is an *ambiguity in the wind direction* determined from a single look direction. Notice that for a given wind speed, a given backscatter could occur for four different look angles. But we also have an *ambiguous wind speed* since if we do not know the direction, we don't know if a low backscatter is because we are looking crosswind, or if the wind is actually weak and we are looking downwind.

To remove the ambiguity, we look at the surface at two different look angles. For an observed value of sigma naught, there will be one possible wind speed for every possible wind direction, given by the curve in figure 4. Now, if we have a second measurement at a different angle relative to the wind direction, we must have a different curve (since the backscatter will change as the look angle changes). These two curves will intersect at four points. Since the wind speed and direction must satisfy both curves, the possible values are given only by the intersection point. Note that the wind speed does not vary much among intersection points, but it does for wind direction. This defines the wind speed and direction, with an ambiguity of four possible values for the wind direction. With a third look, the ambiguity can be further reduced.

There will always be an wind direction ambiguity of 180 degrees, since the sea surface looks the same whether we are looking upwind or crosswind. This ambiguity can be removed by comparing to other meteorological data or a numerical model.

7

Infrared (IR) measurement of sea surface temperature (SST)

This chapter considers another class of passive radiometers, those operating in the thermal-IR part of the spectrum, which can measure the temperature of the sea surface. IR sensors have been deployed for over 25 years on operational meteorological satellites to provide routine images of cloud top temperatures, and when there is no cloud they observe SST patterns. This was the first of the methods of remote sensing to gain widespread acceptance by the oceanographic community. The chapter examines the different factors that must be taken into consideration if SST is to be measured precisely with a quantifiable accuracy. It starts with the fundamentals of IR physics, explains the various processes required to unravel ocean information from a signal measured at the top of the atmosphere, and then explores the thermal behaviour of the ocean surface as it affects the oceanographic interpretation of satellite data. Having outlined the generic issues for measuring SST from space, there is then a description of the IR systems presently in operation, and the global SST data products that they deliver. Finally, mention is made of the types of oceanographic phenomena that have a thermal signature which can be observed from space, although it is left to *Understanding the Ocean from Space* (Robinson, 2005) to present examples of how satellite SST data are being used to study mesoscale and large scale ocean processes and global change.

7.1 PHYSICS OF IR RADIOMETRY

7.1.1 Thermal emission

The fundamental basis of IR remote sensing of SST is that all surfaces emit radiation, the strength of which depends on the surface temperature. The higher

An explanation of all abbreviated space agencies, satellites, and sensors can be found on p. xxxi of the front matter.

the temperature, the greater is the radiant energy. By measuring the emitted radiation, the temperature can be calculated, provided the physics of the process is well defined.

The spectral characteristics of thermal emission from a body at temperature T in K are described by Planck's radiation law:

$$M(\lambda, T) = \frac{C_1}{\lambda^5[\exp(C_2/\lambda T) - 1]} \tag{7.1}$$

where λ is the wavelength in metres; and M is the spectral exitance, sometimes called emittance (i.e., the radiant flux density of radiation per unit bandwidth centred at λ leaving unit area of surface, irrespective of direction (the equivalent of irradiance discussed in Section 6.2 but for energy leaving rather than falling on a surface)). C_1 and C_2 are constants with the values:

$$C_1 = 3.74 \times 10^{-16}\,\mathrm{W\,m^2}$$

$$C_2 = 1.44 \times 10^{-2}\,\mathrm{m\,K}$$

giving an estimate of M_λ in $\mathrm{W\,m^{-2}\,m^{-1}}$. To obtain M_λ in $\mathrm{W\,m^{-2}\,\mu m^{-1}}$ as is more customary, Equation (7.1) should be multiplied by 10^{-6}.

Integration of Equation (7.1) over all wavelengths gives the total exitance of a black body:

$$M = \sigma T^4 \tag{7.2}$$

where $\sigma = 5.669 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ (Stefan's constant).

Equation (7.1) represents ideal or black body radiation because it is based on ideal thermodynamic principles which apply only if the surface is a perfect emitter. The emitting properties of a real surface are described by its spectral emissivity, $\varepsilon(\lambda)$

$$\varepsilon(\lambda) = \frac{\text{Exitance at wavelength } \lambda \text{ from real surface at temperature } T}{M_\lambda(\text{perfect emitter at temperature } T)} \tag{7.3}$$

The logarithmic form of Equation (7.1) across a wide range of wavelengths has already been illustrated in Figure 2.3, for several temperatures, including those of the Sun and the Earth. Figure 7.1 is a linear plot of the Planck function in the IR part of the spectrum, for temperatures spanning the range of SST. Note that the area under each curve is given by Equation (7.2), while the spectral peak for each temperature is found at wavelength λ_{max} given by Wien's displacement law:

$$\lambda_{max} T = C_3 \tag{7.4}$$

where $C_3 = 2897\,\mu\mathrm{m\,K^{-1}}$.

The consequences of the spectral dependence of the Planck function were already discussed briefly in Chapter 2 where it was noted that solar emitted energy has a peak in the visible part of the spectrum, because the sun is so hot. In contrast, at typical SSTs the emission peak lies between about 9 μm and 11 μm. This makes the thermal-IR an optimal region for monitoring SST since the emitted radiance is a maximum there and it varies rapidly with temperature changes, almost doubling between 0 and 40°C. By measuring the radiance, an IR radiometer can detect the

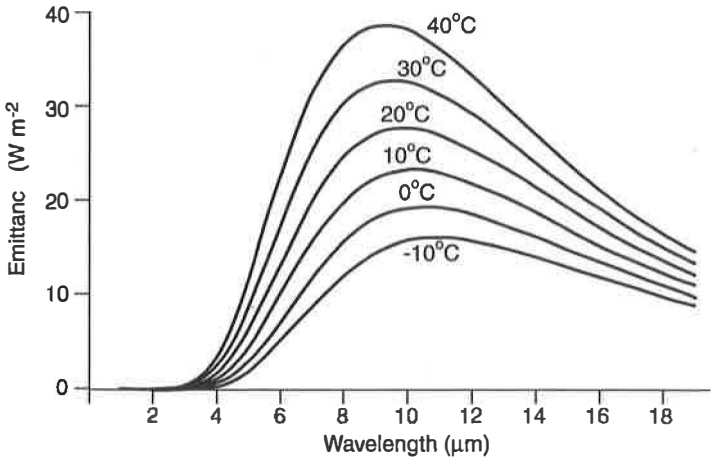


Figure 7.1. IR emission spectra of black bodies at temperatures between −10°C and 40°C.

brightness temperature of the radiation, defined as the temperature of the black body which would emit the measured radiance across the waveband of the detector.

In principle this can be achieved by inverting Equation (7.1), and recalling that for a Lambertian plane surface, the measured radiance L_λ is related to the exitance M_λ by $L_\lambda = M_\lambda/\pi$. In practice a simple inversion is not possible since the bandwidth of a detector is finite and its spectral response is tapered (see Section 7 .2.2). The measured brightness temperature differs from the actual temperature of the observed surface because of its non-unit emissivity, and because of the effects of the interven-ing atmosphere. The latter restricts IR radiometry of the sea surface to two spectral windows in the approximate ranges 3.5–4.1 μm and 10.0–12.5 μm.

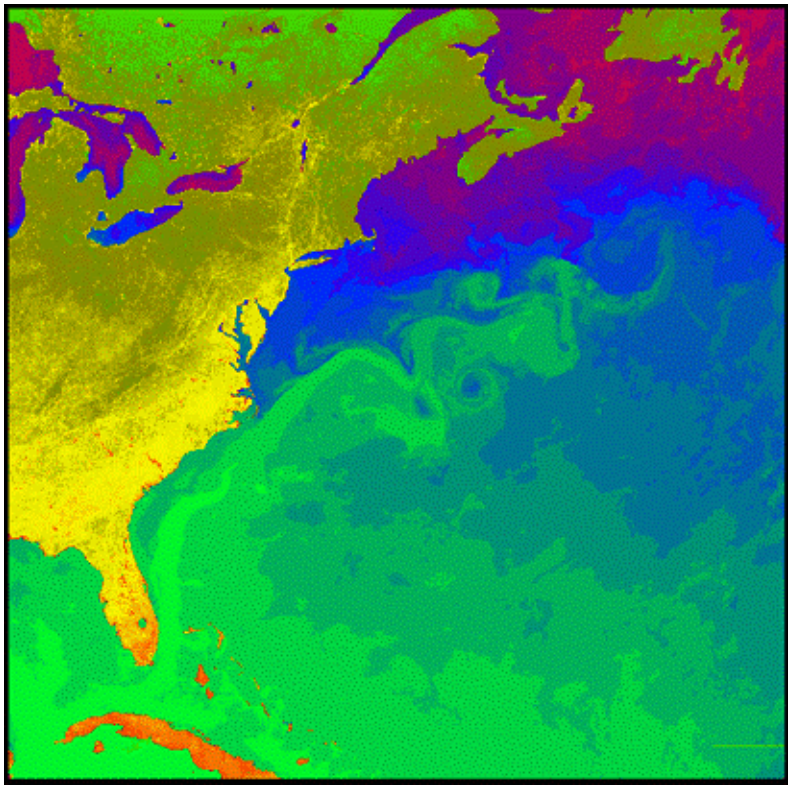
Gulf Stream Temperatures

Ocean currents such as the Gulf Stream are responsible for moving excess heat gained in the tropics to the poles, thus maintaining the Earth's thermal equilibrium. On average, the atmosphere and the ocean are equal partners in the amount of heat they transfer poleward. Sea-surface temperatures are used to determine how much heat is transferred between the atmosphere and the ocean.

The temperature of the ocean also determines how much carbon dioxide can be absorbed from the atmosphere. Knowing how much i absorbed is important because carbon dioxide is one of the major greenhouse gases that may be responsible for global warming.

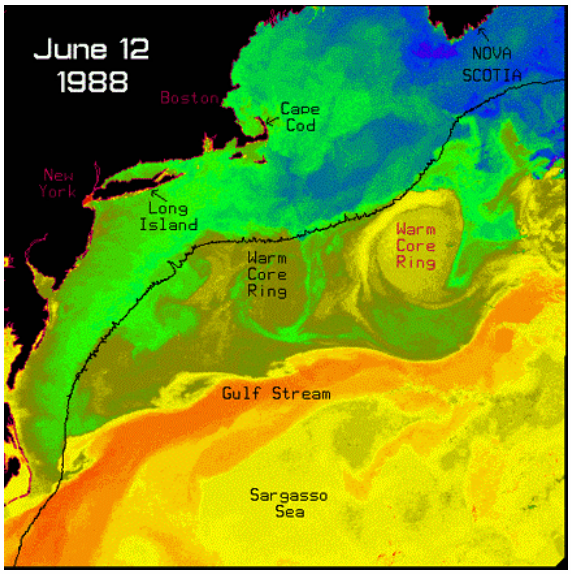
This thermal infrared image of the northwest Atlantic Ocean was taken from a NOAA satellite using the Advanced Very High Resolution Radiometer (AVHRR) instrument, which is sensitive to changes in the temperature of the ocean' s urface. The warmest temperatures (25 degrees C) are represented by red tones, and the coldest temperatures (2 degrees C) by blue and purple tones. The Gulf Stream is clearly visible as a core of warm water moving along the east coast of the United States and turning eastward into the Atlantic near Cape Hatteras, North Carolina.

As the Gulf Stream moves toward the central Atlantic, it releases heat to the atmosphere, so that by the time the Stream reaches the central Atlantic, it has lost its warm core, and its surface waters are no longer distinguishable from the surrounding waters. The area just east of Cape Hatteras sees the largest sustained transfer of heat from the ocean to the atmosphere. Because of this large heat transfer, atmospheric storms tend to intensify in this region.



Here is an image of the North Atlantic from June of 1984. Blue and purple represent cold water while green represents warmer water. Just as Canada is cooler than Florida, water off the coast of Canada is cooler than water off the coast of Florida. This north-south gradient is seen clearly in the image. This image also shows a current that brings warm water from the south up to the north. This current is called the Gulf Stream; it moves north along the coast of Florida and then turns eastward off of North Carolina flowing to the north-east across the Atlantic. It is one of the strongest currents in the ocean with an average velocity of 1 m/s (3 ft/s).

This image also shows that the Gulf Stream does not follow a straight path. It has many twists and turns called meanders. A meander is charecterised by its wavelength (the distance along the stream from one wave crest to the next), and its amplitude (the distance perpendicular to the stream between the wave crest and trough). If a meander becomes really sharp, it may pinch off and form what is called a ring. This is much like the formation of an oxbow lake by a river. Rings can be formed either to the north or to the south of the stream. For those rings formed to the north, the water in the center of the ring is warmer than the surrounding water and thus such rings are called warm core rings. For those rings formed to the south of the stream, the center contains water that is cooler than the surrounding water and they are called cold core rings. In this image you can see one warm core ring and two cold core rings. If you are unsure what a ring looks like, look at the next image from a different time in which the rings are labeled.



This image is a close up of part of the Gulf Stream. The rest of the images will cover this same region and will be color-coded in the same way. In this image the core of the Gulf Stream ranges between 25 and 28 deg C (77 and 82 F). The yellow water below the stream is about 23 deg C (73 F) and the green water off Long Island is about 14 deg C (57 F). The blue water around Nova Scotia is about 5 deg C (41 F)! The black line shows where the ocean is 1000m deep, (water shoreward of this line is less than 1000 meters deep and water seaward of this line is more than 1000 meters deep). This is usualy taken to be the edge of the continental shelf.

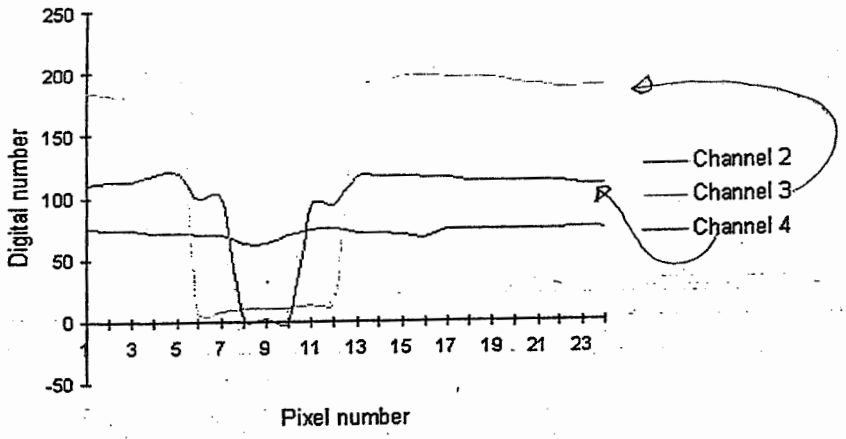
In most of the images to follow we will see a problem with collecting data from a satellite. A thick cloud blocks all the the radiation and we get no data (shown on these images as white). But perhaps more confusing, a thin cloud only blocks part of the radiation, making the water appear cooler than it actually is. Since there are usually clouds in the sky, we usually lose some data. One way of dealing with this problem is to combine 2 images taken close together in time. Since clouds always make the water appear cooler, a third image is made using the warmer of the two values at each point. If the two images contain clouds at different spots, the third image will contain few clouds. More than two images may be composited in the same fashion.

Vulcano Krafla (Islanda)
Immagini AVHRR

Case Study - Classification

Despite their poor quality however, the Channel 3 images were most useful in detecting the eruption. Channel 3 has been shown in several studies to be extremely sensitive to high temperature sources. The DN values for the eruption areas are much lower than those in the surrounding area and the contrast with the surrounding land much higher than in channels 4 and 5.

Transects were also taken across the middle of the lava flow on the images for channels 2, 3 and 4 to obtain DN profiles, as shown on the figure below. This figure shows how the profiles provide a very clear way of detecting lava. In particular, notice the large variation in the DNs of channel 3.



Before classification, the data were geometrically rectified and then a simple classification procedure was applied manually to determine areas of fresh lava cover. This involved a two stage process of training and classification.

Training

In training, areas of known cover type are identified. In this case, training areas were chosen to define land areas not contaminated by the signal from lava, water or cloud. The maximum and minimum pixel brightness values of these areas were used to define the upper and lower limits of a lava-free land *feature space*. A feature space is a range of digital number (DN) values which are characteristic of a land cover type on an image.

Classification

DN values below the minimum value for the feature space were classified as land covered by lava and DN values above the maximum value for the feature space were classified as areas of cloud or snow and ice.

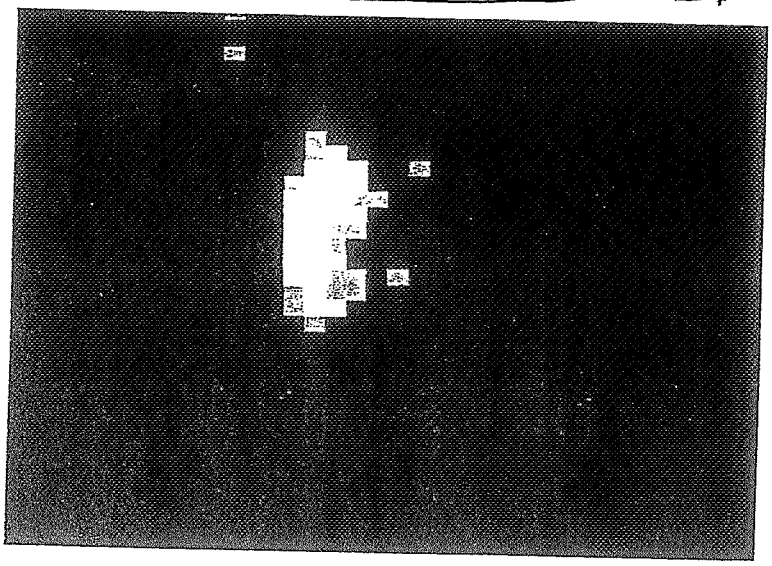


Figure 5. Contrast stretched channel 2 night-time image acquired on 5 September at 04:47 am, the Krafla area is at maximum magnification showing a surface area of around 100 km in width (east to west). The lava flow field is a source of sufficiently intense heat to radiate at these short wavelengths, creating a 'glow' from an area of bright pixels against a uniformly dark background. Occasional light pixels in the background are due to noise.

thermal radiance has been detected at similar and shorter wavelengths on remotely sensed data only at much higher spatial resolution. For example, the November 1979 lava flow from Sierra Negra, Galapagos Islands, was radiant in bands 5, 6 and 7 (which cover the 0·6 to 1·1 μm range) of the Landsat MSS (Rothery *et al.* 1988), and a 700 m length of the 1992 Etna lava flow was radiant in band 4 (0·76 to 0·90 μm) of the Landsat-TM (Rothery *et al.* 1992).

4.4. *Channel 1 (0·58–0·68 μm)*

The Krafla lava flow field was not detectable on any channel 1 image. By day this channel was swamped by reflected radiation, and by night the source was not sufficiently radiant in this channel to be detectable.

5. Estimation of the size of the lava flow field using AVHRR thermal data

Dimensions of the lava flow field defined by the threshold in channels 4 and 5 varied from image to image, varying between 3 and 7 pixels east to west and 11 and 16 pixels north to south. The ground area represented by an AVHRR pixel increases with distance from nadir, increasing from an area of 1·1 km at nadir to a maximum of 2·5 km by 5·6 km off-nadir (Holben and Fraser, figure 1, 1984). Taking these distortions into account the dimensions of the lava flow field were calculated. Using images where Krafla was close to nadir, size estimates approaching those given by air-photo mapping (Smithsonian Institution 1989) were obtained. On two images Krafla was within 80 pixels of nadir, on these images we estimate lava dimensions as

17·6 km by 3·3 km. This compares well with air photo mapping which gave maximum dimensions of 15 km by 4 km.

However, lava flow field dimensions estimated using the AVHRR data are unlikely to be precise for, two reasons:

5.1. *The effect of mixed pixels*

It is extremely unlikely that the edge of the lava flow field will coincide with the edge of a pixel. Instead pixels are likely to overlap the edge of the lava flow field and will therefore be mixed land-lava pixels, which may contain only a small portion of fresh lava flow field. Radiance from the small fraction of the pixel occupied by hot lava will cause the whole pixel to be anomalous. Inclusion, or exclusion, of the whole of such mixed pixels by imposition of a threshold will result in over- or under-estimate of the dimensions.

5.2. *The effect of optical blurring*

Optical blurring will result in radiation being bled from high radiance pixels into adjacent lower radiance pixels, spreading the anomaly along and across scan lines. Oppenheimer (1989) and Rothery and Oppenheimer (1994) identified such sensor-related problems (streaking out of the anomaly along a scan line, and bleeding of the anomaly between scan lines) on a channel 3 image of Mount Erebus. Here the effect is most obvious in all channel 3 images, where the size of the anomaly (figure 4) was much larger than that detected in channels 4 and 5 (figure 3) and the lava area of 24 km^2 given by Tryggvason (1986), but spreading of the anomaly also affected all channel 4 and 5 images. This effect will cause pixels that do not contain areas of fresh lava to appear anomalously radiant, and will result in over-estimation of the size of the lava-flow field.

6. Estimation of lava temperatures using AVHRR thermal data

Spectral radiance measured by the sensor can be converted to a pixel brightness temperature using the inverse of Planck's formula. Since this temperature will be an integral of all the temperature sources which occupy the pixel, this will give a pixel-integrated temperature (Rothery *et al.* 1988). For surfaces where temperatures are fairly uniform over large areas, pixel-integrated temperatures approximate the true surface temperature. However, in areas of volcanic activity uniform temperatures are unlikely. Instead, temperatures will vary greatly over a small area, and a 1·1 km by 1·1 km pixel will contain a variety of temperature sources. The resulting pixel-integrated temperature will be meaningless since it will fail to identify the true variations in surface temperature.

An active lava flow will consist of hot, incandescent, molten material in cracks or open channels, surrounded by a chilled crust. In the simplest case the thermal surface occupying the pixel will be made up of two distinct components (Rothery *et al.* 1988, Oppenheimer *et al.* 1993, Flynn *et al.* 1993): a hot molten lava component at temperature T_h which occupies portion p of the pixel, and a cool crust component at temperature T_c which will occupy the remaining portion of the pixel $(1-p)$.

Using the dual-band method proposed by Dozier (1981) and Matson and Dozier (1981) the temperature and size of these two sub-pixel heat sources can be calculated. If any one of the three parameters T_h, T_c, or p is known then the method

allows the remaining two parameters to be calculated by graphical or numerical solution of the following simultaneous equations:

$$L_i = p\,L_i(T_h) + (1-p)\,L_i(T_c) \quad (1)$$

$$L_j = p\,L_j(T_h) + (1-p)\,L_j(T_c) \quad (2)$$

Where L_i and L_j are the at-satellite spectral radiances in channels i and j (if atmospheric attenuation can be ignored), p is the portion of the pixel occupied by the hot lava temperature source, $L_i(T_h)$ and $L_j(T_h)$ are the spectral radiances for the hot lava temperature source in channels i and j, and $L_i(T_c)$ and $L_j(T_c)$ are the spectral radiances for the cool crust temperature source in channels i and j.

Rothery *et al.* (1988) adapted this technique to estimate sub-pixel temperatures at several volcanoes using the short-wavelength infrared bands of the Landsat Thematic Mapper, and the potential of this method to monitor volcanic thermal activity has been exploited by several other studies, including Glaze *et al.* (1989), Oppenheimer (1991), Reddy *et al.* (1993), and Oppenheimer *et al.* (1993).

In this study, the corrupt DNs referred to previously prevented us from deriving meaningful spectral radiances in channel 3. Furthermore, the wavelengths of channels 4 and 5 are too similar to each other to be useful in the dual-band method. Therefore, in the example demonstrated here, we sought to solve the simultaneous equations (1 and 2) for AVHRR channels 2 and 4 using data from the 5 September 1984 night-time image. Radiant DN in channels 2 and 4 respectively were 44 and 32, leading, via the calibration formula given by Kidwell (1986), to spectral radiant intensities (spectral radiance expressed in units of wavenumber rather than wavelength) of 0·403 mW m^2 sr^{-1} cm^{-1} and 148 mW m^2 sr^{-1} cm^{-1} respectively. As there are three independent unknowns in these equations (T_h, T_c and p) it is necessary to assume a value for one of them in order to solve the other two. Following the method of Oppenheimer (1991) we elected to choose a realistic value for T_h. In the case of Krafla the sources of T_h were molten basaltic lava (Gronvold 1987) in fountains along the eruptive fissure, flowing in channels, and in cracks in a chilled crust (see figure 2). Cas and Wright (1988) estimate the typical eruption temperature of basalt to be in the range 1000 to 1200° C and Macdonald (1972) gives lava fountain temperatures for Hawaiian volcanoes of between 1050 and 1190° C. Multiple component modelling of spectroradiometer data for a basaltic lava lake (L. Flynn, personal communication) in Hawaii by Flynn *et al.* (1993) gave a hot component temperature of 1100° C in most cases. In this study we use a value of 1050° C for T_h as a reasonable starting point from which to estimate the values of T_c and p.

A peculiarity of attempting to employ widely-separated wavebands such as AVHRR channels 2 and 4 in the dual band method is that the peak sensitivities of each correspond to rather different temperature ranges. By attempting to solve these equations for $T_h = 1050$° C (or any comparable magmatic temperature) it becomes apparent that the model T_c must be too low for *any* thermal radiance to be emanating from the crust in detectable amounts in channel 2. The solution converges near a value of 35° C for T_c occupying 0·996 (99·6 per cent) of a pixel, with T_h (at 1050° C) occupying the remaining 0·004 (0·4 per cent) of the pixel. Since we have not attempted to correct for atmospheric attenuation of upwelling radiance all of these values must be regarded as approximations. More importantly, we regard the value of T_c as untrustworthy because it is likely to represent the integrated spectral radiance in channel 4 from a crust that probably has a range of temperatures, perhaps extending

up to a few hundreds of degrees Celsius, as calculated by Flynn *et al.* (1993) from field spectroradiometer data, in which case the true value of p must be less, and also from some areas of lava-free ground beyond the edge of the flow field, where, at Krafla, pixel-integrated temperatures spans a range from $-4°$ C to $+3°$ C.

All that these data really allow us to determine, is that the *maximum* amount of each pixel that could be occupied by lava at 1050° C is 0·4 per cent. If the crust includes a significant amount of surface area at above about 200° C then the 1050° C molten lava has to be confined to smaller fractions of each pixel.

Since the value of T_c is higher than that expected from a lava-free background but lower than typical crustal temperatures, T_c can be regarded as an integrated value for the cool portion of the pixel, occupied by at-least two temperature sources: (1) an area of lava flow with a chilled crust at temperature T_{c1}, and (2) an area of lava-free ground at temperature T_{c2}, where both T_{c1} and T_{c2} are much less than T_h (1050° C). If T_h occupies portion p of a pixel and T_{c2} occupies portion p_2 of the pixel, then T_{c1} will occupy the remaining portion $(1-p-p_2)$, making it really a three-component or multi-component surface, and invalidating the simple two-component approximation.

The existence of portions of lava-free ground (T_{c2}) in the pixel was confirmed by overlaying a map of the lava flow field produced by air-photo mapping (Smithsonian Institution, 1989) with a pixel grid. Pixel dimensions in the Krafla portion of the 5 September 1984 night-time image were 1·7 km (N to S) by 2·8 km (E to W). A pixel grid of these dimensions, even when optimally placed over the map, showed that all pixels would have contained a portion of lava-free land, and that the maximum portion of any pixel that could be expected to be occupied by lava is only around 0·75. Therefore, since T_{c2} will occupy at-least a 0·25 portion of any pixel, the contribution of T_{c2} to the integrated T_c value is likely to be significant.

Using the dual-band method an attempt was made to estimate the total area of lava flow field occupied by material at a temperature of 1050° C in all pixels. All pixels within the channel 2 thermal anomaly that were above noise levels were considered. The thermal anomaly in channel 2 covered 26 pixels. However, six of these pixels were excluded from this estimation since they were believed to be radiant to the bleeding of radiance from adjacent high radiance pixels. It was therefore unreasonable to assume that they contained a portion of lava flow field. The dual-band method was applied to each of the remaining 20 pixels of the channel 2 anomaly. This gave a total area covered by material at 1050° C of $2{\cdot}4 \times 10^5$ m^2. T_c values for all pixels varied between 33° C and 42° C. These T_c values are significantly less than those expected for a cool pixel portion occupied entirely by a chilled crust, i.e., a few hundred degrees Celcius. This indicates that all pixels within the anomaly contained a portion of lava-free ground, and that the contribution of the lava free land (at temperature T_{c2}) to the integrated T_c value was significant. This confirms the predictions made using the pixel grid over-lay.

7. Image sharpening to monitor the development of the eruption

To aid visual interpretation, an image sharpening method was devised and applied to produce a series of thematic maps for the flow field. This weighted average method is based on the assumption that the feature type at any particular point within a mixed pixel is most likely to be similar to the feature type occupying any pixels adjacent to that point. Application of the method is intended for aesthetic purposes only, improving image clarity and highlighting spatial variations in radiance across the flow-field, and the values produced cannot be used for any quantitative analysis.

AVI

Figure 6. The group
sub-pixel DNs
equally sized su
to 11 (see text).

Pixels were divid
calculated for each
their neighbouring p

SP
SP
SP
SP.
SP.

SP:
SP
SP
SP

Where $Px(A)$ to P
and $SPx(1)$ to $SPx(9)$
which occupies one-ni
applied to calculate ea
the four corner sub-pi
3, 5, 9 and 11; then the
using formulae 4, 6,
(figure 6) is calculated

Sub-pixel DNs we
thermal anomaly at K
each of the original di

3. Fire detection methodology

Matson and Dozier (1981), Matson *et al.* (1984) and Muirhead and Cracknell (1984, 1985) have demonstrated that the use of the 3·8 μm thermal infrared channel on board the NOAA polar-orbiting satellites provides the capability to detect high temperature sources such as steel plants, waste gas flares and fires. Figure 1 shows why this is so. For high temperature targets, the maximum blackbody radiance shifts away from the 11 μm channel and toward the 3·8 μm channel. Figure 2 shows a typical satellite-derived temperature plot over two high temperature sources located in Idaho. Target 1 is a small controlled forest burn and target 2 is a phosphorous plant operated by the FMC Corporation in Pocatello, Idaho. At these sources, the 3·8 μm temperatures are 16·2 deg K and 33·9 deg K higher than the corresponding 11 μm temperatures. Typical temperature differences between the two channels over land surfaces are usually 1–2 deg K. Using an algorithm developed by Matson and Dozier (1981), it is possible to use the temperature difference in the two channels to estimate the areal extent and temperature of the high temperature source causing the 3·8 μm response. For the two sources in figure 2, the area and temperature were 0·28 ha and 430 deg K for target 1, and 1·7 ha and 483 deg K for target 2. As evidenced by the calculated target sizes and the detection of the phosphorous plant, it does not take a 1·1 km target to cause a response in the 3·8 μm channel. Thus, small *subresolution* scale high temperature sources, such as fires, can be detected by the 3·8 μm channel. The size of the subresolution scale high temperature sources which can be detected depends on the target temperature and area. Although the actual lower limit of detectability is not

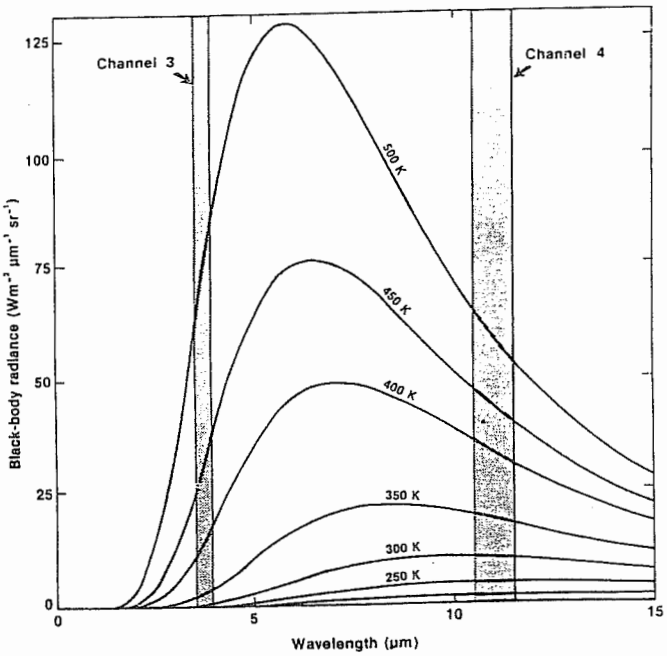


Figure 1. Planck radiances for temperatures from 200 K to 500 K. For a given increase in temperature the increase in area under the 3·8 μm segment (channel 3) of the curve is much greater than under the 11 μm segment (channel 4).

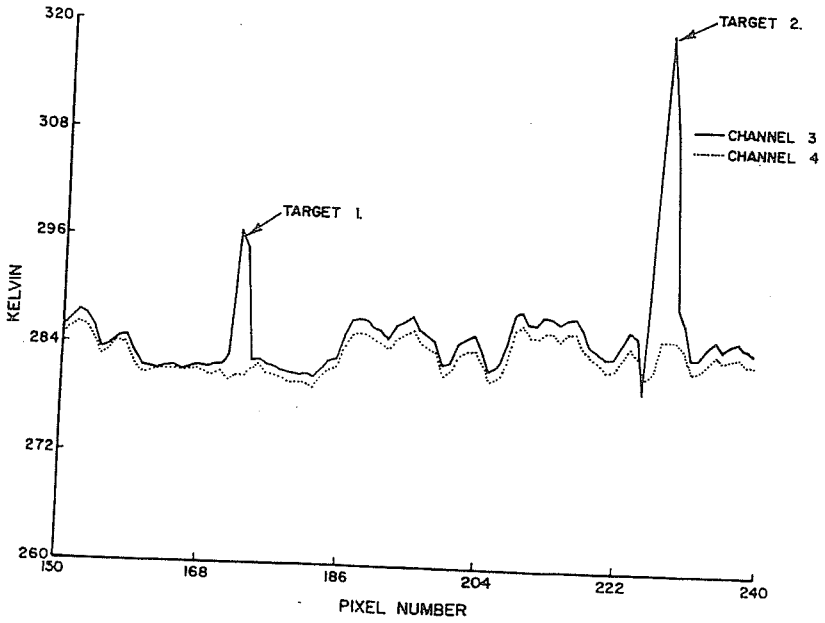


Figure 2. 3·8 μm (channel 3) and 11 μm (channel 4) brightness temperature plot of 'hot' targets 1 and 2.

known, very hot and small targets such as waste gas flares from offshore oil platforms have been detected (Matson and Dozier 1981, Muirhead and Cracknell 1984) as well as small fires from straw burning (Muirhead and Cracknell 1985).

4. Case Studies

Figure 3 is a NOAA-7 visible band image taken of southern Mexico and northern Guatemala on 18 April 1984. A number of large smoke plumes and smoke-covered areas are visible (light grey areas, labelled 'S'). Figure 4 is the 3·8 μm image of the same area. The well-defined white spots throughout the image are fires. Most of the fires seen here are associated with land clearing for agriculture. Scientists familiar with this area of Mexico have identified several distinct areas (Breedlove 1981) and the type of burning associated with each. The Grijalva Basin, the white area near the centre of the picture (labelled 'GB'), is a major agricultural centre; most of the burning here is for clearing of previously cultivated land. In contrast, the area around Veracruz, in the upper left (labelled 'V'), and the area around the Guatemala border (labelled 'G'), in the lower right, are primarily virgin forest, being cleared for agricultural use at great ecological cost. The reader should note how the 3·8 μm channel penetrates the smoke to reveal the underlying fire activity. The 3·8 μm channel radiative response is not significantly attenuated by water vapour. Smoke is largely composed of water vapour, thus the 3·8 μm channel 'sees through' the smoke.

Tucker *et al.* (1984) used NOAA-7 data taken on 9 July 1982 to identify a large (100 km by 400 km) area in Rondonia, Brazil, where massive deforestation is occurring. Figure 5 is a 3·8 μm image of this area taken on 18 August 1984. The image

Figure 3. Visible ba
Guatema

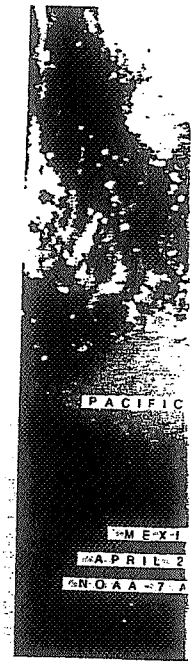


Figure 4. 3·8 μm AVH
1984. Areas labe
Guatemala portio

in the frequency domain generally involves computing the FFT of the original image, multiplying the FFT of a convolution mask of the analyst's choice (e.g., a low pass filter) with the FFT, and inverting the resultant image with the IFFT; that is,

$$f(x,y) \xrightarrow{\text{FFT}} F(u,v) \rightarrow F(u,v)G(u,v)$$
$$\rightarrow F'(u,v) \xrightarrow{\text{IFFT}} f'(x,y)$$

The convolution theorem states that the convolution of two images is equivalent to the multiplication of their Fourier transformations. If

$$f'(x, y) = f(x, y)^* g(x, y) \quad (7\text{-}60)$$

where * represents the operation of convolution, $f(x, y)$ is the original image and $g(x,y)$ is a convolution mask filter, then

$$F'(u, v) = F(u, v)G(u, v) \quad (7\text{-}61)$$

where F', F, and G are Fourier transforms of f', f, and g, respectively.

Two examples of such manipulation are shown in Figures 7-29 and 7-30. A low-pass filter (mask B) and a high-pass filter (mask D) were used to construct the filter function $g(x, y)$ in Figures 7-29 and 7-30, respectively. In practice, one problem must be solved. Usually, the dimensions of $f(x, y)$ and $g(x, y)$ are different; for example, the low-pass filter in Figure 7-29 only has nine elements, while the image is composed of 128×128 pixels. Operation in the frequency domain requires that the sizes of $F(u, v)$ and $G(u, v)$ be the same. This means the sizes of f and g must be made the same because the Fourier transform of an image has the same size as the original image. The solution of this problem is to construct $g(x, y)$ by putting the convolution mask at the center of a zero-value image that has the same size as f. Note that in the Fourier transforms of the two convolution masks the low pass convolution mask has a bright center (Figure 7-29), while the high-pass filter has a dark center (Figure 7-30). The multiplication of Fourier transforms $F(u, v)$ and $G(u, v)$ results in a new Fourier transform, $F'(u, v)$. Computing the inverse fast Fourier transformation yields $f'(x, y)$, a filtered version of the original image. Thus, spatial filtering can be performed both in the spatial and frequency domain.

As demonstrated, filtering in the frequency domain involves one multiplication and two transformations. For general applications, convolution in the spatial domain may be more cost effective. Only when the size of $g(x, y)$ is very large, does the Fourier method become cost effective. However, with the frequency domain method we can also do some filtering that is not easy to do in spatial domain. We may construct a frequency domain filter $G(u, v)$ specifically designed to remove certain frequency components in the image. Numerous articles describe how to construct frequency filters (Al-Hinai et al., 1991; Pan and Chang, 1992; Khan, 1992). Watson (1993) describes how the two-dimensional FFT may be applied to image mosaicking, enlargement, and registration.

Special Transformations

Principal Components Analysis

Principal components analysis (often called PCA, or Karhunen–Loeve analysis) has proved to be of value in the analysis of multispectral remotely sensed data (Press et al., 1992; Wang, 1993). The transformation of the raw remote sensor data using PCA can result in new principal component images that may be more interpretable than the original data (Singh and Harrison, 1985). PCA analysis may also be used to compress the information content of a number of bands of imagery (e.g., seven Thematic Mapper bands) into just two or three transformed principal component images. The ability to reduce the *dimensionality* (i.e., the number of bands in the dataset that must be analyzed to produce usable results) from n to two or three bands is an important economic consideration, especially if the potential information recoverable from the transformed data is just as good as the original remote sensor data. A form of PCA may also be useful for reducing the dimensionality of hyperspectral datasets. Satellite remote sensing datasets of the future may be hyperspectral, containing hundreds of bands (e.g., MODIS). For example, Lee et al. (1990) used a modified PCA transformation (i.e., the maximum noise fraction, or MNF) for data compression and noise reduction of 64-channel hyperspectral scanner data in Australia. Noise was removed from the multispectral data by transforming to the MNF space, smoothing or rejecting the most noisy components, and then retransforming to the original space.

To perform principal component analysis we apply a transformation to a *correlated* set of multispectral data. For example, the Charleston, S.C. TM scene is a likely candidate since bands 1, 2, and 3 are highly correlated, as are bands 5 and 7 (Table 7-5). The application of the transformation to the correlated remote sensor data will result in another *uncorrelated* multispectral dataset that has certain ordered variance properties (Singh and Harrison, 1985). This transformation is conceptualized by considering the two-dimensional distribution of pixel values obtained in two TM bands, which we

Table 7-5. Charleston, South Carolina Thematic Mapper Scene Statistics Used in the Principal Components Analysis

Band Number:	1	2	3	4	5	7	6
μm:	0.45–0.52	0.52–0.60	0.63–0.69	0.76–0.90	1.55–1.75	2.08–2.35	10.4–12.5
Univariate Statistics							
Mean	64.80	25.60	23.70	27.30	32.40	15.00	110.60
Standard Deviation	10.05	5.84	8.30	15.76	23.85	12.45	4.21
Variance	100.93	34.14	68.83	248.40	568.84	154.92	17.78
Minimum	51	17	14	4	0	0	90
Maximum	242	115	131	105	193	128	130
Variance–Covariance Matrix							
1	100.93						
2	56.60	34.14					
3	79.43	46.71	68.83				
4	61.49	40.68	69.59	248.40			
5	134.27	85.22	141.04	330.71	568.84		
7	90.13	55.14	86.91	148.50	280.97	154.92	
6	23.72	14.33	22.92	43.62	78.91	42.65	17.78
Correlation Matrix							
1	1.00						
2	0.96	1.00					
3	0.95	0.96	1.00				
4	0.39	0.44	0.53	1.00			
5	0.56	0.61	0.71	0.88	1.00		
7	0.72	0.76	0.84	0.76	0.95	1.00	
6	0.56	0.58	0.66	0.66	0.78	0.81	1.00

will label simply X_1 and X_2. A scatterplot of all the brightness values associated with each pixel in each band is shown in Figure 7-31a, along with the location of the respective means, μ_1 and μ_2. The spread or variance of the distribution of points is an indication of the correlation and quality of information associated with both bands. If all the data points clustered in an extremely tight zone in the two-dimensional space, these data would probably provide very little information.

The initial measurement coordinate axes (X_1 and X_2) may not be the best arrangement in multispectral feature space to analyze the remote sensor data associated with these two bands. The goal is to use principal components analysis to *translate* and/or *rotate* the original axes so that the original brightness values on axes X_1 and X_2 are redistributed (reprojected) onto a new set of axes or dimensions, X'_1 and X'_2 (Wang, 1993). For example, the best *translation* for the original data points from X_1 to X'_1 and from X_2 to X'_2 coordinate systems might be the simple relationship $X'_1 = X_1 - \mu_1$ and $X'_2 = X_2 - \mu_2$. Thus, the origin of the new coordinate system (X'_1 and X'_2) now lies at the location of both means in the original scatter of points (Figure 7-31b).

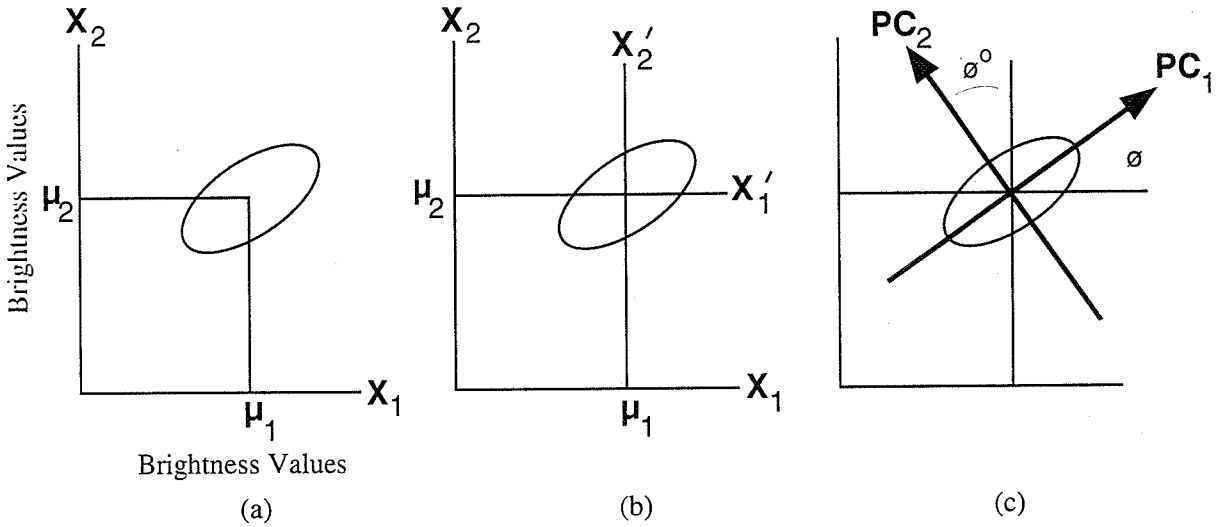


Figure 7-31 Diagrammatic representation of the spatial relationship between the first two principal components: (a) Scatterplot of data points collected from two remotely bands labeled X_1 and X_2 with the means of the distribution labeled μ_1 and μ_2. (b) A new coordinate system is created by shifting the axes to an X' system. The values for the new data points are found by the relationship $X_1' = X_1 - \mu_1$ and $X_2' = X_2 - \mu_2$. (c) The X' axis system is then rotated about its origin (μ_1, μ_2) so that PC_1 is projected through the semimajor axis of the distribution of points and the variance of PC_1 is a maximum. PC_2 must be perpendicular to PC_1. The PC axes are the principal components of this two-dimensional data space. Component 1 usually accounts for approximately 90% of the variance, with component 2 accounting for approximately 5%.

The X' coordinate system might then be *rotated* about its new origin (μ_1, μ_2) in the new coordinate system some ϕ degrees so that the first axis X'_1 is associated with the maximum amount of variance in the scatter of points (Figure 7-31c). This new axis is called the first *principal component* ($PC_1 = \lambda_1$). The second principal component ($PC_2 = \lambda_2$) is perpendicular (orthogonal) to PC_1. Thus, the major and minor axes of the ellipsoid of points in bands X_1 and X_2 are called the principal components. The third, fourth, fifth, and so on, components contain decreasing amounts of the variance found in the data set.

To transform (reproject) the original data on the X_1 and X_2 axes onto the PC_1 and PC_2 axes, we must obtain certain transformation coefficients that we can apply in a linear fashion to the original pixel values. The linear transformation required is derived from the covariance matrix of the original data set. Thus, this is a data-dependent process with each new data set yielding different transformation coefficients.

The transformation is computed from the original spectral statistics as follows (Short, 1982):

1. The $n \times n$ *covariance* matrix, Cov, of the n-dimensional remote sensing data set to be transformed is computed (Table 7-5). Use of the covariance matrix results in an unstandardized PCA, whereas use of the correlation matrix results in a standardized PCA (Eastman and Fulk, 1993).

2. The eigenvalues, $E = [\lambda_{1,1}, \lambda_{2,2}, \lambda_{3,3}, \ldots, \lambda_{n,n}]$, and eigenvectors $EV = [a_{kp} \ldots$ for $k = 1$ to n bands, and $p = 1$ to n components] of the covariance matrix are computed such that

$$EV \quad \text{Cov} \quad EV^T = E$$
$$[n \times n][n \times n][n \times n] = \begin{bmatrix} \lambda_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{4,4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{5,5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{6,6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{n,n} \end{bmatrix} \quad (7\text{-}62)$$

where EV^T is the transpose of the eigenvector matrix, EV, and E is a diagonal covariance matrix whose elements λ_{ii}, called *eigenvalues*, are the variances of the pth *principal components*, where $p = 1$ to n components. The nondiagonal eigenvalues, λ_{ij}, are equal to zero and therefore can be ignored. The number of nonzero eigenvalues in an $n \times n$ covariance matrix always equals n, the number of bands examined. The eigenvalues are often called components (i.e., eigenvalue 1 may be referred to as principal component 1).

Table 7-6. Eigenvalues Computed for the Covariance Matrix

	Component p						
	1	2	3	4	5	6	7
Eigenvalues, λ_p	1010.92	131.20	37.60	6.73	3.95	2.17	1.24
Difference	879.72	93.59	30.88	2.77	1.77	.93	--

Total Variance = 1193.81

Percent of total variance in the data explained by each component:

Computed as $\%_p = \dfrac{\text{eigenvalue } \lambda_p \times 100}{\sum_{p=1} \text{eigenvalue } \lambda_p}$

For example,

$$\sum_{p=1}^{7} \lambda_p = 1010.92 + 131.20 + 37.60 + 6.73 + 3.95 + 2.17 + 1.24 = 1193.81$$

Percentage of variance explained by first component $= \dfrac{1010.92 \times 100}{1193.81} = 84.68$

Percentage:	84.68	10.99	3.15	0.56	0.33	0.18	0.10
Cumulative:	84.68	95.67	98.82	99.38	99.71	99.89	99.99

Table 7-7. Eigenvectors (a_{kp}) (Factor Scores) Computed for the Covariance Matrix found in Table 7-4

	Component p						
	1	2	3	4	5	6	7
band$_k$ 1	0.205	0.637	0.327	−0.054	0.249	−0.611	−0.079
2	0.127	0.342	0.169	−0.077	0.012	0.396	0.821
3	0.204	0.428	0.159	−0.076	−0.075	0.649	−0.562
4	0.443	−0.471	0.739	0.107	−0.153	−0.019	−0.004
5	0.742	−0.177	−0.437	−0.300	0.370	0.007	0.011
7	0.376	0.197	−0.309	−0.312	−0.769	−0.181	0.051
6	0.106	0.033	−0.080	0.887	0.424	0.122	0.005

Eigenvalues and eigenvectors were computed for the Charleston, S.C., TM scene (Tables 7-6 and 7-7). Such computations can be performed using most statistical analysis packages, such as SAS or SPSS.

The eigenvalues contain important information. For example, it is possible to determine the percent of total variance explained by each of the principal components, $\%_p$ using the equation

$$\%_p = \frac{\text{eigenvalue } \lambda_p \times 100}{\sum_{p=1}^{n} \text{eigenvalue } \lambda_p} \tag{7-63}$$

Table 7-8. Degree of Correlation, R_{kp}, between Each Band k and Each Principal Component p

Computed as: $$R_{kp} = \frac{a_{kp} \times \sqrt{\lambda_p}}{\sqrt{\text{Var}_k}}$$

For example:

$$R_{1,1} = \frac{0.205 \times \sqrt{1010.92}}{\sqrt{100.93}} = \frac{0.205 \times 31.795}{10.046} = 0.649$$

$$R_{5,1} = \frac{0.742 \times \sqrt{1010.92}}{\sqrt{568.84}} = \frac{0.742 \times 31.795}{23.85} = 0.989$$

$$R_{2,2} = \frac{0.342 \times \sqrt{131.20}}{\sqrt{34.14}} = \frac{0.342 \times 11.45}{5.842} = 0.670$$

	Component p						
	1	2	3	4	5	6	7
Band$_k$ 1	0.649	0.726	0.199	–0.014	0.049	–0.089	–0.008
2	0.694	0.670	0.178	–0.034	0.004	0.099	0.157
3	0.785	0.592	0.118	–0.023	–0.018	0.115	–0.075
4	0.894	–0.342	0.287	0.017	–0.019	–0.002	–0.000
5	0.989	–0.084	–0.112	–0.032	0.030	0.000	0.000
7	0.961	0.181	–0.152	0.065	–0.122	–0.021	0.004
6	0.799	0.089	–0.116	0.545	0.200	0.042	0.001

where λ_p is the pth eigenvalue out of the possible n eigenvalues. For example, the first principal component (eigenvalue λ_1) of the Charleston TM scene accounts for 84.68% of the variance in the entire multispectral dataset (Table 7-6). Component 2 accounts for 10.99% of the remaining variance. Cumulatively, these first two principal components account for 95.67% of the variance. The third component accounts for another 3.15% bringing the total to 98.82% of the variance explained by the first three components (Table 7-6). Thus, the seven-band TM dataset of Charleston might be compressed into just three new principal component images (or bands) that explain 98.82% of the variance.

But what do these new components represent? For example, what does component 1 stand for? By computing the correlation of each band k with each component p, it is possible to determine how each band "loads" or is associated with each principal component. The equation is

$$R_{kp} = \frac{a_{kp} \times \sqrt{\lambda_p}}{\sqrt{\text{Var}_k}} \qquad (7\text{-}64)$$

where

a_{kp} = eigenvector for band k and component p
λ_p = pth eigenvalue
Var_k = variance of band k in the covariance matrix

This computation results in a new $n \times n$ matrix (Table 7-8) filled with *factor loadings*. For example, the highest correlations (i.e., factor loadings) for principal component 1 were for bands 4, 5, and 7 (0.894, 0.989, and 0.961, respectively, Table 7-8). This suggests that this component is a near- and middle-infrared reflectance band. This makes sense because the golf courses and other vegetation are particularly bright in this image. Conversely, principal component 2 has high loadings only in the visible bands 1, 2, and 3 (0.726, 0.670, and 0.592), and vegetation is noticeably dark in this image. This is a visible spectrum component. Component 3 loads heavily in the near-infrared (0.287) and appears to provide some unique vegetation information. Component 4 accounts for little of the variance but is easy to label since it loads heavily (0.545) on the thermal-infrared band 6. Components 5, 6, and 7 provide no useful information and con-

tain most of the systematic noise. They account for very little of the variance and should probably not be used further.

Now that we understand what information each component contributes, it is useful to see what the images of these components look like. To do this it is necessary to first identify the brightness values ($BV_{i,j,k}$) associated with a given pixel. In this case we will evaluate the first pixel in a hypothetical image at row 1, column 1 for each of seven bands. We will represent this as the vector X, such that

$$X = \begin{bmatrix} BV_{1,1,1} = 20 \\ BV_{1,1,2} = 30 \\ BV_{1,1,3} = 22 \\ BV_{1,1,4} = 60 \\ BV_{1,1,5} = 70 \\ BV_{1,1,7} = 62 \\ BV_{1,1,6} = 50 \end{bmatrix} \quad (7\text{-}65)$$

We will now apply the appropriate transformation to this data such that it is projected onto the first principal component's axes. In this way we will find out what the new brightness value (new $BV_{i,j,p}$) will be for this component, p. It is computed according to the formula

$$\text{new}BV_{i,j,p} = \sum_{k=1}^{n} a_{kp} BV_{i,j,k} \quad (7\text{-}66)$$

where a_{kp} = eigenvectors, BV_{ijk} = brightness value in band k for the pixel at row i, column j, and n = number of bands. In our hypothetical example, this yields

$$\begin{aligned} \text{new}BV_{1,1,1} &= a_{1,1}(BV_{1,1,1}) + a_{2,1}(BV_{1,1,2}) + a_{3,1}(BV_{1,1,3}) + a_{4,1}(BV_{1,1,4}) \\ &\quad + a_{5,1}(BV_{1,1,5}) + a_{6,1}(BV_{1,1,7}) + a_{7,1}(BV_{1,1,6}) \\ &= 0.205(20) + 0.127(30) + 0.204(22) + 0.443(60) \\ &\quad + 0.742(70) + 0.376(62) + 0.106(50) \\ &= 119.53 \end{aligned}$$

This pseudomeasurement is a linear combination of original brightness values and factor scores (eigenvectors). The new brightness value for row 1, column 1 in principal component 1 after truncation to an integer is new$BV_{1,1,1}$ = 119.

This procedure takes place for every pixel in the original image data to produce the principal component 1 image dataset. Then p is incremented by 1 and principal component 2 is created pixel by pixel. This is the method used to produce the principal component images shown in Figure 7-32. If desired, any two or three of the principal components can be placed in the blue, green, and/or red image planes to create a principal component color composite. These displays often depict more subtle differences in color shading and distribution than traditional color-infrared color composite images.

If components 1, 2, and 3 account for most of the variance in the dataset, perhaps the original seven bands of TM data can be set aside, and the remainder of the image enhancement or classification can be performed using just these three principal component images. This greatly reduces the amount of data to be analyzed and completely bypasses the expensive and time-consuming process of feature selection so often necessary when classifying remotely sensed data (discussed in Chapter 8).

Fung and LeDrew (1987) and Eastman and Fulk (1993) suggest that *standardized PCA* (based on the computation of eigenvalues from correlation matrices) is superior to unstandardized PCA (computed from covariance matrices) when analyzing change in multitemporal image datasets. Standardized PCA forces each band to have equal weight in the derivation of the new component images and is identical to converting all image values to standard scores (by subtracting the mean and dividing by the standard deviation) and computing unstandardized PCA of the results. Eastman and Fulk processed 36 monthly AVHRR-derived normalized difference vegetation index (NDVI) images of Africa for the years 1986 to 1988. They found the first component was always highly correlated with NDVI regardless of season, while the second, third, and fourth, components related to seasonal changes in NDVI.

There are other uses for principle components analysis. For example, Gillespie (1992) used PCA to perform decorrelation contrast stretching of multispectral thermal-infrared data. The technique involved transformation of the multiple bands of thermal-infrared data to principal components (e.g., decorrelation), independent contrast stretching of decorrelated PCA bands, and retransformation of the stretched data back to the approximate original axes, based on the inverse of the principle component rotation.

Vegetation Indexes

The collection of accurate, timely information on the world's food and fiber crops will always be important (Groten, 1993). The collection of such information using *in situ* techniques is expensive, time consuming, and often impossible (Eastman and Fulk, 1993). An alternative is the measurement of vegetative amount and condition based on an anal-

to the poor utilization of colour by the original correlated data, as seen in the illustration of Fig. 6.9 and as demonstrated in Fig. 6.6c.

6.1.7 Other Applications of Principal Components Analysis

Owing to the information compression properties of the principal components transformation it lends itself to reduced representation of image data for storage or transmission. In such a situation only the uppermost significant components are retained as a representation of an image, with the information content so lost being indicated by the sum of the eigenvalues corresponding to the components ignored. Thereafter if the original image is to be restored, either on reception through a communications channel or on retrieval from memory, then the inverse of the transformation matrix is used to reconstruct the image from the reduced set of components. Since the matrix is orthogonal its inverse is simply its transpose. This technique is known as bandwidth compression in the field of telecommunications; however it has not found great application in remote sensing image processing, presumably because hitherto image transmission has not been a consideration and available memory has not placed stringent limits on image storage.

An interesting application of principal components analysis is in the detection of features that change with time between images of the same region. This is described by example in Chap. 11.

6.2 The Kauth-Thomas Tasseled Cap Transformation

The principal components transformation treated in the previous section yields a new co-ordinate description of multispectral remote sensing image data by establishing a diagonal form of the global covariance matrix. The new co-ordinates (components) are linear combinations of the original spectral bands. Other linear transformations are of course possible. One is a procedure referred to as canonical analysis, treated in Chap. 10. Another, to be developed below, is application-specific in that the new axes in which data are described have been devised to maximise information of importance, in this case, to agriculture. Other similar special transformations would also be possible.

The so-called "tasseled cap" transformation developed by Kauth and Thomas (1976) is a means for highlighting the most important (spectrally observable) phenomena of crop development in a way that allows discrimination of specific crops, and crops from other vegetative cover, in Landsat multitemporal imagery. Its basis lies in an observation of crop trajectories in band 6 versus band 5, and band 5 versus band 4 subspaces. Consider the former as shown in Fig. 6.10a.

A first observation that can be made is that the variety of soil types on which specific crops might be planted appear as points along a diagonal in the band 6, band 5 space as shown. This is well-known and can be assessed from an observation of the spectral reflectance characteristics for soils. (See for example Chap. 5 of Swain and Davis 1978.) Darker soils lie nearer the origin and lighter soils at higher values of both bands. The actual slope of this line of soils will depend upon global external variables such as atmospheric haze and soil moisture effects. If the transformation to be derived is to be used quantitatively these effects need to be modelled and the data calibrated or corrected beforehand.

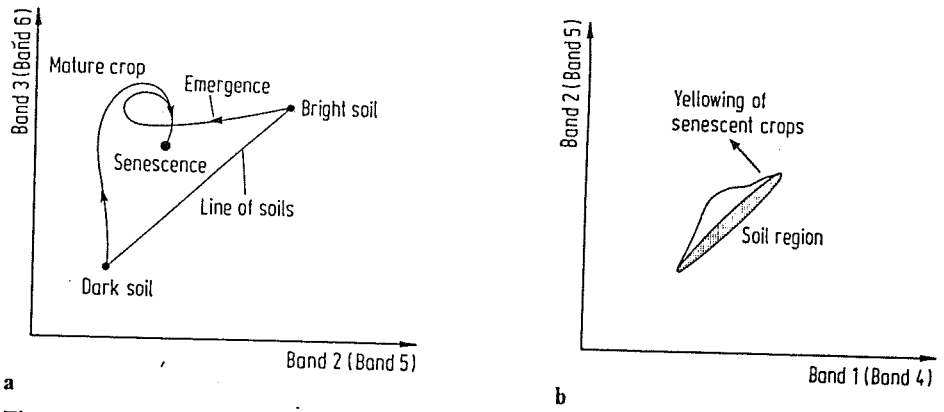


Fig. 6.10. a Band 6 versus band 5 Landsat multispectral scanner subspace showing trajectories of crop development; b Band 5 versus band 4 subspace also depicting crop development

Consider now the trajectories followed in the band 6 versus band 5 subspace for crop pixels corresponding to growth on different soils – in this case take the extreme light and dark soils as depicted in Fig. 6.10a. For both regions at planting the multispectral response is dominated by soil types, as expected. As the crops emerge the shadows cast over the soil dominate any green matter response. As a result there is considerable darkening of the response of the lighter soil crop field and only a slight darkening of that on dark soil. When both crops reach maturity their trajectories come together implying closure of the crop canopies over the soil. The response is then dominated by the green biomass, being in a high band 6 and low band 5 region, as is well known. When the crops senesce and turn yellow their trajectories remain together and move away from the green biomass point in the manner depicted in the diagram. However whereas the development to maturity takes place almost totally in the same plane, the yellowing development in fact moves out of this plane, as can be assessed by how the trajectories develop in the band 5 versus band 4 subspace during senescence as illustrated in Fig. 6.10b.

Should the crops then be harvested the trajectories beyond senescence move, in principle, back towards their original soil positions.

Having made these observations, the two diagrams of Fig. 6.10 can now be combined into a single three dimensional version in which the stages of the crop trajectories can be described according to the parts of a cap, with tassels, from which the name of the subsequent transformation is derived. This is shown in Fig. 6.11. The first point to note is that the line of soils used in Fig. 6.10a is shown now as a plane of soils. Its maximum spread is along the three dimensional diagonal as indicated; however it has a scatter about this line consistent with the spread in band 5 versus band 4 as shown in Fig. 6.10b. Kauth and Thomas note that this plane of soils forms the brim and base of the cap. As crops develop on any soil type their trajectories converge essentially towards the crown of the cap at maturity whereupon they fold over and continue to yellowing as indicated. Thereafter they break up to return ultimately to various soil positions, forming tassels on the cap as shown.

The behaviour observable in Fig. 6.11 led Kauth and Thomas to consider the development of a linear transformation that would be useful in crop discrimination. As with the principal

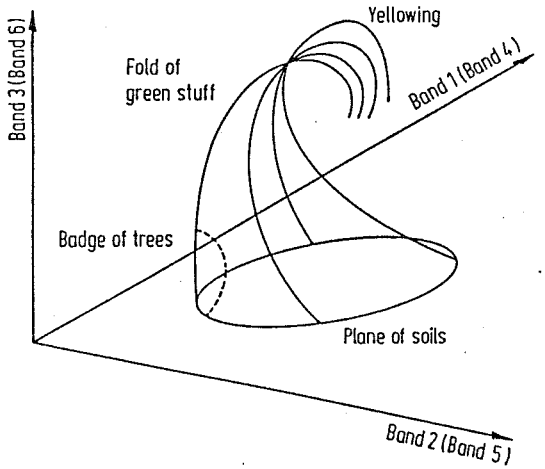


Fig. 6.11. Crop trajectories in Landsat multispectral scanner band 4, 5, 6 space, having the appearance of a tasseled cap

gonal axes. However the axis directions are chosen according to the behaviour seen in Fig. 6.11.

Three major orthogonal directions of significance in agriculture can be identified. The first is the principal diagonal along which soils are distributed. This was chosen by Kauth and Thomas as the first axis in the tasseled cap transformation. The development of green biomass as crops move towards maturity appears to occur orthogonal to the soil major axis. This direction was then chosen as the second axis, with the intention of providing a greeness indicator. Crop yellowing takes place in a different plane to maturity. Consequently choosing a third axis orthogonal to the soil line and greeness axis will give a yellowness measure. Finally a fourth axis is required to account for data variance not substantially associated with differences in soil brightness or vegetative greeness or yellowness. Again this needs to be orthogonal to the previous three. It was called "non-such" by Kauth and Thomas in contrast to the names "soil brightness", "green-stuff" and "yellow-stuff" they applied to the previous three.

The transformation that produces the new description of the data may be expressed as

$$u = Rx + c \tag{6.10}$$

where x is the original Landsat multispectral scanner pixel vector, and u is the vector of transformed brightness values. This has soil brightness as its first component, greeness as its second and yellowness as its third. These can therefore be used as indices, respectively. R is the transformation matrix and c is a constant vector chosen (arbitrarily) to avoid negative values in u.

The transformation matrix R is the transposed matrix of column unit vectors along each of the transformed axes (compare with the principal components transformation matrix). For a particular agricultural region Kauth and Thomas chose the first unit vector as a line of best fit through a set of soil classes. The subsequent unit vectors were generated by using a Gram-Schmidt orthogonalization procedure in the directions required. The transformation matrix generated was

$$R = \begin{bmatrix} 0.433 & 0.632 & 0.586 & 0.264 \\ -0.290 & -0.562 & 0.600 & 0.491 \\ -0.829 & 0.522 & -0.039 & 0.194 \\ 0.223 & 0.012 & -0.543 & 0.810 \end{bmatrix}$$

From this it can be seen, at least for the region investigated by Kauth and Thomas, that the soil brightness is a weighted sum of the original four Landsat bands with approximately equal emphasis. The greeness measure is the difference between the infrared and visible responses. In a sense therefore this is more a biomass index. The yellowness measure can be seen to be substantially the difference between the Landsat visible red and green bands.

Just as new images can be synthesised to correspond to various principal components so can the actual transformed images be created for this approach. By applying (6.10) to every pixel in a Landsat multispectral scanner image, soil brightness, greeness, yellowness and non-such images can be produced. These can then be used to assess stages in crop development.

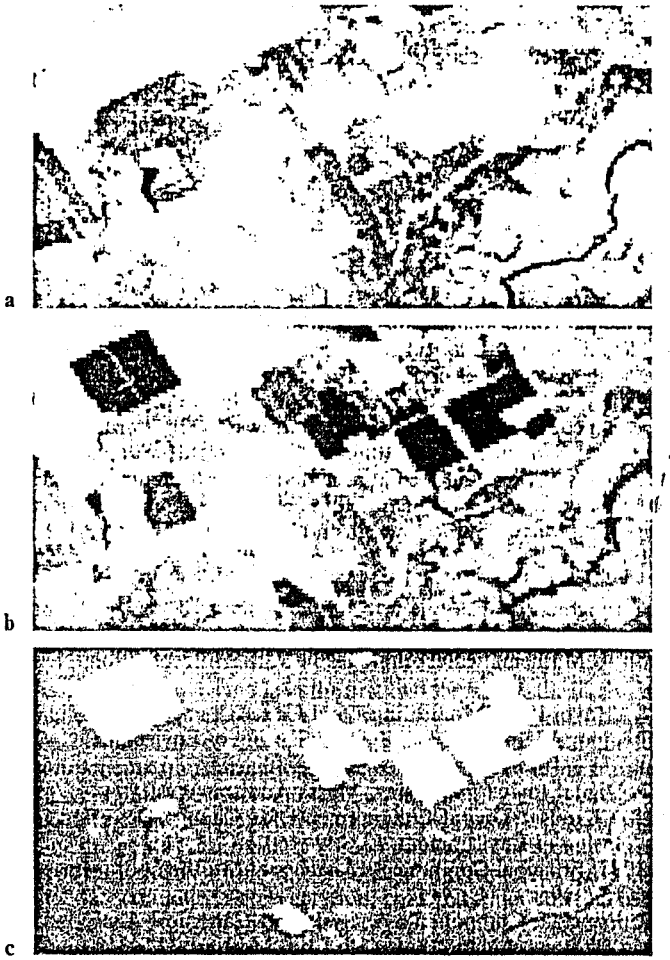


Fig. 6.12. Landsat multispectral scanner band 7 **a** and band 5, **b** images of an arid region containing irrigated crop fields. The ratio of these two images **c** shows vegetated regions as bright, soils as mid to dark grey and water as black

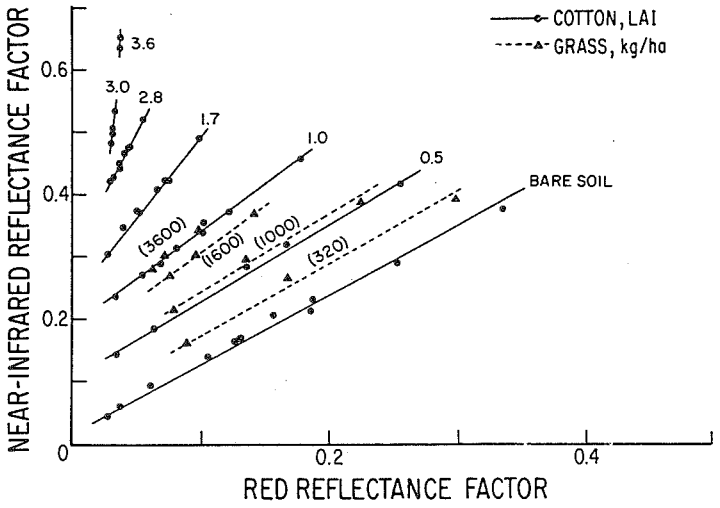


Figure 5 Observed vegetation isolines in NIR–red wavelength space for various canopy densities of cotton and grass with differing soil-background conditions. [Reprinted by permission of the publisher from Huete (1988). Copyright 1988 by Elsevier Science Publishing Co., Inc.]

tern of increasing slopes and NIR intercepts with higher vegetation densities (Figure 5 and Table 4). This same soil-vegetation spectral behavior has been verified with various canopy radiant transfer models including the SAIL model (S. Goward, personal communication), the two-stream hemispherical model of Sellers (1985), and the directional radiative transfer model utilized by Choudhury (1987).

III.B Solar-Angle Interactions

Understanding the spectral behavior of partial vegetation canopies at different solar zenith angles is necessary in many remote-sensing applications such as seasonal and phenological vegetation analyses and productivity studies (Kimes et al., 1980). The change

TABLE 4 Geometric Properties of Cotton-Canopy Isolines in NIR–Red Wavelength Space

Green Cover (%)	Slope	NIR Intercept	Length	n	Correlation Coefficient (r)
0	1.06	0.026	0.396	8	0.997
20	1.24	0.106	0.301	5	0.998
25	1.30	0.131	0.278	4	0.999
40	1.54	0.186	0.224	8	0.997
55	2.01	0.253	0.156	4	0.999
60	2.50	0.240	0.153	8	0.997
75	3.93	0.299	0.094	8	0.994
90	15.26	0.006	0.052	4	0.977
95	2.19	0.005	0.021	4	0.134
100	55.33	−0.015	0.017	2	—

Source: Reprinted by permission of the publisher from Huete et al. (1985). Copyright 1985 by Elsevier Science Publishing Co., Inc.

IV.A.1 Ratio-Based Indices

Vegetation indices or greenness measures developed thus far can be classified into two broad categories: the ratio indices and orthogonal indices. The ratio vegetation index (RVI = NIR/red) and normalized difference vegetation index [NDVI = (NIR − red)/(NIR + red) = (RVI − 1)/(RVI + 1)] are the most common of the ratio transformations used as vegetation measures. They involve ratioing a linear combination of the NIR and red bands by another linear set of the same bands. In the two-dimensional NIR–red space, these indices are graphically displayed by vegetation isolines of increasing slopes diverging out from the origin (Figure 9).

The ratio indices essentially rely on the existence of the soil-line concept in normalizing soil behavior and discriminating vegetation spectra. Since most soil spectra fall on or close to the soil line, and since the intercept of such a line is close to the origin, RVI and NDVI values of bare soils (ratios) will be nearly identical for a wide range of soil conditions. These indices have been found effective in normalizing soil-background spectral variations (Colwell, 1974), and variations in irradiance conditions (Tucker, 1979). Tucker (1977) and Ripple (1985) found that the NDVI was the best estimator of low amounts of blue grama (*Bouteloua gracilis*) and tall fescue (*Festuca arundinacea*) grass phytomass. Colwell (1974) and Pearson et al. (1976) found that the RVI was unreliable in grass canopies with low green covers (<30%). Weiser et al. (1984) reported a direct correlation between RVI and tall grass prairie phytomass, however, they found their results to be site-dependent, year-specific, and sensitive to the presence of senescent vegetation.

IV.A.2 Orthogonal-Based Indices

The second broad category of spectral vegetation measures are orthogonal transformations that include the two-band perpendicular-vegetation index (PVI) of Richardson and Wiegand (1977) and the four-band green-vegetation index (GVI) of Kauth and Thomas

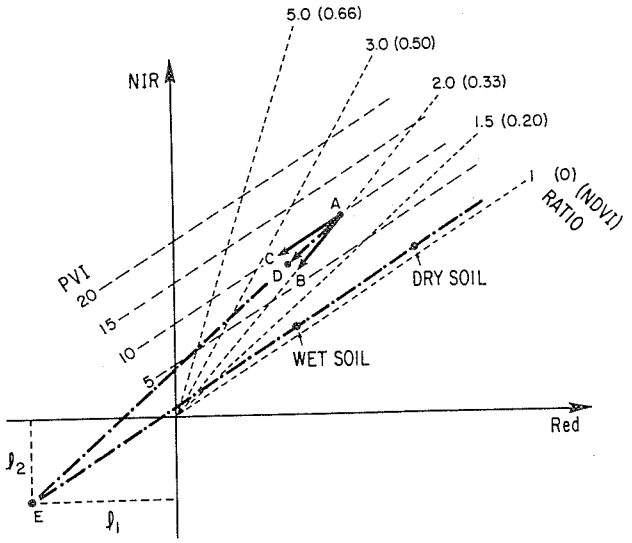


Figure 9 Vegetation spectra isolines and convergence points in NIR–red wavelength space as predicted by the ratio, normalized difference, and perpendicular-vegetation indices. [Reprinted by permission of the publisher from Huete (1988). Copyright 1988 by Elsevier Science Publishing Co., Inc.]

(1976). The six-band (Crist and Cicone, 1984) and *n*-wavelength band (Jackson, 1983) GVI also represent orthogonal indices. The orthogonal indices are distinct from the ratio indices in that isolines of equal "greenness" do not converge at the origin, but instead remain parallel to the principal axis of soil spectral variation, i.e., the soil line (Figure 9). A greenness vector, orthogonal to the soil line, is computed to maximally include green-vegetation signals while holding soil background constant. The projection of composite spectra onto this vector is subsequently used as the measure of vegetation.

The explanation offered by the two types of indices are contradictory with each other in describing soil-vegetation spectral behavior. As an example, a partial canopy over a dry-soil background (A) is shown in Figure 9. If the soil background were to become wet, a vegetation isoline bounded by dry- and wet-soil conditions would be formed. In order for the RVI and NDVI to effectively normalize such a background change, the vegetated pixel would have to shift directly toward the origin (B), following an isoline of constant RVI and NDVI values. The PVI, however, would require the pixel to shift along an isoline parallel to the soil line (C), so that both the wet- and dry-soil vegetated pixels maintain a constant PVI value (equidistant to the soil line). Another vegetation index, called the soil-adjusted vegetation index [SAVI = [(NIR − red)/(NIR + red + 0.5)]1.5] was developed by Huete (1988) to describe observed grass- and cotton-vegetation isolines (Figure 5). The SAVI assumes a shift (D) that lies between those predicted by the ratio and orthogonal approaches (Figure 9). This index takes on properties of both the NDVI and PVI and is more fully described later in this chapter.

There is a lack of detailed analyses concerning the limitations of these vegetation indices in assessing greenness and in monitoring of the plant canopy (Tucker, 1979). For the most part, quantitative information regarding the performance of various spectral measures have been collected over uniform soil backgrounds with ground-based radiometers. Multispectral data collected from space- or airborne sensors, on the other hand, will quite often include different soil types under several soil-moisture conditions and with varying quantities of dead organic material. The usefulness of vegetation spectral indices depend, in part, on how well they minimize these soil-background spectral variations. In wavelength space, it is the migration of a vegetated pixel away from the soil line in relation to vegetation density that must be adequately modeled.

IV.B Soil-Brightness Effects

Soil-brightness influences have been noted in numerous studies, where, for a given amount of vegetation, darker- or lighter-soil substrates resulted in higher vegetation index values when the NDVI, RVI, PVI, and GVI were used as vegetation measures. Colwell (1974) found dark-soil canopies to have the highest RVI values at low- to intermediate-percent oat covers. Elvidge and Lyon (1985) additively mixed vegetation spectra with a series of arid-region soil substrates and reported higher RVI and NDVI greenness values for darker substrates, but found no effect on the PVI. Jackson et al. (1983) found the PVI and GVI to be sensitive to soil-moisture condition on a uniform soil type. Huete et al. (1985) found RVI and NDVI values to decrease while PVI and GVI values increased with increasing soil brightness under a constant amount of vegetation. All of these studies show spectral indices to be partially correlated with soil brightness over certain ranges of vegetation density. Thus, in areas where there are considerable soil-brightness variations resulting from moisture differences, roughness variations, shadow, or organic-matter differences, there are soil-induced influences on the vegetation index values.

Soil-brightness influences are prevalent in partially vegetated canopies because the ratio-based and orthogonal-based vegetation indices fail to predict the behavior of vegetated pixels as they migrate away from the soil line. If vegetation isoline behavior does not agree with that predicted by spectral vegetation indices, then different soil backgrounds under constant vegetation amounts will produce different spectral index values. The isolines presented in Figure 5 neither converged at the origin (as required by the RVI and NDVI) nor maintained slopes identical with the soil line (a PVI requirement). At low-vegetation densities, the isolines are nearly parallel with the soil line, while at very high densities, they nearly approach "zero" intercepts. Over most of the range of vegetation densities, the isolines do not obey the pattern expected from the ratio and orthogonal indices.

Figure 10 shows the soil-brightness problem inherent in the NDVI. Decreases in red-canopy reflectance, due to darker-soil substrates, cause significant increases in the NDVI. The NDVI appears to be as sensitive to soil darkening as to vegetation development. Thus, a very low amount 320 kg/ha, of grass phytomass on a dark-soil substrate has the same NDVI (~ 0.3) as 1000 kg/ha on a bright sandy substrate. The NDVI values for a 20% green cotton cover (LAI = 0.5) ranged from 0.24–0.60 and approached the values for a 60% green cover (LAI = 1.7). The NDVI also does not account for the orientation of the bare-soil line (Figure 10). Since the soil line has a positive NIR intercept in NIR–red wavelength space, darker soils have higher NDVI values than the lighter-colored soils because their vector slope to the origin is steeper. As a result, the NDVI would not be able to differentiate these darker (bare) soils from a 320 kg/ha grass canopy over lighter soils.

Figure 11 demonstrates the nonparallel nature of cotton GVI isolines. From these data one can see: (1) soil-brightness influences, where brighter soils produce the highest GVI values for equivalent vegetation densities, and (2) secondary soil variations within the isolines. Secondary soil noise variations increase with the additional use of more wavelengths in the computation of orthogonal indices. In Figure 11, the GVI values of

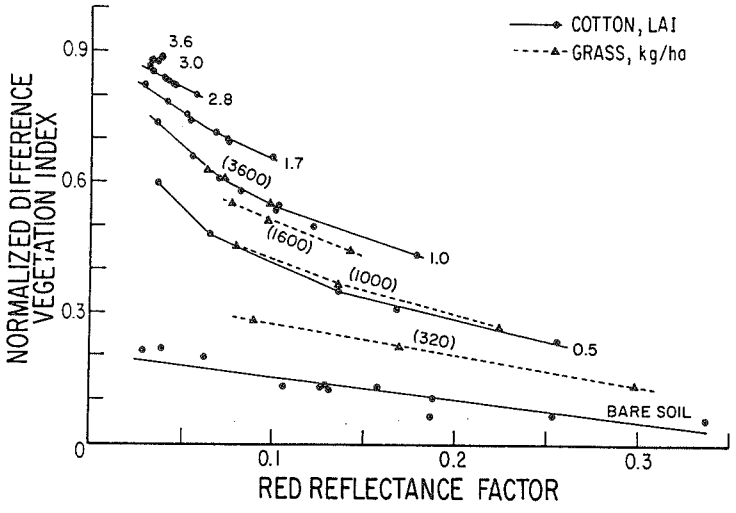


Figure 10 Relationship between the normalized difference and bare-soil red reflectance for various canopy densities of cotton and grass. Numbers in parentheses denote grass biomass. [Reprinted by permission of the publisher from Huete (1988). Copyright 1988 by Elsevier Science Publishing Co., Inc.]

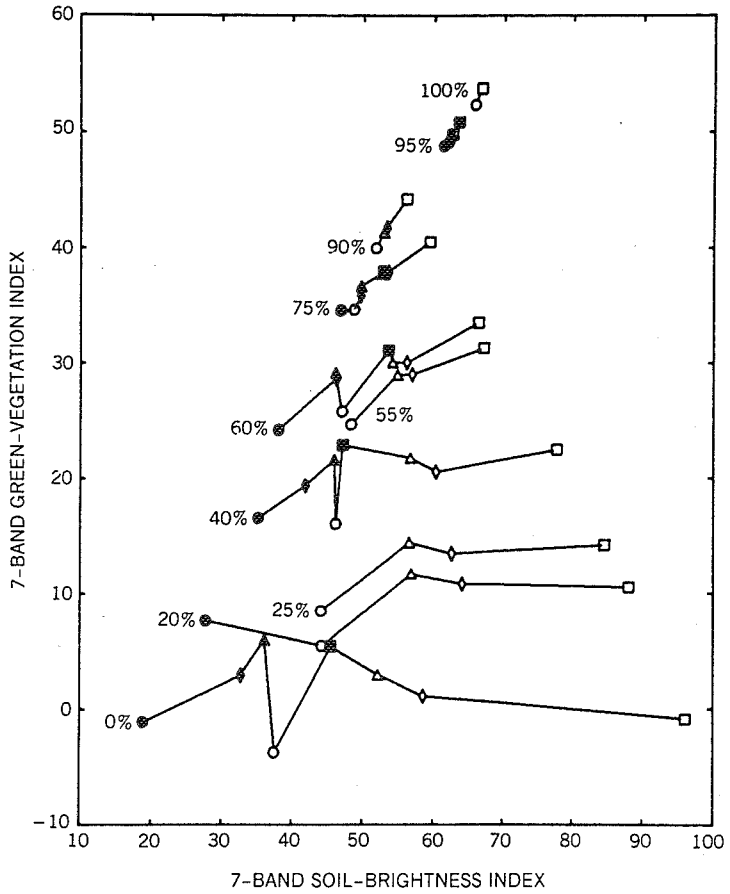


Figure 11 Relationship between a seven-band green vegetation index and seven-band soil-brightness index for various green cotton covers. See Figure 1 caption for symbols; solid symbols denote wet-soil backgrounds. [Reprinted by permission of the publisher from Huete et al. (1985). Copyright 1985 by Elsevier Science Publishing Co., Inc.]

bare soil overlap with those from the 20% green canopy (LAI = 0.5). As expected, secondary soil influences decrease in magnitude with increases in vegetation density due to the decreased soil signal.

Irrespective of how bare-soil spectra are normalized (ratios or rotation), soil-brightness influences become greater with increasing vegetation densities up to 40–60% green-canopy cover (LAI = 1–2). The spectral index values from the 75% canopy cover (LAI = 2.8) are as sensitive to soil background as those from the 20% canopy covers (LAI = 0.5) and soil-induced variations in the NDVI for a constant canopy of LAI = 1.0 encompassed half the total NDVI dynamic range (zero to full cover) of the green canopy.

The SAVI, which was designed to minimize soil-brightness influences, produces vegetation isolines more nearly independent of the soil background (Figure 12). Soil variations are reduced by the SAVI in both the narrow-leaf grass (erectophile) and the broad-leaf cotton (planophile) canopies. In Figure 13, soil-induced spectral index variations are shown as a function of green cotton LAI for the NDVI, PVI, and SAVI. The relationship between NDVI and PVI with LAI is very soil-dependent, as seen by the con-

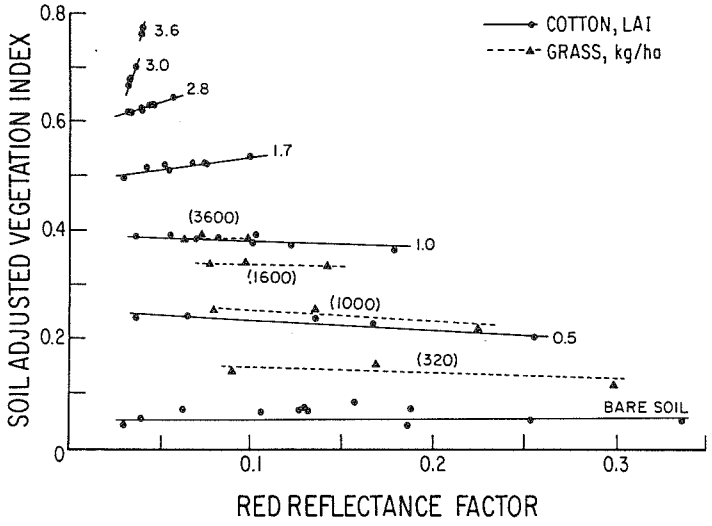


Figure 12 Relationship between the soil-adjusted vegetation index and bare-soil red reflectance for various canopy densities of cotton and grass. Numbers in parentheses denote grass biomass. [Reprinted by permission of the publisher from Huete (1988). Copyright 1988 by Elsevier Science Publishing Co., Inc.]

siderable range in NDVI or PVI values for a constant vegetation density. Note the opposite soil-brightness effects between these two indices. The PVI values for a constant amount of vegetation are greatest with the light-soil background, whereas it is the darker soils that result in highest NDVI values. In both cases, the soil-brightness influence becomes more serious at intermediate levels of vegetation density than at either very

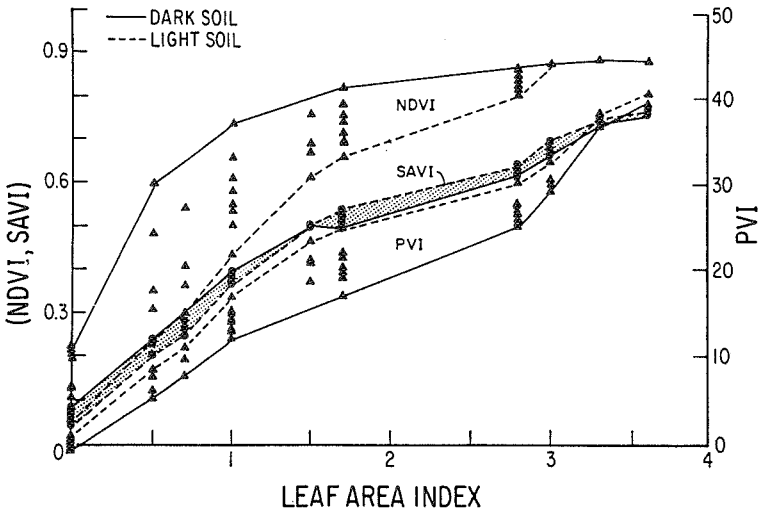


Figure 13 Vegetation index response and soil variations of the NDVI, SAVI, and PVI as a function of cotton leaf-area index. [Reprinted by permission of the publisher from Huete (1988). Copyright 1988 by Elsevier Science Publishing Co., Inc.]

low or very high densities (Figure 13). The SAVI, by comparison, substantially reduces soil-induced variations, and it improves the linearity between the index and LAI.

IV.C Solar-Angle Influences

Jackson et al. (1980) showed how sunlit- and shaded-soil backgrounds might affect spectral indices, primarily through differences in red and NIR flux penetration through the canopy. The changes in red- and NIR-canopy reflectances with sun angle (Figure 6) were mostly a result of differences in sunlit- and shaded-soil contributions. Figure 14 shows the NDVI of a 40% cotton canopy plotted against bare-soil NIR reflectance for the two solar zenith angles. As discussed in the previous section, NDVI values increase as the soil becomes darker for constant vegetation conditions (Figure 10). The NDVI values are also higher at the larger solar zenith angle, partly due to a greater proportion of plant-material irradiance and partly as a result of greater soil darkening (shadow) at this angle. Not only are NDVI values greater, but the soil-induced effects are also smaller at the larger sun angle. The NDVI of the 40% green canopy is greatest with darker-soil types and larger solar zenith angles (shaded soil). This corresponded to minimal soil-component contributions (darkest soil surface) and maximum vegetation irradiance conditions.

Figure 14 shows that NDVI differences with solar zenith angle are primarily a soil-induced effect since (a) they become greater with lighter-colored soils and (b) they are minimal with very dark soils. With the extrapolated "zero" soil background, the 40% canopy only varies from 0.79–0.77 units. The soil-dependent nature of the NDVI is of concern to temporal data sets due to NDVI sensitivity to differences in sunlit- and shaded-soil components resulting from variations in solar zenith angle. Huete (1987a) analyzed sun-angle influences on the PVI and found a strong soil-component influence on diurnal canopy spectral responses.

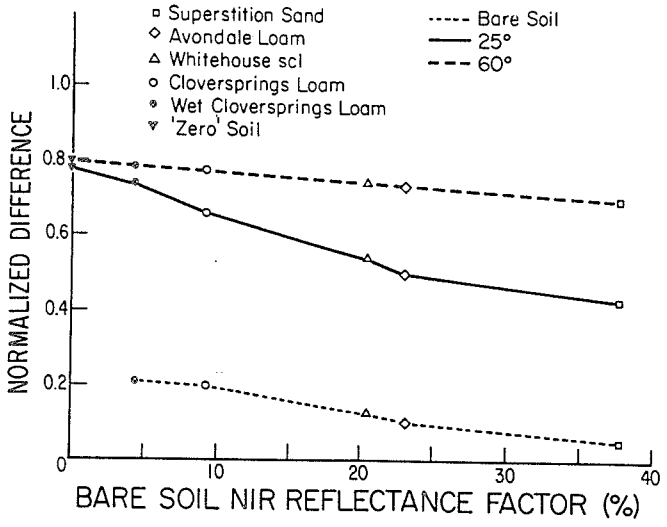


Figure 14 Normalized difference vegetation index values of a 40% cotton canopy at two solar zenith angles and different soil backgrounds. [From Huete (1987a).]

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).

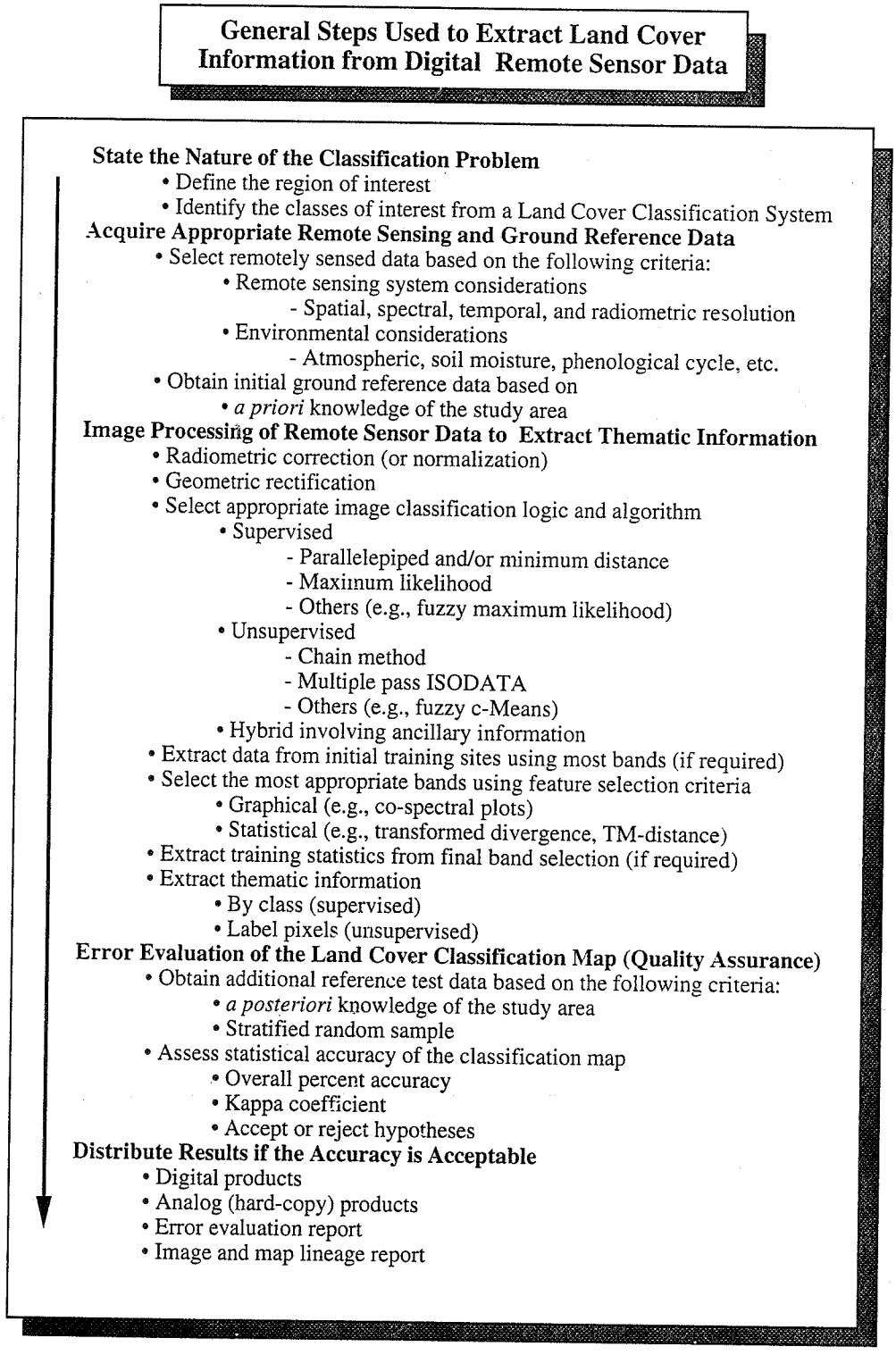


Figure 8-1 General steps required to extract land-cover information from digital remote sensor data.

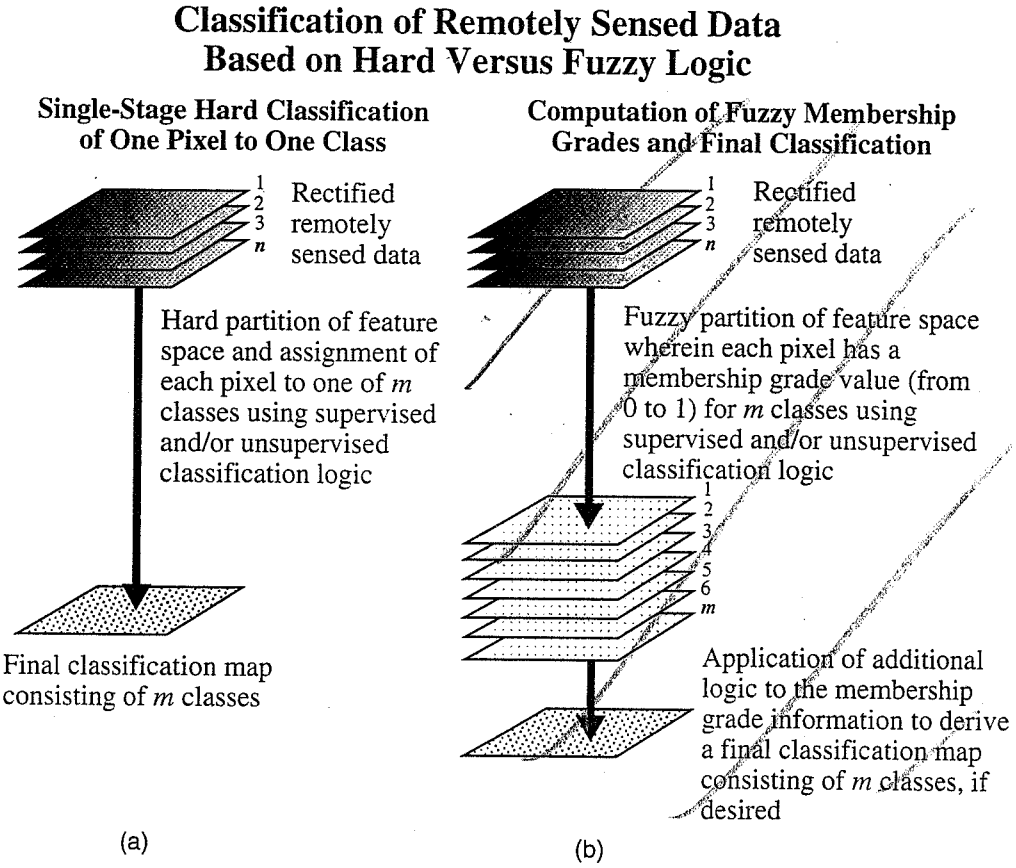


Figure 8-2 Difference between a traditional single-stage hard classification using supervised or unsupervised classification logic and classification using fuzzy logic.

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.

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

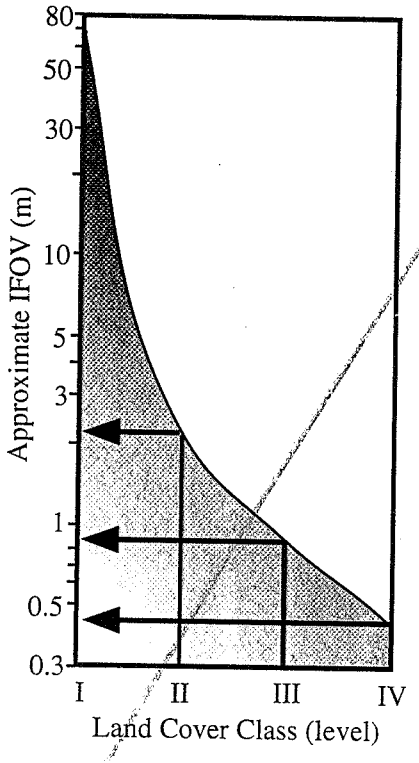


Figure 8-5 Spatial resolution (IFOV) requirements as a function of the mapping requirements for levels I to IV land-cover classes in the United States (based on Anderson et al., 1976). Levels I, II, III, and IV information are normally derived from satellite, high-, medium-, and low-altitude image data,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

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

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 $\leq 30 \times 30$ 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

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

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:

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

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

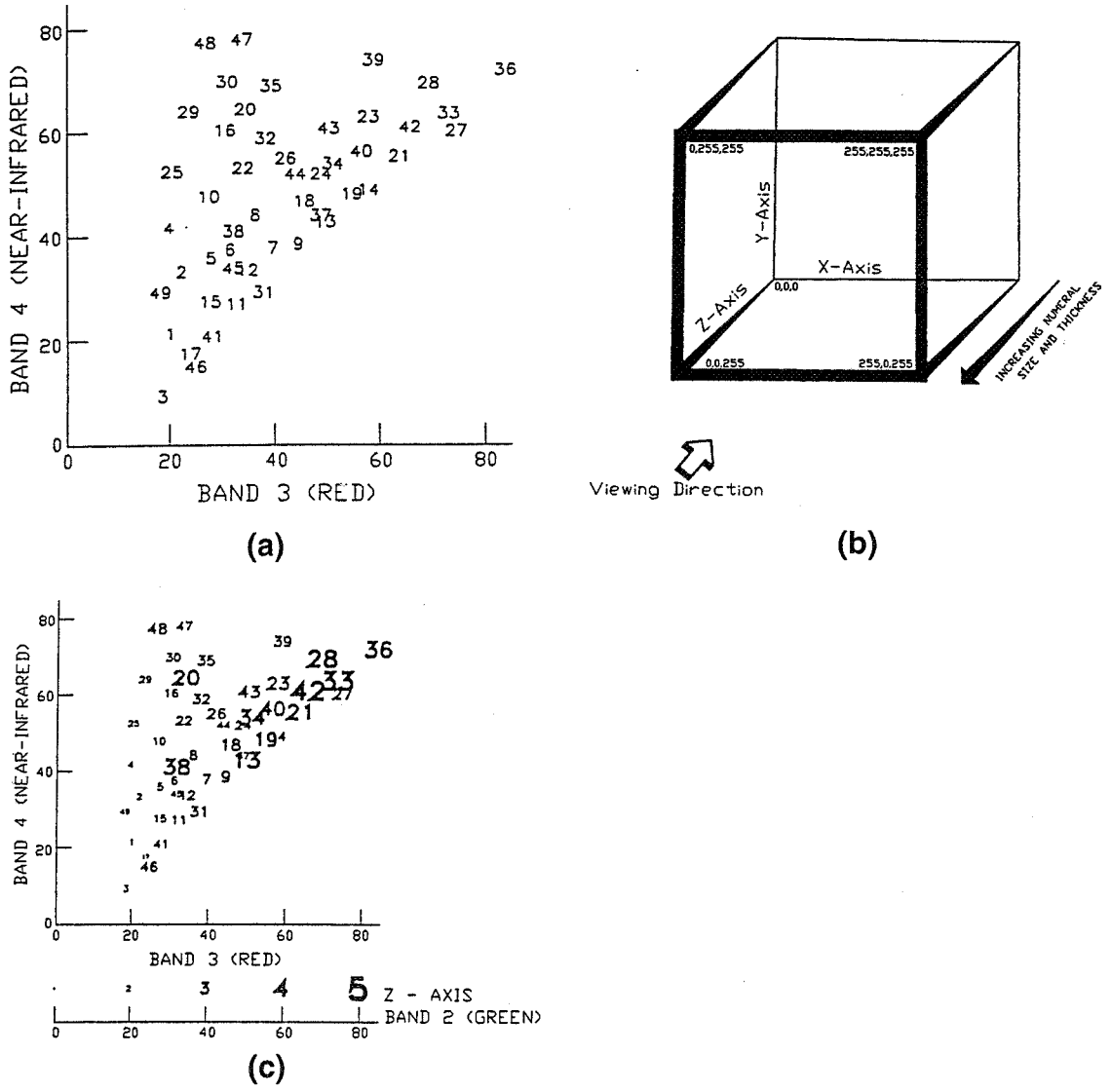


Figure 8-8 (a) Cospectral mean vector plots of 49 clusters from Charleston, S.C., TM data in bands 3 and 4. (b) The logic for increasing numeral size and thickness along the *z* axis. (c) The introduction of band 2 information scaled according to size and thickness along the *z* axis (Hodgson and Plews,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

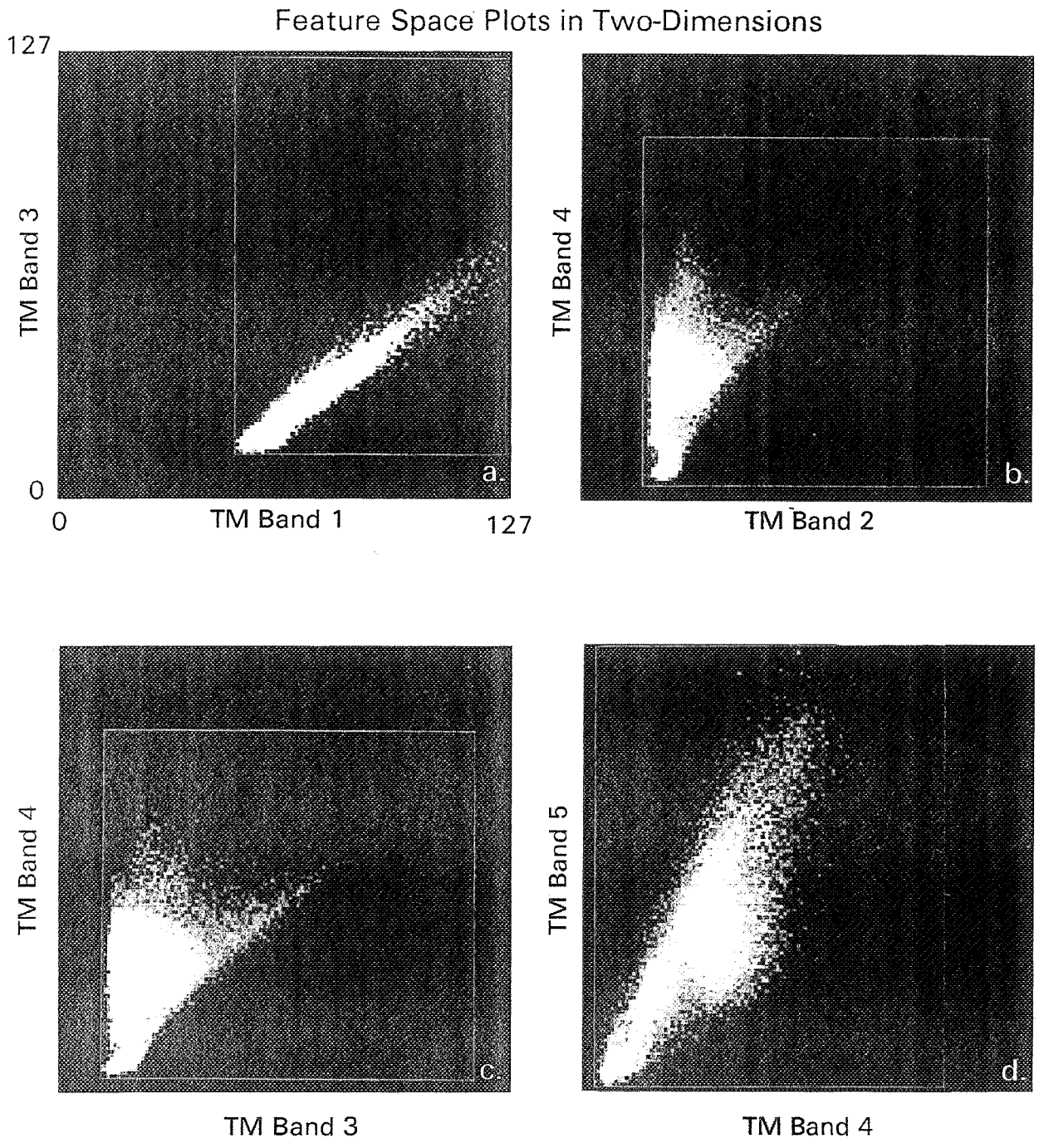


Figure 8-9 Two-dimensional feature space plots of four pairs of Landsat TM data of Charleston, S.C. (a) TM Bands 1 and 3, (b) TM bands 2 and 4, (c) TM bands 3 and 4, and (d) TM bands 4 and 5. The brighter a particular pixel is in the display, the more pixels within the scene having that unique combination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

$$[X,Y,Z,1] = [BVx,BVy,BVz,1]^{*} \begin{matrix} P^{T'} \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ P^{T} \\ \\ \\ \\ \end{matrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi & 0 \\ 0 & \sin\phi & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8\text{-}5)$$

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

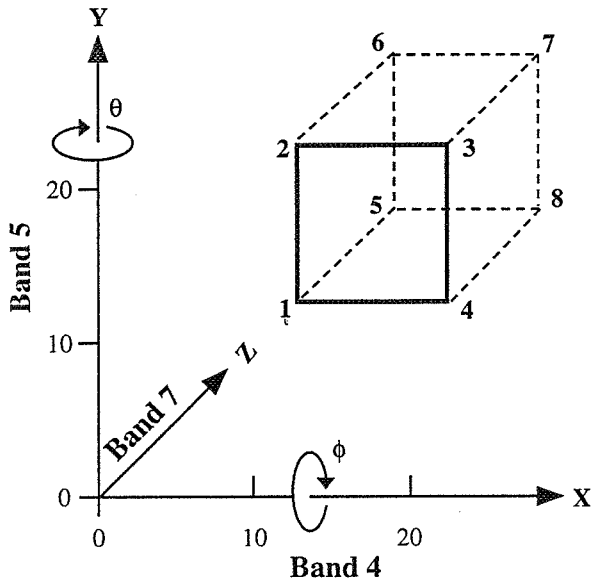


Figure 8-11 Simple parallelepiped displayed in pseudo three-dimensional space. Each of the eight corners represents a unique x, y, z coordinate corresponding to a lower or upper threshold value of the training data. For example, the original coordinates of point 4 are associated with (1) the upper threshold 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 $P^{T'}$ 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

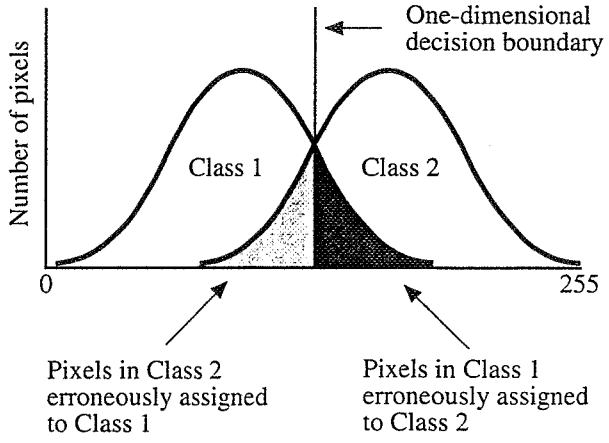


Figure 8-13 The basic problem in remote sensing pattern recognition classification is, given a spectral distribution of data in n bands (here just 1 band), to find an n-dimensional decision boundary that will allow the separation of the major classes (just 2 in this example) with a minimum of error and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

		Divergence (upper number) and Transformed Divergence (lower number)									
		Class Combinations[a]									
Band Combinations	Average Divergence	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)

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

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}

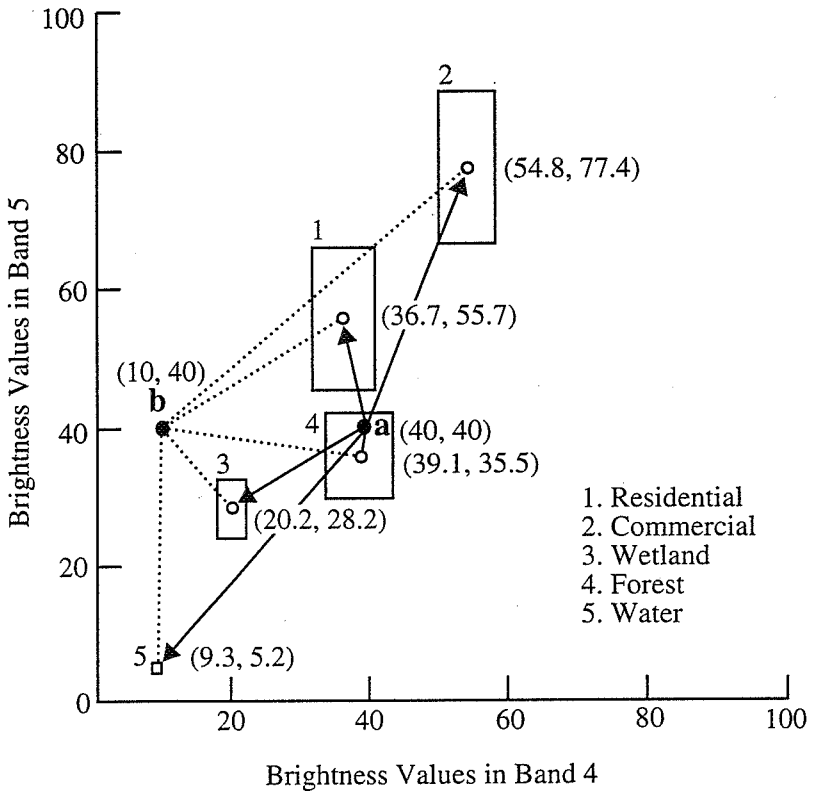


Figure 8-14 Points *a* and *b* are pixels in the image to be classified. Pixel *a* has a brightness value of 40 in band 4 and 40 in band 5. Pixel *b* has a brightness value of 10 in band 4 and 40 in band 5. The boxes represent the *parallelepiped* decision rule associated with a ±1 standard deviation 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? $L_{ck} \leq a \leq H_{ck}$	Does pixel *b* (10, 40) satisfy criteria for this class in this band? $L_{ck} \leq b \leq H_{ck}$
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.

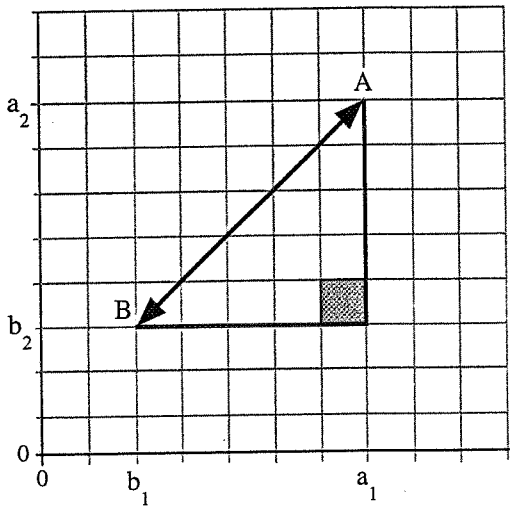
dard deviation is subtracted and added to the mean of each class and for each band to identify the lower (L_{ck}) and upper (H_{ck}) edge of the parallelepiped. In this case only two bands are used, 4 and 5, resulting in a two-dimensional box. This could be extended to *n* dimensions or bands. With the 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., $31.27 \leq 40 \leq 41.23$), 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., $33.99 \leq 40 \leq 44.21$) and band 5 ($29.09 \leq 40 \leq 41.91$). 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,

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

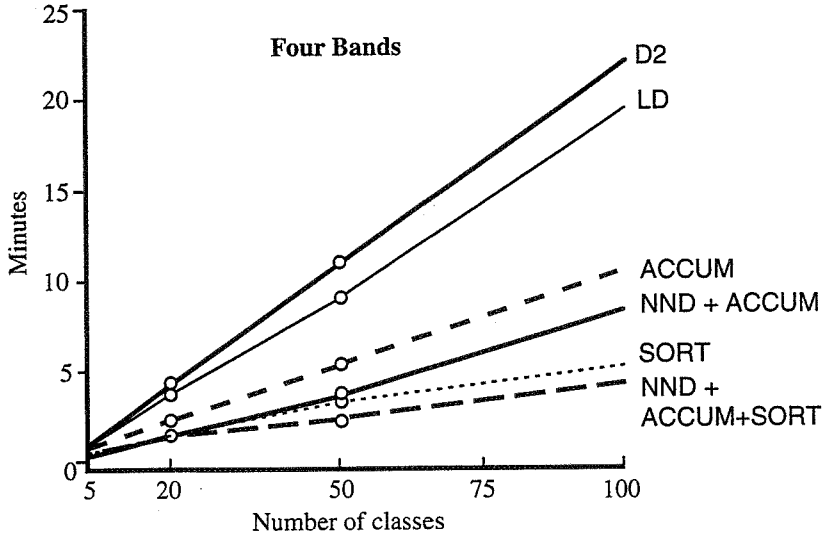


Figure 8-16 Results of applying six improved minimum distance to means classification algorithms to Charleston, S.C., TM data (from Hodgson, 1988).

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

It is of interest to note that SSE has a theoretical minimum of zero, which corresponds to all clusters containing only a single data point. As a result, if an iterative method is used to seek the natural clusters or spectral classes in a set of data then it has a guaranteed termination point, at least in principle. In practice it may be too expensive to allow natural termination. Instead, iterative procedures are often stopped when an acceptable degree of clustering has been achieved.

It is possible now to consider the implementation of an actual clustering algorithm. Whilst it should depend upon a progressive minimisation (and thus calculation) of SSE this is impracticable since it requires an enormous number of values of SSE for the evaluation of all candidate clusterings. For example, there are $C^P/C!$ ways of placing P patterns into C clusters. This number of SSE values would require computation at each stage of clustering to allow a minimum to be chosen. Rather than embark upon such a rigorous and computationally expensive approach the heuristic procedure of the following section is usually adopted in practice.

9.3 The Iterative Optimization (Migrating Means) Clustering Algorithm

The iterative optimization clustering procedure, also called the migrating means technique, is essentially the isodata algorithm presented by Ball and Hall (1965). It is based upon estimating some reasonable assignment of the pixel vectors into candidate clusters and then moving them from one cluster to another in such a way that the SSE measure of the preceding section is reduced.

9.3.1 The Basic Algorithm

The iterative optimization algorithm is implemented by the following set of basic steps:

1. The procedure is initialised by selecting C points in multispectral space to serve as candidate cluster centres. Let these be called

$$\hat{m}_i,\ i = 1, \ldots C.$$

The selection of the $\hat{m}_i$ at this stage is arbitrary with the exception that no two may be the same. To avoid anomolous cluster generation with unusual data sets it is generally wise to space the initial cluster means uniformly over the data. This can also serve to enhance convergence.
Besides choosing the $\hat{m}_i$, the number of clusters C, must be specified beforehand by the user.

2. The location x of each pixel in the segment of the image to be clustered is examined and the pixel is assigned to the nearest candidate cluster. This assignment would be made on the basis of the Euclidean or even $L1$ distance measure.
3. The new set of means that result from the grouping produced in step (2) are computed. Let these be denoted

$$m_i,\ i = 1, \ldots C.$$

4. If $m_i = \hat{m}_i$ for all i, the procedure is terminated. Otherwise $\hat{m}_i$ is redefined as the current value of m_i and the procedure returns to step (2).

The iterative optimization procedure is illustrated for a simple set of two dimensional patterns in Fig. 9.2.

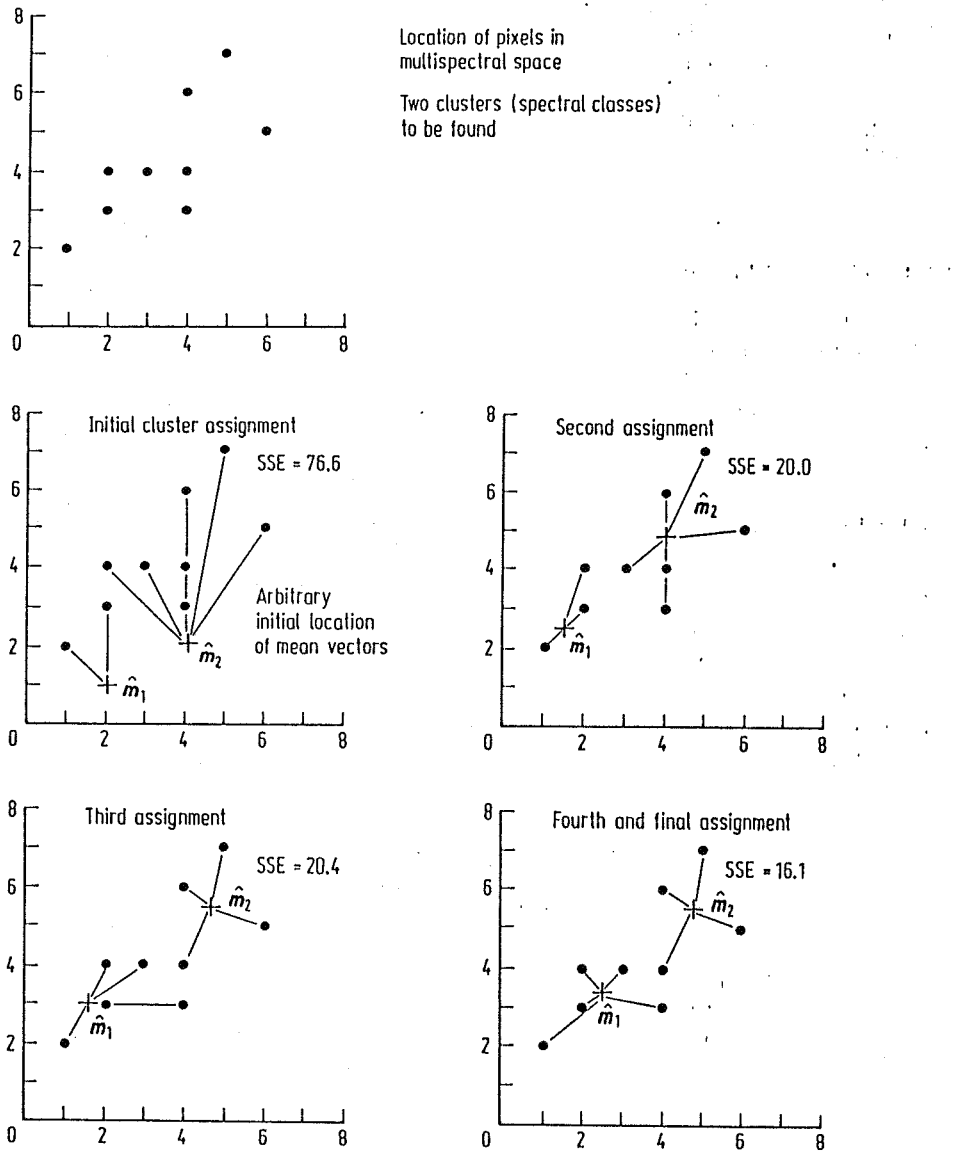


Fig. 9.2. An illustration of clustering by iterative optimization (or the isodata method). As noted, the method leads to a progressive reduction in SSE

9.3.2 Mergings and Deletions

Once clustering is completed, or at any suitable intervening stage, the clusters can be examined to see whether

(i) any clusters contain so few points as to be meaningless (e. g. that they would not give acceptable statistics estimates if used in training a maximum likelihood classifier), or

(ii) some clusters are so close together that they represent an unnecessary or indeed an injudicious division of the data, and thus that they should be merged.

In view of the material of Sect. 8.2.6 a guideline exists for (i), viz that a cluster would be of little value for training a maximum likelihood classifier if it did not contain about $10N$ points where N is the number of spectral components. In Chap. 10, which deals with separability and divergence, means for deciding whether clusters should be merged can also be devised.

9.3.3 Splitting Elongated Clusters

Another stage that can be inserted into the isodata algorithm is to separate elongated clusters into two new clusters. Usually this is done by prespecifying a standard deviation in each spectral band beyond which a cluster should be halved. Again this can be done after a set number of iterations, also specified by the user.

9.3.4 Choice of Initial Cluster Centres

Initialisation in the iterative optimization procedure requires specification of the number of clusters expected, along with their starting positions. In practice the actual or optimum number of clusters to choose will not be known. Therefore it is often chosen conservatively high, having in mind that resulting inseparable clusters can be consolidated after the process is completed, or at intervening iterations if a merging operation is available.

The choice of the initial locations of the cluster centres is not critical although evidently it will have an influence on the time it takes to reach a final, acceptable clustering. Since no guidance is available in general the following is a logical procedure and one which is adopted in LARSYS (Phillips 1973). The initial cluster centres are chosen uniformly spaced along the multidimensional diagonal of the multispectral pixel space. This is a line from the origin to the point corresponding to the maximum brightness value in each spectral component (corresponding to 127, 127, 127, 63 for Landsat multispectral scanner data). This choice can be refined if the user has some idea of the actual range of brightness values in each spectral component, say by having previously computed histograms. In that case the cluster centres would be initialised along a diagonal through the actual multidimensional extremities of the data.

Choice of the initial locations of clusters in the manner described is a reasonable and effective one since they are then well spread over the multispectral space in a region in which many spectral classes occur, especially for correlated data such as that corresponding to soils, rocks, concretes, etc.

9.3.5 Clustering Cost

Obviously the major limitation of the isodata technique is the need to prespecify the number of cluster centres. If this specification is too high then a *posteriori* merging can be used; however this is an expensive strategy. On the other hand, if too few are chosen initially then some multimodal spectral classes will result which, in turn, will prejudice ultimate classification accuracy.

Irrespective of whether too many or too few clusters are used, the isodata approach is computationally expensive since, at each iteration, every pixel must be checked against all cluster centres. Thus for C clusters and P pixels, PC distances have to be computed at each iteration and the smallest found. For 4 band data, each Euclidean distance calculation will require 4 multiplications and 4 additions; ignoring the square root operation in (9.1) since that need not be carried out. Thus for 20 classes and 10,000 pixels, 100 iterations of isodata clustering would take approximately 27 minutes of computation if a multiplication and addition requires approximately 20 μs, just for the distance computations.

9.4 Unsupervised Classification and Cluster Maps

At the completion of clustering, pixels within a given group are usually given a symbol to indicate that they belong to the same cluster or spectral class. Using these symbols a cluster map can be produced; this is a map corresponding to the image which has been clustered, but in which the pixels are represented by their symbol rather than by the original multispectral data. The availability of a cluster map allows a classification to be made. If some pixels with a given label can be identified with a particular ground cover type (by means of maps, site visits or other forms of reference data) then all pixels with the same label can be associated with that class. This method of image classification, depending as it does on *a posteriori* recognition of the classes, is called unsupervised classification since the analyst plays no part until the computational aspects are complete. Often unsupervised classification is used on its own, particularly when reliable training data for supervised classification cannot be obtained or is too expensive to do so. However, it is also of value, as noted earlier, to determine the spectral classes that should be considered in a subsequent supervised approach. This approach is pursued in detail in Chap. 11.

9.5 A Clustering Example

To illustrate the nature of the results produced by the iterative optimization algorithm a simple example with Landsat multispectral scanner data is presented. Fig. 9.3 a shows a small image segment (band 7 only for illustration) which consists of regions of crops and background soils. Figure 9.3 b shows a scatter diagram for the image. In this band 5 versus band 7 brightnesses of the pixels have been plotted. This is a subspace of the full four dimensional multispectral space of the image and gives an illustration of how the data points are distributed.

The data was clustered using the iterative optimization procedure as implemented by Kelly (1983). Only five iterations were used and the algorithm was asked to determine

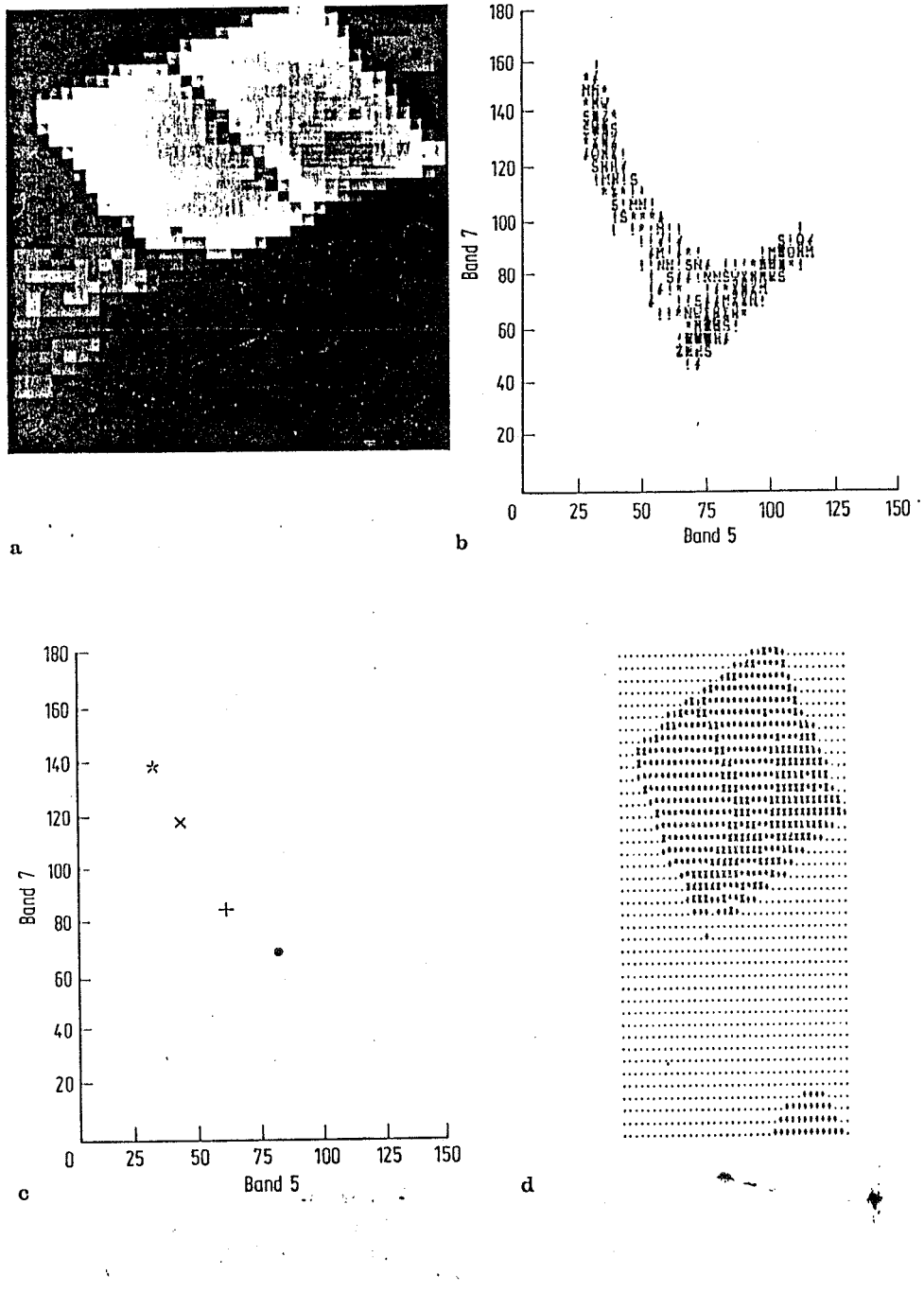


Fig. 9.3. a Image segment used in the clustering illustration; b Band 5 versus band 7 scatter diagram for the image; c Cluster centres on a band 5 versus band 7 diagram; d Cluster map produced by the isodata algorithm

five clusters. Merging and splitting options were employed at the end of each iteration leading ultimately to the four clusters shown on the plot of cluster means in Fig. 9.3 c and the cluster map shown in Fig. 9.3 d. Comparison with Fig. 9.3 a shows that the vegetation classes have been segmented more finely than the background soils in this case. Nevertheless the cluster map displays acceptable spatial homogeneity. Numerical details of the clusters established are given in Table 9.1.

Table 9.1. Cluster means and standard deviations for Fig. 9.3. generated by the iterative optimization procedure

Cluster	Symbol	Band	Mean	St. Dev.
1	·	4	74.4	9.6
		5	85.5	13.7
		6	89.9	14.2
		7	69.8	12.1
2	*	4	45.0	2.0
		5	32.4	2.2
		6	127.4	6.3
		7	136.8	5.9
3	+	4	60.0	3.2
		5	59.5	4.0
		6	94.4	6.9
		7	83.7	7.5
4	×	4	48.9	3.8
		5	39.1	6.5
		6	114.0	5.9
		7	116.3	8.4

It is important to realise that the results generated in this example are not unique but depend upon the clustering parameters chosen. In practice the user may need to apply the algorithm several times with different parameter values to generate the desired segmentation.

9.6 A Single Pass Clustering Technique

In order to reduce the cost of clustering image data, alternatives to iterative optimization have been proposed and are widely implemented in software packages for remote sensing image analysis. Often what they gain in speed they may lose in accuracy; however if the user is aware of their characteristics they can usually be employed effectively. One fast clustering procedure which requires only a single pass through the data is described in the following sub-section.

9.6.1 Single Pass Algorithm

Not all of the region to be clustered must be used in developing cluster centres but rather, for cost reduction, a randomly selected sample may be chosen and the samples arranged into a two dimensional array. The first row of samples is then used to obtain a

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

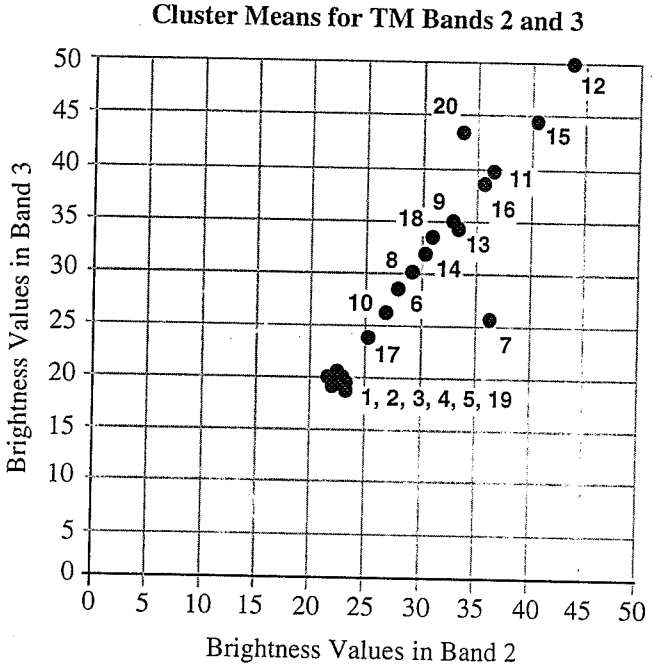


Figure 8-23 The mean vectors of the 20 clusters displayed in Figure 8-22 are shown here using only bands 2 and 3. The mean vector values are summarized in Table 8-9. Notice the substantial amount of overlap among clusters 1 through 5 and 19.

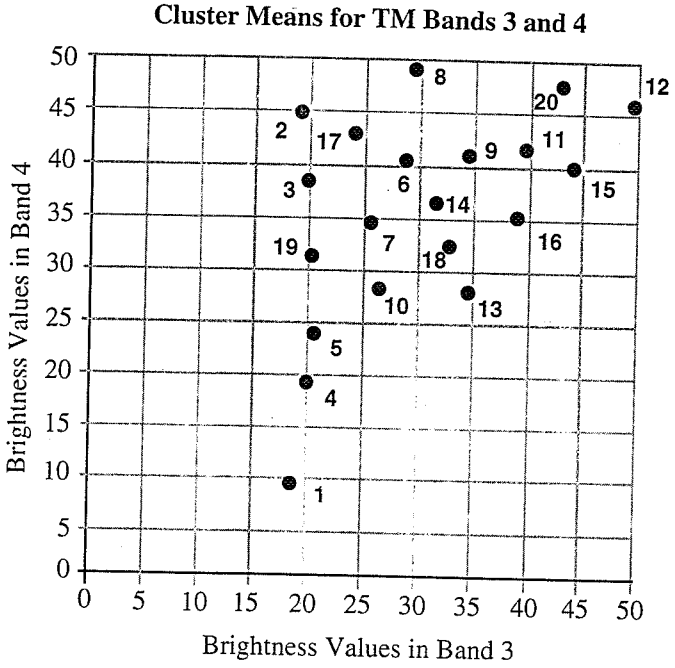


Figure 8-24 The mean vectors of the 20 clusters displayed in Figure 8-22 are shown here using only band 3 and 4 data. The mean vectors values are summarized in Table 8-9. Compare the spatial distribution of these 20 clusters in the red and near-infrared feature space with what is expected in a typical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

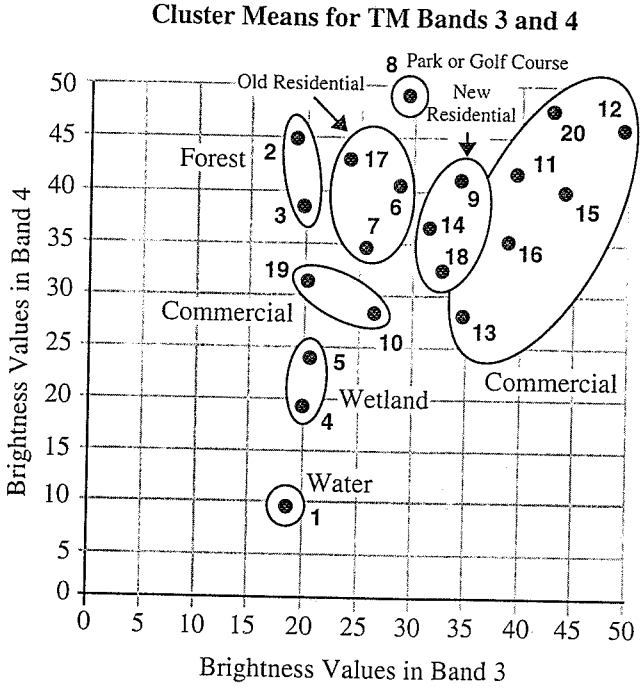


Figure 8-25 Grouping (relabeling) of the original 20 spectral clusters into information classes. The relabeling was performed by analyzing the mean vector locations in bands 3 and 4.

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

- *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

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

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
62 Nonforested Wetland
7 Barren Land
71 Dry Salt Flats
72 Beaches
73 Sandy Areas Other Than Beaches
74 Bare Exposed Rock
75 Strip Mines, Quarries, and Gravel Pits
76 Transitional Areas
77 Mixed Barren Land
8 Tundra
81 Shrub and Brush Tundra
82 Herbaceous Tundra
83 Bare Ground Tundra
84 Wet Tundra
85 Mixed Tundra
9 Perennial Snow or Ice
91 Perennial Snowfields
92 Glaciers

[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.