1 CONCEPTS AND FOUNDATIONS OF REMOTE SENSING

1.1 INTRODUCTION

Remote sensing is the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation. As you read these words, you are employing remote sensing. Your eyes are acting as sensors that respond to the light reflected from this page. The "data" your eyes acquire are impulses corresponding to the amount of light reflected from the dark and light areas on the page. These data are analyzed, or interpreted, in your mental computer to enable you to explain the dark areas on the page as a collection of letters forming words. Beyond this, you recognize that the words form sentences, and you interpret the information that the sentences convey.

In many respects, remote sensing can be thought of as a reading process. Using various sensors, we remotely collect *data* that may be analyzed to obtain *information* about the objects, areas, or phenomena being investigated. The remotely collected data can be of many forms, including variations in force distributions, acoustic wave distributions, or electromagnetic energy distributions. For example, a gravity meter acquires data on variations in the distribution of the

force of gravity. Sonar, like a bat's navigation system, obtains data on variations in acoustic wave distributions. Our eyes acquire data on variations in electromagnetic energy distributions.

Overview of the Electromagnetic Remote Sensing Process

This book is about *electromagnetic* energy sensors that are operated from airborne and spaceborne platforms to assist in inventorying, mapping, and monitoring earth resources. These sensors acquire data on the way various earth surface features emit and reflect electromagnetic energy, and these data are analyzed to provide information about the resources under investigation.

Figure 1.1 schematically illustrates the generalized processes and elements involved in electromagnetic remote sensing of earth resources. The two basic processes involved are *data acquisition* and *data analysis*. The elements of the data acquisition process are energy sources (*a*), propagation of energy through the atmosphere (*b*), energy interactions with earth surface features (*c*), retransmission of energy through the atmosphere (*d*), airborne and/or spaceborne sensors (*e*), resulting in the generation of sensor data in pictorial and/or digital form (*f*). In short, we use sensors to record variations in the way earth surface features reflect and emit electromagnetic energy. The data analysis process (*g*) involves examining the data using various viewing and interpretation devices to analyze pictorial data and/or a computer to analyze digital sensor data. Reference data about the resources being studied (such as soil maps, crop statistics, or field-check data) are used

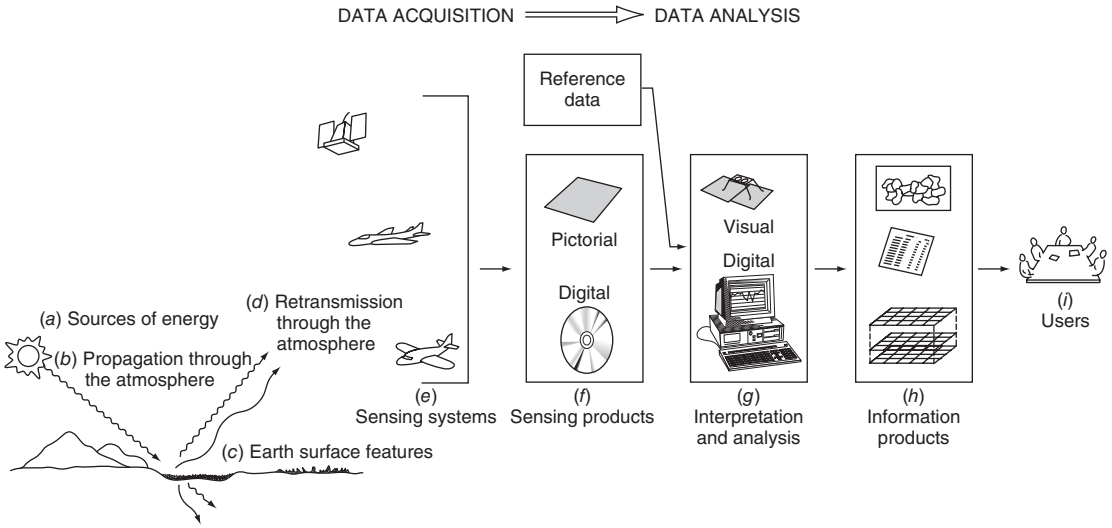


Figure 1.1 Electromagnetic remote sensing of earth resources.

when and where available to assist in the data analysis. With the aid of the reference data, the analyst extracts information about the type, extent, location, and condition of the various resources over which the sensor data were collected. This information is then compiled (*h*), generally in the form of maps, tables, or digital spatial data that can be merged with other "layers" of information in a *geographic information system* (*GIS*). Finally, the information is presented to users (*i*), who apply it to their decision-making process.

Organization of the Book

In the remainder of this chapter, we discuss the basic principles underlying the remote sensing process. We begin with the fundamentals of electromagnetic energy and then consider how the energy interacts with the atmosphere and with earth surface features. Next, we summarize the process of acquiring remotely sensed data and introduce the concepts underlying digital imagery formats. We also discuss the role that reference data play in the data analysis procedure and describe how the spatial location of reference data observed in the field is often determined using *Global Positioning System* (*GPS*) methods. These basics will permit us to conceptualize the strengths and limitations of "real" remote sensing systems and to examine the ways in which they depart from an "ideal" remote sensing system. We then discuss briefly the rudiments of GIS technology and the spatial frameworks (coordinate systems and datums) used to represent the positions of geographic features in space. Because visual examination of imagery will play an important role in every subsequent chapter of this book, this first chapter concludes with an overview of the concepts and processes involved in visual interpretation of remotely sensed images. By the end of this chapter, the reader should have a grasp of the foundations of remote sensing and an appreciation for the close relationship among remote sensing, GPS methods, and GIS operations.

Chapters 2 and 3 deal primarily with photographic remote sensing. Chapter 2 describes the basic tools used in acquiring aerial photographs, including both analog and digital camera systems. Digital videography is also treated in Chapter 2. Chapter 3 describes the photogrammetric procedures by which precise spatial measurements, maps, digital elevation models (DEMs), orthophotos, and other derived products are made from airphotos.

Discussion of nonphotographic systems begins in Chapter 4, which describes the acquisition of airborne multispectral, thermal, and hyperspectral data. In Chapter 5 we discuss the characteristics of spaceborne remote sensing systems and examine the principal satellite systems used to collect imagery from reflected and emitted radiance on a global basis. These satellite systems range from the Landsat and SPOT series of moderate-resolution instruments, to the latest generation of high-resolution commercially operated systems, to various meteorological and global monitoring systems.

Chapter 6 is concerned with the collection and analysis of radar and lidar data. Both airborne and spaceborne systems are discussed. Included in this latter category are such systems as the ALOS, Envisat, ERS, JERS, Radarsat, and ICESat satellite systems.

In essence, from Chapter 2 through Chapter 6, this book progresses from the simplest sensing systems to the more complex. There is also a progression from short to long wavelengths along the electromagnetic spectrum (see Section 1.2). That is, discussion centers on photography in the ultraviolet, visible, and near-infrared regions, multispectral sensing (including thermal sensing using emitted long-wavelength infrared radiation), and radar sensing in the microwave region.

The final two chapters of the book deal with the manipulation, interpretation, and analysis of images. Chapter 7 treats the subject of digital image processing and describes the most commonly employed procedures through which computer-assisted image interpretation is accomplished. Chapter 8 presents a broad range of applications of remote sensing, including both visual interpretation and computer-aided analysis of image data.

Throughout this book, the International System of Units (SI) is used. Tables are included to assist the reader in converting between SI and units of other measurement systems.

Finally, a Works Cited section provides a list of references cited in the text. It is not intended to be a compendium of general sources of additional information. Three appendices provided on the publisher's website (http://www.wiley.com/college/lillesand) offer further information about particular topics at a level of detail beyond what could be included in the text itself. Appendix A summarizes the various concepts, terms, and units commonly used in radiation measurement in remote sensing. Appendix B includes sample coordinate transformation and resampling procedures used in digital image processing. Appendix C discusses some of the concepts, terminology, and units used to describe radar signals.

1.2 ENERGY SOURCES AND RADIATION PRINCIPLES

Visible light is only one of many forms of electromagnetic energy. Radio waves, ultraviolet rays, radiant heat, and X-rays are other familiar forms. All this energy is inherently similar and propagates in accordance with basic wave theory. As shown in Figure 1.2, this theory describes electromagnetic energy as traveling in a harmonic, sinusoidal fashion at the "velocity of light" c. The distance from one wave peak to the next is the *wavelength* λ, and the number of peaks passing a fixed point in space per unit time is the wave *frequency* ν.

From basic physics, waves obey the general equation

$$c = \nu\lambda \tag{1.1}$$

Because c is essentially a constant $\left(3 \times 10^8\,\text{m/sec}\right)$, frequency ν and wavelength λ for any given wave are related inversely, and either term can be used to

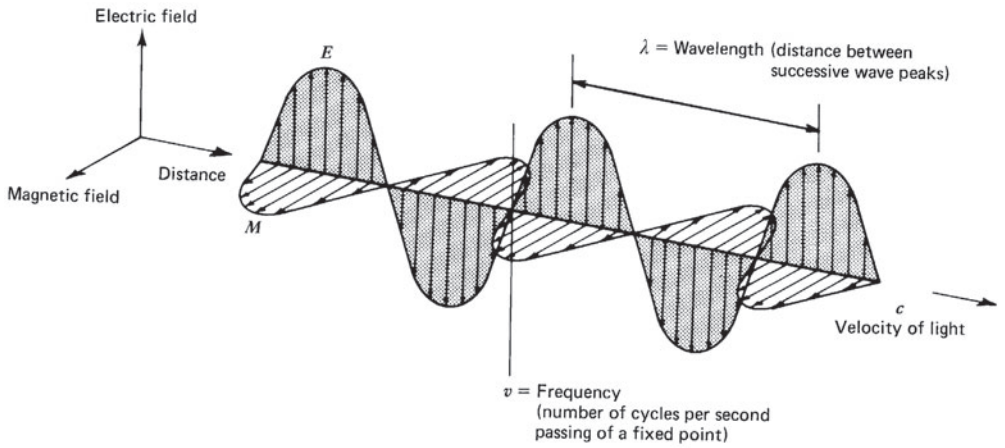


Figure 1.2 Electromagnetic wave. Components include a sinusoidal electric wave (E) and a similar magnetic wave (M) at right angles, both being perpendicular to the direction of propagation.

characterize a wave. In remote sensing, it is most common to categorize electromagnetic waves by their wavelength location within the *electromagnetic spectrum* (Figure 1.3). The most prevalent unit used to measure wavelength along the spectrum is the *micrometer* (μm). A micrometer equals 1×10^{-6} m.

Although names (such as "ultraviolet" and "microwave") are generally assigned to regions of the electromagnetic spectrum for convenience, there is no clear-cut dividing line between one nominal spectral region and the next. Divisions of the spectrum have grown from the various methods for sensing each type of radiation more so than from inherent differences in the energy characteristics of various wavelengths. Also, it should be noted that the portions of the

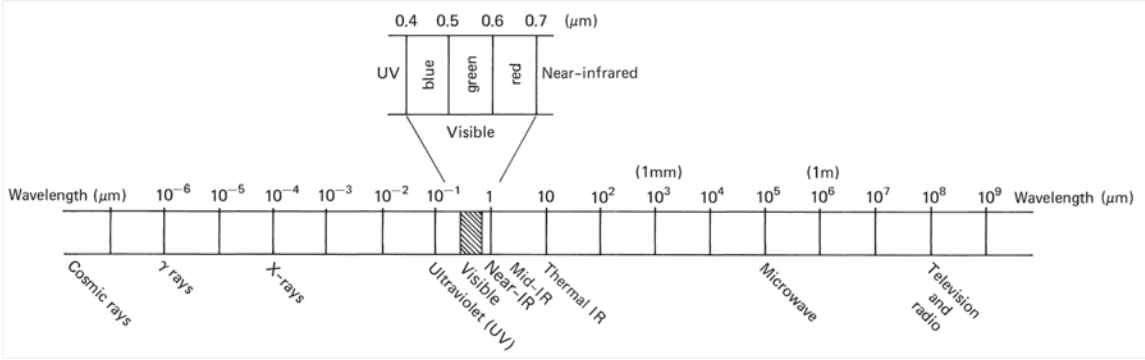


Figure 1.3 Electromagnetic spectrum.

electromagnetic spectrum used in remote sensing lie along a continuum characterized by magnitude changes of many powers of 10. Hence, the use of logarithmic plots to depict the electromagnetic spectrum is quite common. The "visible" portion of such a plot is an extremely small one, because the spectral sensitivity of the human eye extends only from about 0.4 μm to approximately 0.7 μm. The color "blue" is ascribed to the approximate range of 0.4 to 0.5 μm, "green" to 0.5 to 0.6 μm, and "red" to 0.6 to 0.7 μm. *Ultraviolet* (*UV*) energy adjoins the blue end of the visible portion of the spectrum. Beyond the red end of the visible region are three different categories of *infrared* (*IR*) waves: *near IR* (from 0.7 to 1.3 μm), *mid IR* (from 1.3 to 3 μm; also referred to as *shortwave IR* or *SWIR*), and *thermal IR* (beyond 3 to 14 μm, sometimes referred to as *longwave IR*). At much longer wavelengths (1 mm to 1 m) is the *microwave* portion of the spectrum.

Most common sensing systems operate in one or several of the visible, IR, or microwave portions of the spectrum. *Within the IR portion of the spectrum, it should be noted that only thermal-IR energy is directly related to the sensation of heat; near- and mid-IR energy are not*.

Although many characteristics of electromagnetic radiation are most easily described by wave theory, another theory offers useful insights into how electromagnetic energy interacts with matter. This theory—the particle theory—suggests that electromagnetic radiation is composed of many discrete units called *photons* or *quanta*. The energy of a quantum is given as

$$Q = hv \tag{1.2}$$

where

Q = energy of a quantum, joules (J)
h = Planck's constant, 6.626×10^{-34} J sec
v = frequency

We can relate the wave and quantum models of electromagnetic radiation behavior by solving Eq. 1.1 for v and substituting into Eq. 1.2 to obtain

$$Q = \frac{hc}{\lambda} \tag{1.3}$$

Thus, we see that the energy of a quantum is inversely proportional to its wavelength. *The longer the wavelength involved, the lower its energy content*. This has important implications in remote sensing from the standpoint that naturally emitted long wavelength radiation, such as microwave emission from terrain features, is more difficult to sense than radiation of shorter wavelengths, such as emitted thermal IR energy. The low energy content of long wavelength radiation means that, in general, systems operating at long wavelengths must "view" large areas of the earth at any given time in order to obtain a detectable energy signal.

The sun is the most obvious source of electromagnetic radiation for remote sensing. However, *all* matter at temperatures above absolute zero (0 K, or -273°C) continuously emits electromagnetic radiation. Thus, terrestrial objects are also

sources of radiation, although it is of considerably different magnitude and spectral composition than that of the sun. How much energy any object radiates is, among other things, a function of the surface temperature of the object. This property is expressed by the *Stefan–Boltzmann law*, which states that

$$M = \sigma T^4 \tag{1.4}$$

where

M = total radiant exitance from the surface of a material, watts (W) m^{-2}
σ = *Stefan–Boltzmann constant*, 5.6697×10^{-8} W m^{-2} K^{-4}
T = absolute temperature (K) of the emitting material

The particular units and the value of the constant are not critical for the student to remember, yet it is important to note that the total energy emitted from an object varies as T^4 and therefore increases very rapidly with increases in temperature. Also, it should be noted that this law is expressed for an energy source that behaves as a *blackbody*. A blackbody is a hypothetical, ideal radiator that totally absorbs and reemits all energy incident upon it. Actual objects only approach this ideal. We further explore the implications of this fact in Chapter 4; suffice it to say for now that the energy emitted from an object is primarily a function of its temperature, as given by Eq. 1.4.

Just as the total energy emitted by an object varies with temperature, the spectral distribution of the emitted energy also varies. Figure 1.4 shows energy distribution curves for blackbodies at temperatures ranging from 200 to 6000 K. The units on the ordinate scale $(W\,m^{-2}\mu m^{-1})$ express the radiant power coming from a blackbody per 1-μm spectral interval. Hence, the *area* under these curves equals the total radiant exitance, M, and the curves illustrate graphically what the Stefan–Boltzmann law expresses mathematically: The higher the temperature of the radiator, the greater the total amount of radiation it emits. The curves also show that there is a shift toward shorter wavelengths in the peak of a blackbody radiation distribution as temperature increases. The *dominant wavelength*, or wavelength at which a blackbody radiation curve reaches a maximum, is related to its temperature by *Wien's displacement law*,

$$\lambda_m = \frac{A}{T} \tag{1.5}$$

where

λ_m = wavelength of maximum spectral radiant exitance, μm
A = 2898 μm K
T = temperature, K

Thus, for a blackbody, the wavelength at which the maximum spectral radiant exitance occurs varies inversely with the blackbody's absolute temperature. We observe this phenomenon when a metal body such as a piece of iron is heated. As the object becomes progressively hotter, it begins to glow and its color

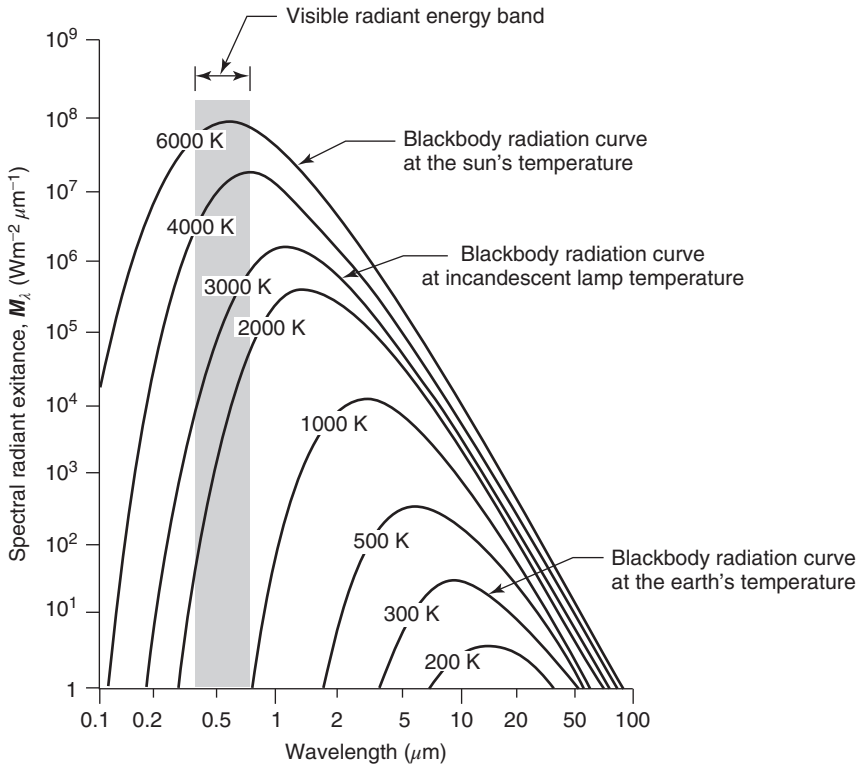


Figure 1.4 Spectral distribution of energy radiated from blackbodies of various temperatures. (Note that spectral radiant exitance M_λ is the energy emitted per unit wavelength interval. Total radiant exitance M is given by the area under the spectral radiant exitance curves.)

changes successively to shorter wavelengths—from dull red to orange to yellow and eventually to white.

The sun emits radiation in the same manner as a blackbody radiator whose temperature is about 6000 K (Figure 1.4). Many incandescent lamps emit radiation typified by a 3000 K blackbody radiation curve. Consequently, incandescent lamps have a relatively low output of blue energy, and they do not have the same spectral constituency as sunlight.

The earth's ambient temperature (i.e., the temperature of surface materials such as soil, water, and vegetation) is about 300 K (27°C). From Wien's displacement law, this means the maximum spectral radiant exitance from earth features occurs at a wavelength of about 9.7 μm. Because this radiation correlates with terrestrial heat, it is termed "thermal infrared" energy. This energy can neither be seen nor photographed, but it can be sensed with such thermal devices as radiometers and scanners (described in Chapter 4). By comparison, the sun has a much higher energy peak that occurs at about 0.5 μm, as indicated in Figure 1.4.

Our eyes—and photographic sensors—are sensitive to energy of this magnitude and wavelength. Thus, when the sun is present, we can observe earth features by virtue of *reflected* solar energy. Once again, the longer wavelength energy *emitted* by ambient earth features can be observed only with a nonphotographic sensing system. The general dividing line between reflected and emitted IR wavelengths is approximately 3 μm. Below this wavelength, reflected energy predominates; above it, emitted energy prevails.

Certain sensors, such as radar systems, supply their own source of energy to illuminate features of interest. These systems are termed "active" systems, in contrast to "passive" systems that sense naturally available energy. A very common example of an active system is a camera utilizing a flash. The same camera used in sunlight becomes a passive sensor.

1.3 ENERGY INTERACTIONS IN THE ATMOSPHERE

Irrespective of its source, all radiation detected by remote sensors passes through some distance, or *path length*, of atmosphere. The path length involved can vary widely. For example, space photography results from sunlight that passes through the full thickness of the earth's atmosphere twice on its journey from source to sensor. On the other hand, an airborne thermal sensor detects energy emitted directly from objects on the earth, so a single, relatively short atmospheric path length is involved. The net effect of the atmosphere varies with these differences in path length and also varies with the magnitude of the energy signal being sensed, the atmospheric conditions present, and the wavelengths involved.

Because of the varied nature of atmospheric effects, we treat this subject on a sensor-by-sensor basis in other chapters. Here, we merely wish to introduce the notion that the atmosphere can have a profound effect on, among other things, the intensity and spectral composition of radiation available to any sensing system. These effects are caused principally through the mechanisms of atmospheric *scattering* and *absorption*.

Scattering

Atmospheric scattering is the unpredictable diffusion of radiation by particles in the atmosphere. *Rayleigh scatter* is common when radiation interacts with atmospheric molecules and other tiny particles that are much smaller in diameter than the wavelength of the interacting radiation. The effect of Rayleigh scatter is inversely proportional to the fourth power of wavelength. Hence, there is a much stronger tendency for short wavelengths to be scattered by this mechanism than long wavelengths.

A "blue" sky is a manifestation of Rayleigh scatter. In the absence of scatter, the sky would appear black. But, as sunlight interacts with the earth's atmosphere,

it scatters the shorter (blue) wavelengths more dominantly than the other visible wavelengths. Consequently, we see a blue sky. At sunrise and sunset, however, the sun's rays travel through a longer atmospheric path length than during midday. With the longer path, the scatter (and absorption) of short wavelengths is so complete that we see only the less scattered, longer wavelengths of orange and red.

Rayleigh scatter is one of the primary causes of "haze" in imagery. Visually, haze diminishes the "crispness," or "contrast," of an image. In color photography, it results in a bluish-gray cast to an image, particularly when taken from high altitude. As we see in Chapter 2, haze can often be eliminated or at least minimized by introducing, in front of the camera lens, a filter that does not transmit short wavelengths.

Another type of scatter is *Mie scatter*, which exists when atmospheric particle diameters essentially equal the wavelengths of the energy being sensed. Water vapor and dust are major causes of Mie scatter. This type of scatter tends to influence longer wavelengths compared to Rayleigh scatter. Although Rayleigh scatter tends to dominate under most atmospheric conditions, Mie scatter is significant in slightly overcast ones.

A more bothersome phenomenon is *nonselective scatter*, which comes about when the diameters of the particles causing scatter are much larger than the wavelengths of the energy being sensed. Water droplets, for example, cause such scatter. They commonly have a diameter in the range 5 to 100 μm and scatter all visible and near- to mid-IR wavelengths about equally. Consequently, this scattering is "nonselective" with respect to wavelength. In the visible wavelengths, equal quantities of blue, green, and red light are scattered; hence fog and clouds appear white.

Absorption

In contrast to scatter, atmospheric absorption results in the effective loss of energy to atmospheric constituents. This normally involves absorption of energy at a given wavelength. The most efficient absorbers of solar radiation in this regard are water vapor, carbon dioxide, and ozone. Because these gases tend to absorb electromagnetic energy in specific wavelength bands, they strongly influence the design of any remote sensing system. The wavelength ranges in which the atmosphere is particularly transmissive of energy are referred to as *atmospheric windows*.

Figure 1.5 shows the interrelationship between energy sources and atmospheric absorption characteristics. Figure 1.5*a* shows the spectral distribution of the energy emitted by the sun and by earth features. These two curves represent the most common sources of energy used in remote sensing. In Figure 1.5*b*, spectral regions in which the atmosphere blocks energy are shaded. Remote sensing data acquisition is limited to the nonblocked spectral regions, the atmospheric windows. Note in Figure 1.5*c* that the spectral sensitivity range of the eye (the

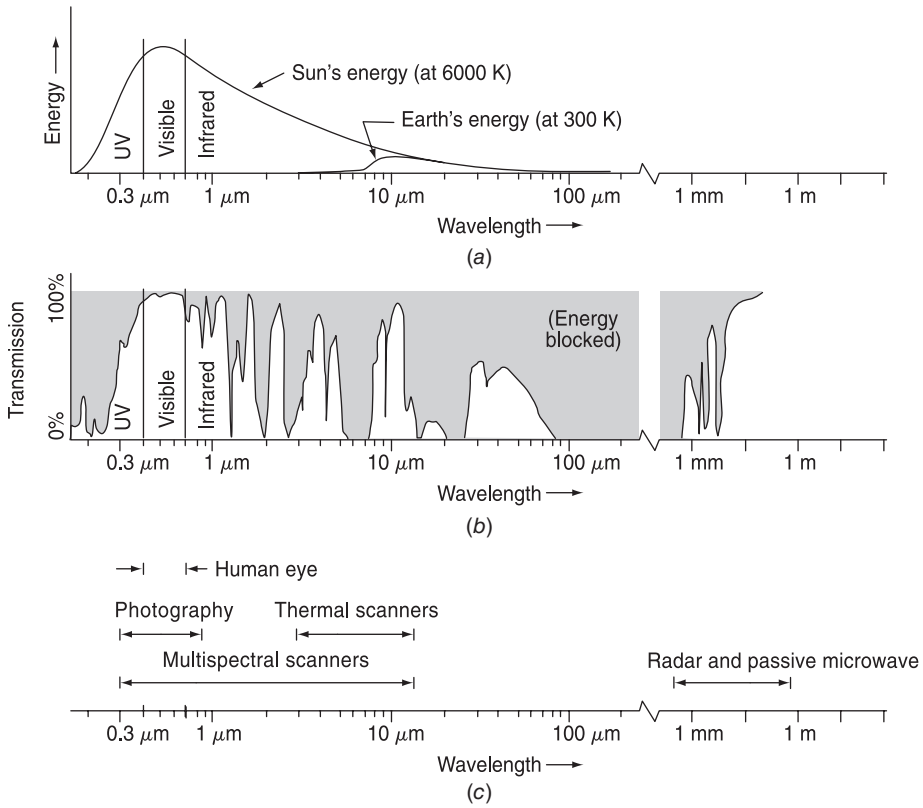


Figure 1.5 Spectral characteristics of (*a*) energy sources, (*b*) atmospheric transmittance, and (*c*) common remote sensing systems. (Note that wavelength scale is logarithmic.)

"visible" range) coincides with both an atmospheric window and the peak level of energy from the sun. Emitted "heat" energy from the earth, shown by the small curve in (*a*), is sensed through the windows at 3 to 5 μm and 8 to 14 μm using such devices as *thermal sensors*. *Multispectral sensors* observe simultaneously through multiple, narrow wavelength ranges that can be located at various points in the visible through the thermal spectral region. *Radar* and *passive microwave systems* operate through a window in the region 1 mm to 1 m.

The important point to note from Figure 1.5 is the interaction and the interdependence between the primary sources of electromagnetic energy, the atmospheric windows through which source energy may be transmitted to and from earth surface features, and the spectral sensitivity of the sensors available to detect and record the energy. One cannot select the sensor to be used in any given remote sensing task arbitrarily; one must instead consider (1) the spectral sensitivity of the sensors available, (2) the presence or absence of atmospheric windows in the spectral range(s) in which one wishes to sense, and (3) the source, magnitude, and

spectral composition of the energy available in these ranges. Ultimately, however, the choice of spectral range of the sensor must be based on the manner in which the energy interacts with the features under investigation. It is to this last, very important, element that we now turn our attention.

1.4 ENERGY INTERACTIONS WITH EARTH SURFACE FEATURES

When electromagnetic energy is incident on any given earth surface feature, three fundamental energy interactions with the feature are possible. These are illustrated in Figure 1.6 for an element of the volume of a water body. Various fractions of the energy incident on the element are *reflected*, *absorbed*, and/or *transmitted*. Applying the principle of conservation of energy, we can state the interrelationship among these three energy interactions as

$$E_{\mathrm{I}}(\lambda) = E_{\mathrm{R}}(\lambda) + E_{\mathrm{A}}(\lambda) + E_{\mathrm{T}}(\lambda) \tag{1.6}$$

where

E_{I} = incident energy
E_{R} = reflected energy
E_{A} = absorbed energy
E_{T} = transmitted energy

with all energy components being a function of wavelength λ.

Equation 1.6 is an energy balance equation expressing the interrelationship among the mechanisms of reflection, absorption, and transmission. Two points concerning this relationship should be noted. First, the proportions of energy reflected, absorbed, and transmitted will vary for different earth features, depending on their material type and condition. These differences permit us to distinguish different features on an image. Second, the wavelength dependency means that, even within a given feature type, the proportion of reflected, absorbed, and

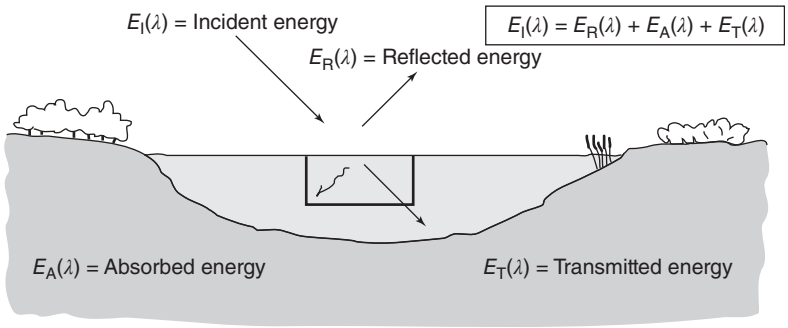


Figure 1.6 Basic interactions between electromagnetic energy and an earth surface feature.

transmitted energy will vary at different wavelengths. Thus, two features may be indistinguishable in one spectral range and be very different in another wavelength band. Within the visible portion of the spectrum, these spectral variations result in the visual effect called *color*. For example, we call objects "blue" when they reflect more highly in the blue portion of the spectrum, "green" when they reflect more highly in the green spectral region, and so on. Thus, the eye utilizes spectral variations in the magnitude of reflected energy to discriminate between various objects. Color terminology and color mixing principles are discussed further in Section 1.12.

Because many remote sensing systems operate in the wavelength regions in which reflected energy predominates, the reflectance properties of earth features are very important. Hence, it is often useful to think of the energy balance relationship expressed by Eq. 1.6 in the form

$$E_R(\lambda) = E_I(\lambda) - [E_A(\lambda) + E_T(\lambda)] \tag{1.7}$$

That is, the reflected energy is equal to the energy incident on a given feature reduced by the energy that is either absorbed or transmitted by that feature.

The reflectance characteristics of earth surface features may be quantified by measuring the portion of incident energy that is reflected. This is measured as a function of wavelength and is called *spectral reflectance*, ρ_λ. It is mathematically defined as

$$\begin{aligned} \rho_\lambda &= \frac{E_R(\lambda)}{E_I(\lambda)} \\ &= \frac{\text{energy of wavelength } \lambda \text{ reflected from the object}}{\text{energy of wavelength } \lambda \text{ incident upon the object}} \times 100 \end{aligned} \tag{1.8}$$

where ρ_λ is expressed as a percentage.

A graph of the spectral reflectance of an object as a function of wavelength is termed a *spectral reflectance curve*. The configuration of spectral reflectance curves gives us insight into the spectral characteristics of an object and has a strong influence on the choice of wavelength region(s) in which remote sensing data are acquired for a particular application. This is illustrated in Figure 1.7, which shows highly generalized spectral reflectance curves for deciduous versus coniferous trees. Note that the curve for each of these object types is plotted as a "ribbon" (or "envelope") of values, not as a single line. This is because spectral reflectances vary somewhat within a given material class. That is, the spectral reflectance of one deciduous tree species and another will never be identical, nor will the spectral reflectance of trees of the same species be exactly equal. We elaborate upon the variability of spectral reflectance curves later in this section.

In Figure 1.7, assume that you are given the task of selecting an airborne sensor system to assist in preparing a map of a forested area differentiating deciduous versus coniferous trees. One choice of sensor might be the human eye. However, there is a potential problem with this choice. The spectral reflectance curves for each tree

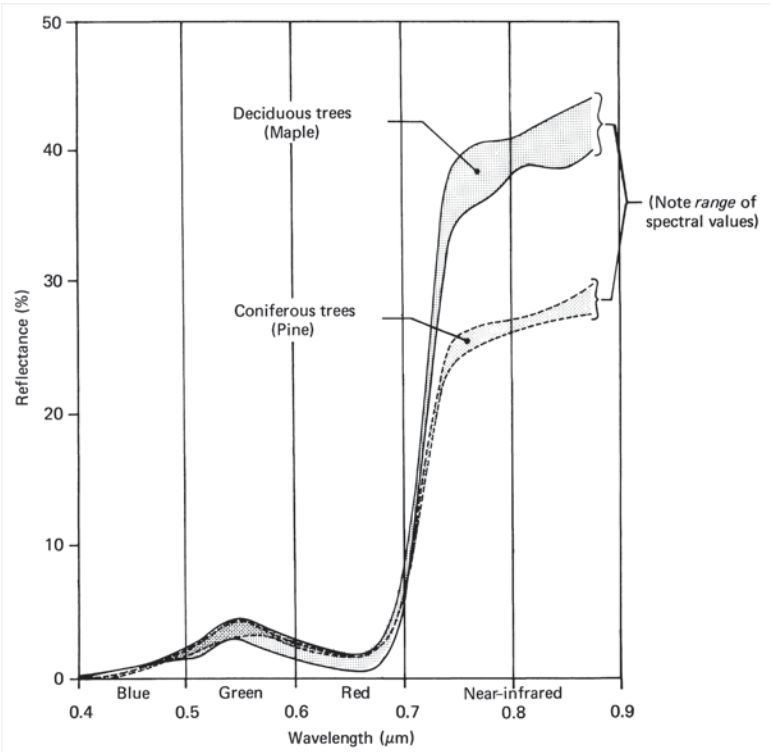


Figure 1.7 Generalized spectral reflectance envelopes for deciduous (broad-leaved) and coniferous (needle-bearing) trees. (Each tree type has a range of spectral reflectance values at any wavelength.) (Adapted from Kalensky and Wilson, 1975.)

type overlap in most of the visible portion of the spectrum and are very close where they do not overlap. Hence, the eye might see both tree types as being essentially the same shade of "green" and might confuse the identity of the deciduous and coniferous trees. Certainly one could improve things somewhat by using spatial clues to each tree type's identity, such as size, shape, site, and so forth. However, this is often difficult to do from the air, particularly when tree types are intermixed. How might we discriminate the two types on the basis of their spectral characteristics alone? We could do this by using a sensor that records near-IR energy. A specialized digital camera whose detectors are sensitive to near-IR wavelengths is just such a system, as is an analog camera loaded with black and white IR film. On near-IR images, deciduous trees (having higher IR reflectance than conifers) generally appear much lighter in tone than do conifers. This is illustrated in Figure 1.8, which shows stands of coniferous trees surrounded by deciduous trees. In Figure 1.8*a* (visible spectrum), it is virtually impossible to distinguish between tree types, even though the conifers have a distinctive conical shape whereas the deciduous trees have rounded crowns. In Figure 1.8*b* (near IR), the coniferous trees have a

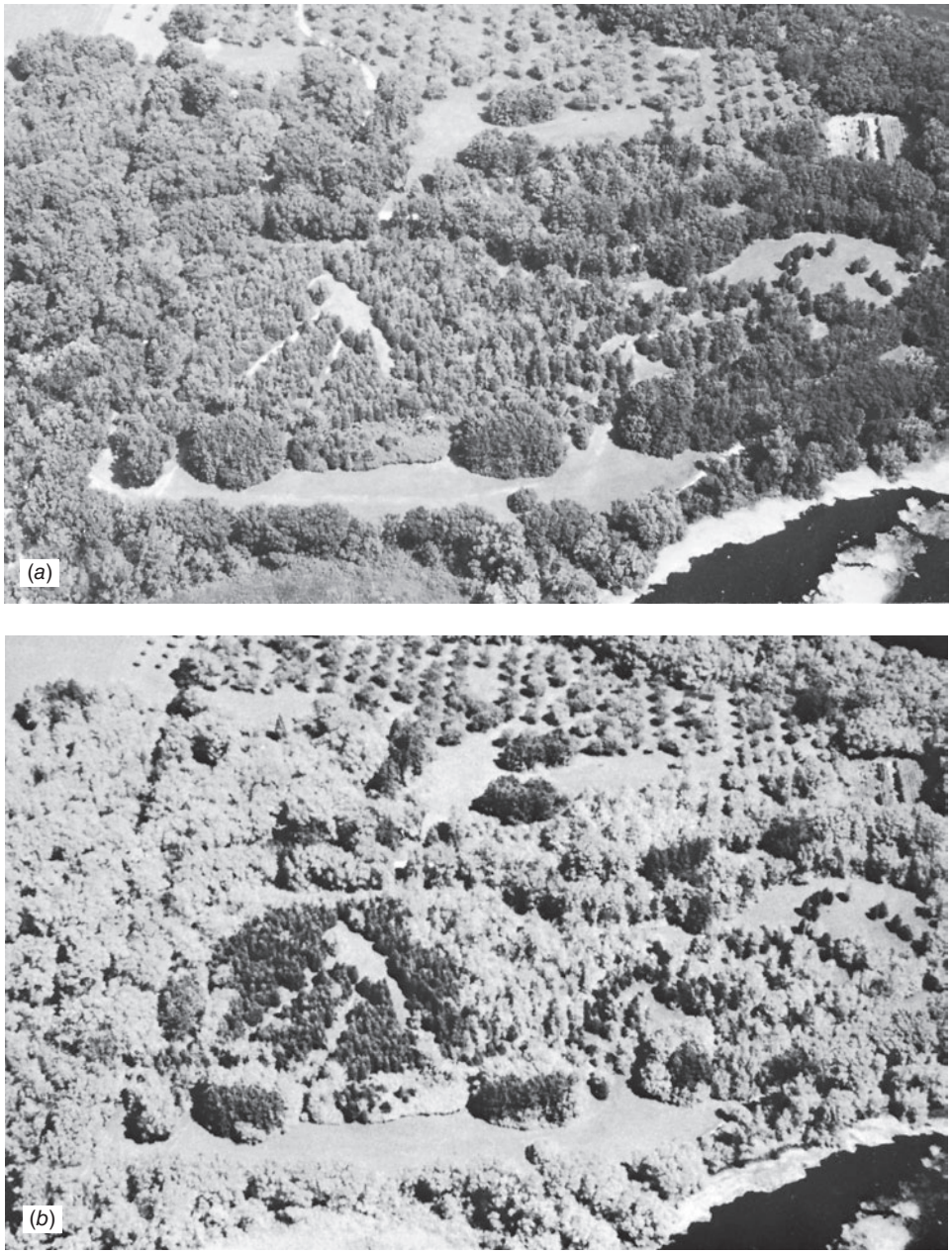


Figure 1.8 Low altitude oblique aerial photographs illustrating deciduous versus coniferous trees. (*a*) Panchromatic photograph recording reflected sunlight over the wavelength band 0.4 to 0.7 μm. (*b*) Black-and-white infrared photograph recording reflected sunlight over 0.7 to 0.9 μm wavelength band. (Author-prepared figure.)

distinctly darker tone. On such an image, the task of delineating deciduous versus coniferous trees becomes almost trivial. In fact, if we were to use a computer to analyze digital data collected from this type of sensor, we might "automate" our entire mapping task. Many remote sensing data analysis schemes attempt to do just that. For these schemes to be successful, the materials to be differentiated must be spectrally separable.

Experience has shown that many earth surface features of interest can be identified, mapped, and studied on the basis of their spectral characteristics. Experience has also shown that some features of interest cannot be spectrally separated. Thus, to utilize remote sensing data effectively, one must know and understand the spectral characteristics of the particular features under investigation in any given application. Likewise, one must know what factors influence these characteristics.

Spectral Reflectance of Earth Surface Feature Types

Figure 1.9 shows typical spectral reflectance curves for many different types of features: healthy green grass, dry (non-photosynthetically active) grass, bare soil (brown to dark-brown sandy loam), pure gypsum dune sand, asphalt, construction concrete (Portland cement concrete), fine-grained snow, clouds, and clear lake water. The lines in this figure represent *average* reflectance curves compiled by measuring a large sample of features, or in some cases *representative* reflectance measurements from a single typical example of the feature class. Note how distinctive the curves are for each feature. In general, the configuration of these curves is an indicator of the type and condition of the features to which they apply. Although the reflectance of individual features can vary considerably above and below the lines shown here, these curves demonstrate some fundamental points concerning spectral reflectance.

For example, spectral reflectance curves for healthy green vegetation almost always manifest the "peak-and-valley" configuration illustrated by green grass in Figure 1.9. The valleys in the visible portion of the spectrum are dictated by the pigments in plant leaves. Chlorophyll, for example, strongly absorbs energy in the wavelength bands centered at about 0.45 and 0.67 μm (often called the "chlorophyll absorption bands"). Hence, our eyes perceive healthy vegetation as green in color because of the very high absorption of blue and red energy by plant leaves and the relatively high reflection of green energy. If a plant is subject to some form of stress that interrupts its normal growth and productivity, it may decrease or cease chlorophyll production. The result is less chlorophyll absorption in the blue and red bands. Often, the red reflectance increases to the point that we see the plant turn yellow (combination of green and red). This can be seen in the spectral curve for dried grass in Figure 1.9.

As we go from the visible to the near-IR portion of the spectrum, the reflectance of healthy vegetation increases dramatically. This spectral feature, known as the *red edge*, typically occurs between 0.68 and 0.75 μm, with the exact position depending on the species and condition. Beyond this edge, from about

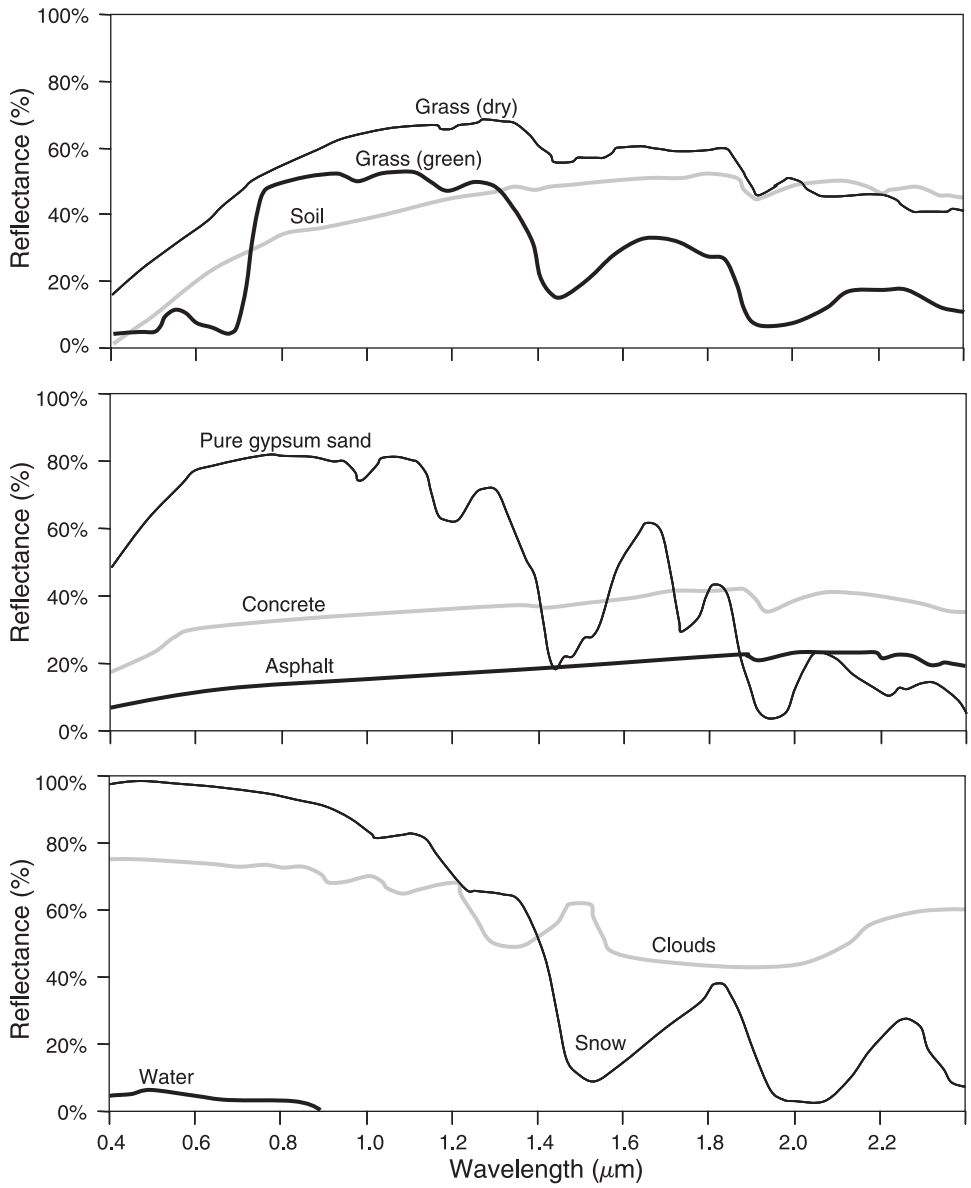


Figure 1.9 Spectral reflectance curves for various features types. (Original data courtesy USGS Spectroscopy Lab, Johns Hopkins University Spectral Library, and Jet Propulsion Laboratory [JPL]; cloud spectrum from Bowker et al., after Avery and Berlin, 1992. JPL spectra © 1999, California Institute of Technology.)

0.75 to 1.3 μm (representing most of the near-IR range), a plant leaf typically reflects 40 to 50% of the energy incident upon it. Most of the remaining energy is transmitted, because absorption in this spectral region is minimal (less than 5%). Plant reflectance from 0.75 to 1.3 μm results primarily from the internal structure of

plant leaves. Because the position of the red edge and the magnitude of the near-IR reflectance beyond the red edge are highly variable among plant species, reflectance measurements in these ranges often permit us to discriminate between species, even if they look the same in visible wavelengths. Likewise, many plant stresses alter the reflectance in the red edge and the near-IR region, and sensors operating in these ranges are often used for vegetation stress detection. Also, multiple layers of leaves in a plant canopy provide the opportunity for multiple transmissions and reflections. Hence, the near-IR reflectance increases with the number of layers of leaves in a canopy, with the maximum reflectance achieved at about eight leaf layers (Bauer et al., 1986).

Beyond 1.3 μm, energy incident upon vegetation is essentially absorbed or reflected, with little to no transmittance of energy. Dips in reflectance occur at 1.4, 1.9, and 2.7 μm because water in the leaf absorbs strongly at these wavelengths. Accordingly, wavelengths in these spectral regions are referred to as *water absorption bands*. Reflectance peaks occur at about 1.6 and 2.2 μm, between the absorption bands. Throughout the range beyond 1.3 μm, leaf reflectance is approximately inversely related to the total water present in a leaf. This total is a function of both the moisture content and the thickness of a leaf.

The soil curve in Figure 1.9 shows considerably less peak-and-valley variation in reflectance. That is, the factors that influence soil reflectance act over less specific spectral bands. Some of the factors affecting soil reflectance are moisture content, organic matter content, soil texture (proportion of sand, silt, and clay), surface roughness, and presence of iron oxide. These factors are complex, variable, and interrelated. For example, the presence of moisture in soil will decrease its reflectance. As with vegetation, this effect is greatest in the water absorption bands at about 1.4, 1.9, and 2.7 μm (clay soils also have hydroxyl absorption bands at about 1.4 and 2.2 μm). Soil moisture content is strongly related to the soil texture: Coarse, sandy soils are usually well drained, resulting in low moisture content and relatively high reflectance; poorly drained fine-textured soils will generally have lower reflectance. Thus, the reflectance properties of a soil are consistent only within particular ranges of conditions. Two other factors that reduce soil reflectance are surface roughness and content of organic matter. The presence of iron oxide in a soil will also significantly decrease reflectance, at least in the visible wavelengths. In any case, it is essential that the analyst be familiar with the conditions at hand. Finally, because soils are essentially opaque to visible and infrared radiation, it should be noted that soil reflectance comes from the uppermost layer of the soil and may not be indicative of the properties of the bulk of the soil.

Sand can have wide variation in its spectral reflectance pattern. The curve shown in Figure 1.9 is from a dune in New Mexico and consists of roughly 99% gypsum with trace amounts of quartz (Jet Propulsion Laboratory, 1999). Its absorption and reflectance features are essentially identical to those of its parent

material, gypsum. Sand derived from other sources, with differing mineral compositions, would have a spectral reflectance curve indicative of its parent material. Other factors affecting the spectral response from sand include the presence or absence of water and of organic matter. Sandy soil is subject to the same considerations listed in the discussion of soil reflectance.

As shown in Figure 1.9, the spectral reflectance curves for asphalt and Portland cement concrete are much flatter than those of the materials discussed thus far. Overall, Portland cement concrete tends to be relatively brighter than asphalt, both in the visible spectrum and at longer wavelengths. It is important to note that the reflectance of these materials may be modified by the presence of paint, soot, water, or other substances. Also, as materials age, their spectral reflectance patterns may change. For example, the reflectance of many types of asphaltic concrete may increase, particularly in the visible spectrum, as their surface ages.

In general, snow reflects strongly in the visible and near infrared, and absorbs more energy at mid-infrared wavelengths. However, the reflectance of snow is affected by its grain size, liquid water content, and presence or absence of other materials in or on the snow surface (Dozier and Painter, 2004). Larger grains of snow absorb more energy, particularly at wavelengths longer than 0.8 μm. At temperatures near 0°C, liquid water within the snowpack can cause grains to stick together in clusters, thus increasing the effective grain size and decreasing the reflectance at near-infrared and longer wavelengths. When particles of contaminants such as dust or soot are deposited on snow, they can significantly reduce the surface's reflectance in the visible spectrum.

The aforementioned absorption of mid-infrared wavelengths by snow can permit the differentiation between snow and clouds. While both feature types appear bright in the visible and near infrared, clouds have significantly higher reflectance than snow at wavelengths longer than 1.4 μm. Meteorologists can also use both spectral and bidirectional reflectance patterns (discussed later in this section) to identify a variety of cloud properties, including ice/water composition and particle size.

Considering the spectral reflectance of water, probably the most distinctive characteristic is the energy absorption at near-IR wavelengths and beyond. In short, water absorbs energy in these wavelengths whether we are talking about water features per se (such as lakes and streams) or water contained in vegetation or soil. Locating and delineating water bodies with remote sensing data are done most easily in near-IR wavelengths because of this absorption property. However, various conditions of water bodies manifest themselves primarily in visible wavelengths. The energy–matter interactions at these wavelengths are very complex and depend on a number of interrelated factors. For example, the reflectance from a water body can stem from an interaction with the water's surface (specular reflection), with material suspended in the water, or with the bottom of the depression containing the water body. Even with

deep water where bottom effects are negligible, the reflectance properties of a water body are a function of not only the water per se but also the material in the water.

Clear water absorbs relatively little energy having wavelengths less than about 0.6 μm. High transmittance typifies these wavelengths with a maximum in the blue-green portion of the spectrum. However, as the turbidity of water changes (because of the presence of organic or inorganic materials), transmittance—and therefore reflectance—changes dramatically. For example, waters containing large quantities of suspended sediments resulting from soil erosion normally have much higher visible reflectance than other "clear" waters in the same geographic area. Likewise, the reflectance of water changes with the chlorophyll concentration involved. Increases in chlorophyll concentration tend to decrease water reflectance in blue wavelengths and increase it in green wavelengths. These changes have been used to monitor the presence and estimate the concentration of algae via remote sensing data. Reflectance data have also been used to determine the presence or absence of tannin dyes from bog vegetation in lowland areas and to detect a number of pollutants, such as oil and certain industrial wastes.

Figure 1.10 illustrates some of these effects, using spectra from three lakes with different bio-optical properties. The first spectrum is from a clear, oligotrophic lake with a chlorophyll level of 1.2 μg/l and only 2.4 mg/l of dissolved organic carbon (DOC). Its spectral reflectance is relatively high in the blue-green portion of the spectrum and decreases in the red and near infrared. In contrast,

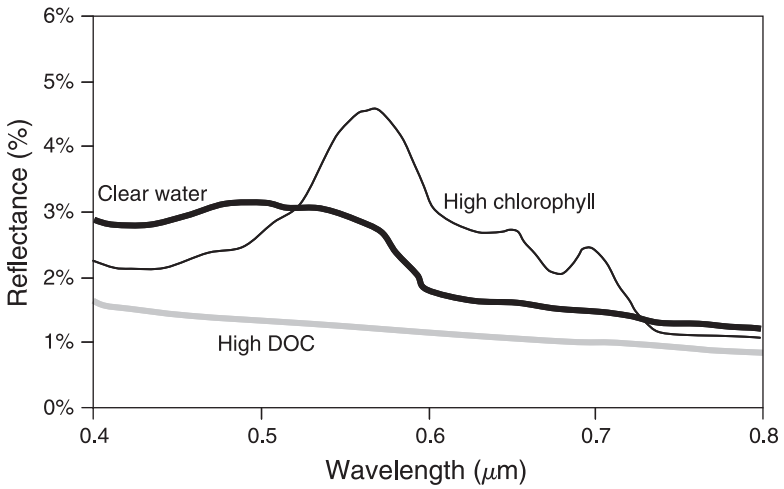


Figure 1.10 Spectral reflectance curves for lakes with clear water, high levels of chlorophyll, and high levels of dissolved organic carbon (DOC).

the spectrum from a lake experiencing an algae bloom, with much higher chlorophyll concentration (12.3 μg/l), shows a reflectance peak in the green spectrum and absorption in the blue and red regions. These reflectance and absorption features are associated with several pigments present in algae. Finally, the third spectrum in Figure 1.10 was acquired on an ombrotrophic bog lake, with very high levels of DOC (20.7 mg/l). These naturally occurring tannins and other complex organic molecules give the lake a very dark appearance, with its reflectance curve nearly flat across the visible spectrum.

Many important water characteristics, such as dissolved oxygen concentration, pH, and salt concentration, cannot be observed directly through changes in water reflectance. However, such parameters sometimes correlate with observed reflectance. In short, there are many complex interrelationships between the spectral reflectance of water and particular characteristics. One must use appropriate reference data to correctly interpret reflectance measurements made over water.

Our discussion of the spectral characteristics of vegetation, soil, and water has been very general. The student interested in pursuing details on this subject, as well as factors influencing these characteristics, is encouraged to consult the various references contained in the Works Cited section located at the end of this book.

Spectral Response Patterns

Having looked at the spectral reflectance characteristics of vegetation, soil, sand, concrete, asphalt, snow, clouds, and water, we should recognize that these broad feature types are often spectrally separable. However, the degree of separation between types varies among and within spectral regions. For example, water and vegetation might reflect nearly equally in visible wavelengths, yet these features are almost always separable in near-IR wavelengths.

Because spectral responses measured by remote sensors over various features often permit an assessment of the type and/or condition of the features, these responses have often been referred to as *spectral signatures*. Spectral reflectance and spectral emittance curves (for wavelengths greater than 3.0 μm) are often referred to in this manner. The physical radiation measurements acquired over specific terrain features at various wavelengths are also referred to as the spectral signatures for those features.

Although it is true that many earth surface features manifest very distinctive spectral reflectance and/or emittance characteristics, these characteristics result in spectral "response patterns" rather than in spectral "signatures." The reason for this is that the term *signature* tends to imply a pattern that is absolute and unique. This is not the case with the spectral patterns observed in the natural world. As we have seen, spectral response patterns measured by remote sensors may be

quantitative, but they are not absolute. They may be distinctive, but they are not necessarily unique.

We have already looked at some characteristics of objects that influence their spectral response patterns. *Temporal effects* and *spatial effects* can also enter into any given analysis. Temporal effects are any factors that change the spectral characteristics of a feature over time. For example, the spectral characteristics of many species of vegetation are in a nearly continual state of change throughout a growing season. These changes often influence when we might collect sensor data for a particular application.

Spatial effects refer to factors that cause the same types of features (e.g., corn plants) at a given point in *time* to have different characteristics at different geographic *locations*. In small-area analysis the geographic locations may be meters apart and spatial effects may be negligible. When analyzing satellite data, the locations may be hundreds of kilometers apart where entirely different soils, climates, and cultivation practices might exist.

Temporal and spatial effects influence virtually all remote sensing operations. These effects normally complicate the issue of analyzing spectral reflectance properties of earth resources. Again, however, temporal and spatial effects might be the keys to gleaning the information sought in an analysis. For example, the process of *change detection* is premised on the ability to measure temporal effects. An example of this process is detecting the change in suburban development near a metropolitan area by using data obtained on two different dates.

An example of a useful spatial effect is the change in the leaf morphology of trees when they are subjected to some form of stress. For example, when a tree becomes infected with Dutch elm disease, its leaves might begin to cup and curl, changing the reflectance of the tree relative to healthy trees that surround it. So, even though a spatial effect might cause differences in the spectral reflectances of the same type of feature, this effect may be just what is important in a particular application.

Finally, it should be noted that the apparent spectral response from surface features can be influenced by shadows. While an object's spectral reflectance (a ratio of reflected to incident energy, see Eq. 1.8) is not affected by changes in illumination, the absolute amount of energy reflected does depend on illumination conditions. Within a shadow, the total reflected energy is reduced, and the spectral response is shifted toward shorter wavelengths. This occurs because the incident energy within a shadow comes primarily from Rayleigh atmospheric scattering, and as discussed in Section 1.3, such scattering primarily affects short wavelengths. Thus, in visible-wavelength imagery, objects inside shadows will tend to appear both darker and bluer than if they were fully illuminated. This effect can cause problems for automated image classification algorithms; for example, dark shadows of trees on pavement may be misclassified as water. The effects of illumination geometry on reflectance are discussed in more detail later in this section, while the impacts of shadows on the image interpretation process are discussed in Section 1.12.

Atmospheric Influences on Spectral Response Patterns

In addition to being influenced by temporal and spatial effects, spectral response patterns are influenced by the atmosphere. Regrettably, the energy recorded by a sensor is always modified to some extent by the atmosphere between the sensor and the ground. We will indicate the significance of this effect on a sensor-by-sensor basis throughout this book. For now, Figure 1.11 provides an initial frame of reference for understanding the nature of atmospheric effects. Shown in this figure is the typical situation encountered when a sensor records reflected solar energy. The atmosphere affects the "brightness," or *radiance*, recorded over any given point on the ground in two almost contradictory ways. First, it attenuates (reduces) the energy illuminating a ground object (and being reflected from the object). Second, the atmosphere acts as a reflector itself, adding a scattered, extraneous *path radiance* to the signal detected by the sensor. By expressing these two atmospheric effects mathematically, the total radiance recorded by the sensor may be related to the reflectance of the ground object and the incoming radiation or *irradiance* using the equation

$$L_{\text{tot}} = \frac{\rho E \text{T}}{\pi} + L_p \tag{1.9}$$

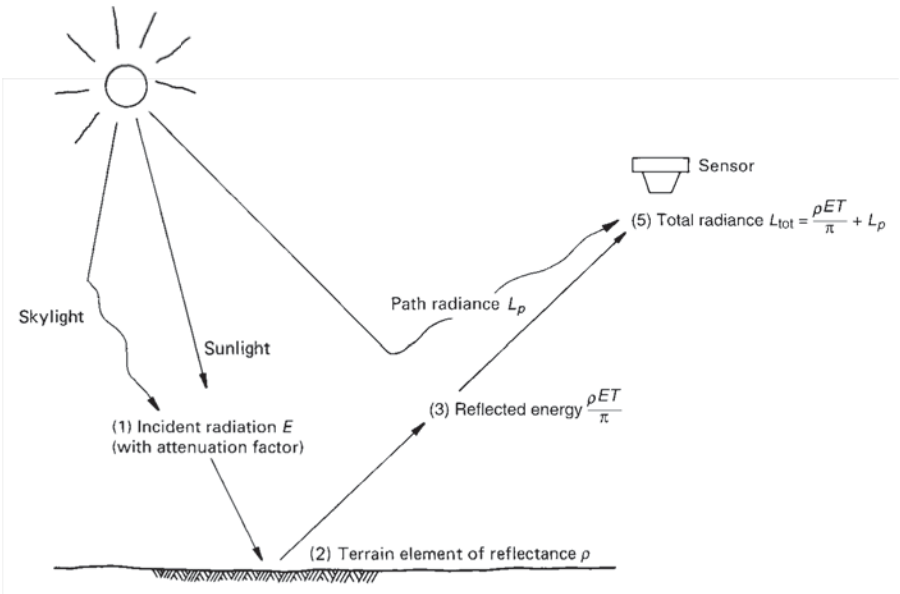


Figure 1.11 Atmospheric effects influencing the measurement of reflected solar energy. Attenuated sunlight and skylight (E) is reflected from a terrain element having reflectance ρ. The attenuated radiance reflected from the terrain element ($\rho ET/\pi$) combines with the path radiance (L_p) to form the total radiance (L_{tot}) recorded by the sensor.

where

L_{tot} = total spectral radiance measured by sensor
ρ = reflectance of object
E = irradiance on object, incoming energy
T = transmission of atmosphere
L_p = path radiance, from the atmosphere and not from the object

It should be noted that all of the above factors depend on wavelength. Also, as shown in Figure 1.11, the irradiance (E) stems from two sources: (1) directly reflected "sunlight" and (2) diffuse "skylight," which is sunlight that has been previously scattered by the atmosphere. The relative dominance of sunlight versus skylight in any given image is strongly dependent on weather conditions (e.g., sunny vs. hazy vs. cloudy). Likewise, irradiance varies with the seasonal changes in solar elevation angle (Figure 7.4) and the changing distance between the earth and sun.

For a sensor positioned close to the earth's surface, the path radiance L_p will generally be small or negligible, because the atmospheric path length from the surface to the sensor is too short for much scattering to occur. In contrast, imagery from satellite systems will be more strongly affected by path radiance, due to the longer atmospheric path between the earth's surface and the spacecraft. This can be seen in Figure 1.12, which compares two spectral response patterns from the same area. One "signature" in this figure was collected using a handheld field spectroradiometer (see Section 1.6 for discussion), from a distance of only a few cm above the surface. The second curve shown in Figure 1.12 was collected by the Hyperion hyperspectral sensor on the EO-1 satellite (hyperspectral systems

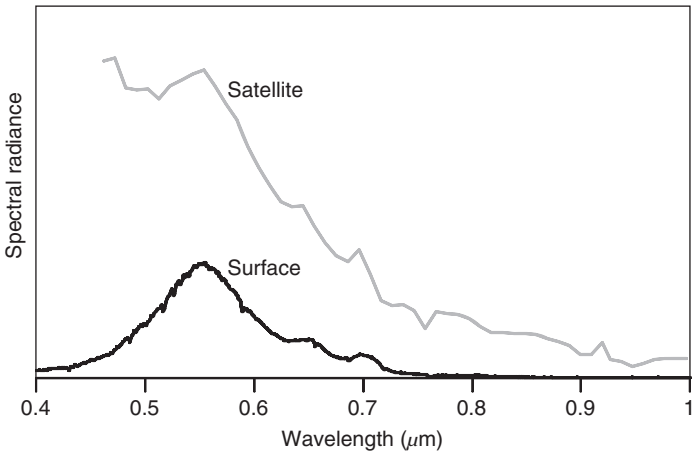


Figure 1.12 Spectral response patterns measured using a field spectroradiometer in close proximity to the earth's surface, and from above the top of the atmosphere (via the Hyperion instrument on EO-1). The difference between the two "signatures" is caused by atmospheric scattering and absorption in the Hyperion image.

are discussed in Chapter 4, and the Hyperion instrument is covered in Chapter 5). Due to the thickness of the atmosphere between the earth's surface and the satellite's position above the atmosphere, this second spectral response pattern shows an elevated signal at short wavelengths, due to the extraneous path radiance.

In its raw form, this near-surface measurement from the field spectroradiometer could not be directly compared to the measurement from the satellite, because one is observing *surface reflectance* while the other is observing the so-called *top of atmosphere* (*TOA*) reflectance. Before such a comparison could be performed, the satellite image would need to go through a process of *atmospheric correction*, in which the raw spectral data are modified to compensate for the expected effects of atmospheric scattering and absorption. This process, discussed in Chapter 7, generally does not produce a perfect representation of the spectral response curve that would actually be observed at the surface itself, but it can produce a sufficiently close approximation to be suitable for many types of analysis.

Readers who might be interested in obtaining additional details about the concepts, terminology, and units used in radiation measurement may wish to consult Appendix A.

Geometric Influences on Spectral Response Patterns

The geometric manner in which an object reflects energy is an important consideration. This factor is primarily a function of the surface roughness of the object. *Specular* reflectors are flat surfaces that manifest mirror-like reflections, where the angle of reflection equals the angle of incidence. *Diffuse* (or *Lambertian*) reflectors are rough surfaces that reflect uniformly in all directions. Most earth surfaces are neither perfectly specular nor perfectly diffuse reflectors. Their characteristics are somewhat between the two extremes.

Figure 1.13 illustrates the geometric character of specular, near-specular, near-diffuse, and diffuse reflectors. The category that describes any given surface is dictated by the surface's roughness *in comparison to the wavelength of the energy being sensed*. For example, in the relatively long wavelength radio range, a sandy beach can appear smooth to incident energy, whereas in the visible portion of the spectrum, it appears rough. In short, when the wavelength of incident energy is much smaller than the surface height variations or the particle sizes that make up a surface, the reflection from the surface is diffuse.

Diffuse reflections contain spectral information on the "color" of the reflecting surface, whereas specular reflections generally do not. *Hence, in remote sensing, we are most often interested in measuring the diffuse reflectance properties of terrain features*.

Because most features are not perfect diffuse reflectors, however, it becomes necessary to consider the viewing and illumination geometry. Figure 1.14 illustrates the relationships that exist among *solar elevation*, *azimuth angle*, and *viewing angle*. Figure 1.15 shows some typical geometric effects that can influence the

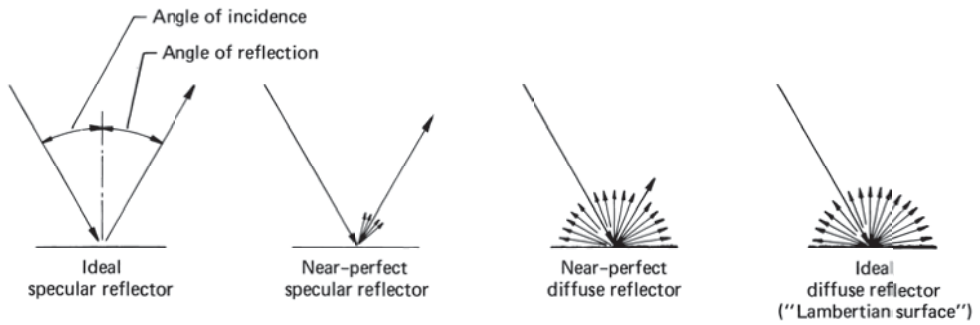


Figure 1.13 Specular versus diffuse reflectance. (We are most often interested in measuring the diffuse reflectance of objects.)

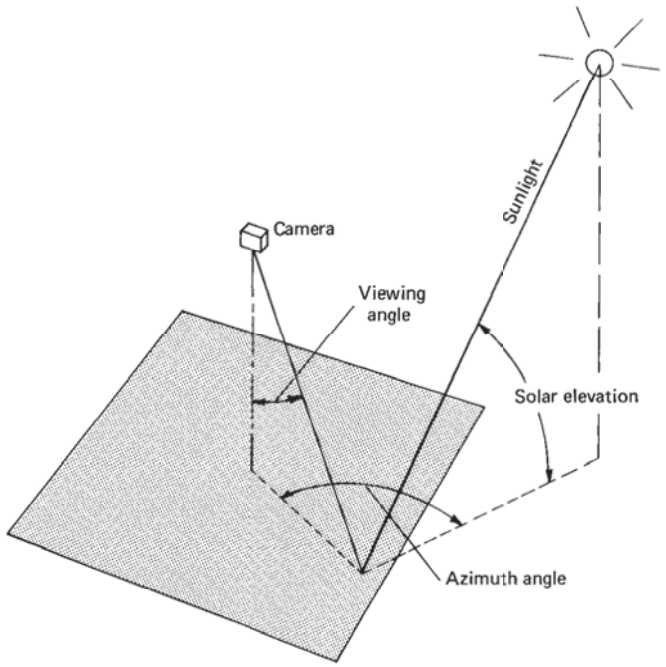


Figure 1.14 Sun-object-image angular relationship.

apparent reflectance in an image. In (*a*), the effect of *differential shading* is illustrated in profile view. Because the sides of features may be either sunlit or shaded, variations in brightness can result from identical ground objects at different locations in the image. The sensor receives more energy from the sunlit side of the tree at *B* than from the shaded side of the tree at *A*. Differential shading is clearly a function of solar elevation and object height, with a stronger effect at

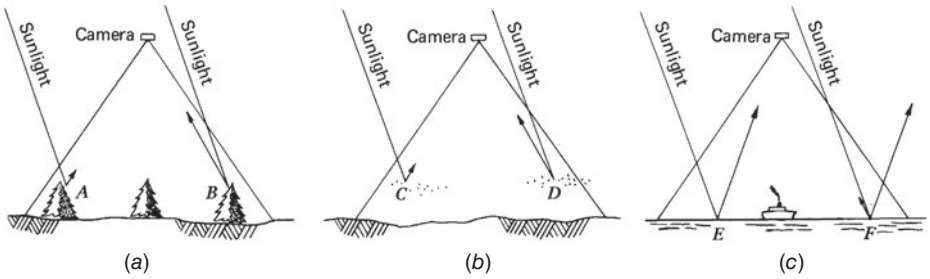


Figure 1.15 Geometric effects that cause variations in focal plane irradiance: (*a*) differential shading, (*b*) differential scattering, and (*c*) specular reflection.

low solar angles. The effect is also compounded by differences in slope and aspect (slope orientation) over terrain of varied relief.

Figure 1.15*b* illustrates the effect of *differential atmospheric scattering*. As discussed earlier, backscatter from atmospheric molecules and particles adds light (path radiance) to that reflected from ground features. The sensor records more atmospheric backscatter from area *D* than from area *C* due to this geometric effect. In some analyses, the variation in this path radiance component is small and can be ignored, particularly at long wavelengths. However, under hazy conditions, differential quantities of path radiance often result in varied illumination across an image.

As mentioned earlier, specular reflections represent the extreme in directional reflectance. When such reflections appear, they can hinder analysis of the imagery. This can often be seen in imagery taken over water bodies. Figure 1.15*c* illustrates the geometric nature of this problem. Immediately surrounding point *E* on the image, a considerable increase in brightness would result from specular reflection. A photographic example of this is shown in Figure 1.16, which includes areas of specular reflection from the right half of the large lake in the center of the image. These mirrorlike reflections normally contribute little information about the true character of the objects involved. For example, the small water bodies just below the larger lake have a tone similar to that of some of the fields in the area. Because of the low information content of specular reflections, they are avoided in most analyses.

The most complete representation of an object's geometric reflectance properties is the *bidirectional reflectance distribution function* (*BRDF*). This is a mathematical description of how reflectance varies for all combinations of illumination and viewing angles at a given wavelength (Schott, 2007). The BRDF for any given feature can approximate that of a Lambertian surface at some angles and be non-Lambertian at other angles. Similarly, the BRDF can vary considerably with wavelength. A variety of mathematical models (including a provision for wavelength dependence) have been proposed to represent the BRDF (Jupp and Strahler, 1991).

Figure 1.16 Aerial photograph containing areas of specular reflection from water bodies. This image is a portion of a summertime photograph taken over Green Lake, Green Lake County, WI. Scale 1:95,000. Cloud shadows indicate direction of sunlight at time of exposure. Reproduced from color IR original. (NASA image.)

Figure 1.17*a* shows graphic representations of the BRDF for three objects, each of which can be visualized as being located at the point directly beneath the center of one of the hemispheres. In each case, illumination is from the south (located at the back right, in these perspective views). The brightness at any point on the hemisphere indicates the relative reflectance of the object at a given viewing angle. A perfectly diffuse reflector (Figure 1.17*a*, top) has uniform reflectance in all directions. At the other extreme, a specular reflector (bottom) has very high reflectance in the direction directly opposite the source of illumination, and very low reflectance in all other directions. An intermediate surface (middle) has somewhat elevated reflectance at the specular angle but also shows some reflectance in other directions.

Figure 1.17*b* shows a geometric reflectance pattern dominated by backscattering, in which reflectance is highest when viewed from the same direction as the source of illumination. (This is in contrast to the intermediate and specular examples from Figure 1.17*a*, in which forward scattering predominates.) Many natural surfaces display this pattern of backscattering as a result of the differential shading (Figure 1.16*a*). In an image of a relatively uniform surface, there may be a localized area of increased brightness (known as a "hotspot"), located where the azimuth and zenith angles of the sensor are the same as those of the

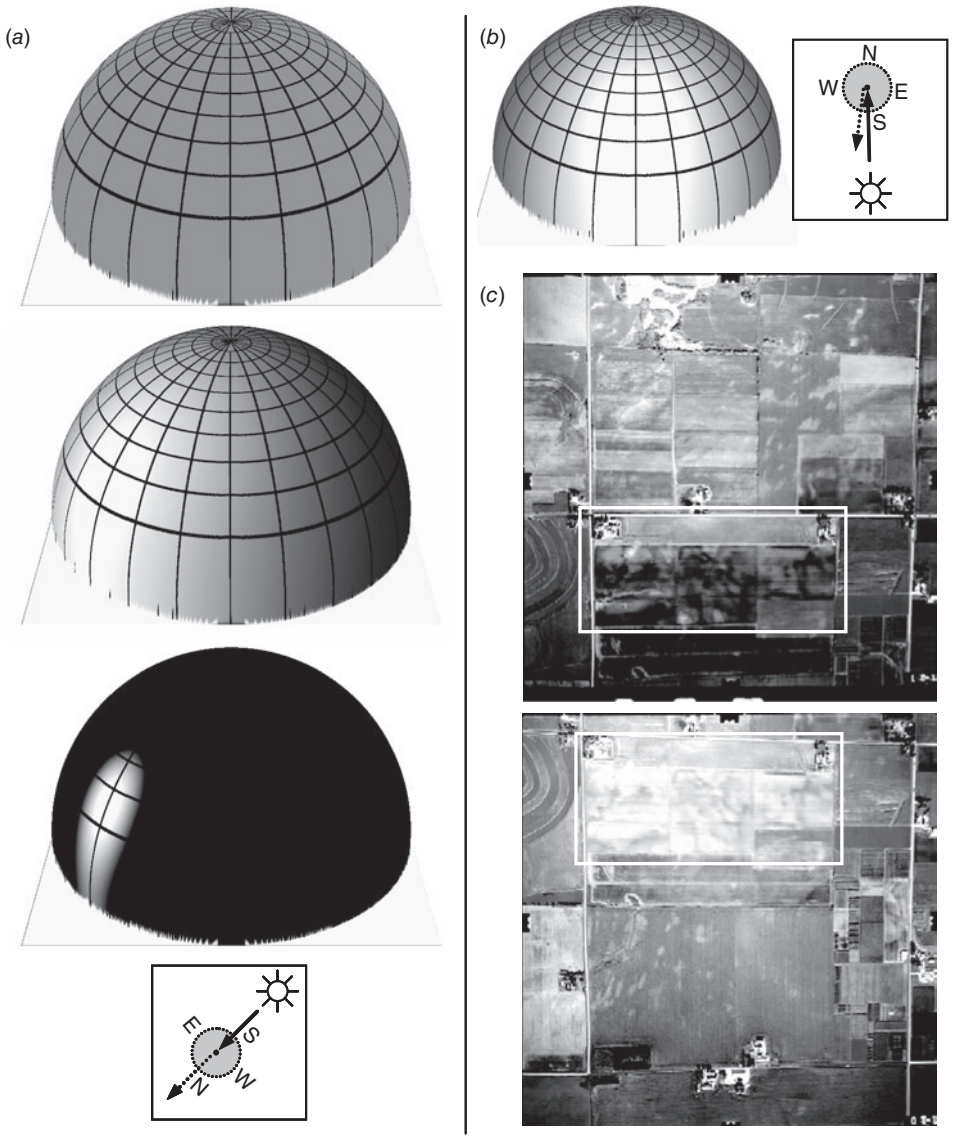


Figure 1.17 (*a*) Visual representation of bidirectional reflectance patterns, for surfaces with Lambertian (top), intermediate (middle), and specular (bottom) characteristics (after Campbell, 2002). (*b*) Simulated bidirectional reflectance from an agricultural field, showing a "hotspot" when viewed from the direction of solar illumination. (*c*) Differences in apparent reflectance in a field, when photographed from the north (top) and the south (bottom). (Author-prepared figure.)

sun. The existence of the hotspot is due to the fact that the sensor is then viewing only the sunlit portion of all objects in the area, without any shadowing.

An example of this type of hotspot is shown in Figure 1.17*c*. The two aerial photographs shown were taken mere seconds apart, along a single north–south

flight line. The field delineated by the white box has a great difference in its apparent reflectance in the two images, despite the fact that no actual changes occurred on the ground during the short interval between the exposures. In the top photograph, the field was being viewed from the north, opposite the direction of solar illumination. Roughness of the field's surface results in differential shading, with the camera viewing the shadowed side of each small variation in the field's surface. In contrast, the bottom photograph was acquired from a point to the south of the field, from the same direction as the solar illumination (the hotspot), and thus appears quite bright.

To summarize, variations in bidirectional reflectance—such as specular reflection from a lake, or the hotspot in an agricultural field—can significantly affect the appearance of objects in remotely sensed images. These effects cause objects to appear brighter or darker solely as a result of the angular relationships among the sun, the object, and the sensor, without regard to any actual reflectance differences on the ground. Often, the impact of directional reflectance effects can be minimized by advance planning. For example, when photographing a lake when the sun is to the south and the lake's surface is calm, it may be preferable to take the photographs from the east or west, rather than from the north, to avoid the sun's specular reflection angle. However, the impact of varying bidirectional reflectance usually cannot be completely eliminated, and it is important for image analysts to be aware of this effect.

1.5 DATA ACQUISITION AND DIGITAL IMAGE CONCEPTS

To this point, we have discussed the principal sources of electromagnetic energy, the propagation of this energy through the atmosphere, and the interaction of this energy with earth surface features. These factors combine to produce energy "signals" from which we wish to extract information. We now consider the procedures by which these signals are detected, recorded, and interpreted.

The *detection* of electromagnetic energy can be performed in several ways. Before the development and adoption of electronic sensors, analog film-based cameras used chemical reactions on the surface of a light-sensitive film to detect energy variations within a scene. By developing a photographic film, we obtained a *record* of its detected signals. Thus, the film acted as both the detecting and the recording medium. These pre-digital photographic systems offered many advantages: They were relatively simple and inexpensive and provided a high degree of spatial detail and geometric integrity.

Electronic sensors generate an electrical signal that corresponds to the energy variations in the original scene. A familiar example of an electronic sensor is a handheld digital camera. Different types of electronic sensors have different designs of detectors, ranging from charge-coupled devices (CCDs, discussed in Chapter 2) to the antennas used to detect microwave signals (Chapter 6). Regardless of the type

of detector, the resulting data are generally recorded onto some magnetic or optical computer storage medium, such as a hard drive, memory card, solid-state storage unit or optical disk. Although sometimes more complex and expensive than film-based systems, electronic sensors offer the advantages of a broader spectral range of sensitivity, improved calibration potential, and the ability to electronically store and transmit data.

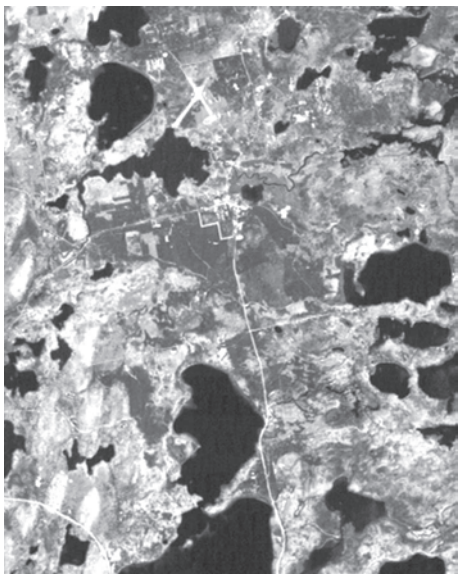
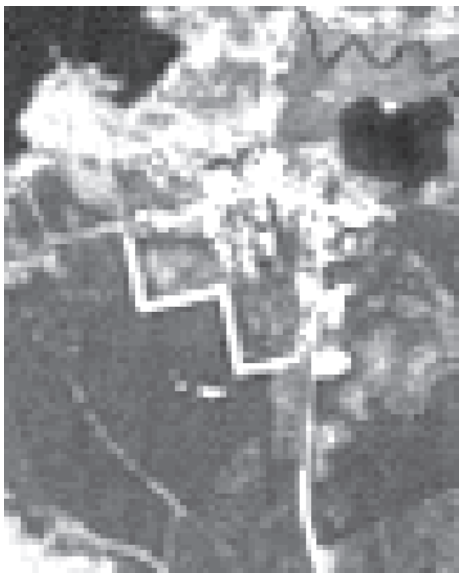
In remote sensing, the term *photograph* historically was reserved exclusively for images that were *detected* as well as recorded on film. The more generic term *image* was adopted for any pictorial representation of image data. Thus, a pictorial record from a thermal scanner (an electronic sensor) would be called a "thermal image," *not* a "thermal photograph," because film would not be the original detection mechanism for the image. Because the term *image* relates to any pictorial product, all photographs are images. Not all images, however, are photographs.

A common exception to the above terminology is use of the term *digital photography*. As we describe in Section 2.5, digital cameras use electronic detectors rather than film for image detection. While this process is not "photography" in the traditional sense, "digital photography" is now the common way to refer to this technique of digital data collection.

We can see that the data interpretation aspects of remote sensing can involve analysis of pictorial (image) and/or digital data. *Visual interpretation* of pictorial image data has long been the most common form of remote sensing. Visual techniques make use of the excellent ability of the human mind to qualitatively evaluate spatial patterns in an image. The ability to make subjective judgments based on selected image elements is essential in many interpretation efforts. Later in this chapter, in Section 1.12, we discuss the process of visual image interpretation in detail.

Visual interpretation techniques have certain disadvantages, however, in that they may require extensive training and are labor intensive. In addition, *spectral characteristics* are not always fully evaluated in visual interpretation efforts. This is partly because of the limited ability of the eye to discern tonal values on an image and the difficulty of simultaneously analyzing numerous spectral images. In applications where spectral patterns are highly informative, it is therefore preferable to analyze *digital*, rather than pictorial, image data.

The basic character of digital image data is illustrated in Figure 1.18. Although the image shown in (*a*) appears to be a continuous-tone photograph, it is actually composed of a two-dimensional array of discrete *picture elements*, or *pixels*. The intensity of each pixel corresponds to the average brightness, or radiance, measured electronically over the ground area corresponding to each pixel. A total of 500 rows and 400 columns of pixels are shown in Figure 1.18*a*. Whereas the individual pixels are virtually impossible to discern in (*a*), they are readily observable in the enlargements shown in (*b*) and (*c*). These enlargements correspond to sub-areas located near the center of (*a*). A 100 row $\times$ 80 column enlargement is shown in (*b*) and a 10 row $\times$ 8 column enlargement is included in (*c*). Part (*d*) shows the individual *digital number (DN)*—also referred to as the

(*a*)   (*b*)

(*c*) 

34	36	35	36	39	46	35	37
36	38	34	31	32	38	32	30
32	31	33	35	37	37	35	30
32	35	44	57	56	38	30	28
57	58	55	60	65	42	26	26
50	47	39	37	60	48	28	27
26	27	25	23	51	51	32	26
26	28	28	22	43	54	36	25
27	26	28	23	37	54	37	24
28	26	26	24	31	53	40	25

(*d*)

Figure 1.18 Basic character of digital image data. (*a*) Original 500 row × 400 column digital image. Scale 1:200,000. (*b*) Enlargement showing 100 row × 80 column area of pixels near center of (*a*). Scale 1:40,000. (*c*) 10 row × 8 column enlargement. Scale 1:4,000. (*d*) Digital numbers corresponding to the radiance of each pixel shown in (*c*). (Author-prepared figure.)

"brightness value" or "pixel value"—corresponding to the average radiance measured in each pixel shown in (*c*). These values result from quantizing the original electrical signal from the sensor into positive integer values using a process called *analog-to-digital (A-to-D) signal conversion*. (The A-to-D conversion process is discussed further in Chapter 4.)

Whether an image is acquired electronically or photographically, it may contain data from a single spectral band or from multiple spectral bands. The image shown in Figure 1.18 was acquired using a single broad spectral band, by integrating all energy measured across a range of wavelengths (a process analogous to photography using "black-and-white" film). Thus, in the digital image, there is a single DN for each pixel. It is also possible to collect "color" or *multispectral* imagery, whereby data are collected simultaneously in several spectral bands. In the case of a color photograph, three separate sets of detectors (or, for analog cameras, three layers within the film) each record radiance in a different range of wavelengths.

In the case of a digital multispectral image, each pixel includes multiple DNs, one for each spectral band. For example, as shown in Figure 1.19, one pixel in a digital image might have values of 88 in the first spectral band, perhaps representing blue wavelengths, 54 in the second band (green), 27 in the third (red), and so on, all associated with a single ground area.

When viewing this multi-band image, it is possible to view a single band at a time, treating it as if it were a discrete image, with brightness values proportional to DN as in Figure 1.18. Alternatively, and more commonly, three bands from the image can be selected and displayed simultaneously in shades of red, green, and blue, to create a *color composite* image, whether on a computer monitor or in a hard copy print. If the three bands being displayed were originally detected by the sensor in the red, green, and blue wavelength ranges of the visible spectrum, then this composite will be referred to as a *true-color* image, because it will approximate the natural combination of colors that would be seen by the human eye. Any other combination of bands—perhaps involving bands acquired in wavelengths outside the visible spectrum—will be referred to as a *false-color* image. One common false-color combination of spectral bands involves displaying near-IR, red, and green bands (from the sensor) in red, green, and blue, respectively, on the display device. Note that, in all cases, these three-band composite images involve displaying some combination of bands from the sensor in red, green, and blue on the display device because the human eye perceives color as a mixture of these three primary colors. (The principles of color perception and color mixing are described in more detail in Section 1.12.)

With multi-band digital data, the question arises of how to organize the data. In many cases, each band of data is stored as a separate file or as a separate block

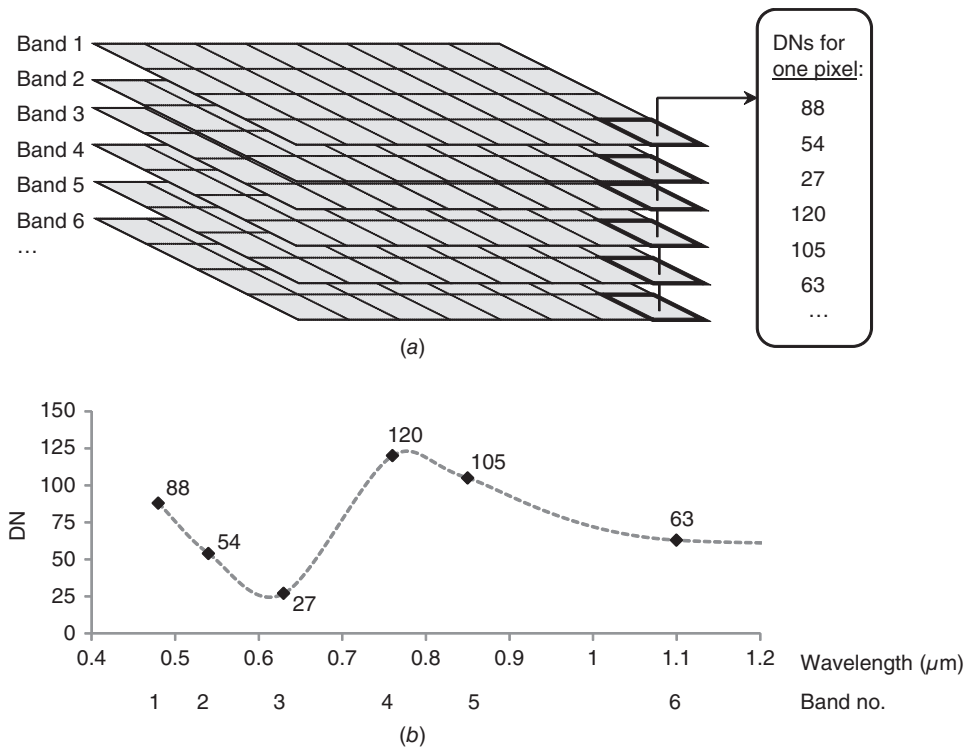


Figure 1.19 Basic character of multi-band digital image data. (*a*) Each band is represented by a grid of cells or pixels; any given pixel has a set of DNs representing its value in each band. (*b*) The spectral signature for the pixel highlighted in (*a*), showing band number and wavelength on the X axis and pixel DN on the Y axis. Values between the wavelengths of each spectral band, indicated by the dashed line in (*b*), are not measured by this sensor and would thus be unknown.

of data within a single file. This format is referred to as *band sequential* (*BSQ*) format. It has the advantage of simplicity, but it is often not the optimal choice for efficient display and visualization of data, because viewing even a small portion of the image requires reading multiple blocks of data from different "places" on the computer disk. For example, to view a true-color digital image in BSQ format, with separate files used to store the red, green, and blue spectral bands, it would be necessary for the computer to read blocks of data from three locations on the storage medium.

An alternate method for storing multi-band data utilizes the *band interleaved by line* (*BIL*) format. In this case, the image data file contains first a line of data from band 1, then the same line of data from band 2, and each subsequent band. This block of data consisting of the first line from each band is then followed by the second line of data from bands 1, 2, 3, and so forth.

The third common data storage format is *band interleaved by pixel* (*BIP*). This is perhaps the most widely used format for three-band images, such as those from

most consumer-grade digital cameras. In this format, the file contains each band's measurement for the first pixel, then each band's measurement for the next pixel, and so on. The advantage of both BIL and BIP formats is that a computer can read and process the data for small portions of the image much more rapidly, because the data from all spectral bands are stored in closer proximity than in the BSQ format.

Typically, the DNs constituting a digital image are recorded over such numerical ranges as 0 to 255, 0 to 511, 0 to 1023, 0 to 2047, 0 to 4095 or higher. These ranges represent the set of integers that can be recorded using 8-, 9-, 10-, 11-, and 12-bit binary computer coding scales, respectively. (That is, $2^8 = 256$, $2^9 = 512$, $2^{10} = 1024$, $2^{11} = 2048$, and $2^{12} = 4096$.) The technical term for the number of bits used to store digital image data is *quantization level* (or *color depth*, when used to describe the number of bits used to display a color image). As discussed in Chapter 7, with the appropriate calibration coefficients these integer DNs can be converted to more meaningful physical units such as spectral reflectance, radiance, or normalized radar cross section.

Elevation Data

Increasingly, remote sensing instruments are used to collect three-dimensional spatial data, in which each observation has a Z coordinate representing elevation, along with the X and Y coordinates used to represent the horizontal position of the pixel's column and row. Particularly when collected over broad areas, these elevation data may represent the *topography*, the three-dimensional shape of the land surface. In other cases (usually, at finer spatial scales), these elevation data may represent the three-dimensional shapes of objects on or above the ground surface, such as tree crowns in a forest, or buildings in a city. Elevation data may be derived from the analysis of raw measurements from many types of remote sensing instruments, including photographic systems, multispectral sensors, radar systems, and lidar systems.

Elevation data may be represented in many different formats. Figure 1.20*a* shows a small portion of a traditional contour map, from the U.S. Geological Survey's 7.5-minute (1:24,000-scale) quadrangle map series. In this map, topographic elevations are indicated by contour lines. Closely spaced lines indicate steep terrain, while flat areas like river floodplains have more widely spaced contours.

Figure 1.20*b* shows a *digital elevation model* (*DEM*). Note that the white rectangle in (*b*) represents the much smaller area shown in (*a*). The DEM is similar to a digital image, with the DN at each pixel representing a surface elevation rather than a radiance value. In (*b*), the brightness of each pixel is represented as being proportional to its elevation, so light-toned areas are topographically higher and dark-toned areas are lower. The region shown in this map consists of highly dissected terrain, with a complex network of river valleys; one major valley runs

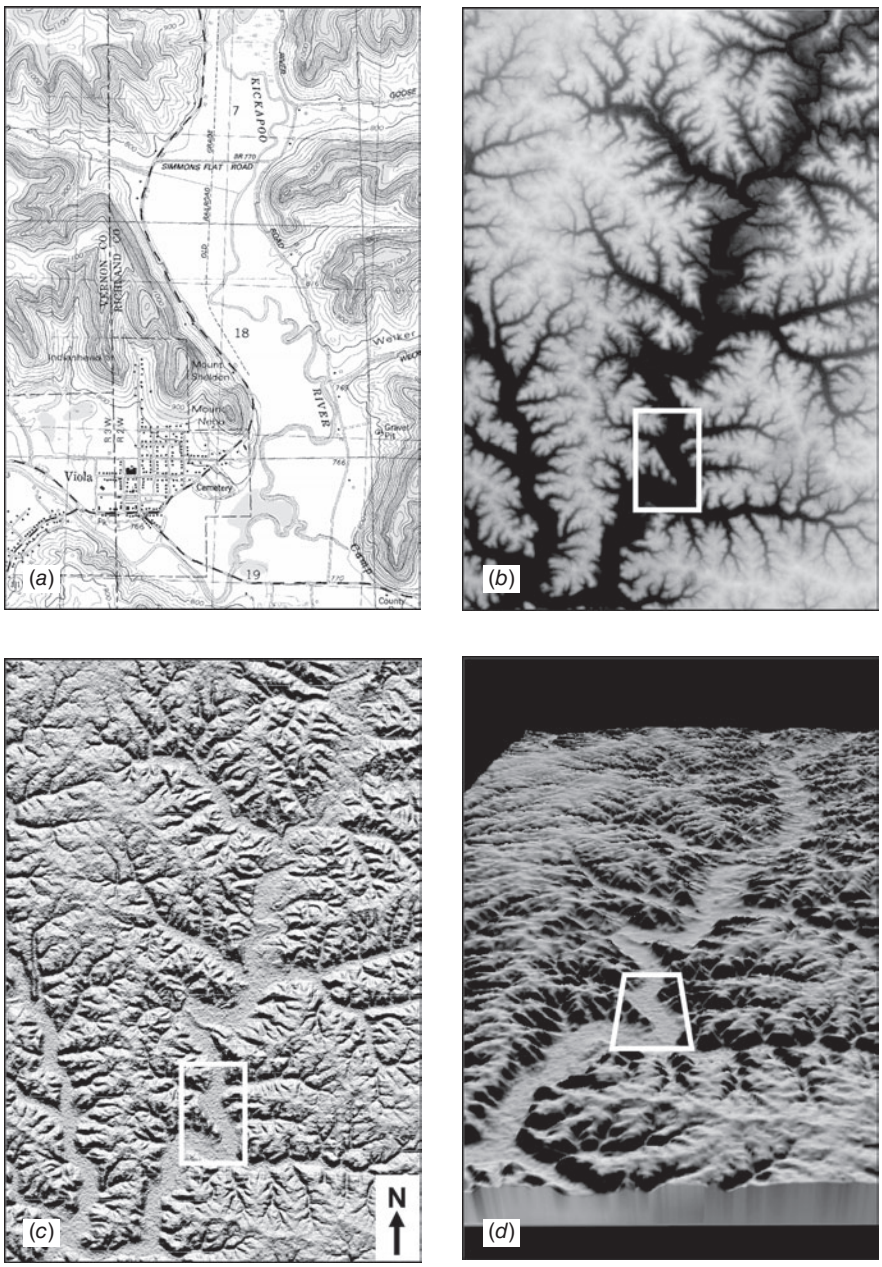


Figure 1.20 Representations of topographic data. (*a*) Portion of USGS 7.5-minute quadrangle map, showing elevation contours. Scale 1:45,000. (*b*) Digital elevation model, with brightness proportional to elevation. Scale 1:280,000. (*c*) Shaded-relief map derived from (*b*), with simulated illumination from the north. Scale 1:280,000. (*d*) Three-dimensional perspective view, with shading derived from (*c*). Scale varies in this projection. White rectangles in (*b*), (*c*), and (*d*) indicate area enlarged in (*a*). (Author-prepared figure.)

from the upper right portion of (*b*) to the lower center, with many tributary valleys branching off from each side.

Figure 1.20*c* shows another way of visualizing topographic data using *shaded relief*. This is a simulation of the pattern of shading that would be expected from a three-dimensional surface under a given set of illumination conditions. In this case, the simulation includes a primary source of illumination located to the north, with a moderate degree of diffuse illumination from other directions to soften the intensity of the shadows. Flat areas will have uniform tone in a shaded relief map. Slopes facing toward the simulated light source will appear bright, while slopes facing away from the light will appear darker.

To aid in visual interpretation, it is often preferable to create shaded relief maps with illumination from the top of the image, regardless of whether that is a direction from which solar illumination could actually come in the real world. When the illumination is from other directions, particularly from the bottom of the image, an untrained analyst may have difficulty correctly perceiving the landscape; in fact, the topography may appear inverted. (This effect is illustrated in Figure 1.29.)

Figure 1.20*d* shows yet another method for visualizing elevation data, a *three-dimensional perspective view*. In this example, the shaded relief map shown in (*c*) has been "draped" over the DEM, and a simulated view has been created based on a viewpoint located at a specified position in space (in this case, above and to the south of the area shown). This technique can be used to visualize the appearance of a landscape as seen from some point of interest. It is possible to "drape" other types of imagery over a DEM; perspective views created using an aerial photograph or high-resolution satellite image may appear quite realistic. Animation of successive perspective views created along a user-defined flight line permits the development of simulated "fly-throughs" over an area.

The term "digital elevation model" or DEM can be used to describe any image where the pixel values represent elevation (Z) coordinates. Two common subcategories of DEMs are a *digital terrain model* (*DTM*) and a *digital surface model* (*DSM*). A DTM (sometimes referred to as a "bald-earth DEM") records the elevation of the bare land surface, without any vegetation, buildings, or other features above the ground. In contrast, a DSM records the elevation of whatever the uppermost surface is at every location; this could be a tree crown, the roof of a building, or the ground surface (where no vegetation or structures are present). Each of these models has its appropriate uses. For example, a DTM would be useful for predicting runoff in a watershed after a rainstorm, because streams will flow over the ground surface rather than across the top of the forest canopy. In contrast, a DSM could be used to measure the size and shape of objects on the terrain, and to calculate intervisibility (whether a given point B can be seen from a reference point A).

Figure 1.21 compares a DSM and DTM for the same site, using airborne lidar data from the Capitol Forest area in Washington State (Andersen, McGaughey, and Reutebuch, 2005). In Figure 1.21*a*, the uppermost lidar points have been used

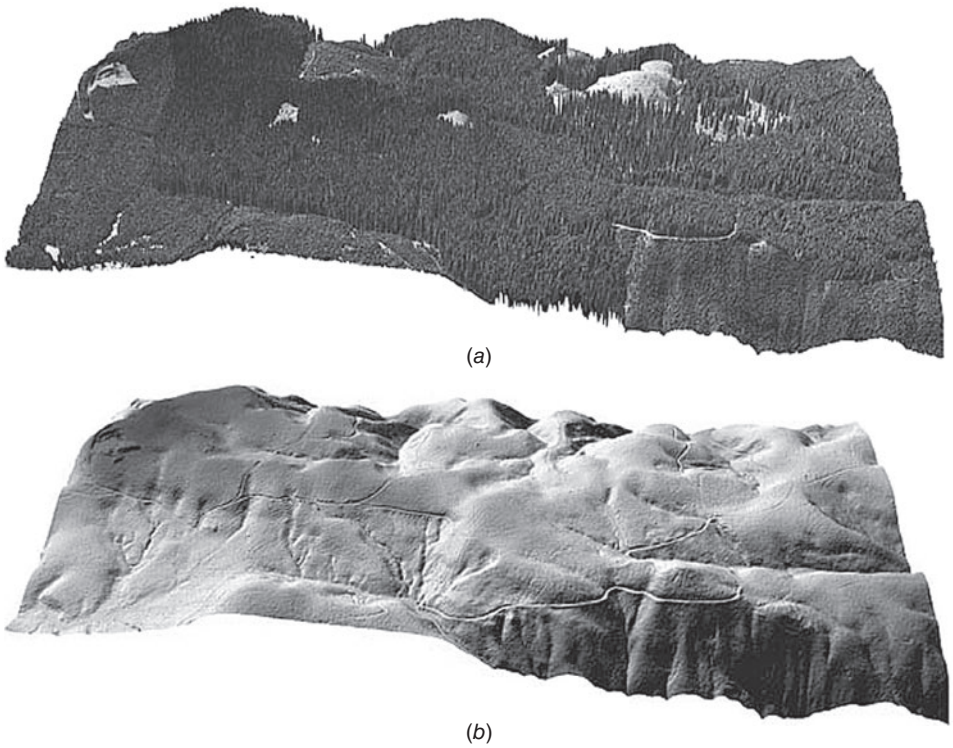


Figure 1.21 Airborne lidar data of the Capitol Forest site, Washington State. (*a*) Digital surface model (DSM) showing tops of tree crowns and canopy gaps. (*b*) Digital terrain model (DTM) showing hypothetical bare earth surface. (From Andersen et al., 2006; courtesy Ward Carlson, USDA Forest Service PNW Research Station.)

to create a DSM showing the elevation of the upper surface of the forest canopy, the presence of canopy gaps, and, in many cases, the shape of individual tree crowns. In Figure 1.21*b*, the lowermost points have been used to create a DTM, showing the underlying ground surface if all vegetation and structures were removed. Note the ability to detect fine-scale topographic features, such as small gullies and roadcuts, even underneath a dense forest canopy (Andersen et al., 2006).

Plate 1 shows a comparison of a DSM (*a*) and DTM (*b*) for a wooded area in New Hampshire. The models were derived from airborne lidar data acquired in early December. This site is dominated by a mix of evergreen and deciduous tree species, with the tallest (pines and hemlocks) exceeding 40 m in height. Scattered clearings in the center and right side are athletic fields, parkland, and former ski slopes now being taken over by shrubs and small trees. With obscuring vegetation removed, the DTM in (*b*) shows a variety of glacial and post-glacial landforms, as

well as small roads, trails, and other constructed features. Also, by subtracting the elevations in (*b*) from those in (*a*), it is possible to calculate the height of the forest canopy above ground level at each point. The result, shown in (*c*), is referred to as a *canopy height model* (*CHM*). In this model, the ground surface has been flattened, so that all remaining variation represents differences in height of the trees relative to the ground. Lidar and other high-resolution 3D data are widely used for this type of canopy height analysis (Clark et al., 2004). (See Sections 6.23 and 6.24 for more discussion.)

Increasingly, elevation data are being used for analysis not just in the form of highly processed DEM, but in the more basic form of a *point cloud*. A point cloud is simply a data set containing many three-dimensional point locations, each representing a single measurement of the (X, Y, Z) coordinates of an object or surface. The positions, spacing, intensity, and other characteristics of the points in this cloud can be analyzed using sophisticated 3D processing algorithms to extract information about features (Rutzinger et al., 2008).

Further discussion of the acquisition, visualization, and analysis of elevation data, including DEMs and point clouds, can be found in Chapters 3 and 6, under the discussion of photogrammetry, interferometric radar, and lidar systems.

1.6 REFERENCE DATA

As we have indicated in the previous discussion, rarely, if ever, is remote sensing employed without the use of some form of *reference data*. The acquisition of reference data involves collecting measurements or observations about the objects, areas, or phenomena that are being sensed remotely. These data can take on any of a number of different forms and may be derived from a number of sources. For example, the data needed for a particular analysis might be derived from a soil survey map, a water quality laboratory report, or an aerial photograph. They may also stem from a "field check" on the identity, extent, and condition of agricultural crops, land uses, tree species, or water pollution problems. Reference data may also involve field measurements of temperature and other physical and/or chemical properties of various features. The geographic positions at which such field measurements are made are often noted on a map base to facilitate their location in a corresponding remote sensing image. Usually, GPS receivers are used to determine the precise geographic position of field observations and measurements (as described in Section 1.7).

Reference data are often referred to by the term *ground truth*. This term is not meant literally, because many forms of reference data are not collected on the ground and can only approximate the truth of actual ground conditions. For example, "ground" truth may be collected in the air, in the form of detailed aerial photographs used as reference data when analyzing less detailed high altitude or satellite imagery. Similarly, the "ground" truth will actually be "water" truth if we

are studying water features. In spite of these inaccuracies, ground truth is a widely used term for reference data.

Reference data might be used to serve any or all of the following purposes:

1. To aid in the analysis and interpretation of remotely sensed data.
2. To calibrate a sensor.
3. To verify information extracted from remote sensing data.

Hence, reference data must be collected in accordance with the principles of statistical sampling design appropriate to the particular application.

Reference data can be very expensive and time consuming to collect properly. They can consist of either *time-critical* and/or *time-stable* measurements. Time-critical measurements are those made in cases where ground conditions change rapidly with time, such as in the analysis of vegetation condition or water pollution events. Time-stable measurements are involved when the materials under observation do not change appreciably with time. For example, geologic applications often entail field observations that can be conducted at any time and that would not change appreciably from mission to mission.

One form of reference data collection is the ground-based measurement of the reflectance and/or emittance of surface materials to determine their spectral response patterns. This might be done in the laboratory or in the field using the principles of *spectroscopy*. Spectroscopic measurement procedures can involve the use of a variety of instruments. Often, a *spectroradiometer* is used in such measurement procedures. This device measures, as a function of wavelength, the energy coming from an object within its view. It is used primarily to prepare spectral reflectance curves for various objects.

In laboratory spectroscopy, artificial sources of energy might be used to illuminate objects under study. In the laboratory, other field parameters such as viewing geometry between object and sensor are also simulated. More often, therefore, in situ field measurements are preferred because of the many variables of the natural environment that influence remote sensor data that are difficult, if not impossible, to duplicate in the laboratory.

In the acquisition of field measurements, spectroradiometers may be operated in a number of modes, ranging from handheld to helicopter or aircraft mounted. Figures 1.10 and 1.12, in Section 1.4 of this chapter, both contain examples of measurements acquired using a handheld field spectroradiometer. Figure 1.22 illustrates a highly portable instrument that is well suited for handheld operation. Through a fiber-optic input, this particular system acquires a continuous spectrum by recording data in over 1000 narrow bands simultaneously (over the range 0.35 to 2.5 μm). The unit is typically transported in a backpack carrier with provision for integrating the spectrometer with a notebook computer. The computer provides for flexibility in data acquisition, display, and storage. For example, reflectance spectra can be displayed in real

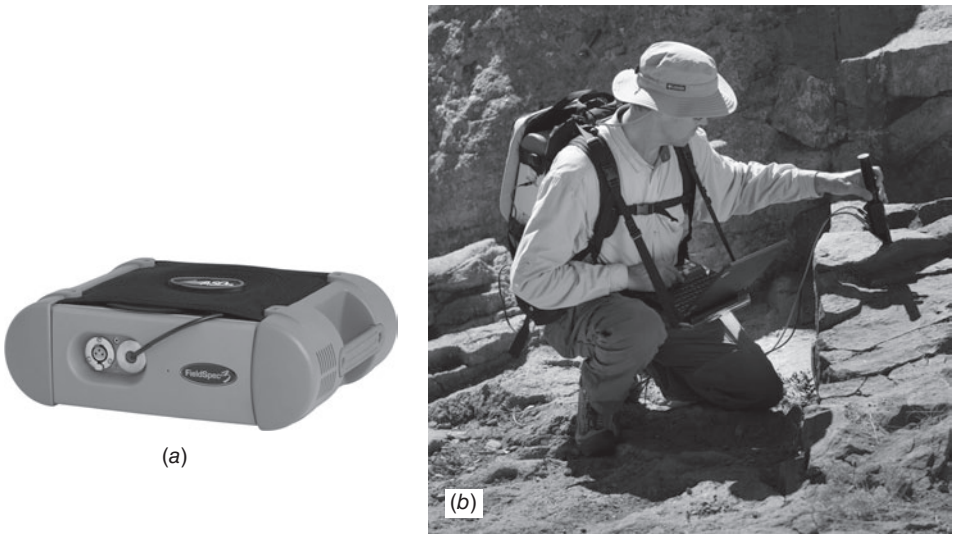


Figure 1.22 ASD, Inc. FieldSpec Spectroradiometer: (*a*) the instrument; (*b*) instrument shown in field operation. (Courtesy ASD, Inc.)

time, as can computed reflectance values within the wavelength bands of various satellite systems. In-field calculation of band ratios and other computed values is also possible. One such calculation might be the normalized difference vegetation index (NDVI), which relates the near-IR and visible reflectance of earth surface features (Chapter 7). Another option is matching measured spectra to a library of previously measured samples. The overall system is compatible with a number of post-processing software packages and also affords Ethernet, wireless, and GPS compatibility as well.

Figure 1.23 shows a versatile all-terrain instrument platform designed primarily for collecting spectral measurements in agricultural cropland environments. The system provides the high clearance necessary for making measurements over mature row crops, and the tracked wheels allow access to difficult landscape positions. Several measurement instruments can be suspended from the system's telescopic boom. Typically, these include a spectroradiometer, a remotely operated digital camera system, and a GPS receiver (Section 1.7). While designed primarily for data collection in agricultural fields, the long reach of the boom makes this device a useful tool for collecting spectral data over such targets as emergent vegetation found in wetlands as well as small trees and shrubs.

Using a spectroradiometer to obtain spectral reflectance measurements is normally a three-step process. First, the instrument is aimed at a *calibration panel* of known, stable reflectance. The purpose of this step is to quantify the

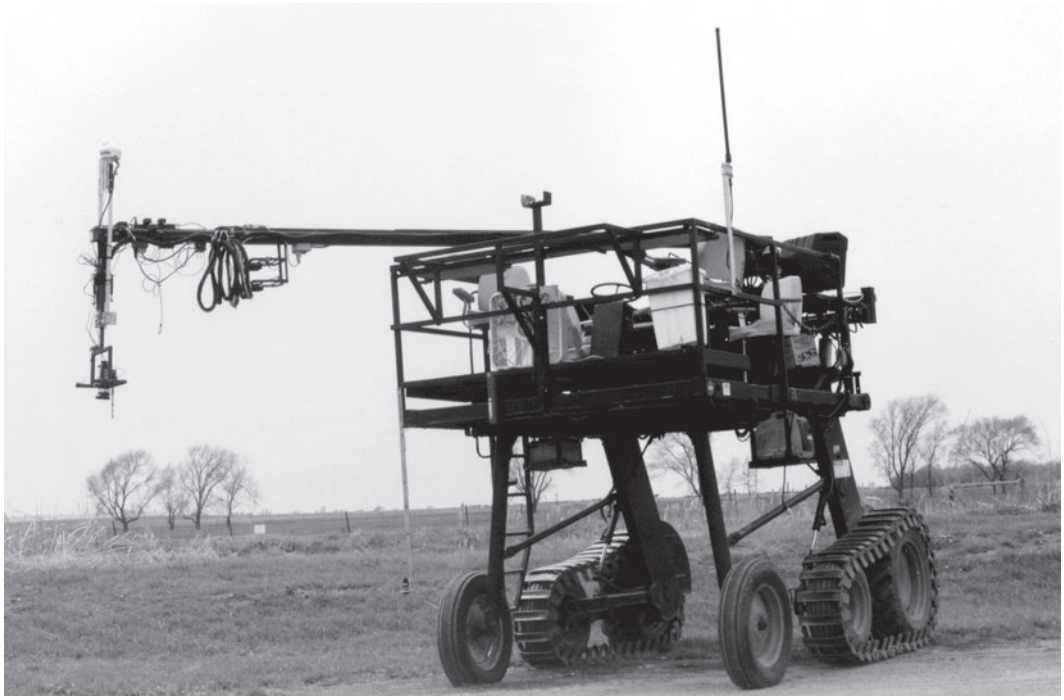


Figure 1.23 All-terrain instrument platform designed for collecting spectral measurements in agricultural cropland environments. (Courtesy of the University of Nebraska-Lincoln Center for Advanced Land Management Information Technologies.)

incoming radiation, or irradiance, incident upon the measurement site. Next, the instrument is suspended over the target of interest and the radiation reflected by the object is measured. Finally, the spectral reflectance of the object is computed by ratioing the reflected energy measurement in each band of observation to the incoming radiation measured in each band. Normally, the term *reflectance factor* is used to refer to the result of such computations. A reflectance factor is defined formally as the ratio of the radiant flux actually reflected by a sample surface to that which would be reflected into the same sensor geometry by an ideal, perfectly diffuse (Lambertian) surface irradiated in exactly the same way as the sample.

Another term frequently used to describe the above type of measurement is *bidirectional reflectance factor*: one direction being associated with the sample viewing angle (usually 0° from normal) and the other direction being that of the sun's illumination (defined by the solar zenith and azimuth angles; see Section 1.4). In the bidirectional reflectance measurement procedure described above, the sample and the reflectance standard are measured sequentially. Other approaches exist in which the incident spectral irradiance and reflected spectral radiance are measured simultaneously.

1.7 THE GLOBAL POSITIONING SYSTEM AND OTHER GLOBAL NAVIGATION SATELLITE SYSTEMS

As mentioned previously, the location of field-observed reference data is usually determined using a *global navigation satellite system* (*GNSS*). GNSS technology is also used extensively in such other remote sensing activities as navigating aircraft during sensor data acquisition and geometrically correcting and referencing raw image data. The first such system, the U.S. *Global Positioning System* (*GPS*) was originally developed for military purposes, but has subsequently become ubiquitous in many civil applications worldwide, from vehicle navigation to surveying, and location-based services on cellular phones and other personal electronic devices. Other GNSS "constellations" have been or are being developed as well, a trend that will greatly increase the accuracy and reliability of GNSS for end-users over the next decade.

The U.S. Global Positioning System includes at least 24 satellites rotating around the earth in precisely known orbits, with subgroups of four or more satellites operating in each of six different orbit planes. Typically, these satellites revolve around the earth approximately once every 12 hours, at an altitude of approximately 20,200 km. With their positions in space precisely known at all times, the satellites transmit time-encoded radio signals that are recorded by ground-based receivers and can be used to aid in positioning and navigation. The nearly circular orbital planes of the satellites are inclined about 60° from the equator and are spaced every 60° in longitude. This means that, in the absence of obstructions from the terrain or nearby buildings, an observer at any point on the earth's surface can receive the signal from at least four GPS satellites at any given time (day or night).

International Status of GNSS Development

Currently, the U.S. Global Positioning System has only one operational counterpart, the Russian *GLONASS* system. The full GLONASS constellation consists of 24 operational satellites, a number that was reached in October 2011. In addition, a fully comprehensive European GNSS constellation, *Galileo*, is scheduled for completion by 2020 and will include 30 satellites. The data signals provided by Galileo will be compatible with those from the U.S. GPS satellites, resulting in a greatly increased range of options for GNSS receivers and significantly improved accuracy. Finally, China has announced plans for the development of its own *Compass* GNSS constellation, to include 30 satellites in operational use by 2020. The future for these and similar systems is an extremely bright and rapidly progressing one.

GNSS Data Processing and Corrections

The means by which GNSS signals are used to determine ground positions is called *satellite ranging*. Conceptually, the process simply involves measuring

the time required for signals transmitted by at least four satellites to reach the ground receiver. Knowing that the signals travel at the speed of light $(3 \times 10^8 \text{ m/sec in a vacuum})$, the distance from each satellite to the receiver can be computed using a form of three-dimensional triangulation. In principle, the signals from only four satellites are needed to identify the receiver's location, but in practice it is usually desirable to obtain measurements from as many satellites as practical.

GNSS measurements are potentially subject to numerous sources of error. These include *clock bias* (caused by imperfect synchronization between the high-precision atomic clocks present on the satellites and the lower-precision clocks used in GNSS receivers), uncertainties in the satellite orbits (known as *satellite ephemeris errors*), errors due to atmospheric conditions (signal velocity depends on time of day, season, and angular direction through the atmosphere), receiver errors (due to such influences as electrical noise and signal-matching errors), and multipath errors (reflection of a portion of the transmitted signal from objects not in the straight-line path between the satellite and receiver).

Such errors can be compensated for (in great part) using *differential* GNSS measurement methods. In this approach, simultaneous measurements are made by a stationary base station receiver (located over a point of precisely known position) and one (or more) roving receivers moving from point to point. The positional errors measured at the base station are used to refine the position measured by the rover(s) at the same instant in time. This can be done either by bringing the data from the base and rover together in a post-processing mode after the field observations are completed or by instantaneously broadcasting the base station corrections to the rovers. The latter approach is termed *real-time differential* GNSS positioning.

In recent years, there have been efforts to improve the accuracy of GNSS positioning through the development of regional networks of high-precision base stations, generally referred to as *satellite-based augmentation systems* (*SBAS*). The data from these stations are used to derive spatially explicit correction factors that are then broadcast in real time, allowing advanced receiver units to determine their positions with a higher degree of accuracy. One such SBAS network, the *Wide Area Augmentation System* (*WAAS*), consists of approximately 25 ground reference stations distributed across the United States that continuously monitor GPS satellite transmissions. Two main stations, located on the U.S. east and west coasts, collect the data from the reference stations and create a composited correction message that is location specific. This message is then broadcast through one of two *geostationary* satellites, satellites occupying a fixed position over the equator. Any WAAS-enabled GPS unit can receive these correction signals. The GPS receiver then determines which correction data are appropriate at the current location.

The WAAS signal reception is ideal for open land, aircraft, and marine applications, but the position of the relay satellites over the equator makes it difficult to receive the signals at high latitudes or when features such as trees and

mountains obstruct the view of the horizon. In such situations, GPS positions can sometimes actually contain more error with WAAS correction than without. However, in unobstructed operating conditions where a strong WAAS signal is available, positions are normally accurate to within 3 m or better.

Paralleling the deployment of the WAAS system in North America are the Japanese *Multi-functional Satellite Augmentation System* (*MSAS*) in Asia, the *European Geostationary Navigation Overlay Service* (*EGNOS*) in Europe, and proposed future SBAS networks such as India's *GPS Aided Geo-Augmented Navigation* (*GAGAN*) system. Like WAAS, these SBAS systems use geostationary satellites to transmit data for real-time differential correction.

In addition to the regional SBAS real-time correction systems such as WAAS, some nations have developed additional networks of base stations that can be used for post-processing GNSS data for differential correction (i.e., high-accuracy corrections made after data collection, rather than in real time). One such system is the U.S. National Geodetic Survey's *Continuously Operating Reference Stations* (*CORS*) network. More than 1800 sites in the cooperative CORS network provide GNSS reference data that can be accessed via the Internet and used in post-processing for differential correction.

With the development of new satellite constellations, and new resources for real-time and post-processed differential correction, GNSS-based location services are expected to become even more widespread in industry, resource management, and consumer technology applications in the coming years.

1.8 CHARACTERISTICS OF REMOTE SENSING SYSTEMS

Having introduced some basic concepts, we now have the elements necessary to characterize a remote sensing system. In so doing, we can begin to appreciate some of the problems encountered in the design and application of the various sensing systems examined in subsequent chapters. In particular, the design and operation of every real-world sensing system represents a series of compromises, often in response to the limitations imposed by physics and by the current state of technological development. When we consider the process from start to finish, users of remote sensing systems need to keep in mind the following factors:

1. **The energy source.** All passive remote sensing systems rely on energy that originates from sources other than the sensor itself, typically in the form of either reflected radiation from the sun or emitted radiation from earth surface features. As already discussed, the spectral distribution of reflected sunlight and self-emitted energy is far from uniform. Solar energy levels obviously vary with respect to time and location, and different earth surface materials emit energy with varying degrees of efficiency. While we have some control over the sources of energy for active systems such as radar and lidar, those sources have their own particular

characteristics and limitations, as discussed in Chapter 6. Whether employing a passive or active system, the remote sensing analyst needs to keep in mind the nonuniformity and other characteristics of the energy source that provides illumination for the sensor.

2. **The atmosphere.** The atmosphere normally compounds the problems introduced by energy source variation. To some extent, the atmosphere always modifies the strength and spectral distribution of the energy received by a sensor. It restricts where we can look spectrally, and its effects vary with wavelength, time, and place. The importance of these effects, like source variation effects, is a function of the wavelengths involved, the sensor used, and the sensing application at hand. Elimination of, or compensation for, atmospheric effects via some form of calibration is particularly important in those applications where repetitive observations of the same geographic area are involved.
3. **The energy–matter interactions at the earth's surface.** Remote sensing would be simple if every material reflected and/or emitted energy in a unique, known way. Although spectral response patterns such as those in Figure 1.9 play a central role in detecting, identifying, and analyzing earth surface materials, the spectral world is full of ambiguity. Radically different material types can have great spectral similarity, making identification difficult. Furthermore, the general understanding of the energy–matter interactions for earth surface features is at an elementary level for some materials and virtually nonexistent for others.
4. **The sensor.** An ideal sensor would be highly sensitive to all wavelengths, yielding spatially detailed data on the absolute brightness (or radiance) from a scene as a function of wavelength, throughout the spectrum, across wide areas on the ground. This "supersensor" would be simple and reliable, require virtually no power or space, be available whenever and wherever needed, and be accurate and economical to operate. At this point, it should come as no surprise that an ideal "supersensor" does not exist. No single sensor is sensitive to all wavelengths or energy levels. All real sensors have fixed limits of spatial, spectral, radiometric and temporal resolution.

   The choice of a sensor for any given task always involves trade-offs. For example, photographic systems generally have very fine spatial resolution, providing a detailed view of the landscape, but they lack the broad spectral sensitivity obtainable with nonphotographic systems. Similarly, many nonphotographic systems are quite complex optically, mechanically, and/or electronically. They may have restrictive power, space, and stability requirements. These requirements often dictate the type of *platform*, or vehicle, from which a sensor can be operated. Platforms can range from stepladders to aircraft (fixed-wing or helicopters) to satellites.

   In recent years, *uninhabited aerial vehicles* (*UAVs*) have become an increasingly important platform for remote sensing data acquisition.

While the development of UAV technology for military applications has received a great deal of attention from the media, such systems are also ideally suited for many civilian applications, particularly in environmental monitoring, resource management, and infrastructure management (Laliberte et al., 2010). UAVs can range from palm-size radio-controlled airplanes and helicopters to large-size aircraft weighing tens of tons and controlled from thousands of km away. They also can be completely controlled through human intervention, or they can be partially or fully autonomous in their operation. Figure 1.24 shows two different types of UAVs used for environmental applications of remote sensing. In 1.24*(a)*, the Ikhana UAV is a fixed-wing aircraft based on the design of a military UAV but operated by NASA for civilian scientific research purposes. (The use of Ikhana for monitoring wildfires is discussed in Section 4.10, and imagery from this system is illustrated in Figure 4.34 and Plate 9.) In contrast, the UAV shown in 1.24*(b)* is a vertical takeoff UAV, designed in the form of a helicopter. In this photo the UAV is carrying a lightweight hyperspectral sensor used to map marine environments such as seagrass beds and coral reefs.

Depending on the sensor–platform combination needed in a particular application, the acquisition of remote sensing data can be a very expensive endeavor, and there may be limitations on the times and places that data can be collected. Airborne systems require detailed flight planning in advance, while data collection from satellites is limited by the platform's orbit characteristics.

5. **The data processing and supply system.** The capability of current remote sensors to generate data far exceeds the capacity to handle these data. This is generally true whether we consider "manual" image interpretation procedures or digital analyses. Processing sensor data into an interpretable format can be—and often is—an effort entailing considerable thought, hardware, time, and experience. Also, many data users would like to receive their data immediately after acquisition by the sensor in order to make the timely decisions required in certain applications (e.g., agricultural crop management, disaster assessment). Fortunately, the distribution of remote sensing imagery has improved dramatically over the past two decades. Some sources now provide in-flight data processing immediately following image acquisition, with near real-time data downloaded over the Internet. In some cases, users may work with imagery and other spatial data in a *cloud computing environment*, where the data and/or software are stored remotely, perhaps even distributed widely across the Internet. At the opposite extreme—particularly for highly specialized types of imagery or for experimental or newly developed remote sensing systems—it may take weeks or months before data are made available, and the user may need to acquire not just the data but highly specialized or custom software for data processing. Finally, as discussed

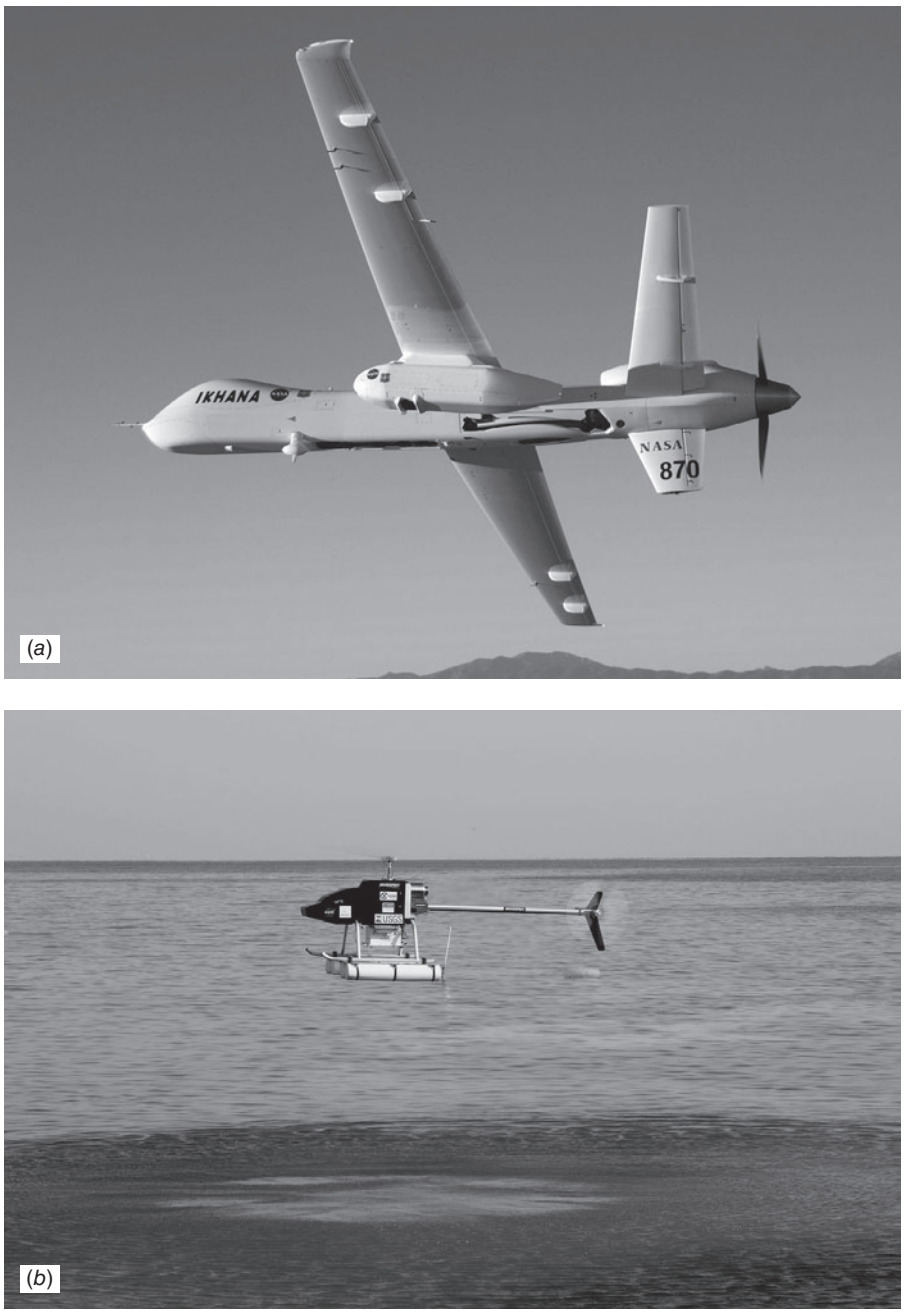


Figure 1.24 Uninhabited aerial vehicles (UAVs) used for environmental applications of remote sensing. (*a*) NASA's Ikhana UAV, with imaging sensor in pod under left wing. (Photo courtesy NASA Dryden Flight Research Center and Jim Ross.) (*b*) A vertical takeoff UAV mapping seagrass and coral reef environments in Florida. (Photo courtesy Rick Holasek and NovaSol.)

in Section 1.6, most remote sensing applications require the collection and analysis of additional reference data, an operation that may be complex, expensive, and time consuming.

6. **The users of remotely sensed data.** Central to the successful application of any remote sensing system is the person (or persons) using the remote sensor data from that system. The "data" generated by remote sensing procedures become "information" only if and when someone understands their generation, knows how to interpret them, and knows how best to use them. *A thorough understanding of the problem at hand is paramount to the productive application of any remote sensing methodology. Also, no single combination of data acquisition and analysis procedures will satisfy the needs of all data users.*

Whereas the interpretation of aerial photography has been used as a practical resource management tool for nearly a century, other forms of remote sensing are relatively new, technically complex, and unconventional means of acquiring information. In earlier years, these newer forms of remote sensing had relatively few satisfied users. Since the late 1990s, however, as new applications continue to be developed and implemented, increasing numbers of users are becoming aware of the potentials, as well as the limitations, of remote sensing techniques. As a result, remote sensing has become an essential tool in many aspects of science, government, and business alike.

One factor in the increasing acceptance of remote sensing imagery by end-users has been the development and widespread adoption of easy-to-use geovisualization systems such as Google Maps and Earth, NASA's WorldWinds, and other web-based image services. By allowing more potential users to become comfortable and familiar with the day-to-day use of aerial and satellite imagery, these and other software tools have facilitated the expansion of remote sensing into new application areas.

1.9 SUCCESSFUL APPLICATION OF REMOTE SENSING

The student should now begin to appreciate that successful use of remote sensing is premised on the *integration* of multiple, interrelated data sources and analysis procedures. No single combination of sensor and interpretation procedure is appropriate to all applications. The key to designing a successful remote sensing effort involves, at a minimum, (1) clear definition of the problem at hand, (2) evaluation of the potential for addressing the problem with remote sensing techniques, (3) identification of the remote sensing data acquisition procedures appropriate to the task, (4) determination of the data interpretation procedures to be employed and the reference data needed, and (5) identification of the criteria by which the quality of information collected can be judged.

All too often, one (or more) of the above components of a remote sensing application is overlooked. The result may be disastrous. Many programs exist with little or no means of evaluating the performance of remote sensing systems in terms of information quality. Many people have acquired burgeoning quantities of remote sensing data with inadequate capability to interpret them. In some cases an inappropriate decision to use (or *not* to use) remote sensing has been made, because the problem was not clearly defined and the constraints or opportunities associated with remote sensing methods were not clearly understood. A clear articulation of the information requirements of a particular problem and the extent to which remote sensing might meet these requirements in a timely manner is paramount to any successful application.

The success of many applications of remote sensing is improved considerably by taking a *multiple-view* approach to data collection. This may involve *multistage* sensing, wherein data about a site are collected from multiple altitudes. It may involve *multispectral* sensing, whereby data are acquired simultaneously in several spectral bands. Or, it may entail *multitemporal* sensing, where data about a site are collected on more than one occasion.

In the multistage approach, satellite data may be analyzed in conjunction with high altitude data, low altitude data, and ground observations (Figure 1.25). Each successive data source might provide more detailed information over smaller geographic areas. Information extracted at any lower level of observation may then be extrapolated to higher levels of observation.

A commonplace example of the application of multistage sensing techniques is the detection, identification, and analysis of forest disease and insect problems. From space images, the image analyst could obtain an overall view of the major vegetation categories involved in a study area. Using this information, the areal extent and position of a particular species of interest could be determined and representative subareas could be studied more closely at a more refined stage of imaging. Areas exhibiting stress on the second-stage imagery could be delineated. Representative samples of these areas could then be field checked to document the presence and particular cause of the stress.

After analyzing the problem in detail by ground observation, the analyst would use the remotely sensed data to extrapolate assessments beyond the small study areas. By analyzing the large-area remotely sensed data, the analyst can determine the severity and geographic extent of the disease problem. Thus, while the question of specifically *what* the problem is can generally be evaluated only by detailed ground observation, the equally important questions of *where*, *how much*, and *how severe* can often be best handled by remote sensing analysis.

In short, more information is obtained by analyzing multiple views of the terrain than by analysis of any single view. In a similar vein, multispectral imagery provides more information than data collected in any single spectral band. When the signals recorded in the multiple bands are analyzed in conjunction with each other, more information becomes available than if only a single band were used or if the multiple bands were analyzed independently. The multispectral approach

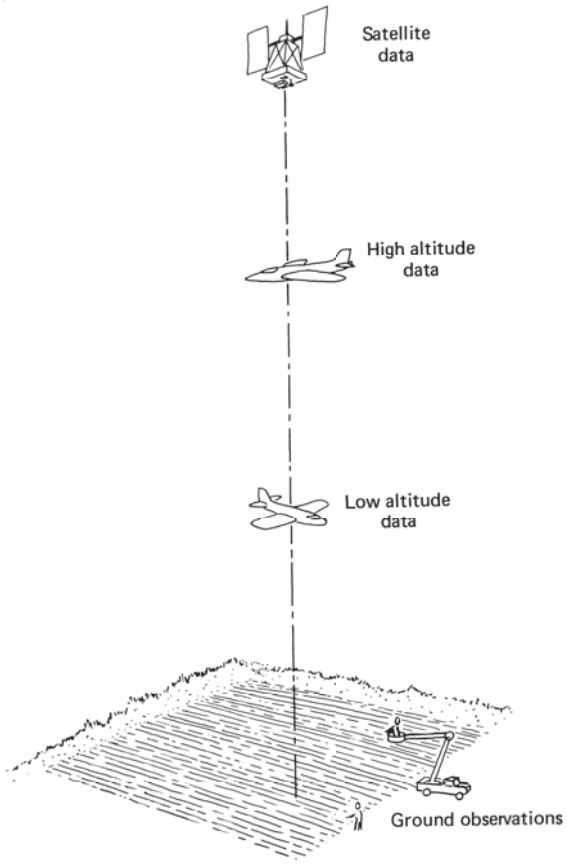


Figure 1.25 Multistage remote sensing concept.

forms the heart of numerous remote sensing applications involving discrimination of earth resource types, cultural features, and their condition.

Again, multitemporal sensing involves sensing the same area at multiple times and using changes occurring with time as discriminants of ground conditions. This approach is frequently taken to monitor land use change, such as suburban development in urban fringe areas. In fact, regional land use surveys might call for the acquisition of multisensor, multispectral, multistage, multitemporal data to be used for multiple purposes!

In any approach to applying remote sensing, not only must the right mix of data acquisition and data interpretation techniques be chosen, but the right mix of remote sensing and "conventional" techniques must also be identified. The student must recognize that remote sensing is a tool best applied in concert with others; it is not an end in itself. In this regard, remote sensing data are currently

being used extensively in computer-based GISs (Section 1.10). The GIS environment permits the synthesis, analysis, and communication of virtually unlimited sources and types of biophysical and socioeconomic data—as long as they can be geographically referenced. Remote sensing can be thought of as the "eyes" of such systems providing repeated, synoptic (even global) visions of earth resources from an aerial or space vantage point.

Remote sensing affords us the capability to literally see the invisible. We can begin to see components of the environment on an ecosystem basis, in that remote sensing data can transcend the cultural boundaries within which much of our current resource data are collected. Remote sensing also transcends disciplinary boundaries. It is so broad in its application that nobody "owns" the field. Important contributions are made to—and benefits derived from—remote sensing by both the "hard" scientist interested in basic research and the "soft" scientist interested in its operational application.

There is little question that remote sensing will continue to play an increasingly broad and important role in the scientific, governmental, and commercial sectors alike. The technical capabilities of sensors, space platforms, data communication and distribution systems, GPSs, digital image processing systems, and GISs are improving on almost a daily basis. At the same time, we are witnessing the evolution of a spatially enabled world society. Most importantly, we are becoming increasingly aware of how interrelated and fragile the elements of our global resource base really are and of the role that remote sensing can play in inventorying, monitoring, and managing earth resources and in modeling and helping us to better understand the global ecosystem and its dynamics.

1.10 GEOGRAPHIC INFORMATION SYSTEMS (GIS)

We anticipate that the majority of individuals using this book will at some point in their educational backgrounds and/or professional careers have experience with geographic information systems. The discussion below is provided as a brief introduction to such systems primarily for those readers who might lack such background.

Geographic information systems are computer-based systems that can deal with virtually any type of information about features that can be referenced by geographical location. These systems are capable of handling both *locational data* and *attribute data* about such features. That is, not only do GISs permit the automated mapping or display of the locations of features, but also these systems provide a capability for recording and analyzing descriptive characteristics ("attributes") of the features. For example, a GIS might contain not only a map of the locations of roads but also a database of descriptors about each road. These attributes might include information such as road width, pavement type, speed limit, number of traffic lanes, date of construction, and so on. Table 1.1 lists other examples of attributes that might be associated with a given point, line, or area feature.

TABLE 1.1 Example Point, Line, and Area Features and Typical Attributes Contained in a GIS[a]

Point feature	Well (depth, chemical constituency)
Line feature	Power line (service capacity, age, insulator type)
Area feature	Soil mapping unit (soil type, texture, color, permeability)

[a]Attributes shown in parentheses.

The data in a GIS may be kept in individual standalone files (e.g., "shapefiles"), but increasingly a *geodatabase* is used to store and manage spatial data. This is a type of *relational database*, consisting of tables with attributes in columns and data records in rows (Table 1.2), and explicitly including locational information for each record. While database implementations vary, there are certain desirable characteristics that will improve the utility of a database in a GIS. These characteristics include flexibility, to allow a wide range of database queries and operations; reliability, to avoid accidental loss of data; security, to limit access to authorized users; and ease of use, to insulate the end user from the details of the database implementation.

One of the most important benefits of a GIS is the ability to spatially interrelate multiple types of information stemming from a range of sources. This concept is illustrated in Figure 1.26, where we have assumed that a hydrologist wishes to use a GIS to study soil erosion in a watershed. As shown, the system contains data from a range of source maps (*a*) that have been geocoded on a cell-by-cell basis to form a series of *layers* (*b*), all in geographic registration. The analyst can then manipulate and overlay the information contained in, or derived from, the various data layers. In this example, we assume that assessing the potential for soil erosion throughout the watershed involves the simultaneous cell-by-cell consideration of three types of data derived from the original data layers: slope, soil erodibility, and surface runoff potential. The slope information can be computed from the elevations in the topography layer. The erodibility, which is an attribute associated with each soil type, can be extracted from a relational database management system incorporated in the GIS. Similarly, the runoff potential is an attribute associated with each land cover type (land cover data can

TABLE 1.2 Relational Database Table Format

ID Number[a]	Street Name	Lanes	Parking	Repair Date	. . .
143897834	"Maple Ct"	2	Yes	2012/06/10	. . .
637292842	"North St"	2	Seasonal	2006/08/22	. . .
347348279	"Main St"	4	Yes	2015/05/15	. . .
234538020	"Madison Ave"	4	No	2014/04/20	. . .

[a]Each data record, or "tuple," has a unique identification, or ID, number.

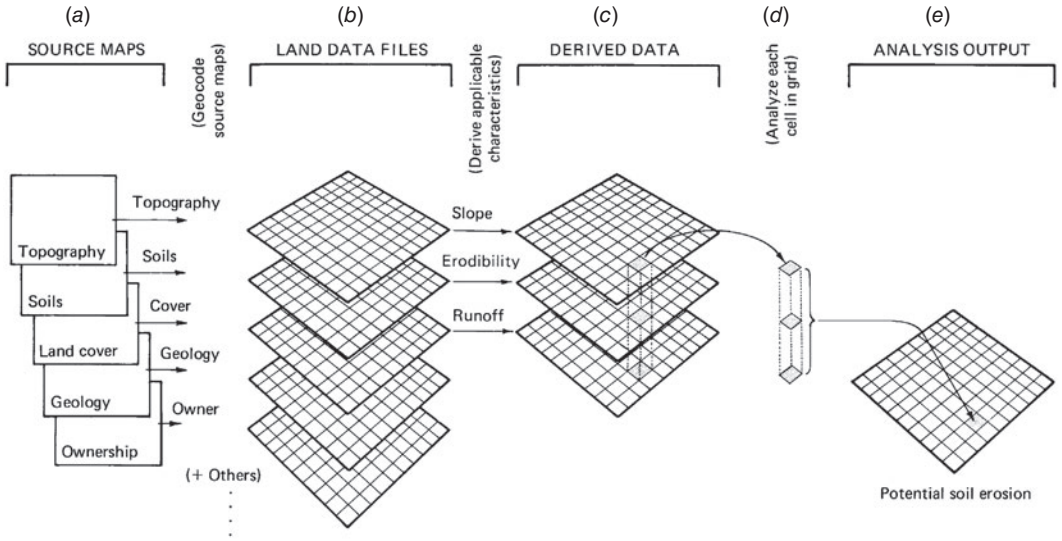


Figure 1.26 GIS analysis procedure for studying potential soil erosion.

be obtained through interpretation of aerial photographs or satellite images). The analyst can use the system to interrelate these three sources of derived data (*c*) in each grid cell and use the result to locate, display, and/or record areas where combinations of site characteristics indicate high soil erosion potential (i.e., steep slopes and highly erodible soil–land cover combinations).

The above example illustrates the GIS analysis function commonly referred to as *overlay analysis*. The number, form, and complexity of other data analyses possible with a GIS are virtually limitless. Such procedures can operate on the system's spatial data, the attribute data, or both. For example, *aggregation* is an operation that permits combining detailed map categories to create new, less detailed categories (e.g., combining "jack pine" and "red pine" categories into a single "pine" category). *Buffering* creates a zone of specified width around one or more features (e.g., the area within 50 m of a stream). *Network analysis* permits such determinations as finding the shortest path through a street network, determining the stream flows in a drainage basin, or finding the optimum location for a fire station. *Intervisibility* operations use elevation data to permit *viewshed mapping* of what terrain features can be "seen" from a specified location. Similarly, many GISs permit the generation of *perspective views* portraying terrain surfaces from a viewing position other than vertical.

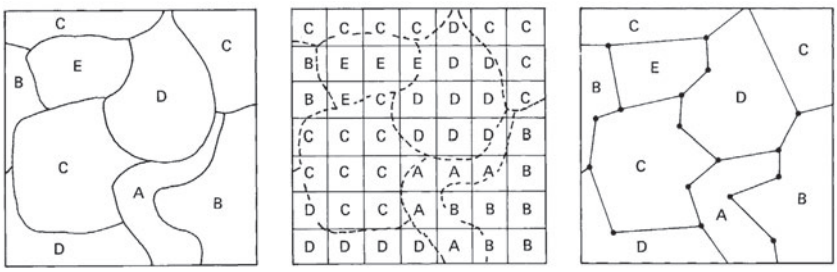
One constraint on the use of multiple layers in a GIS is that the spatial scale at which each of the original source maps was compiled must be compatible with the others. For example, in the analysis shown in Figure 1.26 it would be inappropriate to incorporate both soil data from very high resolution aerial photographs of a single township and land cover digitized from a highly generalized

map of the entire nation. Another common constraint is that the compilation dates of different source maps must be reasonably close in time. For example, a GIS analysis of wildlife habitat might yield incorrect conclusions if it were based on land cover data that are many years out of date. On the other hand, since other types of spatial data are less changeable over time, the map compilation date might not be as important for a layer such as topography or bedrock geology.

Most GISs use two primary approaches to represent the locational component of geographic information: a *raster* (grid-based) or *vector* (point-based) format. The raster data model that was used in our soil erosion example is illustrated in Figure 1.27*b*. In this approach, the location of geographic objects or conditions is defined by the row and column position of the cells they occupy. The value stored for each cell indicates the type of object or condition that is found at that location over the entire cell. Note that the finer the grid cell size used, the more geographic specificity there is in the data file. A coarse grid requires less data storage space but will provide a less precise geographic description of the original data. Also, when using a very coarse grid, several data types and/or attributes may occur in each cell, but the cell is still usually treated as a single homogeneous unit during analysis.

The vector data model is illustrated in Figure 1.27*c*. Using this format, feature boundaries are converted to straight-sided polygons that approximate the original regions. These polygons are encoded by determining the coordinates of their vertices, called *points* or *nodes*, which can be connected to form *lines* or *arcs*. *Topological coding* includes "intelligence" in the data structure relative to the spatial relationship (connectivity and adjacency) among features. For example, topological coding keeps track of which arcs share common nodes and what polygons are to the left and right of a given arc. This information facilitates such spatial operations as overlay analysis, buffering, and network analysis.

Raster and vector data models each have their advantages and disadvantages. Raster systems tend to have simpler data structures; they afford greater computational efficiency in such operations as overlay analysis; and they represent



(*a*) (*b*) (*c*)

Figure 1.27 Raster versus vector data formats: (*a*) landscape patches, (*b*) landscape represented in raster format, and (*c*) landscape represented in vector format.

features having high spatial variability and/or "blurred boundaries" (e.g., between pure and mixed vegetation zones) more effectively. On the other hand, raster data volumes are relatively greater; the spatial resolution of the data is limited to the size of the cells comprising the raster; and the topological relationships among spatial features are more difficult to represent. Vector data formats have the advantages of relatively lower data volumes, better spatial resolution, and the preservation of topological data relationships (making such operations as network analysis more efficient). However, certain operations (e.g., overlay analysis) are more complex computationally in a vector data format than in a raster format.

As we discuss frequently throughout this book, digital remote sensing images are collected in a raster format. Accordingly, digital images are inherently compatible spatially with other sources of information in a raster domain. Because of this, "raw" images can be easily included directly as layers in a raster-based GIS. Likewise, such image processing procedures as automated land cover classification (Chapter 7) result in the creation of interpreted or derived data files in a raster format. These derived data are again inherently compatible with the other sources of data represented in a raster format. This concept is illustrated in Plate 2, in which we return to our earlier example of using overlay analysis to assist in soil erosion potential mapping for an area in western Dane County, Wisconsin. Shown in (*a*) is an automated land cover classification that was produced by processing Landsat Thematic Mapper (TM) data of the area. (See Chapter 7 for additional information on computer-based land cover classification.) To assess the soil erosion potential in this area, the land cover data were merged with information on the intrinsic erodibility of the soil present (*b*) and with land surface slope information (*c*). These latter forms of information were already resident in a GIS covering the area. Hence, all data could be combined for analysis in a mathematical model, producing the soil erosion potential map shown in (*d*). To assist the viewer in interpreting the landscape patterns shown in Plate 2, the GIS was also used to visually enhance the four data sets with topographic shading based on a DEM, providing a three-dimensional appearance.

For the land cover classification in Plate 2*a*, water is shown as dark blue, nonforested wetlands as light blue, forested wetlands as pink, corn as orange, other row crops as pale yellow, forage crops as olive, meadows and grasslands as yellow-green, deciduous forest as green, evergreen forest as dark green, low-intensity urban areas as light gray, and high-intensity urban areas as dark gray. In (*b*), areas of low soil erodibility are shown in dark brown, with increasing soil erodibility indicated by colors ranging from orange to tan. In (*c*), areas of increasing steepness of slope are shown as green, yellow, orange, and red. The soil erosion potential map (*d*) shows seven colors depicting seven levels of potential soil erosion. Areas having the highest erosion potential are shown in dark red. These areas tend to have row crops growing on inherently erodible soils with sloping terrain. Decreasing erosion potential is shown in a spectrum of colors from orange through yellow to green. Areas with the lowest erosion potential are

indicated in dark green. These include forested regions, continuous-cover crops, and grasslands growing on soils with low inherent erodibility, and flat terrain.

Remote sensing images (and information extracted from such images), along with GPS data, have become primary data sources for modern GISs. Indeed, the boundaries between remote sensing, GIS, and GPS technology have become blurred, and these combined fields will continue to revolutionize how we inventory, monitor, and manage natural resources on a day-to-day basis. Likewise, these technologies are assisting us in modeling and understanding biophysical processes at all scales of inquiry. They are also permitting us to develop and communicate cause-and-effect "what-if" scenarios in a spatial context in ways never before possible.

1.11 SPATIAL DATA FRAMEWORKS FOR GIS AND REMOTE SENSING

If one is examining an image purely on its own, with no reference to any outside source of spatial information, there may be no need to consider the type of coordinate system used to represent locations within the image. In many cases, however, analysts will be comparing points in the image to GPS-located reference data, looking for differences between two images of the same area, or importing an image into a GIS for quantitative analysis. In all these cases, it is necessary to know how the column and row coordinates of the image relate to some real-world map coordinate system.

Because the shape of the earth is approximately spherical, locations on the earth's surface are often described in an angular coordinate or *geographical* system, with latitude and longitude specified in degrees (°), minutes (′), and seconds (″). This system originated in ancient Greece, and it is familiar to many people today. Unfortunately, the calculation of distances and areas in an angular coordinate system is complex. More significantly, it is impossible to accurately represent the three-dimensional surface of the earth on the two-dimensional planar surface of a map or image without introducing distortion in one or more of the following elements: shape, size, distance, and direction. Thus, for many purposes we wish to mathematically transform angular geographical coordinates into a planar, or Cartesian $(X-Y)$ coordinate system. The result of this transformation process is referred to as a *map projection*.

While many types of map projections have been defined, they can be grouped into several broad categories based either on the geometric models used or on the spatial properties that are preserved or distorted by the transformation. Geometric models for map projection include cylindrical, conic, and azimuthal or planar surfaces. From a map user's perspective, the spatial properties of map projections may be more important than the geometric model used. A *conformal* map projection preserves angular relationships, or shapes, within local

areas; over large areas, angles and shapes become distorted. An *azimuthal* (or *zenithal*) projection preserves absolute directions relative to the central point of projection. An *equidistant* projection preserves equal distances, for some but not all points—scale is constant either for all distances along meridians or for all distances from one or two points. An *equal-area* (or equivalent) projection preserves equal areas. Because a detailed explanation of the relationships among these properties is beyond the scope of this discussion, suffice it to say that no two-dimensional map projection can accurately preserve all of these properties, but certain subsets of these characteristics can be preserved in a single projection. For example, the azimuthal equidistant projection preserves both direction and distance—but only relative to the central point of the projection; directions and distances between other points are not preserved.

In addition to the map projection associated with a given image, GIS data layer, or other spatial data set, it is also often necessary to consider the *datum* used with that map projection. A datum is a mathematical definition of the three-dimensional solid (generally a slightly flattened ellipsoid) used to represent the surface of the earth. The actual planet itself has an irregular shape that does not correspond perfectly to any ellipsoid. As a result, a variety of different datums have been described; some designed to fit the surface well in one particular region (such as the North American Datum of 1983, or NAD83) and others designed to best approximate the planet as a whole. Most GISs require that both a map projection and a datum be specified before performing any coordinate transformations.

To apply these concepts to the process of collecting and working with remotely sensed images, most such images are initially acquired with rows and columns of pixels aligned with the flight path (or orbit track) of the imaging platform, be it a satellite, an aircraft, or a UAV. Before the images can be mapped, or used in combination with other spatial data, they need to be *georeferenced*. Historically, this process typically involved identification of visible control points whose true geographic coordinates were known. A mathematical model would then be used to transform the row and column coordinates of the raw image into a defined map coordinate system. In recent years, remote sensing platforms have been outfitted with sophisticated systems to record their exact position and angular orientation. These systems, incorporating an *inertial measurement unit* (*IMU*) and/or multiple onboard GPS units, enable highly precise modeling of the viewing geometry of the sensor, which in turn is used for *direct georeferencing* of the sensor data—relating them to a defined map projection without the necessity of additional ground control points.

Once an image has been georeferenced, it may be ready for use with other spatial information. On the other hand, some images may have further geometric distortions, perhaps caused by varying terrain, or other factors. To remove these distortions, it may be necessary to orthorectify the imagery, a process discussed in Chapters 3 and 7.

1.12 VISUAL IMAGE INTERPRETATION

When we look at aerial and space images, we see various objects of different sizes, shapes, and colors. Some of these objects may be readily identifiable while others may not, depending on our own individual perceptions and experience. When we can identify what we see on the images and communicate this information to others, we are practicing *visual image interpretation*. The images contain raw image *data*. These data, when processed by a human interpreter's brain, become usable *information*.

Image interpretation is best learned through the experience of viewing hundreds of remotely sensed images, supplemented by a close familiarity with the environment and processes being observed. Given this fact, no textbook alone can fully train its readers in image interpretation. Nonetheless, Chapters 2 through 8 of this book contain many examples of remote sensing images, examples that we hope our readers will peruse and interpret. To aid in that process, the remainder of this chapter presents an overview of the principles and methods typically employed in image interpretation.

Aerial and space images contain a detailed record of features on the ground at the time of data acquisition. An image interpreter systematically examines the images and, frequently, other supporting materials such as maps and reports of field observations. Based on this study, an interpretation is made as to the physical nature of objects and phenomena appearing in the images. Interpretations may take place at a number of levels of complexity, from the simple recognition of objects on the earth's surface to the derivation of detailed information regarding the complex interactions among earth surface and subsurface features. Success in image interpretation varies with the training and experience of the interpreter, the nature of the objects or phenomena being interpreted, and the quality of the images being utilized. Generally, the most capable image interpreters have keen powers of observation coupled with imagination and a great deal of patience. In addition, it is important that the interpreter have a thorough understanding of the phenomenon being studied as well as knowledge of the geographic region under study.

Elements of Image Interpretation

Although most individuals have had substantial experience in interpreting "conventional" photographs in their daily lives (e.g., newspaper photographs), the interpretation of aerial and space images often departs from everyday image interpretation in three important respects: (1) the portrayal of features from an overhead, often unfamiliar, vertical perspective; (2) the frequent use of wavelengths outside of the visible portion of the spectrum; and (3) the depiction of the earth's surface at unfamiliar scales and resolutions (Campbell and Wynne, 2011). While these factors may be insignificant to the experienced image interpreter, they can represent a substantial challenge to the novice image analyst! However, even

this challenge continues to be mitigated by the extensive use of aerial and space imagery in such day-to-day activities as navigation, GIS applications, and weather forecasting.

A systematic study of aerial and space images usually involves several basic characteristics of features shown on an image. The exact characteristics useful for any specific task and the manner in which they are considered depend on the field of application. However, most applications consider the following basic characteristics, or variations of them: shape, size, pattern, tone (or hue), texture, shadows, site, association, and spatial resolution (Olson, 1960).

Shape refers to the general form, configuration, or outline of individual objects. In the case of stereoscopic images, the object's *height* also defines its shape. The shape of some objects is so distinctive that their images may be identified solely from this criterion. The Pentagon building near Washington, DC, is a classic example. All shapes are obviously not this diagnostic, but every shape is of some significance to the image interpreter.

Size of objects on images must be considered in the context of the image scale. A small storage shed, for example, might be misinterpreted as a barn if size were not considered. Relative sizes among objects on images of the same scale must also be considered.

Pattern relates to the spatial arrangement of objects. The repetition of certain general forms or relationships is characteristic of many objects, both natural and constructed, and gives objects a pattern that aids the image interpreter in recognizing them. For example, the ordered spatial arrangement of trees in an orchard is in distinct contrast to that of natural forest tree stands.

Tone (or *hue*) refers to the relative brightness or color of objects on an image. Figure 1.8 showed how relative photo tones could be used to distinguish between deciduous and coniferous trees on black and white infrared photographs. Without differences in tone or hue, the shapes, patterns, and textures of objects could not be discerned.

Texture is the frequency of tonal change on an image. Texture is produced by an aggregation of unit features that may be too small to be discerned individually on the image, such as tree leaves and leaf shadows. It is a product of their individual shape, size, pattern, shadow, and tone. It determines the overall visual "smoothness" or "coarseness" of image features. As the scale of the image is reduced, the texture of any given object or area becomes progressively finer and ultimately disappears. An interpreter can often distinguish between features with similar reflectances based on their texture differences. An example would be the smooth texture of green grass as contrasted with the rough texture of green tree crowns on medium-scale airphotos.

Shadows are important to interpreters in two opposing respects: (1) The shape or outline of a shadow affords an impression of the profile view of objects (which aids interpretation) and (2) objects within shadows reflect little light and are difficult to discern on an image (which hinders interpretation). For example, the shadows cast by various tree species or cultural features (bridges, silos,

towers, poles, etc.) can definitely aid in their identification on airphotos. In some cases, the shadows cast by large animals can aid in their identification. Figure 1.28 is a large-scale aerial photograph taken under low sun angle conditions that shows camels and their shadows in Saudi Arabia. Note that the camels themselves can be seen at the "base" of their shadows. Without the shadows, the animals could be counted, but identifying them specifically as camels could be difficult. Also, the shadows resulting from subtle variations in terrain elevations, especially in the case of low sun angle images, can aid in assessing natural topographic variations that may be diagnostic of various geologic landforms.

As a general rule, the shape of the terrain is more easily interpreted when shadows fall toward the observer. This is especially true when images are examined monoscopically, where relief cannot be seen directly, as it can be in stereoscopic images. In Figure 1.29*a*, a large ridge with numerous side valleys can be seen in the center of the image. When this image is inverted (i.e., turned such that the shadows fall away from the observer), as in (*b*), the result is a confusing image that almost seems to have a valley of sorts running through the center of the image (from bottom to top). This arises because one "expects" light sources to generally be above objects (ASPRS, 1997, p. 73). The orientation of shadows with respect to the observer is less important for interpreting images of buildings, trees, or animals (as in Figure 1.28) than for interpreting the terrain.

Site refers to topographic or geographic location and is a particularly important aid in the identification of vegetation types. For example, certain tree species would be expected to occur on well-drained upland sites, whereas other tree species would be expected to occur on poorly drained lowland sites. Also, various tree species occur only in certain geographic areas (e.g., redwoods occur in California, but not in Indiana).

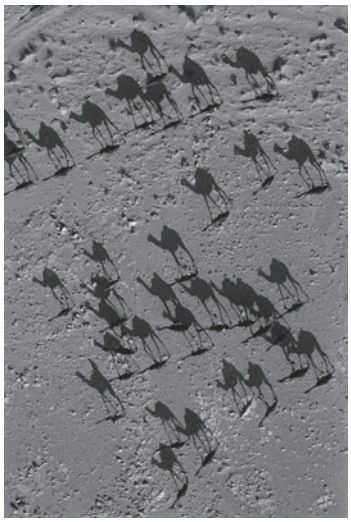


Figure 1.28 Vertical aerial photograph showing camels that cast long shadows under a low sun angle in Saudi Arabia. Black-and-white rendition of color original. (© George Steinmetz/Corbis.)

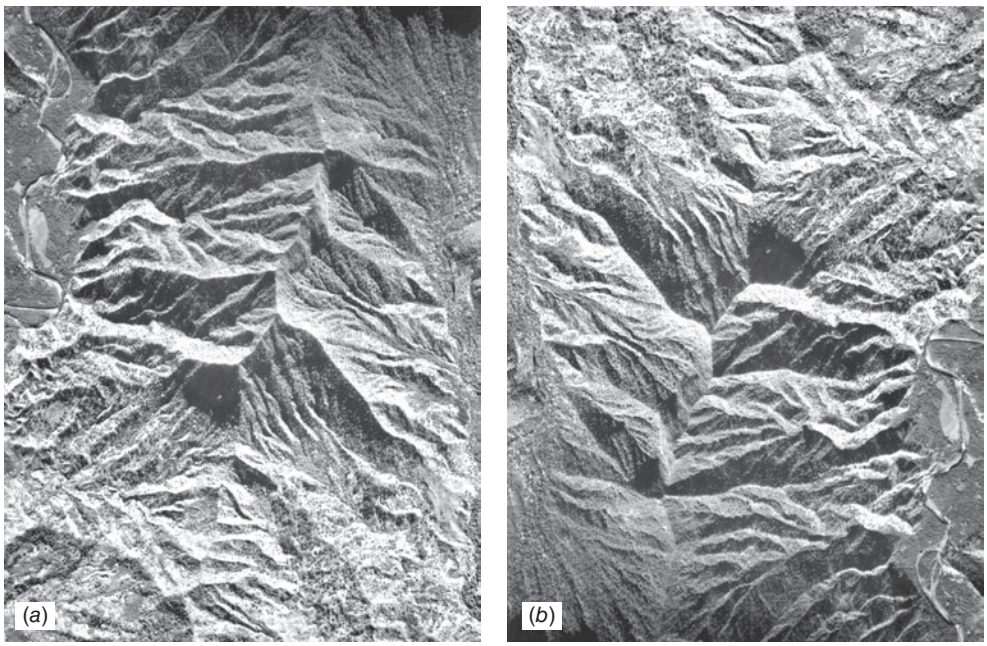


Figure 1.29 Photograph illustrating the effect of shadow direction on the interpretability of terrain. Island of Kauai, Hawaii, mid-January. Scale 1:48,000. (*a*) Shadows falling toward observer, (*b*) same image turned such that shadows are falling away from observer. (Courtesy USDA–ASCS panchromatic photograph.)

Association refers to the occurrence of certain features in relation to others. For example, a Ferris wheel might be difficult to identify if standing in a field near a barn but would be easy to identify if in an area recognized as an amusement park.

Spatial resolution depends on many factors, but it always places a practical limit on interpretation because some objects are too small or have too little contrast with their surroundings to be clearly seen on the image.

Other factors, such as image scale, spectral resolution, radiometric resolution, date of acquisition, and even the condition of images (e.g., torn or faded historical photographic prints) also affect the success of image interpretation activities.

Many of the above elements of image interpretation are illustrated in Figure 1.30. Figure 1.30*a* is a nearly full-scale copy of a 230 $\times$ 230-mm airphoto that was produced at an original scale of 1:28,000 (or 1 cm = 280 m). Parts (*b*) through (*e*) of Figure 1.30 show four scenes extracted and enlarged from this airphoto. Among the land cover types in Figure 1.30*b* are water, trees, suburban houses, grass, a divided highway, and a drive-in theater. Most of the land cover types are easily identified in this figure. The drive-in theater could be difficult for inexperienced interpreters to identify, but a careful study of the elements of image interpretation leads to its identification. It has a unique *shape* and *pattern*. Its *size* is consistent with a drive-in theater (note the relative size of the cars on

Figure 1.30 Aerial photographic subscenes illustrating the elements of image interpretation, Detroit Lakes area, Minnesota, mid-October. (*a*) Portion of original photograph of scale 1:32,000; (*b*) and (*c*) enlarged to a scale of 1:4,600; (*d*) enlarged to a scale of 1:16,500; (*e*) enlarged to a scale of 1:25,500. North is to the bottom of the page. (Courtesy KBM, Inc.)

Figure 1.30 (*Continued*)

the highway and the parking spaces of the theater). In addition to the curved rows of the parking area, the *pattern* also shows the projection building and the screen. The identification of the screen is aided by its *shadow*. It is located in *association* with a divided highway, which is accessed by a short roadway.

Many different land cover types can be seen in Figure 1.30*c*. Immediately noticeable in this photograph, at upper left, is a feature with a superficially similar appearance to the drive-in theater. Careful examination of this feature, and the surrounding grassy area, leads to the conclusion that this is a baseball diamond. The trees that can be seen in numerous places in the photograph are casting shadows of their trunks and branches because the mid-October date of this photograph is a time when deciduous trees are in a leaf-off condition. Seen in the right one-third of the photograph is a residential area. Running top to bottom through the center of the

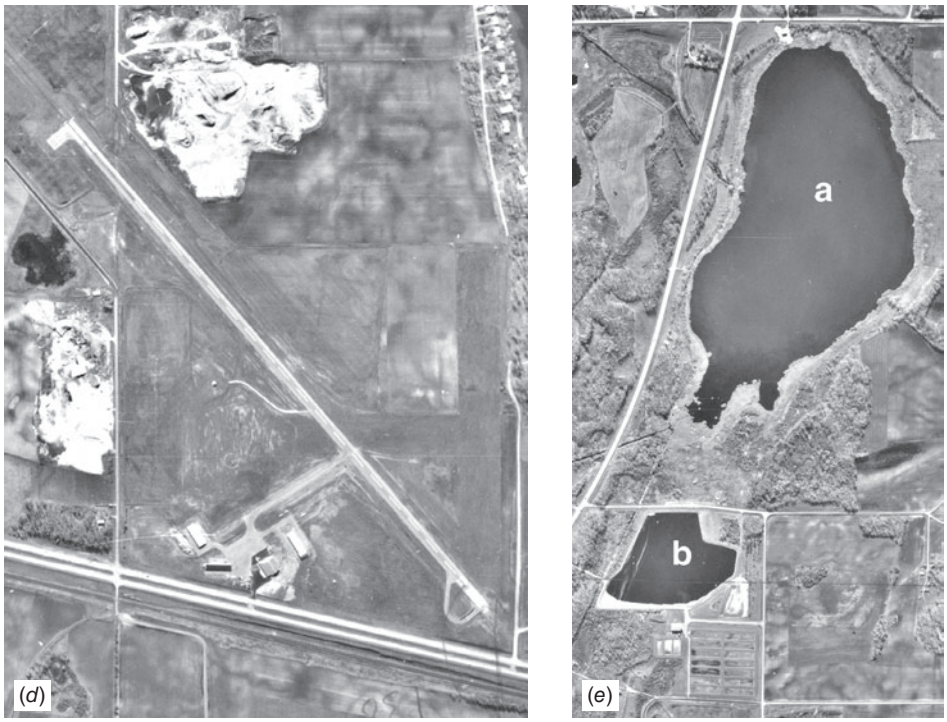


Figure 1.30 (*Continued*)

image is a commercial area with buildings that have a larger size than the houses in the residential area and large parking areas surrounding these larger buildings.

Figure 1.30*d* shows two major linear features. Near the bottom of the photograph is a divided highway. Running diagonally from upper left to lower right is an airport runway 1390 m long (the scale of this figure is 1:16,500, and the length of this linear feature is 8.42 cm at this scale). The terminal area for this airport is near the bottom center of Figure 1.30*d*.

Figure 1.30*e* illustrates natural versus constructed features. The water body at *a* is a natural feature, with an irregular shoreline and some surrounding wetland areas (especially visible at the narrow end of the lake). The water body at *b* is part of a sewage treatment plant; the "shoreline" of this feature has unnaturally straight sections in comparison with the water body shown at *a*.

Image Interpretation Strategies

As previously mentioned, the image interpretation process can involve various levels of complexity, from a simple direct recognition of objects in the scene to the inference of site conditions. An example of direct recognition would be the

identification of a highway interchange. Assuming the interpreter has some experience with the vertical perspective of aerial and space images, recognition of a highway interchange should be a straightforward process. On the other hand, it may often be necessary to infer, rather than directly observe, the characteristics of features based on their appearance on images. In the case of a buried gas pipeline, for example, the actual pipeline cannot be seen, but there are often changes at the ground surface caused by the buried pipeline that are visible on aerial and space images. Soils are typically better drained over the pipeline because of the sand and gravel used for backfill, and the presence of a buried pipeline can often be inferred by the appearance of a light-toned linear streak across the image. Also, the interpreter can take into account the probability of certain ground cover types occurring on certain dates at certain places. Knowledge of the crop development stages (*crop calendar*) for an area would determine if a particular crop is likely to be visible on a particular date. For example, corn, peas, and winter wheat would each have a significant vegetative ground cover on different dates. Likewise, in a particular growing region, one crop type may be present over a geographic area many times larger than that of another crop type; therefore, the probability of occurrence of one crop type would be much greater than another.

In a sense, the image interpretation process is like the work of a detective trying to put all the pieces of evidence together to solve a mystery. For the interpreter, the mystery might be presented in terms of trying to understand why certain areas in an agricultural field look different from the rest of that field. At the most general level, the interpreter must recognize the area under study as an agricultural field. Beyond this, consideration might be made as to whether the crop present in the field is a row crop (e.g., corn) or a continuous cover crop (e.g., alfalfa). Based on the crop calendar and regional growing conditions, a decision might be made that the crop is indeed corn, rather than another row crop, such as soybeans. Furthermore, it might be noted that the anomalously appearing areas in the field are associated with areas of slightly higher topographic relief relative to the rest of the field. With knowledge of the recent local weather conditions, the interpreter might infer that the anomalously appearing areas are associated with drier soil conditions and the corn in these areas is likely drought stressed. Hence, the interpreter uses the process of *convergence of evidence* to successively increase the accuracy and detail of the interpretation.

Image Interpretation Keys

The image interpretation process can often be facilitated through the use of *image interpretation keys*. Keys can be valuable training aids for novice interpreters and provide useful reference or refresher materials for more experienced interpreters. An image interpretation key helps the interpreter evaluate the information presented on aerial and space images in an organized and consistent manner. It provides guidance about the correct identification of features or conditions on the

images. Ideally, a key consists of two basic parts: (1) a collection of annotated or captioned images (preferably stereopairs) illustrative of the features or conditions to be identified and (2) a graphic or word description that sets forth in some systematic fashion the image recognition characteristics of those features or conditions. Two general types of image interpretation keys exist, differentiated by the method of presentation of diagnostic features. A *selective key* contains numerous example images with supporting text. The interpreter selects the example that most nearly resembles the feature or condition found on the image under study.

An *elimination key* is arranged so that the interpretation proceeds step by step from the general to the specific and leads to the elimination of all features or conditions except the one being identified. Elimination keys often take the form of *dichotomous keys* where the interpreter makes a series of choices between two alternatives and progressively eliminates all but one possible answer. Figure 1.31 shows a dichotomous key prepared for the identification of fruit and nut crops in the Sacramento Valley, California. The use of elimination keys can lead to more positive

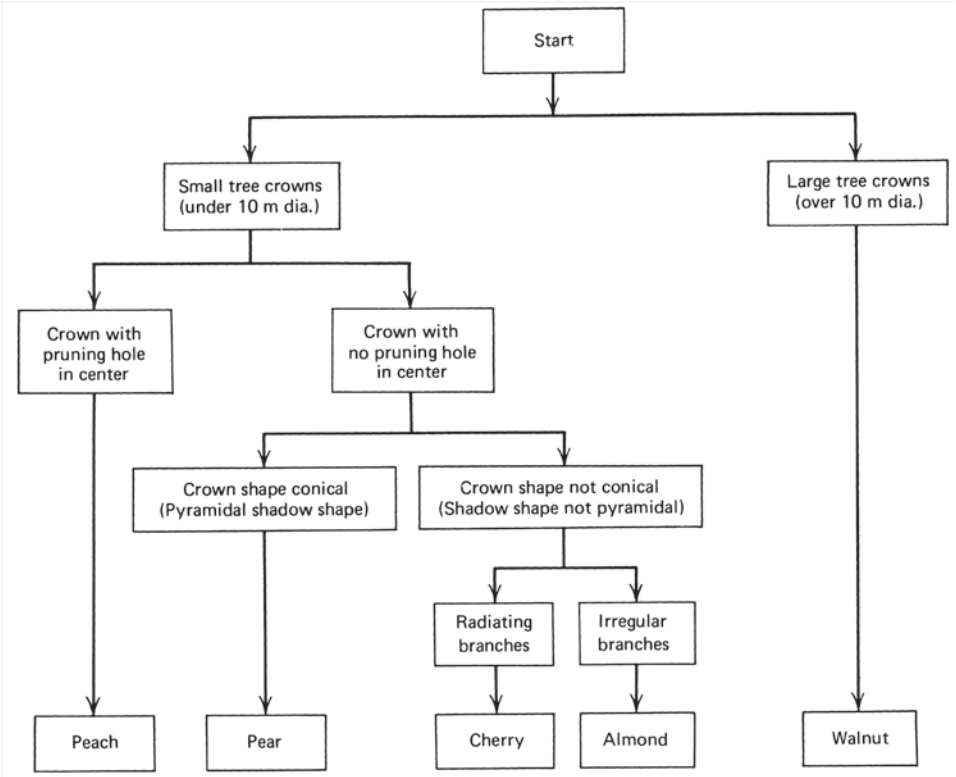


Figure 1.31 Dichotomous airphoto interpretation key to fruit and nut crops in the Sacramento Valley, CA, designed for use with 1:6,000 scale panchromatic aerial photographs. (Adapted from American Society of Photogrammetry, 1983. Copyright © 1975, American Society of Photogrammetry. Reproduced with permission.)

answers than selective keys but may result in erroneous answers if the interpreter is forced to make an uncertain choice between two unfamiliar image characteristics.

As a generalization, keys are more easily constructed and more reliably utilized for cultural feature identification (e.g., houses, bridges, roads, water towers) than for vegetation or landform identification. However, a number of keys have been successfully employed for agricultural crop identification and tree species identification. *Such keys are normally developed and used on a region-by-region and season-by-season basis because the appearance of vegetation can vary widely with location and season*.

Wavelengths of Sensing

The band(s) of the electromagnetic energy spectrum selected for aerial and space imaging affects the amount of information that can be interpreted from the images. Numerous examples of this are scattered throughout this book. The general concepts of multiband imagery were discussed in Section 1.5. To explain how the combinations of colors shown in an image relate to the various bands of data recorded by a sensor, we next turn our attention to the principles of color, and how combinations of colors are perceived.

Color Perception and Color Mixing

Color is among the most important elements of image interpretation. Many features and phenomena in an image can best be identified and interpreted through examination of subtle differences in color. As discussed in Section 1.5, multiwavelength remote sensing images may be displayed in either true- or false-color combinations. Particularly for interpreting false-color imagery, an understanding of the principles of color perception and color mixing is essential.

Light falling on the retina of the human eye is sensed by rod and cone cells. There are about 130 million rod cells, and they are 1000 times more light sensitive than the cone cells. When light levels are low, human vision relies on the rod cells to form images. All rod cells have the same wavelength sensitivity, which peaks at about 0.55 μm. Therefore, human vision at low light levels is monochromatic. It is the cone cells that determine the colors the eye sees. There are about 7 million cone cells; some sensitive to blue energy, some to green energy, and some to red energy. The *trichromatic theory of color vision* explains that when the blue-sensitive, green-sensitive, and red-sensitive cone cells are stimulated by different amounts of light, we perceive color. When all three types of cone cells are stimulated equally, we perceive white light. Other theories of color vision have been proposed. The *opponent process of color vision* hypothesizes that color vision involves three mechanisms, each responding to a pair of so-called opposites: white–black, red–green, and blue–yellow. This theory is based on many psychophysical observations and states that colors are formed by a *hue cancellation*

method. The hue cancellation method is based on the observation that when certain colors are mixed together, the resulting colors are not what would be intuitively expected. For example, when red and green are mixed together, they produce yellow, not reddish green. (For further information, see Robinson et al., 1995.)

In the remainder of this discussion, we focus on the trichromatic theory of color vision. Again, this theory is based on the concept that we perceive all colors by synthesizing various amounts of just three (blue, green, and red).

Blue, green, and red are termed *additive primaries*. Plate 3*a* shows the effect of projecting blue, green, and red light in partial superimposition. Where all three beams overlap, the visual effect is white because all three of the eyes' receptor systems are stimulated equally. Hence, white light can be thought of as the mixture of blue, green, and red light. Various combinations of the three additive primaries can be used to produce other colors. As illustrated, when red light and green light are mixed, yellow light is produced. Mixture of blue and red light results in the production of magenta light (bluish-red). Mixing blue and green results in cyan light (bluish-green).

Yellow, magenta, and cyan are known as the *complementary colors*, or *complements*, of blue, green, and red light. Note that the complementary color for any given primary color results from mixing the remaining two primaries.

Like the eye, color television and computer monitors operate on the principle of additive color mixing through use of blue, green, and red elements on the screen. When viewed at a distance, the light from the closely spaced screen elements forms a continuous color image.

Whereas color television and computer monitors simulate different colors through *additive* mixture of blue, green, and red *lights*, color film photography is based on the principle of *subtractive* color mixture using superimposed yellow, magenta, and cyan *dyes*. These three dye colors are termed the *subtractive primaries*, and each results from subtracting one of the additive primaries from white light. That is, yellow dye absorbs the blue component of white light. Magenta dye absorbs the green component of white light. Cyan dye absorbs the red component of white light.

The subtractive color-mixing process is illustrated in Plate 3*b*. This plate shows three circular filters being held in front of a source of white light. The filters contain yellow, magenta, and cyan dye. The yellow dye absorbs blue light from the white background and transmits green and red. The magenta dye absorbs green light and transmits blue and red. The cyan dye absorbs red light and transmits blue and green. The superimposition of magenta and cyan dyes results in the passage of only blue light from the background. This comes about because the magenta dye absorbs the green component of the white background, and the cyan dye absorbs the red component. Superimposition of the yellow and cyan dyes results in the perception of green. Likewise, superimposition of yellow and magenta dyes results in the perception of red. Where all three dyes overlap, all light from the white background is absorbed and black results.

In color film photography, and in color printing, various proportions of yellow, magenta, and cyan dye are superimposed to control the proportionate

amount of blue, green, and red light that reaches the eye. Hence, the subtractive mixture of yellow, magenta, and cyan dyes on a photograph is used to control the additive mixture of blue, green, and red light reaching the eye of the observer. To accomplish this, color film is manufactured with three emulsion layers that are sensitive to blue, green, and red light but contain yellow, magenta, and cyan dye after processing (see Section 2.4).

In digital color photography, the photosites in the detector array are typically covered with a blue, green, or red filter, resulting in independent recording of the three additive primary colors (see Section 2.5).

When interpreting color images, the analyst should keep in mind the relationship between the color of a feature in the imagery, the color mixing process that would produce that color, and the sensor's wavelength ranges that are assigned to the three primary colors used in that mixing process. It is then possible to work backwards to infer the spectral properties of the feature on the landscape. For example, if a feature in a false-color image has a yellow hue when displayed on a computer monitor, that feature can be assumed to have a relatively high reflectance in the wavelengths that are being displayed in the monitor's red and green color planes, and a relatively low reflectance in the wavelength displayed in blue (because yellow results from the additive combination of red plus green light). Knowing the spectral sensitivity of each of the sensor's spectral bands, the analyst can interpret the color as a spectral response pattern (see Section 1.4) that is characterized by high reflectance at two wavelength ranges and low reflectance at a third. The analyst can then use this information to draw inferences about the nature and condition of the feature shown in the false-color image.

Temporal Aspects of Image Interpretation

The temporal aspects of natural phenomena are important for image interpretation because such factors as vegetative growth and soil moisture vary during the year. For crop identification, more positive results can be achieved by obtaining images at several times during the annual growing cycle. Observations of local vegetation emergence and recession can aid in the timing of image acquisition for natural vegetation mapping. In addition to seasonal variations, weather can cause significant short-term changes. Because soil moisture conditions may change dramatically during the day or two immediately following a rainstorm, the timing of image acquisition for soil studies is very critical.

Another temporal aspect of importance is the comparison of *leaf-off* photography with *leaf-on* photography. Leaf-off conditions are preferable for applications in which it is important to be able to see as much detail as possible underneath trees. Such applications include activities such as topographic mapping and urban feature identification. Leaf-on conditions are preferred for vegetation mapping. Figure 1.32*a* illustrates leaf-on photography. Here considerable detail of the ground surface underneath the tree crowns is obscured. Figure 1.32*b*

Figure 1.32 Comparison of leaf-on photography with leaf-off photography. Gladstone, OR. (*a*) Leaf-on photograph exposed in summer. (*b*) Leaf-off photograph exposed in spring. Scale 1:1,500. (Courtesy Oregon Metro.)

illustrates leaf-off photography. Here, the ground surface underneath the tree crowns is much more visible than in the leaf-on photograph. Because leaf-off photographs are typically exposed in spring or fall, there are longer shadows in the image than with leaf-on photography, which is typically exposed in summer. (Shadow length also varies with time of day.) Also, these photographs illustrate leaf-on and leaf-off conditions in an urban area where most of the trees are deciduous and drop their leaves in fall (leaf-off conditions). The evergreen trees in the images (e.g., lower right) maintain their needles throughout the year and cast dark shadows. Hence, there would not be leaf-off conditions for such trees.

Image Spatial Resolution and Ground Sample Distance

Every remote sensing system has a limit on how small an object on the earth's surface can be and still be "seen" by a sensor as being separate from its surroundings. This limit, called the *spatial resolution* of a sensor, is an indication of how well a sensor can record spatial detail. In some cases, the *ground sample distance* (*GSD*), or ground area represented by a single pixel in a digital image, may correspond closely to the spatial resolution of that sensor. In other cases, the ground sample distance may be larger or smaller than the sensor's spatial resolution, perhaps as a result of the A-to-D conversion process or of digital image manipulation such as resampling (see Chapter 7 and Appendix B). This distinction between spatial resolution and ground sample distance is subtle but important. For the sake of simplicity, the following discussion treats the GSD in a digital image as being equivalent to the spatial resolution of the sensor that produced the image, but note that in actual images the sampling distance may be larger or smaller than the spatial resolution.

Figure 1.33 illustrates, in the context of a digital image, the interplay between the spatial resolution of a sensor and the spatial variability present in a ground scene. In (*a*), a single pixel covers only a small area of the ground (on the order of

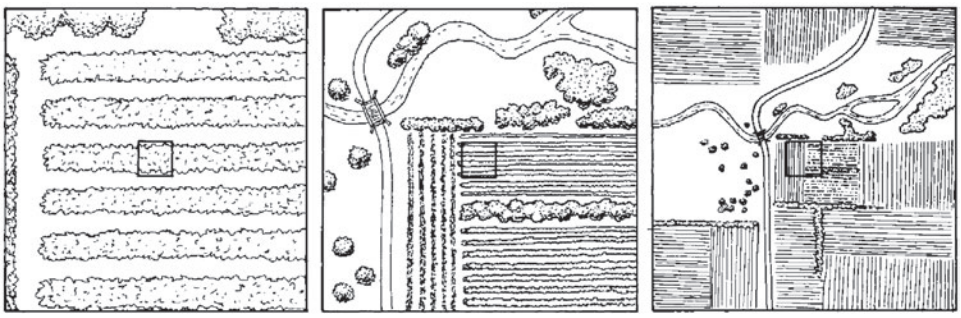


Figure 1.33 Ground resolution cell size effect: (*a*) small, (*b*) intermediate, and (*c*) large ground resolution cell size.

the width of the rows of the crop shown). In (*b*), a coarser ground resolution is depicted and a single pixel integrates the radiance from both the crop rows and the soil between them. In (*c*), an even coarser resolution results in a pixel measuring the average radiance over portions of two fields. Thus, depending on the spatial resolution of the sensor and the spatial structure of the ground area being sensed, digital images comprise a range of "pure" and "mixed" pixels. In general, the larger the percentage of mixed pixels, the more limited is the ability to record and extract spatial detail in an image. This is illustrated in Figure 1.34, in which the same area has been imaged over a range of different ground resolution cell sizes.

Further discussion of the spatial resolution of remote sensing systems—including the factors that determine spatial resolution and the methods used for measuring or calculating a system's resolution—can be found in Chapter 2 (for camera systems), Chapters 4 and 5 (for airborne and spaceborne multispectral and thermal sensors), and Chapter 6 (for radar systems).

Other Forms of Resolution Important in Image Interpretation

It should be noted that there are other forms of resolution that are important characteristics of remote sensing images. These include the following:

Spectral resolution, referring to a sensor's ability to distinguish among different ground features based on their spectral properties. Spectral resolution depends upon the number, wavelength location, and narrowness of the spectral bands in which a sensor collects image data. The bands in which any sensor collects data can range from a single broad band (for *panchromatic* images), a few broad bands (for *multispectral* images), or many very narrow bands (for *hyperspectral* images).

Radiometric resolution, referring to the sensor's ability to differentiate among subtle variations in brightness. Does the sensor divide the range from the "brightest" pixel to "darkest" pixel that can be recorded in an image (the *dynamic range*) into 256, or 512, or 1024 gray level values? The finer the radiometric resolution is, the greater the quality and interpretability of an image. (See also the discussion of quantization and digital numbers in Section 1.5.)

Temporal resolution, referring to the ability to detect changes over shorter or longer periods of time. Most often, this term is used in reference to a sensor that produces a time-series of multiple images. This could be a satellite system with a defined 16-day or 26-day orbital repeat cycle, or a tripod-mounted camera with a timer that collects one image every hour to serve as reference data. The importance of rapid and/or repeated coverage of an area varies dramatically with the application at hand. For example, in disaster response applications, temporal resolution might outweigh the importance of some, or all, of the other types of resolution we have summarized above.

In subsequent chapters, as new remote sensing systems are introduced and discussed, the reader should keep in mind these multiple resolutions that determine whether a given system would be suitable for a particular application.

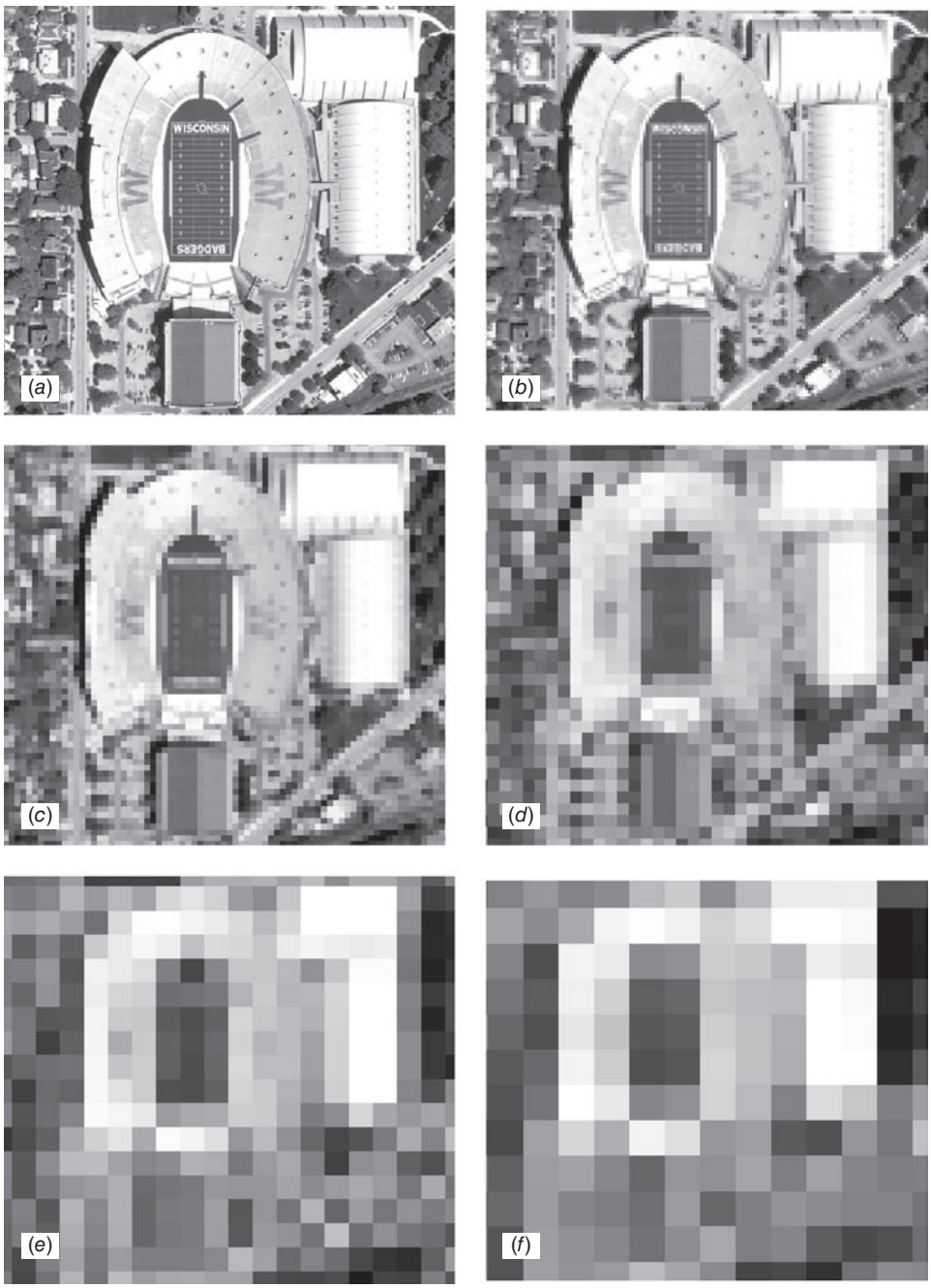


Figure 1.34 Ground resolution cell size effect on ability to extract detail from a digital image. Shown is a portion of the University of Wisconsin-Madison campus, including Camp Randall Stadium and vicinity, at a ground resolution cell size (per pixel) of: (*a*) 1 m, (*b*) 2.5 m, (*c*) 5 m, (*d*) 10 m, (*e*) 20 m, and (*f*) 30 m, and an enlarged portion of the image at (*g*) 0.5 m, (*h*) 1 m, and (*i*) 2.5 m. (Courtesy University of Wisconsin-Madison, Environmental Remote Sensing Center, and NASA Affiliated Research Center Program.)

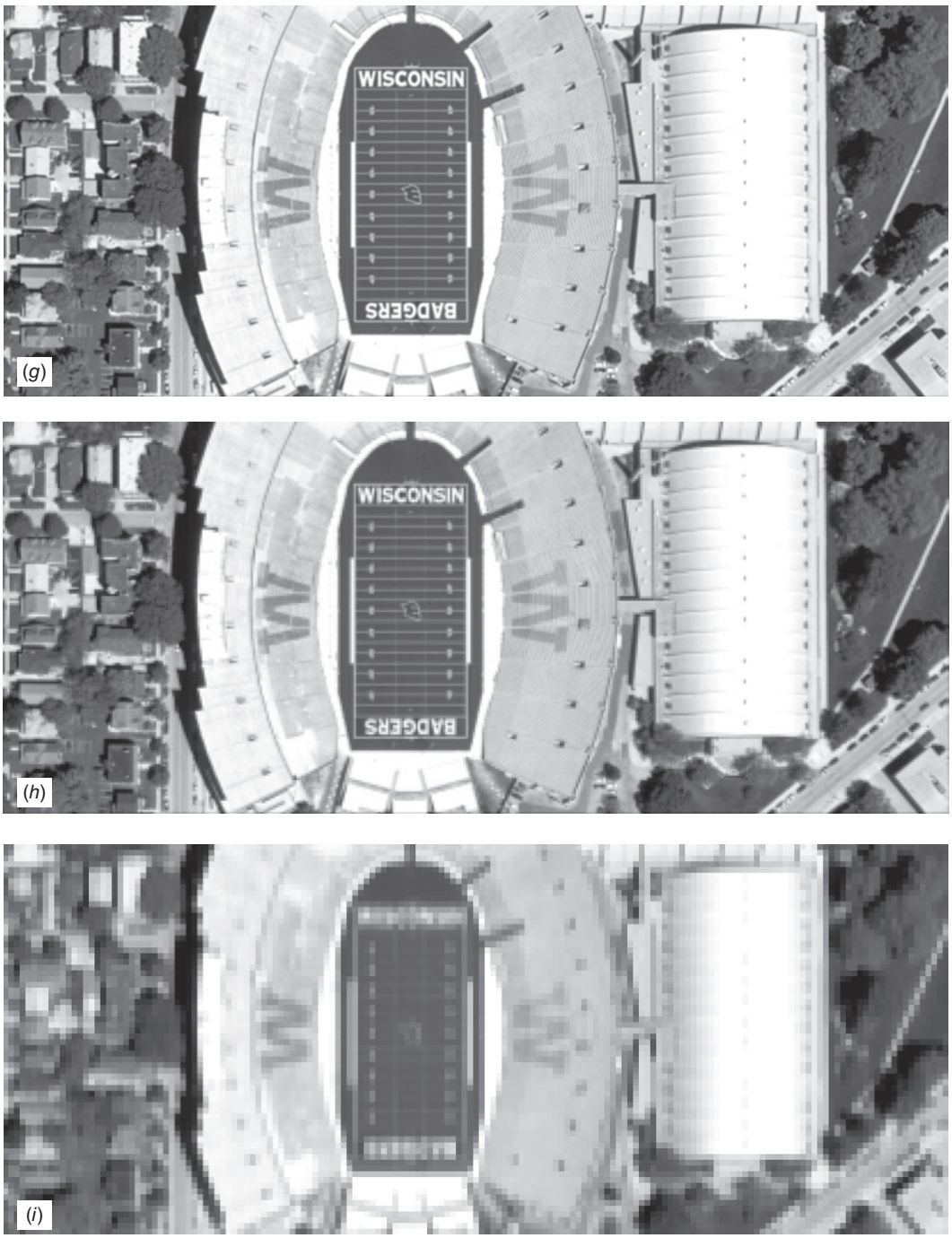


Figure 1.34 (*Continued*)

Image Scale

Image scale, discussed in detail in Section 3.3, affects the level of useful information that can be extracted from aerial and space images. The scale of an image can be thought of as the relationship between a distance measured on the image and the corresponding distance on the ground. Although terminology with regard to image scale has not been standardized, we can consider that *small-scale* images have a scale of 1:50,000 or smaller, *medium-scale* images have a scale between 1:12,000 and 1:50,000, and *large-scale* airphotos have a scale of 1:12,000 or larger.

In the case of digital data, images do not have a fixed scale per se; rather, they have a specific ground sample distance, as discussed previously (and illustrated in Figures 1.33 and 1.34), and can be reproduced at various scales. Thus, one could refer to the *display scale* of a digital image as it is displayed on a computer monitor or as printed in hardcopy.

In the figure captions of this book, we have stated the hardcopy display scale of many images—including photographic, multispectral, and radar images—so that the reader can develop a feel for the degree of detail that can be extracted from images of varying scales.

As generalizations, the following statements can be made about the appropriateness of various image scales for resource studies. Small-scale images are used for regional surveys, large-area resource assessment, general resource management planning, and large-area disaster assessment. Medium-scale images are used for the identification, classification, and mapping of such features as tree species, agricultural crop type, vegetation community, and soil type. Large-scale images are used for the intensive monitoring of specific items such as surveys of the damage caused by plant disease, insects, or tree blowdown. Large-scale images are also used for emergency response to such events as hazardous waste spills and planning search and rescue operations in association with tornadoes, floods, and hurricanes.

In the United States, the National High-Altitude Photography (NHAP) program, later renamed the National Aerial Photography Program (NAPP), was a federal multiagency activity coordinated by the U.S. Geological Survey (USGS). It provided nationwide photographic coverage at nominal scales ranging from 1:80,000 and 1:58,000 (for NHAP) to 1:40,000 (for NAPP). The archive of NHAP and NAPP photos has proven to be an extremely valuable ongoing source of medium-scale images supporting a wide range of applications.

The National Agriculture Imagery Program (NAIP) acquires peak growing season leaf-on imagery in the continental United States and delivers this imagery to U.S. Department of Agriculture (USDA) County Service Centers in order to assist with crop compliance and a multitude of other farm programs. NAIP imagery is typically acquired with GSDs of one to two meters. The one-meter GSD imagery is intended to provide updated digital orthophotography. The two-meter GSD imagery is intended to support USDA programs that require current imagery

acquired during the agricultural growing season but do not require high horizontal accuracy. NAIP photographs are also useful in many non-USDA applications, including real estate, recreation, and land use planning.

Approaching the Image Interpretation Process

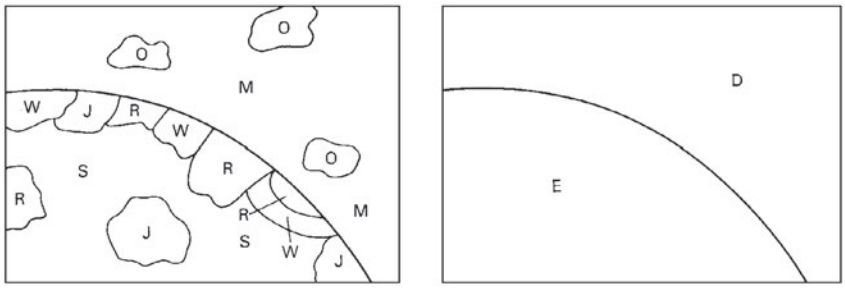
There is no single "right" way to approach the image interpretation process. The specific image products and interpretation equipment available will, in part, influence how a particular interpretation task is undertaken. Beyond these factors, the specific goals of the task will determine the image interpretation process employed. Many applications simply require the image analyst to identify and count various discrete objects occurring in a study area. For example, counts may be made of such items as motor vehicles, residential dwellings, recreational watercraft, or animals. Other applications of the interpretation process often involve the identification of anomalous conditions. For example, the image analyst might survey large areas looking for such features as failing septic systems, sources of water pollution entering a stream, areas of a forest stressed by an insect or disease problem, or evidence of sites having potential archaeological significance.

Many applications of image interpretation involve the delineation of discrete areal units throughout images. For example, the mapping of land use, soil types, or forest types requires the interpreter to outline the boundaries between areas of one type versus another. Such tasks can be problematic when the boundary is not a discrete edge, but rather a "fuzzy edge" or gradation from one type of area to another, as is common with natural phenomena such as soils and natural vegetation.

Two extremely important issues must be addressed before an interpreter undertakes the task of delineating separate areal units on remotely sensed images. The first is the definition of the *classification system* or criteria to be used to separate the various categories of features occurring in the images. For example, in mapping land use, the interpreter must fix firmly in mind what specific characteristics determine if an area is "residential," "commercial," or "industrial." Similarly, the forest type mapping process must involve clear definition of what constitutes an area to be delineated in a particular species, height, or crown density class.

The second important issue in delineation of discrete areal units on images is the selection of the *minimum mapping unit* (MMU) to be employed in the process. This refers to the smallest size areal entity to be mapped as a discrete area. Selection of the MMU will determine the extent of detail conveyed by an interpretation. This is illustrated in Figure 1.35. In (*a*), a small MMU results in a much more detailed interpretation than does the use of a large MMU, as illustrated in (*b*).

Once the classification system and MMU have been determined, the interpreter can begin the process of delineating boundaries between feature types. Experience suggests that it is advisable to delineate the most highly contrasting feature types first and to work from the general to the specific. For example, in a



(*a*) (*b*)

Figure 1.35 Influence of minimum mapping unit size on interpretation detail. (*a*) Forest types mapped using a small MMU: O, oak; M, maple; W, white pine; J, jack pine; R, red pine; S, spruce. (*b*) Forest types mapped using a large MMU: D, deciduous; E, evergreen.

land use mapping effort, it would be better to separate "urban" from "water" and "agriculture" before separating more detailed categories of each of these feature types based on subtle differences.

In certain applications, the interpreter might choose to delineate *photomorphic regions* as part of the delineation process. These are regions of reasonably uniform tone, texture, and other image characteristics. When initially delineated, the feature type identity of these regions may not be known. Field observations or other ground truth can then be used to verify the identity of each region. Regrettably, there is not always a one-to-one correspondence between the appearance of a photomorphic region and a mapping category of interest. However, the delineation of such regions often serves as a stratification tool in the interpretation process and can be valuable in applications such as vegetation mapping (where photomorphic regions often correspond directly to vegetation classes of interest).

Basic Equipment for Visual Interpretation

Visual image interpretation equipment generally serves one of several fundamental purposes: viewing images, making measurements on images, performing image interpretation tasks, and transferring interpreted information to base maps or digital databases. Basic equipment for viewing images and transferring interpreted information is described here. Equipment involved in performing measuring and mapping tasks will be described in Chapter 3.

The airphoto interpretation process typically involves the utilization of stereoscopic viewing to provide a three-dimensional view of the terrain. Some space images are also analyzed stereoscopically. The stereo effect is possible because we have binocular vision. That is, since we have two eyes that are slightly separated, we continually view the world from two slightly different perspectives. Whenever

objects lie at different distances in a scene, each eye sees a slightly different view of the objects. The differences between the two views are synthesized by the mind to provide depth perception. Thus, the two views provided by our separated eyes enable us to see in three dimensions.

Vertical aerial photographs are often taken along flight lines such that successive images overlap by at least 50% (see Figure 3.2). This overlap also provides two views taken from separated positions. By viewing the left image of a pair with the left eye and the right image with the right eye, we obtain a three-dimensional view of the terrain surface. The process of stereoscopic viewing can be done using a traditional *stereoscope*, or using various methods for stereoscopic viewing on computer monitors. This book contains many *stereopairs*, or *stereograms*, which can be viewed in three dimensions using a lens stereoscope such as shown in Figure 1.36. An average separation of about 58 mm between common points has been used in the stereograms in this book. The exact spacing varies somewhat because of the different elevations of the points.

Typically, images viewed in stereo manifest *vertical exaggeration*. This is caused by an apparent difference between the vertical scale and the horizontal scale of the stereomodel. Because the vertical scale appears to be larger than the horizontal scale, objects in the stereomodel appear to be too tall. A related consideration is that slopes in the stereomodel will appear to be steeper than they actually are. (The geometric terms and concepts used in this discussion of vertical exaggeration are explained in more detail in Chapter 3.)

Figure 1.36 Simple lens stereoscope. (Author-prepared figure.)

Many factors contribute to vertical exaggeration, but the primary cause is the lack of equivalence between the original, in-flight, *photographic base–height ratio,* B/H' (Figure 3.24*a*), and the corresponding, in-office, *stereoviewing base–height ratio,* b_e/h_e (Figure 1.37). The perceived vertical exaggeration in the stereomodel is approximately the ratio of these two ratios. The photographic base–height ratio is the ratio of the *air base* distance between the two exposure stations to the flying height above the average terrain elevation. The stereoviewing base–height ratio is the ratio between the viewer's *eye base* (b_e) to the distance from the eyes at which the stereomodel is perceived (h_e). The perceived vertical exaggeration, *VE,* is approximately the ratio of the photographic base–height ratio to the stereoviewing base–height ratio,

$$VE \cong \frac{B/H'}{b_e/h_e} \tag{1.10}$$

where $b_e/h_e = 0.15$ for most observers.

In short, vertical exaggeration varies directly with the photographic base–height ratio.

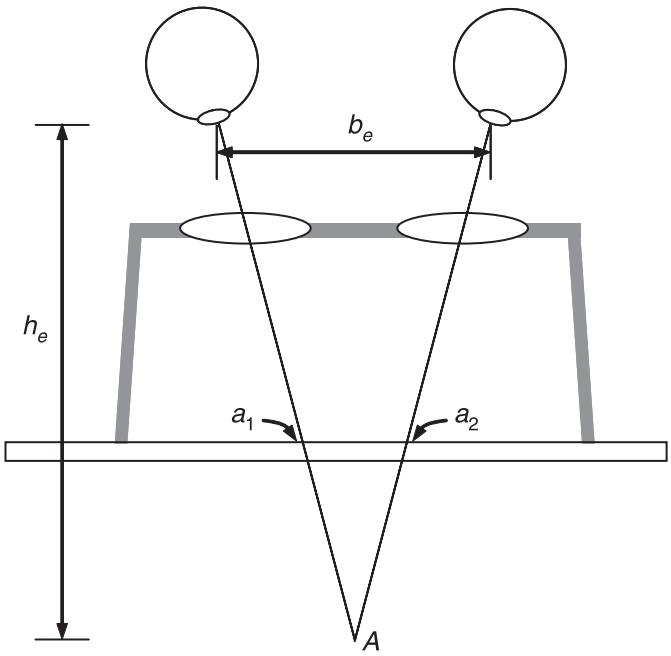


Figure 1.37 Major cause of vertical exaggeration perceived when viewing a stereomodel. There is a lack of equivalence between the photographic base–height ratio when the photographs are taken and the stereoviewing base–height ratio when the photographs are viewed with a stereoscope.

While vertical exaggeration is often misleading to novice image interpreters, it is often very useful in the interpretation process. This is due to the fact that subtle variation in the heights of image features is more readily discriminated when exaggerated. As we note in Chapter 3, large base–height ratios also improve the quality of many photogrammetric measurements made from vertical aerial photographs.

Figure 1.38 can be used to test stereoscopic vision. When this diagram is viewed through a stereoscope, the rings and other objects should appear to be at varying distances from the observer. Your stereovision ability can be evaluated by filling in Table 1.3 (answers are in the second part of the table). People whose eyesight is very weak in one eye may not have the ability to see in stereo. This will preclude three-dimensional viewing of the stereograms in this book. However, many people with essentially monocular vision have become proficient photo interpreters. In fact, many forms of interpretation involve monocular viewing with such basic equipment as handheld magnifying glasses or tube magnifiers (2× to 10× lenses mounted in a transparent stand).

Some people will be able to view the stereograms in this book without a stereoscope. This can be accomplished by holding the book about 25 cm from your eyes and allowing the view of each eye to drift into a straight-ahead viewing position (as when looking at objects at an infinite distance) while still maintaining focus on the stereogram. When the two images have fused into one, the stereogram will be seen in three dimensions. Most persons will find stereoviewing without proper stereoscopes to be a tiring procedure, producing eyestrain. It is, however, a useful technique to employ when stereoscopes are not available.

Several types of analog stereoscopes are available, utilizing lenses or a combination of lenses, mirrors, and prisms. *Lens stereoscopes*, such as the one shown in Figure 1.36, are portable and comparatively inexpensive. Most are small instruments with folding legs. The lens spacing can usually be adjusted from about

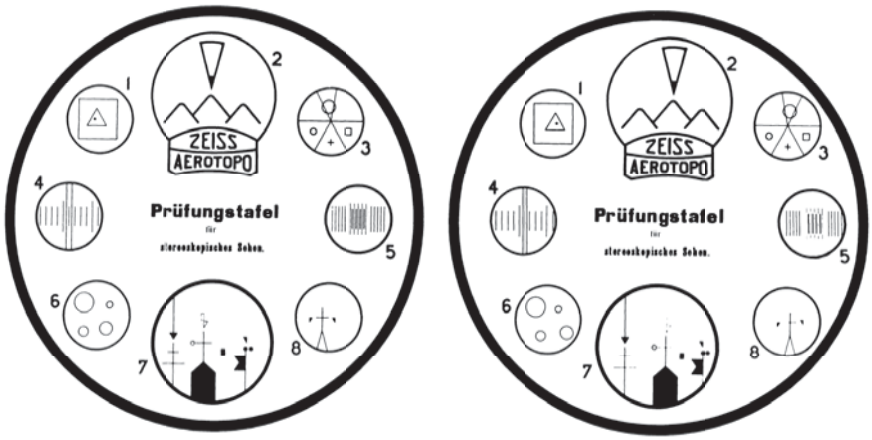


Figure 1.38 Stereoscopic vision test. (Courtesy Carl Zeiss, Inc.)

TABLE 1.3 Stereovision Test for Use with Figure 1.38

PART I

Within the rings marked 1 through 8 are designs that appear to be at different elevations. Using "1" to designate the highest elevation, write down the depth order of the designs. It is possible that two or more designs may be at the same elevation. In this case, use the same number for all designs at the same elevation.

Ring 1		**Ring 6**	
Square	(2)	Lower left circle	( )
Marginal ring	(1)	Lower right circle	( )
Triangle	(3)	Upper right circle	( )
Point	(4)	Upper left circle	( )
		Marginal ring	( )
Ring 7		**Ring 3**	
Black flag with ball	( )	Square	( )
Marginal ring	( )	Marginal ring	( )
Black circle	( )	Cross	( )
Arrow	( )	Lower left circle	( )
Tower with cross	( )	Upper left circle	( )
Double cross	( )		
Black triangle	( )		
Black rectangle	( )		

PART II

Indicate the relative elevations of the rings 1 through 8.

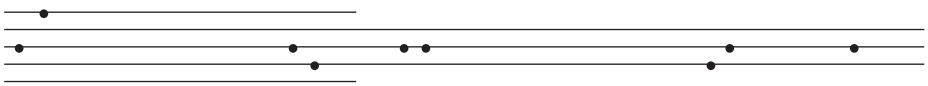
( ) ( ) ( ) ( ) ( ) ( ) ( ) ( )

Highest Lowest

PART III

Draw profiles to indicate the relative elevations of the letters in the words "prufungstafel" and "stereoskopisches sehen."

P R U F U N G S T A F E L S T E R E O S K O P I S C H E S S E H E N



(Answers to Stereovision Test on next page)

45 to 75 mm to accommodate individual eye spacings. Lens magnification is typically 2 power but may be adjustable. The principal disadvantage of small lens stereoscopes is that the images must be quite close together to be positioned properly underneath the lenses. Because of this, the interpreter cannot view the entire stereoscopic area contained by the overlapping aerial photographs without

TABLE 1.3 *(Continued)*

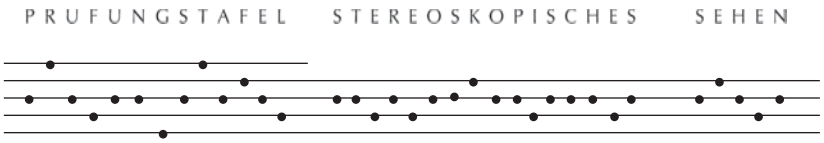
PART I

Ring 1		**Ring 6**	
Square	(2)	Lower left circle	(4)
Marginal ring	(1)	Lower right circle	(5)
Triangle	(3)	Upper right circle	(1)
Point	(4)	Upper left circle	(3)
		Marginal ring	(2)
Ring 7		**Ring 3**	
Black flag with ball	(5)	Square	(4)
Marginal ring	(1)	Marginal ring	(2)
Black circle	(4)	Cross	(3)
Arrow	(2)	Lower left circle	(1)
Tower with cross	(7)	Upper left circle	(5)
Double cross	(2)		
Black triangle	(3)		
Black rectangle	(6)		

PART II

(7) (6) (5) (1) (4) (2)[a] (3)[a] (8)
Highest Lowest

PART III



[a]Rings 2 and 3 are at the same elevation.

raising the edge of one of the photographs. More advanced devices for viewing hardcopy stereoscopic image pairs include *mirror stereoscopes* (larger stereoscopes that use a combination of prisms and mirrors to separate the lines of sight from each of the viewer's eyes) and *zoom stereoscopes*, expensive precision instruments used for viewing stereopairs under variable magnification.

With the proliferation of digital imagery and software for viewing and analyzing digital images, analog stereoscopes have been replaced in the laboratory (if not in the field) by various computer hardware configurations for stereoviewing. These devices are discussed in Section 3.10.

Relationship between Visual Image Interpretation and Computer Image Processing

In recent years, there has been increasing emphasis on the development of quantitative, computer-based processing methods for analyzing remotely sensed data. As will be discussed in Chapter 7, those methods have become increasingly sophisticated and powerful. Despite these advances, computers are still somewhat limited in their ability to evaluate many of the visual "clues" that are readily apparent to the human interpreter, particularly those referred to as image *texture*. Therefore, visual and numerical techniques should be seen as complementary in nature, and consideration must be given to which approach (or combination of approaches) best fits a particular application.

The discussion of visual image interpretation in this section has of necessity been brief. As mentioned at the start of this section, the skill of image interpretation is best learned interactively, through the experience of interpreting many images. In the ensuing chapters of this book, we provide many examples of remotely sensed images, from aerial photographs to synthetic aperture radar images. We hope that the reader will apply the principles and concepts discussed in this section to help interpret the features and phenomena illustrated in those images.