1.4 Classification of remote sensing systems

Remote sensing systems may be classified in a number of ways, but the most useful distinctions we can draw are between *active* and *passive* systems, and between *imaging* and *non-imaging* systems. We may also distinguish sensors by the wavelength of the radiation to which they respond.

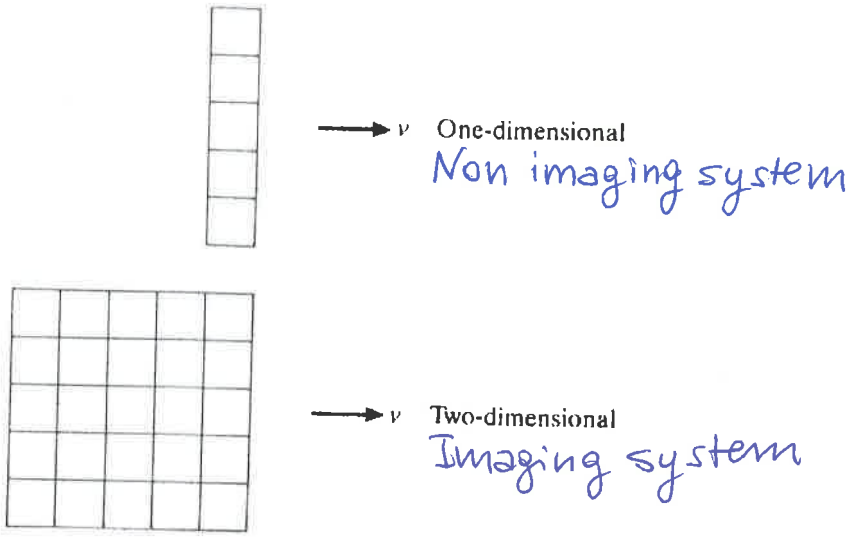
Active systems illuminate the object of study with their own supplied radiation, whereas passive systems sense naturally occurring (emitted thermal or reflected solar) radiation. The choice between active and passive systems is influenced by a number of factors. A passive system will be inappropriate at certain wavelengths at which insignificant amounts of radiation occur naturally. An active system may be technically infeasible if the amount of power which has to be radiated in order to obtain a measurable reflected signal is too great. It may be desirable to have exact knowledge of the nature of the illuminating radiation, which can be tailored to some particular aspect which is to be studied. (For example, we may wish to observe Doppler shifts in the reflected radiation, so as to be able to calculate the relative motion of the target and the sensor.)

An imaging system is slightly harder to define. We shall take the term to mean a system, either active or passive, which measures the intensity of the radiation reaching it and which does so as a function of position on the earth's surface so that a two-dimensional pictorial representation of the intensity can be constructed. A non-imaging system is thus one which either does not measure radiation intensity, or does ~~not~~ do so as a function of ~~position~~ a single coordinate on the earth's surface. We are thus employing a rather restricted definition of the word 'image'. Note that the spatial condition on an imaging system can be reduced to a statement that, for a single location of the system, it should measure the radiation intensity from a number of discrete regions distributed in ~~one~~ two dimensions. One dimension is ~~a~~ not adequate for the production of an image, since motion of the platform (on which the system is supported) ~~in~~ a not the perpendicular direction achieves the necessary two-dimensional scanning (see fig. 1.1).

Fig. 1.2 illustrates the division of remote sensing systems according to whether they are active or passive, imaging or non-imaging. The figure

also attempts to show that the definition provided in the previous paragraph is somewhat fuzzy. Nevertheless, the division is a useful one which will influence the structure of the book.

Fig. 1.1. Imaging systems measure the intensity of the radiation from each cell (resolution element) on the earth's surface. As the field of view is carried forward by the platform, a two-dimensional image is built up.



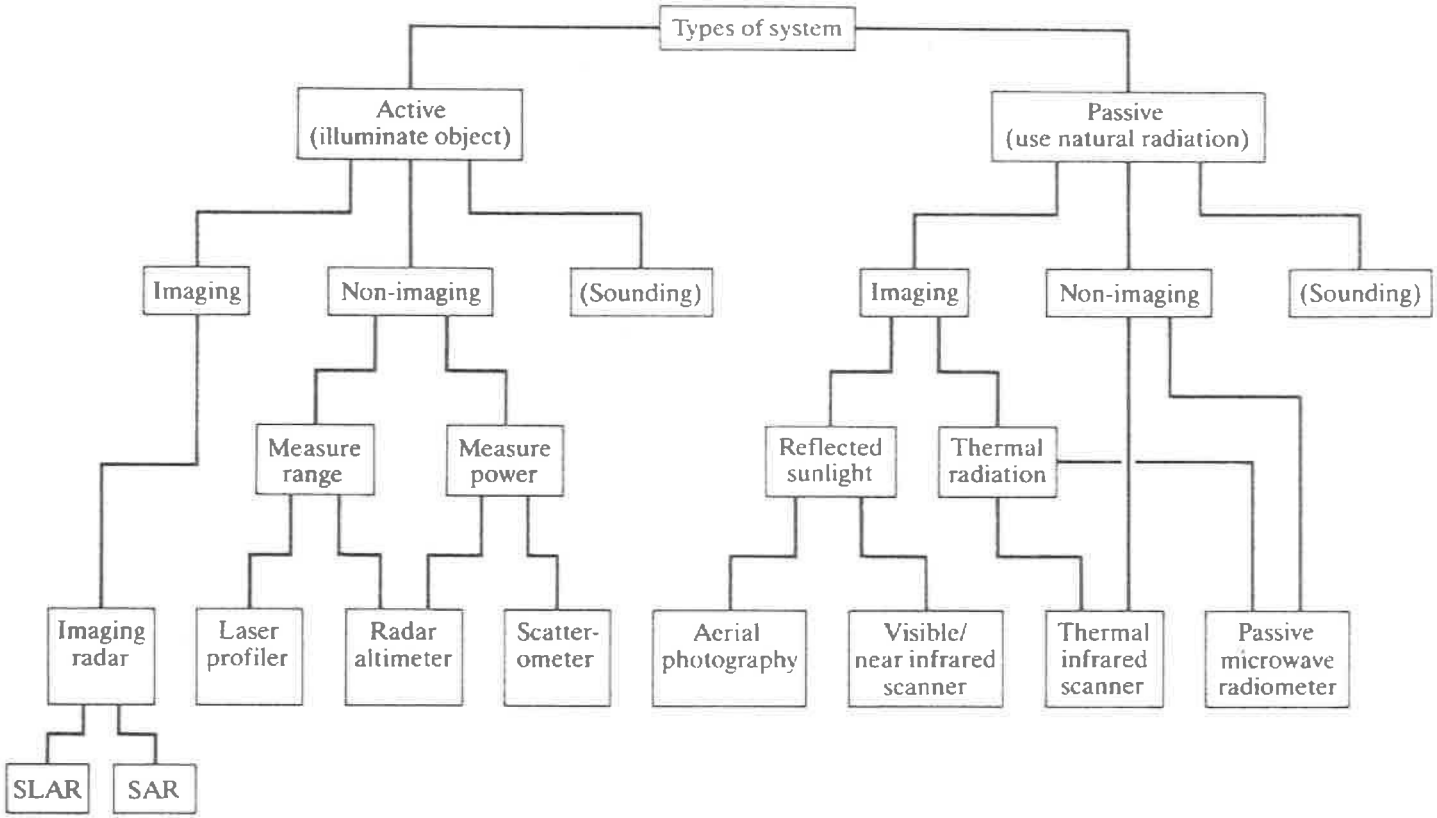


Fig. 1.2. A schematic 'family tree' of the types of remote sensing system which are considered in this book. Sounding systems, which yield profiles of the atmosphere, are not discussed in detail.

2. ELECTROMAGNETIC RADIATION

Electromagnetic radiation and its transfer from sources to objects or from objects to sensors is fundamental to remote sensing, and it is important that certain characteristics of this phenomenon be understood. Electromagnetic radiation is energy and that the energy might be in wave or particulate (i.e., photon or quantum) form. All electromagnetic radiation has wave properties; at all levels, radiation shows interference and diffraction. But studies also indicate that the energy carried by electromagnetic waves may, under certain conditions, be regarded as discontinuous rather than the continuously graded energy that would be expected from a wave.

2.1 Maxwell's Equations

The fundamental description of Electromagnetic radiation begins with Maxwell's equations which describe propagating plane waves. For EM radiation in a vacuum (free space), Maxwell's equations are written:

$$\nabla \bullet \mathbf{E} = 0 \qquad \textbf{(2. 1)}$$

$$\nabla \bullet \mathbf{B} = 0 \qquad \textbf{(2. 2)}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B} / \partial t \qquad \textbf{(2. 3)}$$

$$\nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \partial \mathbf{E} / \partial t \qquad \textbf{(2. 4)}$$

where: $\mathbf{E}$ = electric field
$\mathbf{B}$ = magnetic-induction field
ρ_e = electric charge density
ε_0 = electrical permittivity of free space = 8.85×10^{-12} coulomb^2/Newton-meter^2
μ_0 = magnetic permeability of free space = 12.57×10^{-7} weber/ampere-meter
t = time
∇ = divergence (a spatial vector derivative operator)
$\nabla\times$ = curl (a spatial vector derivative operator)

For our purposes, the key point to be gleaned from these equations is the symmetry between the electric and magnetic fields and the fact that they are always coupled. Electric fields are generated by time varying magnetic fields and magnetic fields are generated by time varying electric fields. By convention we usually only treat electric fields knowing that there will be an associated magnetic field.

2.2 Wave Properties

The general concept of wave transmission involves a source or transmitter, the medium through which the wave propagates, and a receiver. As implied in **Error! Reference source not found.**, one property of a wave is its **frequency**, ***f***, the number of vibrations, oscillations or cycles that the wave makes each second (one cycle/second = one Hertz, abbreviated Hz). A wave's frequency is determined by the source. The time of one vibration, the temporal period or period, T, is related to the frequency by $f = 1/T$. A propagating wave will have a characteristic velocity that is dependent on the medium. If a source is emitting a frequency, f, and if the velocity of wave propagation is v, at the end of one second there will be f waves spread over a distance $v \times (1s)$ (i.e., distance = velocity × time). One wave will occupy a distance λ where $\lambda = v/f$. (Note that $\lambda = vT$.) This distance, the spatial period or **wavelength**, is the shortest distance between consecutive similar

points on the wave (e.g., between consecutive crests or troughs). As an example, c = 2.99792458×10^8 m/s is the speed of light in a vacuum, and this is essentially the speed of light in the atmosphere. (We will use the approximation, c = 3×10^8 m/s.) At a wavelength of 0.5 μm (blue-green light), the frequency of the radiation is:

$$f = v/\lambda = (3 * 10^8 \text{ m/s}) / (0.5 * 10^{-6} \text{ m}) = 6 \times 10^{14} \text{ Hz} = \mathbf{600\ THz} \qquad \mathbf{(2.\ 5)}$$

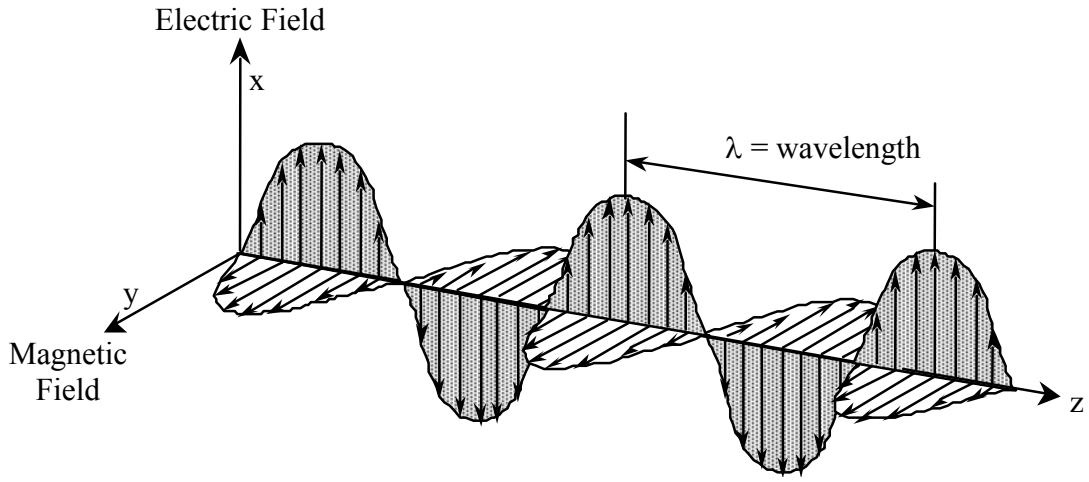


Figure 2.1: Electromagnetic waves consisting of transverse, oscillating electric (**E**) and Magnetic (**B**) fields. The case shown is plane polarized radiation with the **E** and **B** fields both oscillating within a plane.

The wave motion is called **sinusoidal**, **simple harmonic** or **harmonic**. This is the waveform that is most commonly used to represent electromagnetic radiation, if for no other reason than virtually any wave shape can be synthesized by a superposition of sinusoidal waves. A general expression for a sinusoidal wave traveling in the +z direction is:

$$E_x = E_0 \sin(kz + \omega t - \varphi) \qquad \mathbf{(2.\ 6)}$$

$$B_y = (E_0/c) \cos(kz + \omega t - \varphi) \qquad \mathbf{(2.\ 7)}$$

where: E_x = amplitude or displacement of the electric field
E_0 = maximum amplitude or displacement of wave
B_y = amplitude or displacement of the magnetic field
k = wave number = $2\pi/\lambda$
ω = angular frequency = $2\pi f$
t = time
φ = phase angle or phase constant
c = speed of light in a vacuum ≈ 3×10^8 m/s

The area subtended by an incremental solid angle dΩ can be defined as the area on a sphere of radius, r, with dimensions of (r dφ) in the azimuthal direction and of (r sinθ dθ) in the zenith direction. The ratio of the area on the sphere, divided by the square of the radius of the sphere then yields the differential element of solid angle in the direction (θ, φ):

$$d\Omega = \frac{r^2 \sin\theta \, d\theta \, d\varphi}{r^2} = \sin\theta \, d\theta \, d\varphi \tag{2.20}$$

For small angles the solid angle may be approximated by the area subtended by the solid angle on a sphere of radius r divided by the square of the radius.

$$\omega \approx A / r^2 \tag{2.21}$$

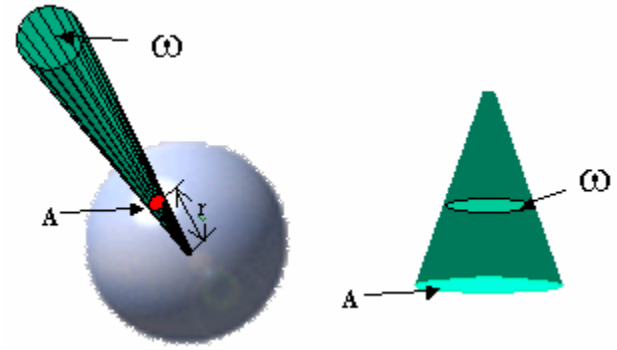


Figure 2.6: Small-angle approximation of the solid angle.

2.9 Basic Radiative Terms

There are many terms commonly (and not so commonly) used to describe radiation. For now we will only consider the three most important terms for our purposes. These are described below in some detail.

Power or flux, Φ [watts = Joule sec^{-1}]

The quantity that is actually measured is the radiant flux, radiant power, or, simply, **power**. Denoted by **Φ** (or **P** in older notation), power is normally measured in watts milliwatts [mW] = 10^{-3} watts; microwatts [μW] = 10^{-6} watts).

Radiance, L [watts m^{-2} $ster^{-1}$]

The power falling onto a surface of area, dA, from the solid angle dΩ, in the direction ξ = (θ,φ), is defined as the radiance, **L**, at that surface. If the surface is perpendicular to the incoming radiation, then the radiance may be defined simply as:

$$L = \frac{d\Phi}{dA \, d\Omega} \tag{2.22}$$

If the surface is oriented at an angle, θ, to the incoming radiation, then we must adjust the relationship accordingly:

$$L = \frac{d\Phi}{dA' d\Omega} = \frac{d\Phi}{\cos\theta \, dA \, d\Omega} \tag{2.23}$$

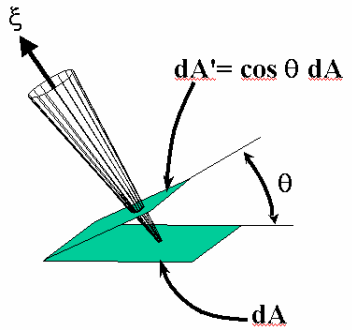


Figure 2.7: Projection of the area onto the direction of the radiation

Typical units for radiance are [w m^{-2} st^{-1}] or [mw cm^{-2} st^{-1}]. Radiance describes power radiating from a defined area through a defined solid angle. Alternatively, radiance may describe the power incident on a defined area through a defined solid angle.

Irradiance, E [w m^2]

Irradiance describes the radiant power received at any surface (e.g., reflecting surface; sensor's aperture or detector). The focus for irradiance is on the surface and the total radiation at that surface. The incident direction is not specified because, for irradiance, it doesn't matter where the radiation comes from, just how much arrives. The surface itself however, is of a fixed size and has some orientation. For example, we might be concerned with how much radiation is incident on a 1 m^2 area of the earth from the sky. In this case the radiation is only arriving from the upward hemisphere, and the effective area of the surface varies as the cosine of the incident angle (as in Equation 2.33). In this case, the irradiance at that surface is just the integral of the radiance incident over the upward hemisphere:

$$\underset{\text{downwelling irradiance}}{E_d} = \int_{\text{upward hemisphere}} L \cos\theta \, d\Omega = \int_{\theta=0}^{\pi/2} \int_{\varphi=0}^{2\pi} L \cos\theta \sin\theta \, d\theta \, d\varphi \tag{2.24}$$

Since the integral is only over the upward hemisphere we are dealing with downwelling irradiance, E_d. If we are interested in the radiation from the lower hemisphere, there is an equivalent radiance that can be defined for the upwelling radiation:

$$\underset{\text{upwelling irradiance}}{E_u} = \int_{\text{downward hemisphere}} L \cos\theta \, d\Omega = \int_{\theta=\pi/2}^{\pi} \int_{\varphi=0}^{2\pi} L \cos\theta \sin\theta \, d\theta \, d\varphi \tag{2.25}$$

When the integral refers to radiation emanating from a surface, it is sometimes referred to as **exitance** or **radiant exitance, M**. It is the same quantity with a different label to indicate the direction in which the radiation is traveling.

There is one extremely important special case. When the radiance is constant with angle (think of a completely overcast sky or a sunlit wall painted a flat white), then the integral is easy to solve:

$$E = \int_{\theta=\pi/2}^{\pi} \int_{\varphi=0}^{2\pi} L \cos\theta \sin\theta \, d\theta \, d\varphi = L \int_{\theta=\pi/2}^{\pi} \int_{\varphi=0}^{2\pi} \cos\theta \sin\theta \, d\theta \, d\varphi$$

$$= 2\pi L \int_{\theta=\pi/2}^{\pi} \cos\theta \sin\theta \, d\theta = \boldsymbol{\pi L} \qquad (2.26)$$

If we are concerned with incoming radiation, then this describes a Lambertian source. If we are looking at reflective surface, then it is a Lambertian reflector or a Lambertian surface.

2.10 Radiance Invariance

One of the fundamental, crucial properties of radiance is that it is invariant over a path in a vacuum. This is simply a statement of conservation of energy, but can be less than obvious when one is computing radiance in a specific problem. Consider Figure 2.8 in which a radiance detector with area A_r and field of view defined by the solid angle, Ω_r is directed toward a source, a distance r away. Area A_S of the source completely fills the field of view of the detector.

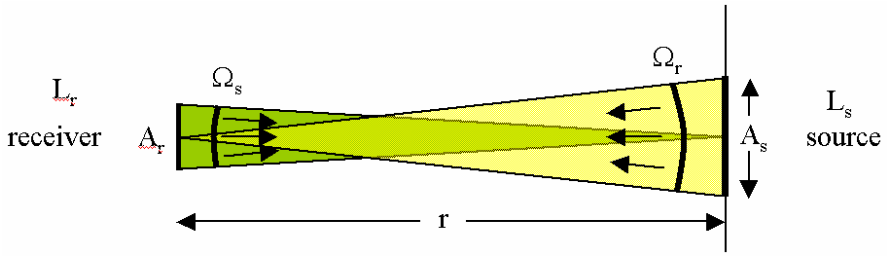


Figure 2.8 : **Geometry used to illustrate radiance invariance law.**

The radiance viewed by the detector is the is the power originating from the source that reaches the detector, Φ_0, through the solid angle defined by the detector field of view, Ω_r, and incident on the detector area, A_r:

$$L_r = \frac{\Phi_o}{A_r \Omega_r} \qquad (2.27)$$

Alternatively, from the point of view of the source, the power emitted that reaches the detector, Φ_r, is emitted from surface A_s into the solid angle Ω_s:

$$L_s = \frac{\Phi_r}{A_s \Omega_s} \qquad (2.28)$$

where $\Phi_r = \Phi_s$. From the definition of the solid angle, we have the relationships:

$$\Omega_r = A_s / r^2 \tag{2.29}$$

$$\Omega_s = A_r / r^2 \tag{2.30}$$

$$\rightarrow A_s\Omega_s = A_r\ \Omega_r \tag{2.31}$$

Using Equation 2.41, together with Equations 2.37 and 2.38, it is clear that $\mathbf{L_r = L_s}$.

Most observations are generally complicated by the fact that we are often interested in the radiance on a surface of area A_r, emanating from a surface A_s neither of which need be perpendicular to the line connecting the two. In this case, consider the source first. The power passing through surface A_s (which is perpendicular to the direction of interest) into solid angle Ω_r must be the same as that leaving surface A'_s into the same solid angle. The radiance invariance law then allows us to assign the distantly measured field radiance L_r to the radiance from surface A'_s. (This is only exactly true if there is no loss along the path r, but we'll deal with that later.) From Equation 2.38,

$$L_s = \Phi_r / A_s\Omega_s \approx \Phi_r / (A'_s\ \cos\theta_s\ \Omega_s) \tag{2.32}$$

. Similarly, the radiance on surface A_r, as given by Equation 2.37, can be written:

$$L_r = \Phi_s / A_r\Omega_r \approx \Phi_s / (A'_r\ \cos\theta_r\ \Omega_r) \tag{2.33}$$

Since $L_s = L_r$ and $\Phi_0 = \Phi_r$, then:

$$(A'_s\ \cos\theta_s\ \Omega_s) = (A'_r\ \cos\theta_r\ \Omega_r) \tag{2.34}$$

One interesting conclusion to be derived from this analysis is that, if the radiance emanating from a surface is the same in all direction then, as long as the surface fills the field of view of the detecting device, Ω_r, then the radiance observed will be independent of the angle of the emitting surface. It will also be independent of the distance, r, again, as long as the surface fills the field of view of the detector.

The import of this fact is made clear by Mobley (1994):

> "Consider the pleasant situation of sitting in front of a fireplace as you read this book. You look into the fire and perceive a certain "brightness" (in essence, radiance), and you feel a certain amount of heat coming from the fire (in essence irradiance; it is energy absorbed per unit time per unit area of your body that warms you). If you move your chair farther away from the fire, it appears smaller but the flames are just as bright; this is the radiance invariance law in action. However, you feel less heat coming from the fire; this is a consequence of the inverse square law for irradiance. This simple situation holds true as long as our radiometer – in this case, our eye – can resolve a finite solid angle for the fire. If we move so far away from the fire that our eye sees it as a true point source of light, then our eye responds to the irradiance received from the source. Thus the fire eventually fades from view as we continue to move farther away."

2.11 Black body radiation

All real objects emit radiation with an intensity and spectral distribution that is determined by their temperature and their ability to emit efficiently. A blackbody is an ideal object that

emits (and absorbs) radiation with perfect efficiency. This is true for all angles of incidence and for radiation of all wavelengths. A blackbody emits the maximum possible radiation that any object can emit in every direction, at every temperature and wavelength. The spectral distribution of blackbody radiation is given by Planck's radiation formula[1]:

$$M_\lambda = \frac{2\pi c_1}{\lambda^5\left[\exp\left(c_2/\lambda T\right)-1\right]} \tag{2.35}$$

where: M_λ = spectral exitance or power emitted per unit area at wavelength λ
$c_1 = hc^2$ k = Boltzman's constant
$c_2 = h/k$ T = absolute temperature
h = Planck's constant λ = wavelength
c = speed of light

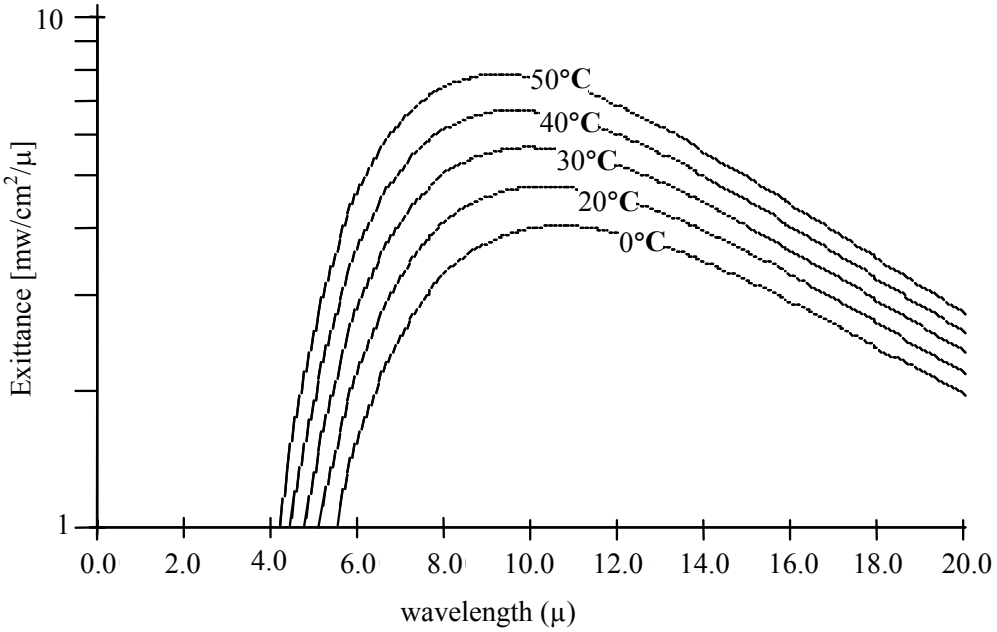


Figure 2.9: Wavelength distribution for a blackbody according to Planck's radiation law

As shown by the graphs of Planck's equation in Figure 2.9, each temperature produces a different curve; the discontinuities between curves follow from the quantum nature of radiation. By integrating Planck's equation over all wavelengths, one obtains the total radiation emitted at a given temperature (i.e., the area under the appropriate exitance vs. temperature curve) and the Stefan-Boltzmann law:

[1] Max Planck initially proposed this formula in October of 1900 based on his intuitive observation that it would both match experimental data and reconcile the difference between two formula, one by Wein which was applicable for the short wavelength limit, and one by Lord Rayleigh which was effective for the long wavelength range. Plank quickly provided a theoretical explanation for this formula and presented his results in December of 1900, noting that he had needed to assume that energy was discontinuous (e.g., quantized). He initially believed that this was simply a necessary mathematical fiction rather than a fundamental requirement. Certainly no one realized that the implications would go far beyond the theory of radiation. (For more on Max Plank, see http://www.max-planck.mpg.de/frameset_e.html)

$$M_{tot} = \sigma T^4 \tag{2.36}$$

where: σ = Stefan-Boltzmann constant (5.67×10^{-8} watts/m^2 K^4), and T = absolute temperature. The total radiant energy emitted by a blackbody is thus a function only of its temperature.

At any temperature or wavelength, a real object (i.e., non-blackbody) will emit only a fraction of the radiation that a blackbody would emit at the same temperature or wavelength. This fraction is the object's **emissivity**, ε, or **spectral emissivity**, ε_λ. Emissivity is a measure of how well an object emits radiation as compared to a blackbody, over all wavelengths,

$$\varepsilon = M_{tot} \text{ of object} / M_{tot} \text{ of blackbody} \tag{2.37}$$

Thus, the radiant emittance of the graybody (a body with a constant emissivity over the range of wavelengths of interest) is expressed as:

$$M_{tot} = \varepsilon \sigma T^4 \tag{2.38}$$

Taking the derivative of Planck's formula (Equation 2.45) with respect to wavelength and setting it equal to zero yields another important formula, called Wein's Displacement Law, which specifies the wavelength at which the emission will reach its maximum value:

$$\lambda_{max} = \frac{A}{T} \tag{2.39}$$

where T is the temperature in °K and A is a constant equal to about 2.898×10^{-3} °K m. A typical earth temperature would be ~300°K, which makes $\lambda_{max} \approx 10\mu$ the wavelength of maximum emission from the earth surface. From Planck's curves (Figure 2.9Figure 2.9), it is apparent that a blackbody and most natural objects at temperatures normally encountered on the earth's surface will emit maximum radiation near 10 μm.

Example 2.1: Computation of solar radiance and irradiance

Much of what has been presented in this chapter can be illustrated by considering the radiation from the sun as it illuminates the earth-moon system. To a good approximation the sun may be considered a gray body with an effective temperature of 5800°K and an emissivity of $\varepsilon_{sun} = 0.99$. Thus, the sun's total radiant exitance is:

$$M_{tot} = \varepsilon \sigma T^4 = 6.35 \times 10^7 \text{ W m}^{-2}$$

The radiation that we see from the sun is emitted in the photosphere, the surface of which will define for us the diameter of the sun, $R_{sun} \approx 6.96 \times 10^8$ m. The total power radiated by the sun is obtained by multiplying the total radiant exitance by the total surface area of the sun:

$$\Phi_{sun} = (4 \pi R_{sun}^2) M_{sun} = 3.87 \times 10^{26} \text{ W}$$

At the earth, a distance of $D_e \approx 1.5 \times 10^{11}$ m, this power is spread out over the surface area of a sphere of area $4\pi D_e^2$, making the irradiance at the earth (the solar constant):

$$E_{sun} = \Phi_{sun} / 4\pi D_e^2 = 1.37 \times 10^3 \text{ W m}^{-2}$$

(Measured values of E_{sun} are presented at http://www.pmodwrc.ch/solar_const/solar_const.html.)

Treating the sun as a disk, the solid angle subtended by the sun as seen from the earth is

$$\Omega_{sun} = \pi (R_{sun})^2 / (D_e)^2 = 6.71 \times 10^{-5} \text{ sr}$$

The corresponding radiance observed by a detector with a field of view filled by the sun is then,

$$L_{sun} = E_{sun} / \Omega_{sun} = \varepsilon\sigma T^4/\pi \approx 2.02 \times 10^7 \text{ W m}^{-2} \text{ sr}^{-1}$$

2.12 Summary of Radiation quantities

Notation in radiometry has a long history, complicated by the fact that developments were going on simultaneously in different fields meaning with the result that the notation has varied somewhat with time as well as discipline. The table below summarizes the terms that we will use in this class and includes some of the old symbols together with the newer symbols.

	Symbols		Units
	new	**old**	
a. **Radiant flux or radiant power** The time rate of flow of radiant energy. We may imagine the flow to be in the form of swarms of tiny colored particles -- photons. Each photon contains a well-defined amount (a quantum, hf) of radiant energy.	**Φ**	**P**	watts (Joules/sec)
b. **Irradiance** (also: exitance, emittance: notated as M) Power per unit area. The radiant flux incident on (irradiance) or emitted by (exitance, emittance) a surface divided by the area of that surface. (See Planck's equation.)	**E, M**	**H, M**	watts/m^2
d. **Radiance** Radiant flux per unit solid angle per unit projected area. – If a source is percieved/measured as an extended source, radiated power is described by radiance, L. – Measure of radiant intensity (I) of many small, but known portions of an extended source.	**L**	**N**	watts/m^2/sr

2.13 The Electromagnetic Spectrum

As shown in Figure 2.4, levels of electromagnetic radiation can be arranged on scales of wavelength, λ, frequency, ν, or energy content, Q ($Q = h\nu$ and $\nu = v/\lambda$; in free space, $v = c$). The electromagnetic spectrum extends from highly energetic cosmic radiation to very low energy oscillations along transmission power lines. Its divisions represent regions within which a common body of experimental technique exists (common sources or detectors). These regions often overlap, for it is possible to produce and sense a given frequency by two or more methods.

The parts of the spectrum, wavelength regions or spectral bands used for electromagnetic remote sensing generally range from the ultraviolet through the microwave, with much activity in the narrow, visible region. The most commonly used wavelength units in the ultraviolet, visible and infrared regions are micrometers (1 μm = 10^{-6} meters), although nanometers

(1 nm = 10^{-9} meters) may be used by certain investigators particularly those working in the visible and near-visible wavelengths. Millimeters (1 mm= 10^{-3} meters) and centimeters (1 cm = 10^{-2} meters) are used to describe wavelengths in the microwave region, but here, frequency designations tend to be more common than wavelength designations. Frequencies may be in kilohertz (1 kHz = 10^{3} Hz), megahertz (1 mHz = 10^{6} Hz) or gigahertz (1 gHz = 10^{9} Hz).

Although sensors will not be reviewed in this monograph, it should be realized that the sensors listed in Table 1.1 are designed or confined to collect radiation in specific spectral bands (Figure 2.5). Photographic films, for example, are not sensitive to wavelengths longer than about 0.9 mm, while glass camera lenses will not transmit wavelengths shorter than about 0.36 μm. Thermal infrared sensing is normally conducted in the 3 to 5 μm or 3 to 14 μm bands.

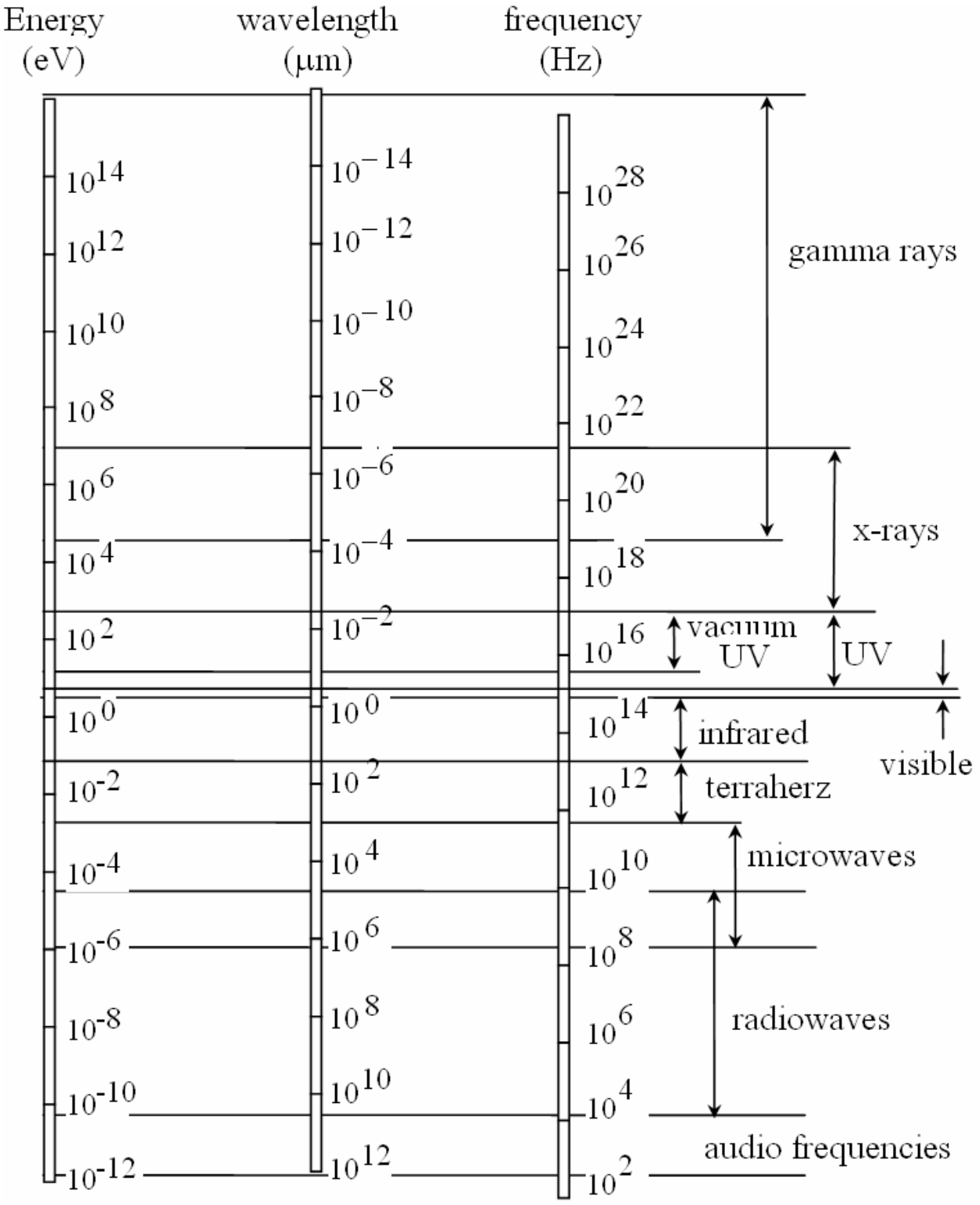


Figure 2.10: Names ranges of the electromagnetic spectrum in energy, wavelength and frequency scales.

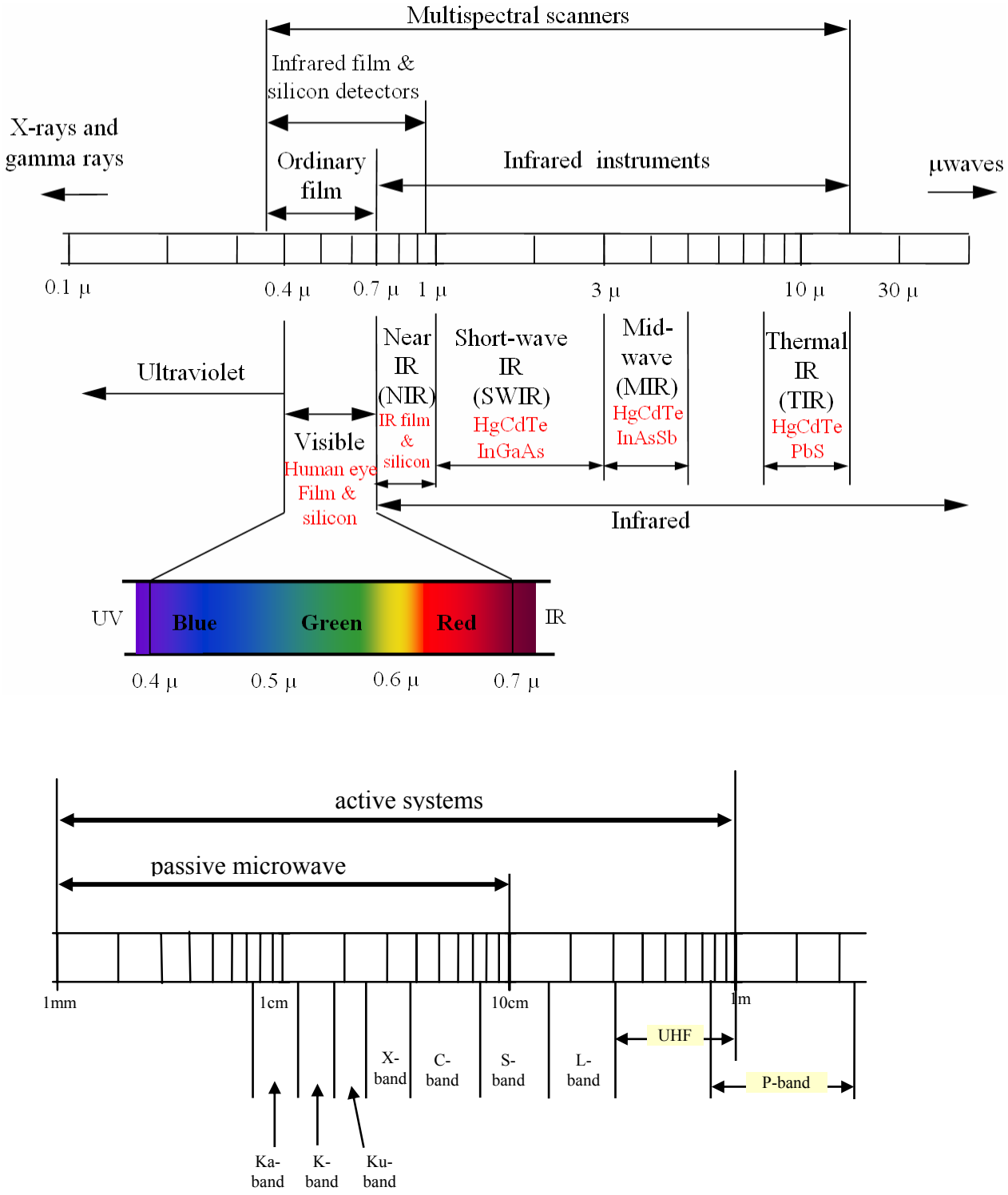


Figure 2.11: Regions of the electromagnetic spectrum used for remote sensing and the principal sensors.

lens to form an image of the scene at the *focal plane,* which is the plane at which the image is sharply defined. A shutter opens at selected intervals to allow light to enter the camera, where the image is recorded on photographic film. A *vidicon* is a type of television camera that records the image on a photosensitive electronically charged surface. An electron beam then sweeps the surface to detect the pattern of charge differences that constitute the image. The electron beam produces a signal that may be transmitted and recorded on magnetic tape for eventual display on film.

Successive frames of camera and vidicon images may be acquired with *forward overlap* (Figure 1-11). The overlapping portions of the two frames may be viewed with a stereoscope to produce a three-dimensional view, as described in Chapter 2. Film is sensitive only to portions of the UV, visible, and reflected IR regions (0.3 to 0.9 μm). Special vidicons are sensitive into the thermal band of the IR region. A framing system can instantaneously image a large area because the system has a dense array of detectors located at the focal plane: The retina of the eye has a network of rods and cones, the emulsion of camera film contains tiny grains of silver halide, and a vidicon surface is coated with sensitive phosphors.

Scanning Systems

A *scanning system* employs a single detector with a narrow field of view that is swept across the terrain to produce an image. When photons of electromagnetic energy radiated or reflected from the terrain encounter the detector, an electrical signal is produced that varies in proportion to the number of photons. The electrical signal is amplified, recorded on magnetic tape, and played back later to produce an image. All scanning systems sweep the detector's field of view across the terrain in a series of parallel scan lines. Figure 1-12 shows the four common scanning modes: cross-track scanning, circular scanning, along-track scanning, and side scanning.

Cross-Track Scanners The widely used *cross-track scanners* employ a faceted mirror that is rotated by an electric motor, with a horizontal axis of rotation aligned parallel with the flight direction (Figure 1-16A). The mirror sweeps across the terrain in a pattern of parallel scan lines oriented *normal* (perpendicular) to the flight direction. Energy radiated or reflected from the ground is focused onto the detector by secondary mirrors (not shown). Images recorded by cross-track scanners, and other scanner systems, are described by two characteristics: spectral resolution and spatial resolution.

Spectral resolution refers to the wavelength interval that is recorded by a detector. In Figure 1-13 the vertical scale shows the response, or signal strength, of a detector as a function of wavelength, shown in the horizontal scale. As the wavelength increases, the detector response increases to a maximum and then decreases. Spectral resolution, or *bandwidth,* is defined as the wavelength interval recorded at 50 percent of the peak

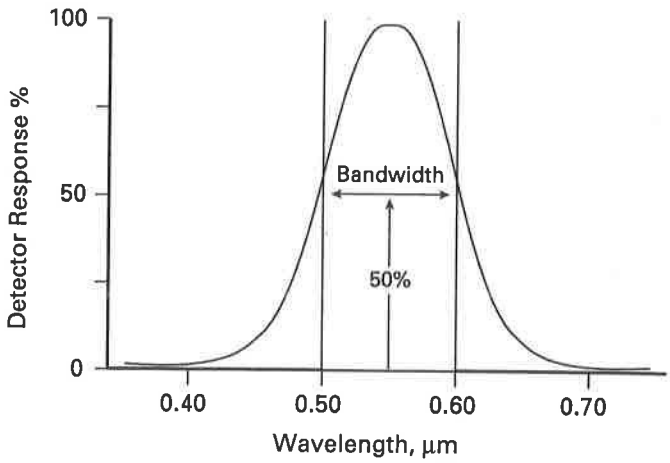


Figure 1-13 Spectral resolution, or bandwidth, of a detector. Bandwidth of this detector is 0.10 μm.

response of a detector. In Figure 1-13 the 50 percent limits occur at 0.50 and 0.60 μm, corresponding to a bandwidth of 10 μm. The section on Reflectance Spectra from Hyperspectral Data describes the effects of different bandwidths on image data.

Spatial resolution was defined earlier, using the eye as an example. The physical dimensions of a detector determine its spatial resolution which is expressed as angular resolving power, measured in milliradians (mrad). Angular resolving power determines the *IFOV* (defined earlier). As shown in Figure 1-16A, the *IFOV* subtends an area on the terrain called a *ground resolution cell.* Dimensions of a ground resolution cell are determined by the detector IFOV and the altitude of the scanning system. A detector with an *IFOV* of 1 mrad at an altitude of 10 km subtends a ground resolution cell of 10 by 10 m.

The *angular field of view* (Figure 1-16A) is that portion of the mirror sweep, measured in degrees, that is recorded as a scan line. The angular field of view and the altitude of the system determine the *ground swath,* which is the width of the

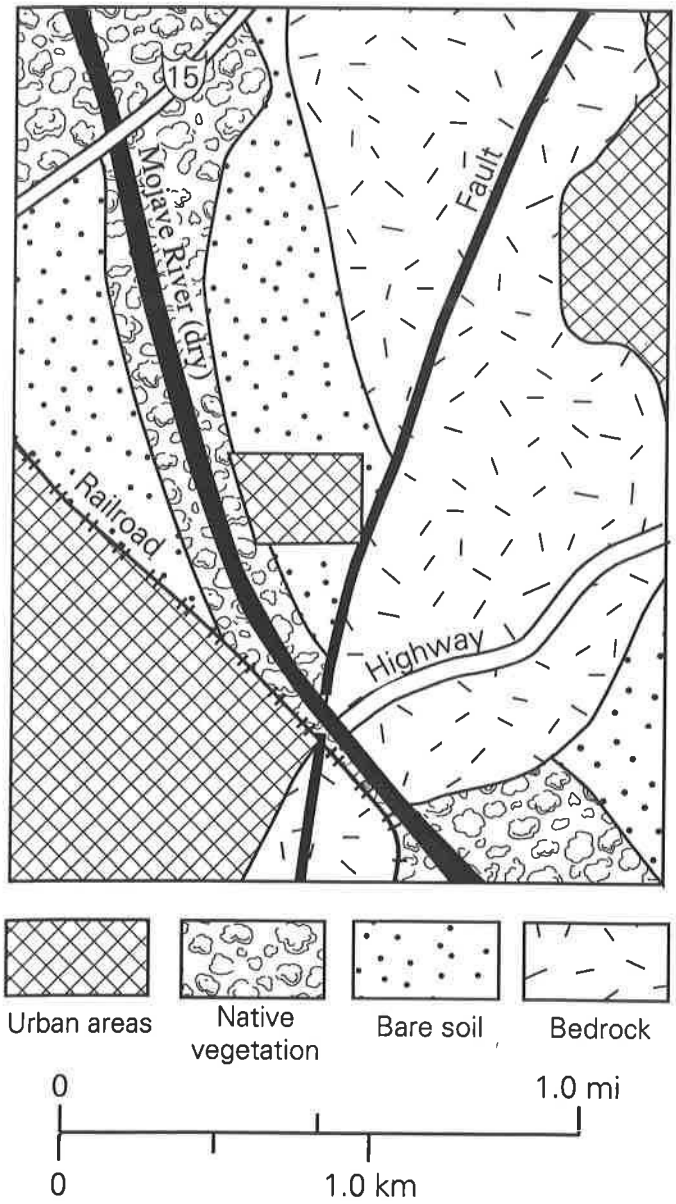


Figure 1-15 Location map of Victorville, California, showing land-cover categories.

terrain strip represented by the image. Ground swath is calculated as

$$\text{Ground swath} = \tan\left(\frac{\text{angular field of view}}{2}\right) \times \text{altitude} \times 2 \qquad (1\text{-}6)$$

The distance between the scanner and terrain is greater at the margins of the ground swath than at the center of the swath. As a result, ground resolution cells are larger toward the margins than at the center, which results in a geometric distortion that is characteristic of cross-track scanner images. This distortion is corrected by digital processing, as shown in Chapter 8.

At the high altitude of satellites, a narrow angular field of view is sufficient to cover a broad swath of terrain. For this reason the rotating mirror is replaced by a flat mirror that oscillates back and forth through an angle of approximately 15°. An example is the scanner of Landsat described in Chapter 3.

The strength of the signal generated by a detector is a function of the following factors:

Energy flux The amount of energy reflected or radiated from terrain is the *energy flux*. For visible detectors, this flux is lower on a dark day than on a sunny day.

Altitude For a given ground resolution cell, the amount of energy reaching the detector is inversely proportional to the square of the distance. At greater altitudes the signal strength is weaker.

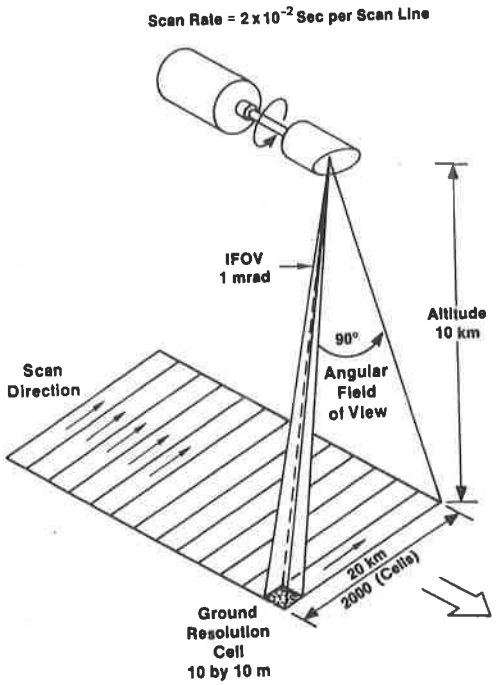
Spectral bandwidth of the detector The signal is stronger for detectors that respond to a broader bandwidth of energy. For example, a detector that is sensitive to the entire visible range will receive more energy than a detector that is sensitive to a narrow band, such as visible red.

Instantaneous field of view Both the physical size of the sensitive element of the detector and the effective focal length of the scanner optics determine the *IFOV*. A small *IFOV* is required for high spatial resolution but also restricts the *signal strength* (amount of energy received by the detector).

Dwell time The time required for the detector *IFOV* to sweep across a ground resolution cell is the *dwell time*. A longer dwell time allows more energy to impinge on the detector, which creates a stronger signal.

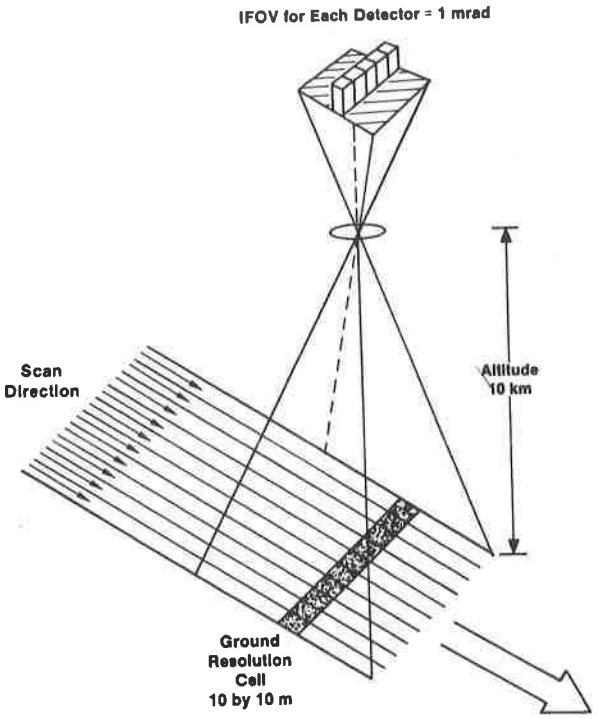
For a cross-track scanner, the dwell time is determined by the detector *IFOV* and by the velocity at which the scan mirror sweeps the *IFOV* across the terrain. As shown in Figure 1-16A a typical airborne scanner with a detector *IFOV* of 1 mrad, a 90° angular field of view, and operating at 2×10^{-2} sec per scan line at an altitude of 10 km has a dwell time of 1×10^{-5} sec per ground resolution cell. It is instructive to compare the dwell time with the ground speed of the aircraft. At a typical ground speed of 720 km · h^{-1}, or 200 m · sec^{-1}, the aircraft crosses the 10 m of a ground resolution cell in 5×10^{-2} sec. The cross-track scanner time of 1×10^{-5} is 5×10^{3} times faster than the ground velocity of the aircraft. The high scanner speed relative to ground speed is required to prevent gaps between adjacent scan lines.

The short dwell time of cross-track scanners imposes constraints on the other factors that determine signal strength. For example, the *IFOV* and spectral bandwidth must be large enough to produce a signal of sufficient strength to overcome the inherent electronic noise of the system. The signal-to-noise ratio must be sufficiently high for the signal to be recognizable.



$$\text{Dwell Time} = \frac{\text{Scan Rate per Line}}{\text{Number Cells per Line}} = \frac{2 \times 10^{-2}\ \text{sec}}{2000\ \text{cells}} = 1 \times 10^{-5}\ \text{sec} \cdot \text{cell}^{-1}$$

A. Cross-track scanner.



$$\text{Dwell Time} = \frac{\text{Cell Dimension}}{\text{Velocity}} = \frac{10\ \text{m} \cdot \text{cell}^{-1}}{200\ \text{m} \cdot \text{sec}^{-1}} = 5 \times 10^{-2}\ \text{sec} \cdot \text{cell}^{-1}$$

B. Along-track scanner.

Figure 1-16 Dwell time calculated for cross-track and along-track scanners.

Along-Track Scanners For scanner systems to achieve finer spatial and spectral resolution, the dwell time for each ground resolution cell must be increased. One method is to eliminate the scanning mirror and provide an individual detector for each ground resolution cell across the ground swath (Figure 1-12C). The detectors are placed in a linear array in the focal plane of the image formed by a lens system.

The long axis of the linear array is oriented normal to the flight path, and the *IFOV* of each detector sweeps a ground resolution cell along the terrain parallel with the flight track direction (Figure 1-12C). *Along-track scanning* refers to this movement of the ground resolution cells. These systems are also called pushbroom scanners because the detectors are analogous to the bristles of a broom pushed along the floor.

For along-track scanners, the dwell time of a ground resolution cell is determined solely by the ground velocity, as Figure 1-16B illustrates. For a jet aircraft flying at 720 km · h^{-1}, or 200 m · sec^{-1}, the along-track dwell time for a 10-m cell is 5×10^{-2} sec, which is 5×10^{3} times greater than the dwell time for a comparable cross-track scanner. The increased dwell time allows two improvements: (1) detectors can have a smaller *IFOV*, which provides finer spatial resolution, and (2) detectors can have a narrower spectral bandwidth, which provides higher spectral resolution. Some experimental airborne along-track scanners operate with a spectral bandwidth of 0.01 μm. Typical cross-track scanners have bandwidths of 0.10 μm, which is a spectral resolution coarser by one order of magnitude.

3.2 Overall Sensor Model

An electro-optical sensor may be modeled by the processes shown in Fig. 3-1. The scanning operation (Chapter 1) converts the spatial at-sensor radiance to a continuous, time-varying optical signal on the detectors. The detectors, in turn, convert that optical signal into a continuous time-varying electronic signal, which is amplified and further processed by the sensor electronics. At the Analog/Digital (A/D) converter, the processed signal is sampled in time and quantized into discrete *DN* values representing the spatial image pixels.

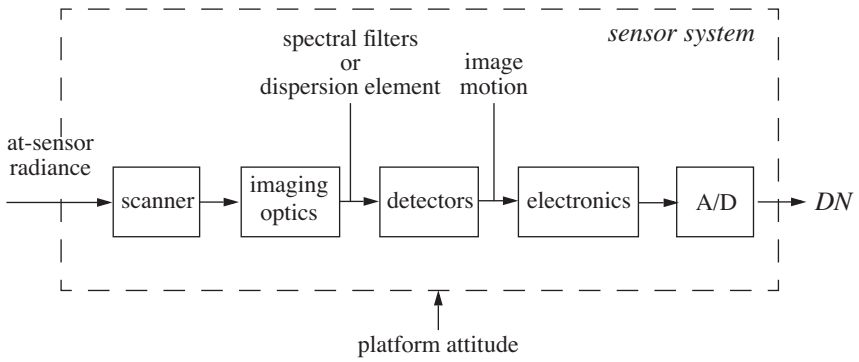


FIGURE 3-1. The primary components in an electro-optical remote-sensing system. A whiskbroom type of system is illustrated here. Although the platform attitude is external to the sensor per se, it has important effects on the final image characteristics and quality.

3.3 Resolution

No property of images is more widely quoted, and simultaneously misused, than *resolution*. It is a term that conveys a strong intuitive meaning, but is difficult to define quantitatively. Remote sensing systems have "resolution" in the spectral, spatial, and temporal measurement domains. Furthermore, there is a numerical resolution associated with the data itself by virtue of radiometric quantization. We will discuss the notion of resolution in this section, and in the process, introduce important concepts such as convolution, mixed pixels, and sampling.

3.3.1 The Instrument Response

No instrument, remote-sensing systems included, can measure a physical signal with infinite precision. If the signal varies in time, the instrument must average over a non-zero integration time; if the signal varies in wavelength, the instrument must average over a non-zero spectral bandwidth; or if the signal varies in space, the instrument must average over a non-zero spatial distance. In general, we can write the output of most instruments as

$$o(z_0) = \int_W i(\alpha) r(z_0 - \alpha) d\alpha \tag{3-1}$$

where

$i(\alpha)$ = input signal,
$r(z_0 - \alpha)$ = instrument response (unit area), inverted and shifted by z_0,
$o(z_0)$ = output signal at $z = z_0$, and
W = range over which the instrument response is significant.

The physical interpretation of Eq. (3-1) is that the instrument weights the input signal in the vicinity (W) of z_0 and integrates the result. If we allow z_0 to have any continuous value, say z, then this relation is known as a *convolution*. A convenient, commonly-used shorthand notation for Eq. (3-1) is,

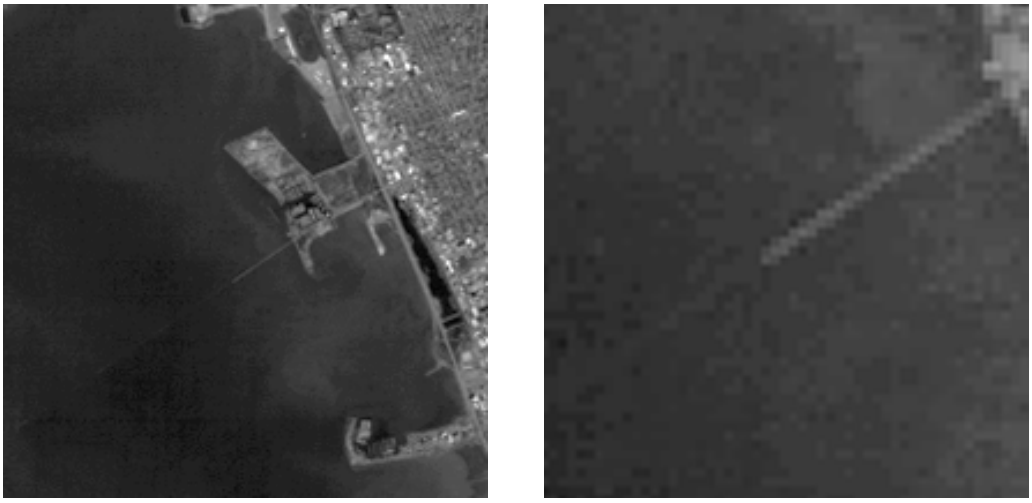
$$o(z) = i(z) * r(z) \tag{3-2}$$

which reads "the output signal equals the input signal convolved with the response function." This mathematical description can be applied to a wide range of instruments.[1] In the following sections, we use it to describe the spatial and spectral response characteristics of a remote-sensing imaging system.

3.3.2 Spatial Resolution

While the spatial "resolution" of a sensor (or its image) is often quoted as the *GSI* or *GIFOV*, it is well known that it is possible to *detect* considerably smaller objects if the contrast with the surrounding background is sufficiently high. Even though such objects may be detectable, they are not necessarily *recognizable*, except by the general context of the image. Thus, in TM imagery one frequently "sees" roads or bridges over water that are 10 m (one third of a pixel) or less wide (Fig. 3-2). Similarly, high contrast, linear features much smaller than a pixel are visible in MODIS and IKONOS imagery (Fig. 3-3).

1. Specifically, systems that are linear (the output for multiple inputs is the sum of their individual outputs) and invariant (the form of $r(z_0 - \alpha)$ does not depend on z_0).



1X *4X*

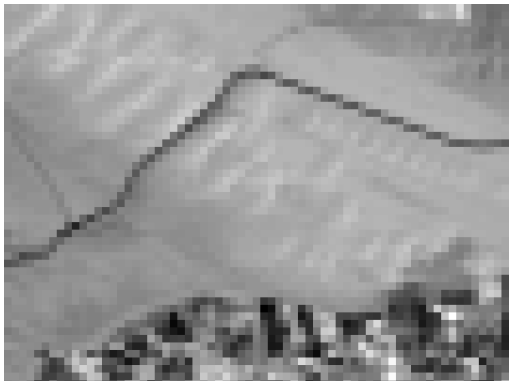
FIGURE 3-2. Example of subpixel object detection. This is part of a TM band 3 image of San Francisco showing the Berkeley Pier. The pier is 7m wide and made of concrete. An older extension of the same width, but made of wood with a lower reflectance than concrete, is barely visible. (Acknowledgments to Joseph Paola for providing details on the size and construction of the pier.)

Suppose we have a sensor that produces a linear *DN* output versus scene reflectance.[2] A ground area that is larger than *GIFOV*-squared[3] and has zero reflectance produces a zero *DN*, and an area with a reflectance of one produces a maximum *DN*, say 255. Now, suppose the area within one *GIFOV* contains two types of materials, with reflectances of zero and one. Surrounding pixels contain only the dark material and represent the background against which the pixel of interest is compared. At that pixel, the signal produced by the sensor will be the integrated effect of the mixture of the two components, target and background, within the *GIFOV*. If their relative proportion within the *GIFOV* is 50%, then the *DN* will be 128 (rounded to the nearest integer *DN*). If the brighter material occupies only 10% of the *GIFOV* area, the *DN* will be 26. Since the sensor presumably can reliably distinguish two *DN*s separated by only one unit, we could theoretically (assuming no image noise) detect a bright object against a dark background even if it occupied only 0.4% of the *GIFOV* area (Fig. 3-4).

2. We simplify the discussion by assuming remote-sensing measurements are a direct function of reflectance, without intervening influences.

3. The product of the in-track and cross-track *GIFOV*s. For brevity, this two-dimensional area will simply be called a *GIFOV* in the rest of this section

TM band 4



MODIS band 2

IKONOS pan band

FIGURE 3-3. More examples of subpixel object detection. At the top are TM and MODIS images of the All American Canal in California. The Canal carries water for irrigation and is about 60m wide (two TM pixels). Because of the high contrast to the Imperial Sand Dunes, it is visible in the 250m MODIS band 2 image. Even the narrower Interstate 8 highway at the top is visible. At the bottom is an IKONOS image of a parking lot and buildings in Tucson, AZ. The parking lot markings are about 10cm wide, or about 1/10 of a IKONOS 1m pan band pixel. (IKONOS image provided by Space Imaging LLC and NASA Scientific Data Purchase Program.)

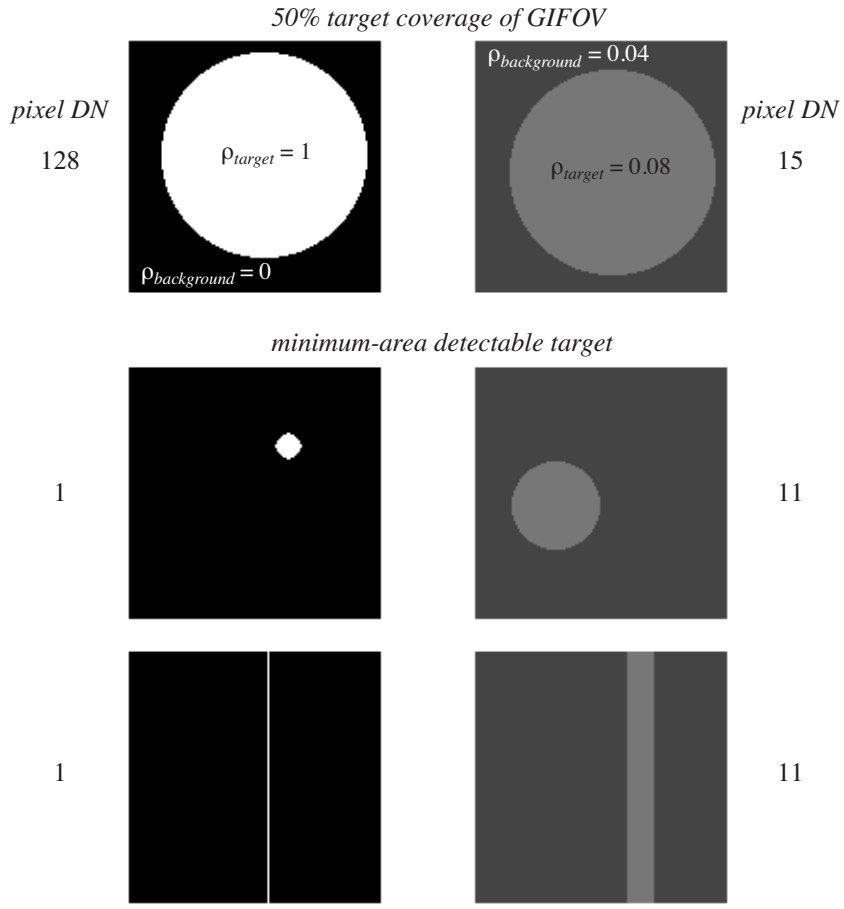


FIGURE 3-4. Detectability analysis for a single target at two different contrasts to the surrounding background and for an idealized sensor. The squares represent the area sampled by a single GIFOV. The minimum-area detectable target results in a one DN difference from the background. Note that the target does not have to be centered in the GIFOV (if the sensor response is uniform across the GIFOV); the spatially-integrated signal will be the same for any internal location. For the same reason, the shape of the target cannot be discerned from a single pixel. If the target is linear, such as a road or bridge, then the shape may be inferred by the context of several pixels; cases in point are Fig. 3-2 and Fig. 3-3.

Now, we will make the object reflectances more realistic. The darker background might have a reflectance of 4% and the lighter target a reflectance of 8%, yielding a *contrast ratio* of 2:1. If the target covers 50% of the *GIFOV*, the *DN* of that pixel will be 15, while the *DN* of pure background pixels will be 10. The target is still detectable since it differs from the background by five *DN*s. If the size of the target is less than 10% of a pixel, however, it falls below the threshold (one *DN*) for detection. We see, therefore, that the radiometric quantization, coupled with the target and background reflectances, and the sensor *GIFOV*, all conspire to determine the "resolution" of the image. If there is any noise in the image, the threshold for detection will be higher, i.e., a greater target-to-background contrast will be required for reliable detection.

An often neglected influence on image resolution is the *sample-scene phase*, i.e., the relative location of the pixels and the target (Park and Schowengerdt, 1982; Schowengerdt *et al*., 1984). This relative spatial phase is unpredictable (and almost always unknown) for any given image acquisition, and varies from acquisition-to-acquisition with a uniform probability distribution between ±1/2 pixel. Two of an infinite number of possible sample-scene phases are shown in Fig. 3-5 for the low-contrast case described previously. On the left, the target appears equally in each of four *GIFOV*s, and on the right, unequally among the four. The resulting *DN*s of the four pixels containing part of the target are shown below the diagram. In the case of equal subdivision of the target, the four pixels would have a *DN* of 11 and the target would be detectable in each. If the image pixel grid happened to partition the target unequally as shown to the right in the figure, the percent coverage by the target in each *GIFOV* would be 30, 5, 2, and 13 (clockwise from the upper left). The four pixels would thus have *DN*s of 13, 11, 10, and 12, respectively. The target would remain detectable in three of the pixels (again, if there were no noise).

Sample-scene phase is important for more complex targets, as well. For example, suppose the target consists of a series of equally-spaced bright bars against a dark background. If the sensor *GIFOV* is equal to the width of one bar, the two extreme sample-scene phases result in a spatial signal that either has maximum contrast or no contrast! If the sensor *GIFOV* is exactly twice the width of one bar, the spatial signal will have zero contrast for *any* sample-scene phase.

An enlargement of the TM image in Fig. 3-2 is shown in Fig. 3-6. It can be seen that the *DN* profile along each scan line across the pier is different. This is the result of sample-scene phase variation, because the pier is not oriented at 90° to the scan. Thus, in one particular scan the pier may appear to be one pixel wide, while in another it appears to be two pixels wide. To estimate the true width of this subpixel object, an interleaved composite of many lines should be made, with careful attention to phasing them correctly to fractions of a pixel. The composite would represent the convolution, Eq. (3-2), of the sensor spatial response function and the pier radiance profile, sampled at an interval much finer than the pixel *GSI*. Just such an analysis is used later in this chapter to evaluate the ALI and QuickBird sensor spatial response functions.

With the inclusion of factors that come into play in real images, namely sensor noise, non-uniform targets and backgrounds, and variable solar angle and topography, it becomes clear that the situation with respect to image resolution is not simple! A common statement such as "the resolution of this image is 30 meters," probably refers to the sensor *GIFOV* or *GSI*, but it is not a precise statement, as we have seen. In this book, we will assume the term resolution refers simply to the *GSI*.

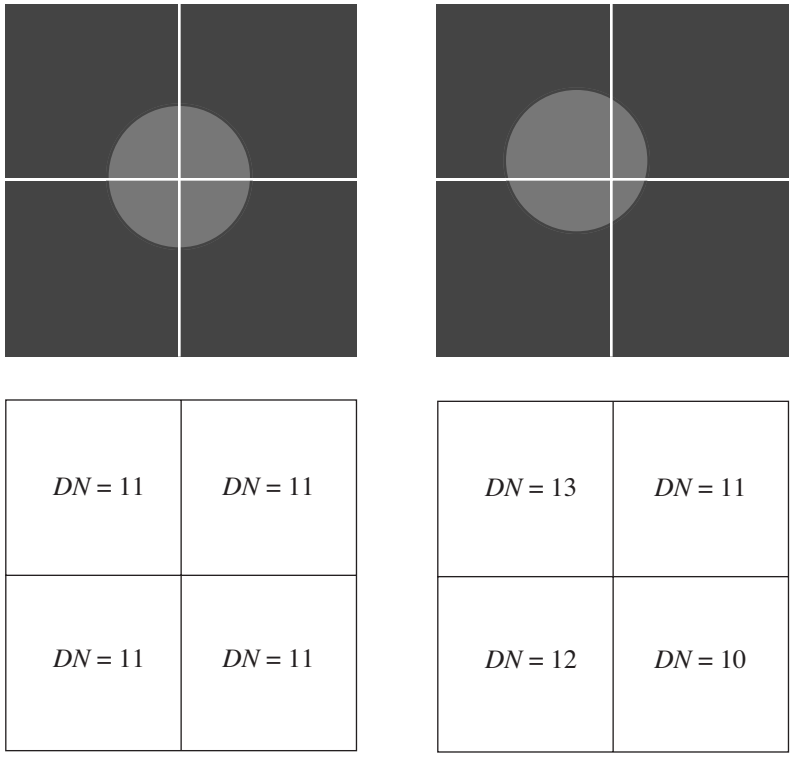


FIGURE 3-5. The effect of spatial phasing between the pixel grid and the ground target. Four adjacent GIFOVs are shown and the target area is 50% of the GIFOV area. On the left, the target fills 12.5% of each GIFOV. On the right, the target occupies 30%, 5%, 2%, and 13% of the four GIFOVs. The location of the grid is unpredictable prior to imaging any given scene; it could be anywhere within ±1/2 pixel interval with equal probability. The imaging of long linear features can sometimes allow precise measurement of the phase to a small fraction of a pixel. The background DN is 10.

3.3.3 Spectral Resolution

The total energy measured in each spectral band of a sensor is a spectrally-weighted sum of the image irradiance over the spectral passband, Eq. (3-26). This weighting by wavelength is the primary determinant of the sensor's capability to *resolve* details in the spectral signal. To see why this is so, we will take the reflectance data for the mineral alunite (Chapter 1) in the vicinity of the OH absorption doublet between 1350 and 1550 nm wavelength, and simulate its measurement by a multispectral sensor. Each half of the doublet is only about 10 to 20 nm wide, and they are separated by about 50 nm. Now, imagine these data are measured with a hyperspectral sensor having many

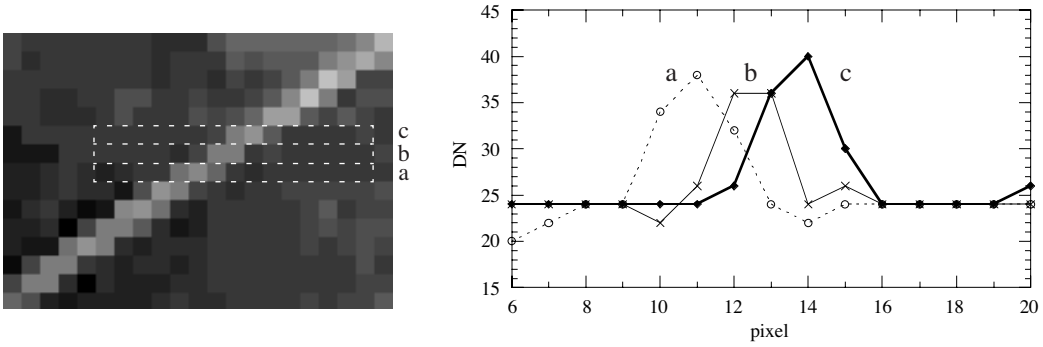


FIGURE 3-6. A contrast-enhanced enlargement of Fig. 3-2 and DN profile plots along three adjacent scanlines near the center of the pier, illustrating the sample-scene phase effect. The pier's profile is different in each line (the linear interpolation between individual pixels is only to aid visualization of the graph).

spectral bands, each 10 nm wide (at 50% of the peak responsivity) and spaced at 10 nm intervals. Each band sees an effective reflectance which is the weighted reflectance over the band; the weighting function is the spectral response of the sensor in each band (Fig. 3-7).[4] We will use a bell-shaped function which approximates actual spectral band responsivities. Since the spectral passband is comparable in width to the details in the signal, they are preserved. The exact reflectance minima are not found, however, because of the location of the bands along the wavelength scale.[5]

If we do a similar exercise with 50 nm-wide spectral bands, the result shows complete loss of information about the doublet; it is literally "averaged away" by the broad spectral bands (Fig. 3-7). Even if the spectral band locations were shifted, the doublet would not be "resolved." If the spectral bands do not overlap as they do in Fig. 3-7, (for example, bands 3, 4, 5, and 7 of TM), the sensor's ability to resolve even coarser features is seriously hampered. The trade-off, of course, is the increased data burden of finely-sampled spectra.

To illustrate how the actual TM spectral response modifies the at-sensor radiance, we will use the Kentucky Bluegrass spectral reflectance shown in Chapter 1 and the atmospheric propagation model of Chapter 2. The net at-sensor radiance is shown in Fig. 3-8. This is multiplied by the spectral response of each of the four VNIR TM bands (Fig. 3-22), to yield the weighted spectral distribution seen by each band. The integral of this function over wavelength then provides the total *effective* radiance in each band.

4. To simplify the illustration, we are not including solar irradiance or atmospheric propagation as described in Chapter 2. These factors would change the input function, but not alter the concept.

5. This is another example of sample-scene phase, but in the spectral dimension rather than the spatial dimension.

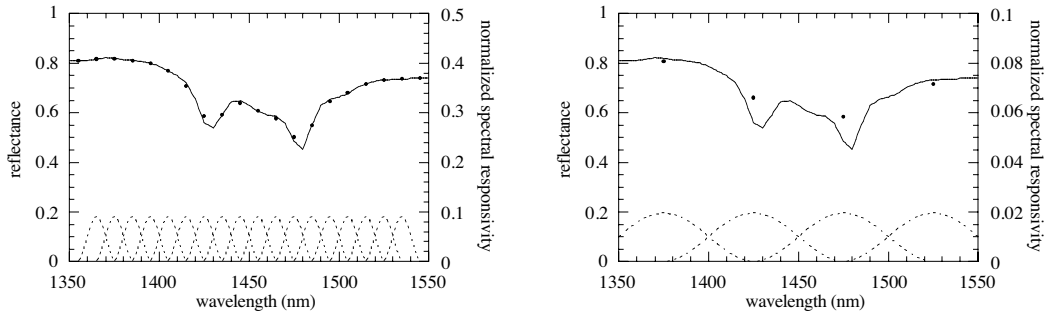


FIGURE 3-7. The effective reflectance of alunite as measured by a multispectral sensor. The solid line is the original reflectance sampled at 1nm and the individual band spectral responsivities are shown as dashed lines. Each solid dot is the output of the corresponding band. The graph with 10nm-wide spectral bands represents the response of a hyperspectral sensor such as AVIRIS or HYDICE. The graph with 50nm-wide spectral bands represents the response of a sensor such as TM (although TM does not actually have any spectral bands in this part of the spectrum).

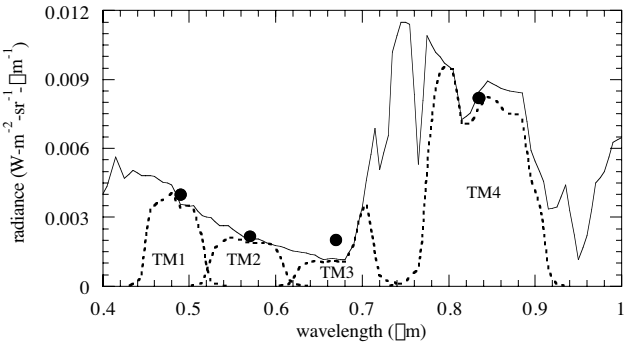


FIGURE 3-8. The at-sensor radiance for Kentucky Bluegrass (solid curve), the weighted spectral distribution seen by each TM band (dotted curves), and the total (integrated over the weighted spectral distribution) effective radiance (solid circles). Note how the broad bandwidth of bands 3 and 4, in particular, averages spectral detail present in the original radiance.

Now, in fact, the spectral reflectances of many natural materials are fairly smooth, without fine absorption lines (see the reflectance curves for soil and vegetation in Chapter 1). So the high spectral resolution of hyperspectral sensors at first seems to be of little value in those cases. With better understanding of the absorption properties and biochemical interactions among natural material components, for example, cellulose, lignin and protein in vegetation, the additional resolution afforded by imaging spectroscopy can be fully utilized (Verdebout *et al*., 1994; Wessman, 1994). As seen in the prior examples, the *placement* of spectral bands is as important as the spectral *bandwidth* to a sensor's ability to resolve spectral features. Hyperspectral sensors offer generally contiguous bands over a wide spectral range, and are, therefore, superior to multispectral sensors with a few, albeit judiciously-placed, spectral bands.

3.4 Spatial Response

The sensor modifies the spatial properties of the scene in two ways: (1) blurring due to the sensor's optics, detectors, and electronics; and (2) distortion of the geometry. In this section, we discuss spatial blurring, which generally occurs on a smaller spatial scale (a few pixels) than distortion.

The image of the scene viewed by the sensor is not a completely faithful reproduction. Small details are blurred relative to larger features; this blurring is characterized by the net (total) sensor *Point Spread Function (PSF_{net})*, which we can accurately view as the *spatial responsivity* of the sensor (just as we described a spectral responsivity in the previous section). PSF_{net} is the weighting function for a spatial convolution (refer to Eq. (3-1)), resulting in the electronic signal, e_b,

$$e_b(x,y) = \int_{\alpha_{min}}^{\alpha_{max}} \int_{\beta_{min}}^{\beta_{max}} s_b(\alpha,\beta) PSF_{net}(x-\alpha, y-\beta)\, d\alpha\, d\beta \,. \tag{3-3}$$

The sensor response function weights the measured physical signal, which is then integrated over the range of the response function to produce the output value. The limits of the integral define the *spatial extent* of the *PSF* about the coordinate (x,y). Note that the left side of Eq. (3-3) still depends on the continuous spatial coordinates (x,y); we do not convert this to discrete pixel coordinates until the signal is sampled.[6]

PSF_{net} consists of several components. First, the optics induce blurring by the *optical PSF*. The image formed by the optics on the detectors may in some cases move during the integration time for each pixel; this introduces an *image motion PSF*. Then the detector adds additional blurring due to the *detector PSF*. The detected signal is further degraded by the *electronics PSF*. The process is outlined in the flow diagram of Fig. 3-1. The following type of analysis has been described for the Landsat MSS (Park *et al*., 1984), TM (Markham, 1985) and AVHRR (Reichenbach *et al*., 1995).

6. Throughout this chapter and the remainder of the book, the coordinates (x,y) will represent the cross-track and in-track directions in image space, respectively. This "path-oriented" coordinate system will suffice for any sensor that points to nadir relative to the platform track. It would need to be reconsidered for describing off-nadir imaging by a pointable sensor.

Using the same logic, it is easy to understand why longer wavelength radars (L-band) dominated the early spacecraft radar imaging systems. Recalling that the effect of range distance is inverse and to the 4th power, the large distances (thousands of kilometers vs. thousands of meters) characteristic of spacecraft operations result in a severe reduction of P_R. The P_T could be increased, but that would require an increase in payload weight, an option only considered when all else fails. With all other parameters held constant the selection of L-band (25 cm) over K-band (1 cm) would mean an increase in P_R by a factor of 25^2 or 625. Combined with the greater penetration capabilities of L-band, the initial selection of longer wavelength for spaceborne radars was logical.

The radar interpreter needs to try to understand the target's influence on P_R. Along with the system-related parameters in the above radar equation, the fundamental parameters which influence surface radar backscatter are related to several additional system parameters, target parameters and the interplay of both (Table 3-1).

Complex interactions and relationships take place between and among the parameters in Table 3-1 when the signal penetrates the surface and volume scattering occurs. In addition to the parameters in Table 3-1, complex volume scattering, subsurface roughness and the index of refraction must be taken into account, often in a very convoluted manner. For example, the amount of radar penetration is a function of wavelength, complex dielectric, incident angle (θ) and polarization coupled with surface and subsurface roughness characteristics and the medium's index of refraction. For illustrations and discussion of these complex relationships applied to specific geoscience areas, the reader is referred to the chapters on vegetation (see Chapters 7 and 9) and soils (see Chapter 8).

TABLE 3-1
Fundamental System and Target Parameters That Influence Radar Power Return (P_R)

System Parameters	Target Parameters
1. Wavelength or Frequency	1. Surface Roughness
2. Polarization	2. Complex Dielectric
3. Look Angle	3. Slope Angle & Orientation
4. Look Direction	
5. Resolution	

Direct Interplay of System and Target Parameters

1. Surface Roughness — defined in terms of system wavelength
2. Look Angle (ϕ) and Slope Angle (α) — combine to determine Incident Angle (θ)
3. Look Direction and Slope (or target) Orientation — influence the area and geometry of the target presented to the radar

3-1.2 DEFINITIONS

3-1.2.1 Incident Angle (θ)

Incident angle (θ), defined as the angle between the radar line-of-sight and the local vertical (Figures 3-1A and 3-1B) with respect to the geoid, is a major factor influencing the radar backscatter and the appearance of objects on the imagery, caused by foreshortening or radar layover. In general, reflectivity from distributed scatterers decreases with increasing incident angles. Figure 3-1A illustrates incident angle incorporating look angle (ϕ) and the curvature of the earth (see also Figure 2-11). This model assumes a level terrain or constant slope angle (α). In contrast, Figure 3-1B illustrates the "local incident angle" and takes into account the local slope angle (α). For example, surface roughness changes as a function of the local incident angle

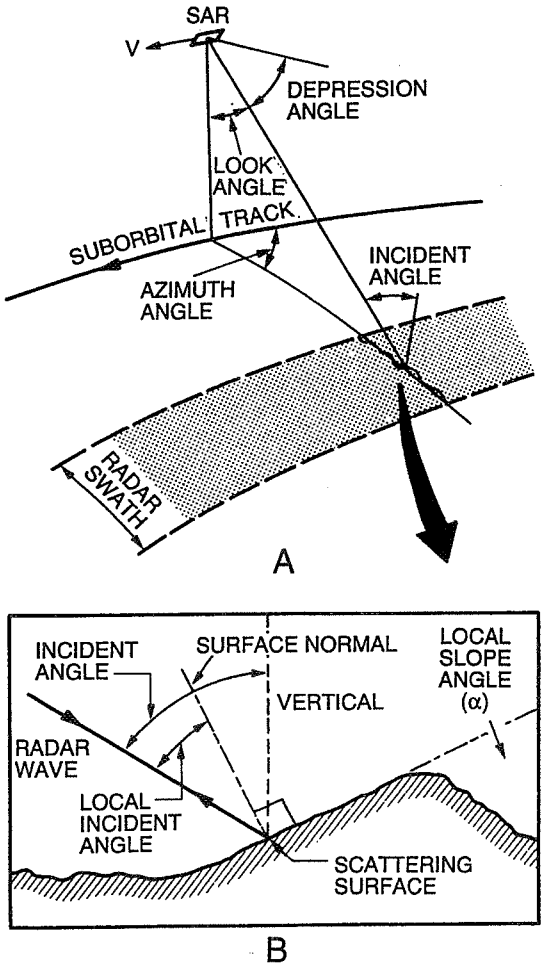


Figure 3-1. Schematic Diagrams of System (3-1A) and Local (3-1B) Incident Angle (θ).

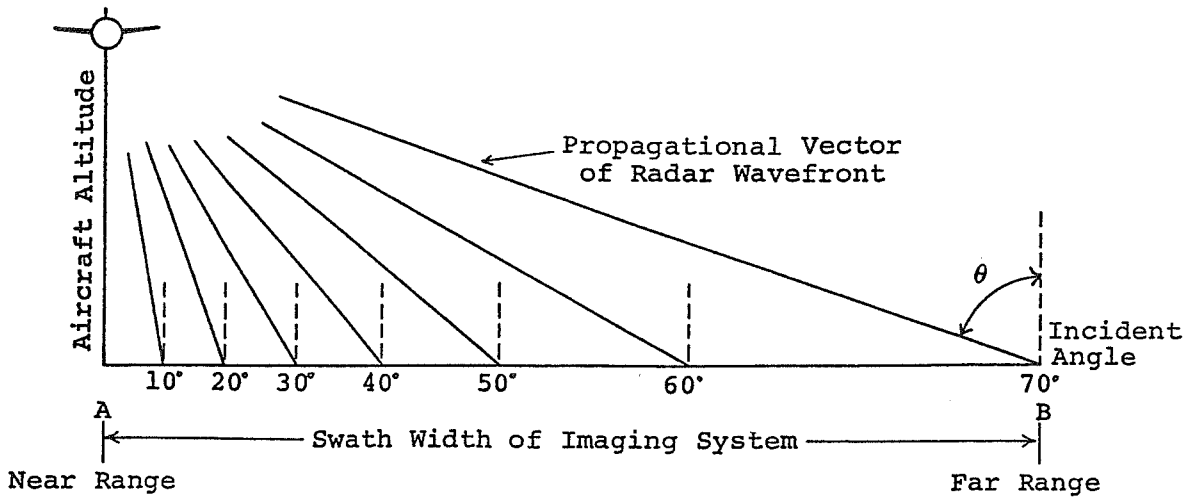


Figure 3-2. Effect of Position in the Range on Incident Angle (θ) assuming level terrain and ignoring the curvature of the Earth. The flat Earth assumption is valid for the reconnaissance-type interpretations of aircraft radar imagery. However, it is not valid in radargrammetric analysis nor when using satellite radar imagery. (From MacDonald, 1969).

(see Figure 2-32). Figures 3-2 and 3-3 demonstrate the system parameter — depression angle (β) or look angle (ϕ) — and the target parameter — terrain slope (α) — that affect incident angle (θ). These figures relate primarily to airborne radar systems. The influence of the earth's curvature must be incorporated for space imaging radar. The difference is illustrated in Figure 2-22. Figure 2-22 provides an example of the differences a change in incident angle can have on image appearance of the landscape. Figure 3-2 demonstrates the increase in incident angle from near to far range when the terrain slope is constant. Note that as the range increases, the incident angle also increases over terrain of a constant slope angle. However, assuming a constant position in range and therefore a constant depression angle (β), a decrease in the local incident angle with increasing terrain slope (α) facing the sensor is evident in Figure 3-3. The effect of changing the local incident angle is evident in Figures 3-3A, 3-3B and 3-3C. Figure 3-3A represents the standard radar presentation, *i.e.*, foreshortening of the slope. Figure 3-3B illustrates the reduction of a terrain slope to a line on the image and Figure 3-3C demonstrates the conditions for radar layover.

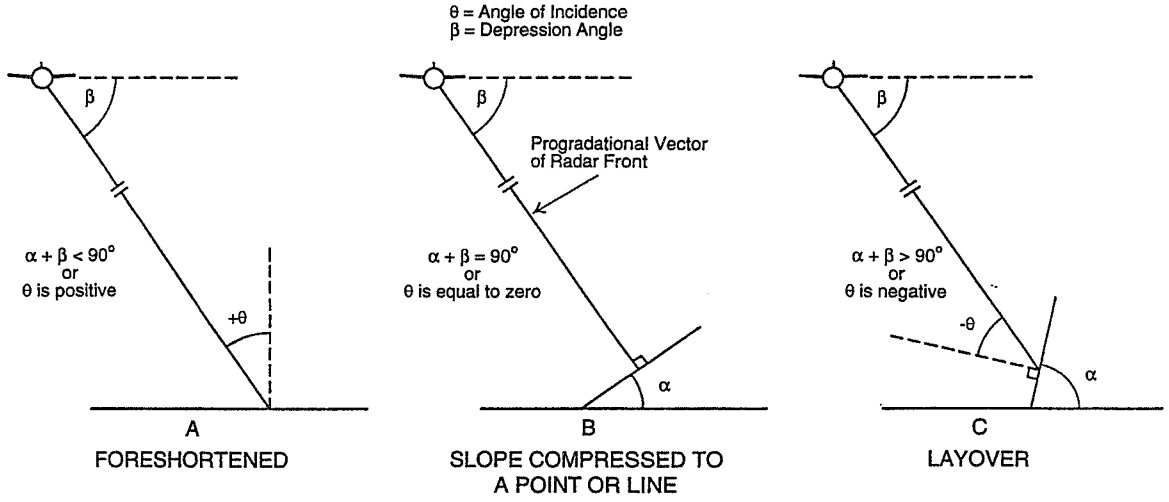


Figure 3-3. The Combined Effect of Terrain Slopes (α) and Depression Angle (β) on Local Incident Angle (θ) and Presentation of the Slope on Radar Imagery. Figure 3-3A represents the conditions for Radar Foreshortening. Figure 3-3B demonstrates the conditions for a terrain slope to be presented as a line. Figure 3-3C illustrates the conditions for Radar Layover. (Modified from Lewis, 1971).

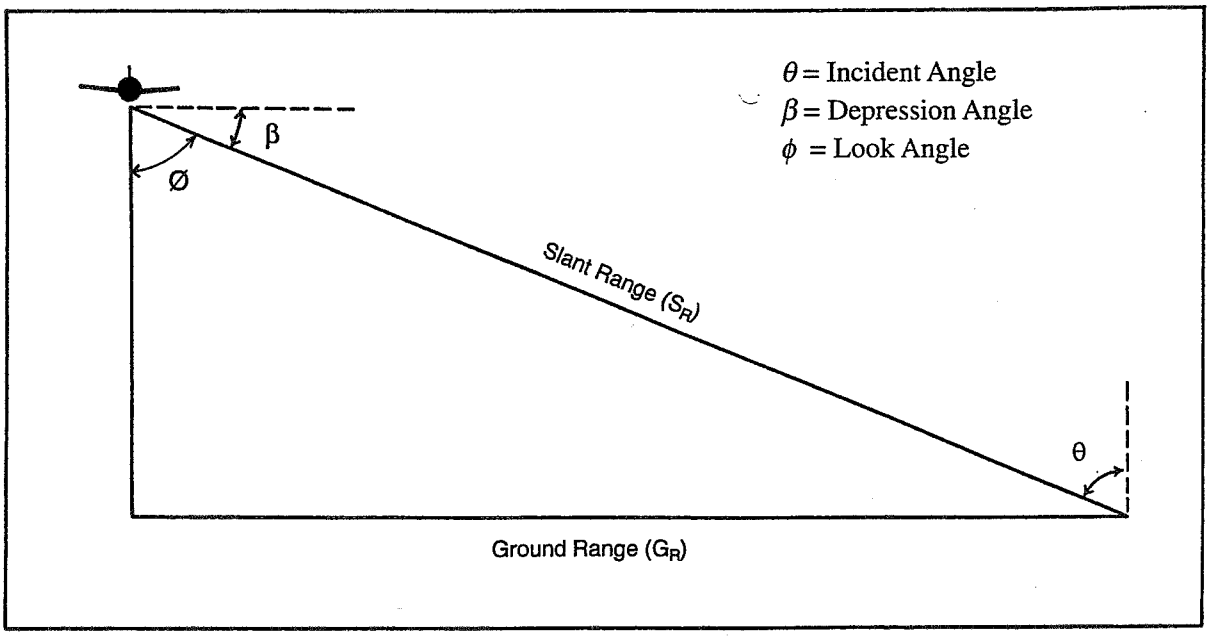


Figure 3-4. Relationship of Look Angle (ϕ), Depression Angle (β) and Incident Angle (θ) when the Terrain is Level and Earth's Curvature is not considered.

These geometric distortions are discussed in more detail later in this chapter. A changing local incident angle may also affect radar backscatter depending on the roughness characteristics of the target and the amount of change.

3-1.2.2 Look Angle (ϕ)

Look angle (ϕ), also referred to as elevation angle, is defined as the angle between the vertical of the antenna to the ground and the transmitted ray at the point of incidence (Figure 3-4). For an aircraft system flying level with the horizon, the look angle is the angle between nadir and the transmitted ray at the point of incidence; however, the angle will change with aircraft or spacecraft roll. Look angle increases from the near to the far range and is the complement of depression angle (90°-β).

3-1.2.3 Depression Angle (β)

Depression angle (β) is defined as the angle between the horizontal line of the antenna and the transmitted ray at the point of incidence (Figure 3-4). In recent interferometric literature, this angle has also been referred to as the "Look Angle." Caution is urged when reading the literature in these areas. As with the look angle, the depression angle (β) will be affected by aircraft or platform roll. When the aircraft is flying level with the horizon, β is the angle from the horizon to the transmitted ray at the point of incidence. Depression angle increases from far range to near range and is the complement of look angle (90°-ϕ). In a rectangular coordinate system, the depression angle (β) would also be the complement of the angle of incidence. However, this is not true for space imaging radars since the earth's curvature must be incorporated and the incident angle will be larger.

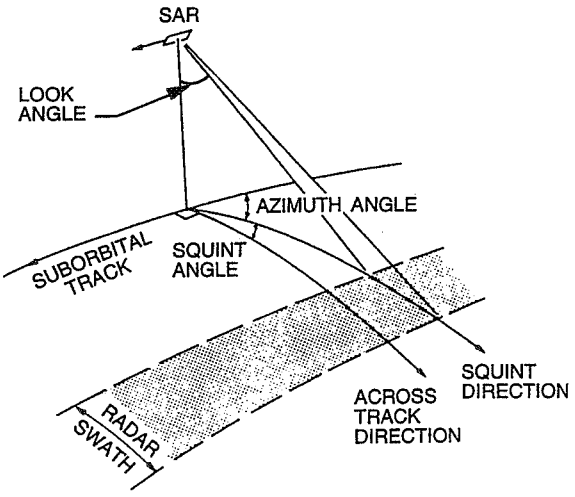


Figure 3-5. The Squint Mode and the Azimuth and Squint Angles. Operating in the Squint Mode occurs when data are collected looking forward or aft of the across track direction. Azimuth Angle is the angle described between the track of the aircraft or spacecraft and the squint direction. Squint Angle is the complement of the Azimuth Angle and is the angle between the across track direction and the squint direction. (From NASA, 1986).

3-1.2.4 Aspect Angle

Aspect angle is best defined as the azimuth viewing angle expressed in a compass direction. This angle describes the angle between the propagating wavefront and the orientation of the target. This use of this term in the early literature is often variable and contradictory to the above definition. Aspect angle has been previously described as 1) the same as incident angle or 2) the complement of depression angle. Obviously from Figure 3-4 it cannot be both, and based on the more conventional usage, it should be neither. The reader is cautioned about the early discrepancy of definitions for aspect angle. The recommended use of aspect angle refers to the compass orientation of the feature to the radar system.

3-1.2.5 Squint and Azimuth Angles

Azimuth angle is defined as the angle between the azimuth direction, which is parallel to the flightline, and the range direction, which is perpendicular to the flightline or the across track direction (Figure 3-5). In most cases the azimuth angle is 90°; it is the complement of squint angle which is normally 0°. By definition squint angle is the angle between across track (range) direction and the squint mode look direction (Figure 3-5). Squint and azimuth angles are angles of importance for synergistic imaging of SAR and other sensors or for day/night imaging at the same azimuth angle (NASA, 1989).

As mentioned, the azimuth angle and the squint angle are complementary and are 90° and 0°, respectively, when imaging in the normal SAR position. However, when the antenna is configured to look forward and/or backward from the traditional right angle look, a squint angle greater than zero results. This procedure alters the terrain/signal relationships and will generate a different perspective of the imaged terrain. Multi-azimuth (or squint) angle images have been combined into false color composite radar images to take advantage of the effect of changing azimuth angle on target return.

Multiple squint mode images may provide additional target information. Techniques for their generation are being developed. Mechanical and electronic maneuvering of the platform in a yaw direction is the primary method of altering and therefore obtaining multiple squint mode radar images. For example, electronic steering was used in SIR-C.

More discussion of the squint mode of operation and its importance are provided in Chapters 2 and 4.

3-1.3 RADAR OPERATION

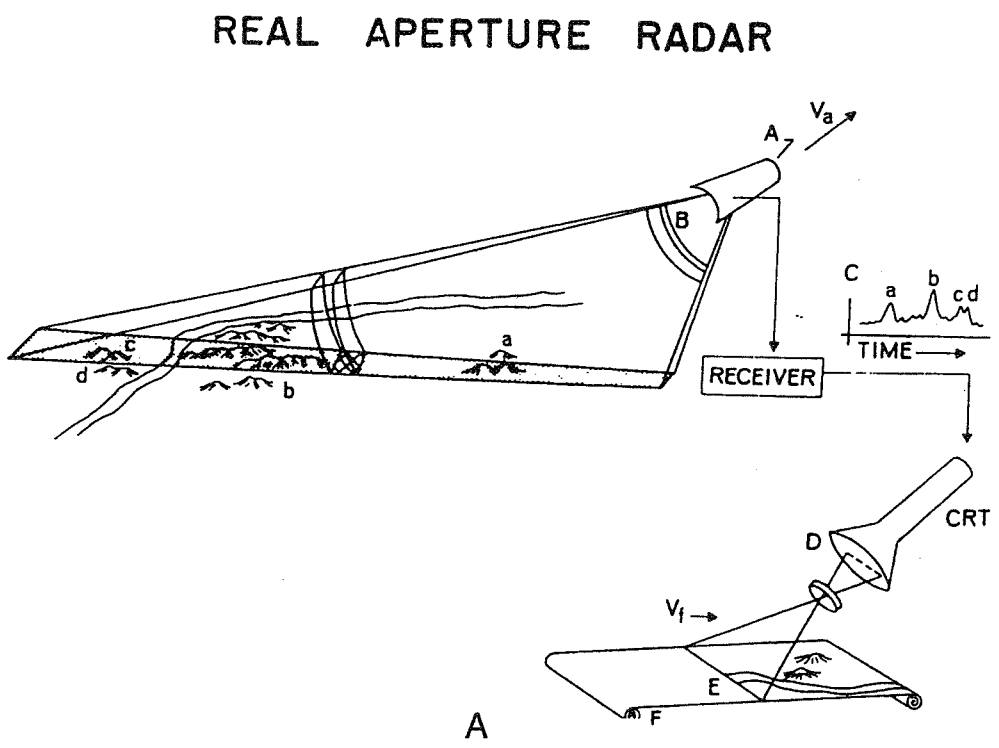
Although more accurate and detailed discussion of side-looking radar is presented in Chapter 2, a brief discussion of a simple side-looking radar is worthwhile. A simple radar has antenna, synchronizer, modulator, transmitter, transmit/receive switch, oscillator, receiver, detector and a recorder/display. In a monostatic radar the antenna acts as both a transmitter and receiver, being regulated by the transmit/receive switch. The antenna serves to obliquely direct the pulsed energy towards the terrain as well as to concentrate or focus the energy. Antenna gain is the ability of the antenna to concentrate the transmitted active microwave energy. The synchronizer is the clocking device that synchronizes the functions of all of the above elements of the imaging radar and produces a chain of accurately measured pulses. This chain of pulses is referred to as the pulse repetition frequency or PRF. The function of the modulator is to determine the pulse duration (τ), an important parameter in controlling range resolution. The transmitter obviously is responsible for transmitting the high frequency energy to be radiated. Between transmission periods or pulses the transmitter is off, and the antenna can serve as a receiver. In this situation, the transmit/receiver switch shunts the incoming reflected or backscattered radiation to the oscillator. The incoming signal is a relatively small portion of the transmitted signal, so the oscillator converts the signal to a lower frequency to make amplification more efficient. The receiver amplifies the signal to a usable level, and the detector recovers the wave form.

A variety of displays have been used in the past to present aircraft radar data. The most common are the scanning antenna and the side-looking antenna. Examples of the scanning antenna are the Plan Position Indicator (PPI) and the B-Scan. A ground mounted PPI is the common format used for TV weather radar, and B-Scan radars are frequently used in commercial aircraft for tracking storms. As mentioned earlier in Chapter 1, the first radar images were photographs taken of an aircraft mounted PPI. The side-looking antenna is the operating mode of imaging radar systems used in geosciences. The antenna is moved along a vector by the platform, aircraft or spacecraft. In the 1960s the radar data were recorded directly onto film and the image was generated. Eventually magnetic tapes and even holograms were used to record the returned radar signal. The collection of radar data also switched from analog to digital allowing machine processing (signal enhancement, rectification, classification, etc.) of the data. The other major change during the switch from analog to digital was the switch from real aperture radar (RAR) to synthetic aperture radar (SAR). The main advantage of the SAR over the RAR, besides the fact that the RAR systems were old and required more maintenance, was the better azimuth resolution. This is discussed in Section 3-1.6 and Sections 2-2 and 2-4.

3-1.3.1 Real Aperture Radar (RAR)

Real Aperture Radars (RAR) are the simpler of the two operating modes. The two most common commercial RAR systems were the K-band, dual polarized radar (AN/APQ-97) operated by Westinghouse, and the X-band radar (APS-94) operated by Motorola (MARS). Outside of the United States, the Westinghouse radar acquired coverage of Panama, Colombia, Ecuador, Nicaragua and Indonesia. Motorola radar coverage included several countries in Africa, Canada and the United Kingdom. Neither system is currently operational.

The operation of a RAR is presented in Figure 3-6A. The antenna (A) is mounted on a moving platform such as an aircraft and looks off one or both sides of the platform. The platform motion (V_a) provides the necessary mechanism to generate an image made up of individual scan lines. A pulse or burst of energy (B) is transmitted towards the terrain. The geometry and resolution of the pulse is determined in large part by the antenna length and the pulse length. As the pulse (B) interacts with the terrain, the signal is transmitted (for example, through a vegetation canopy), absorbed, reflected and/or backscattered. The returned portion of the transmitted signal (generally radar backscatter



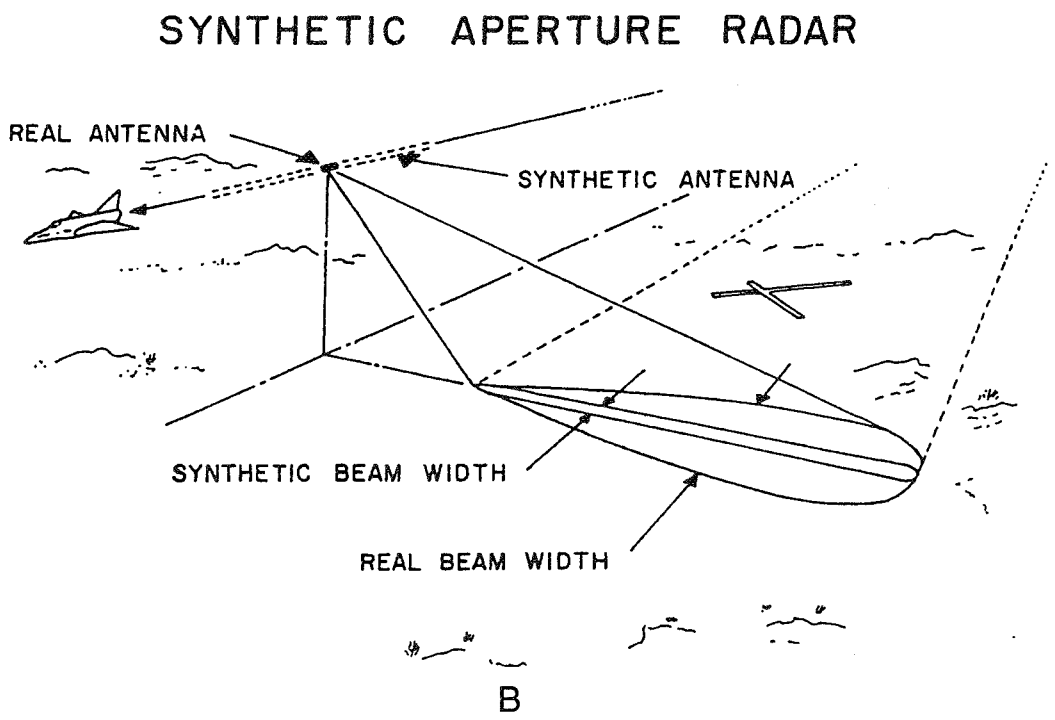


Figure 3-6. Sketch diagrams illustrating the operation of side looking radar systems: (A) Real Aperture Radar (RAR) and (B) Synthetic Aperture Radar (SAR).

except near 0° incident angles) is received as a time-dependent amplified signal. The time element is related to the distance to the target, and the amplification is related to a variety of signal/target interactions presented in Table 3-1 and discussed later. The amplification is also related to the tone on the image; the higher the returned signal, the brighter the portion of the radar image. The returned time-amplified signal is displayed on a cathode ray tube (CRT), as in Figure 3-6A, or recorded on magnetic tape. Each transmitted pulse represents a single scan line on the radar image. The illumination from the CRT is used to expose film moving at a velocity (V_f) proportional to V_a.

3-1.3.2 Synthetic Aperture Radar (SAR)

A simplified sketch of a Synthetic Aperture Radar (SAR) is provided in Figure 3-6B. SAR imaging systems are much more sophisticated and complex than are the RAR systems; however, the advantage of better azimuth resolution outweighs any disadvantage of RAR, especially from spacecraft altitudes. A summary of aircraft SAR systems can be found in Table 2-8.

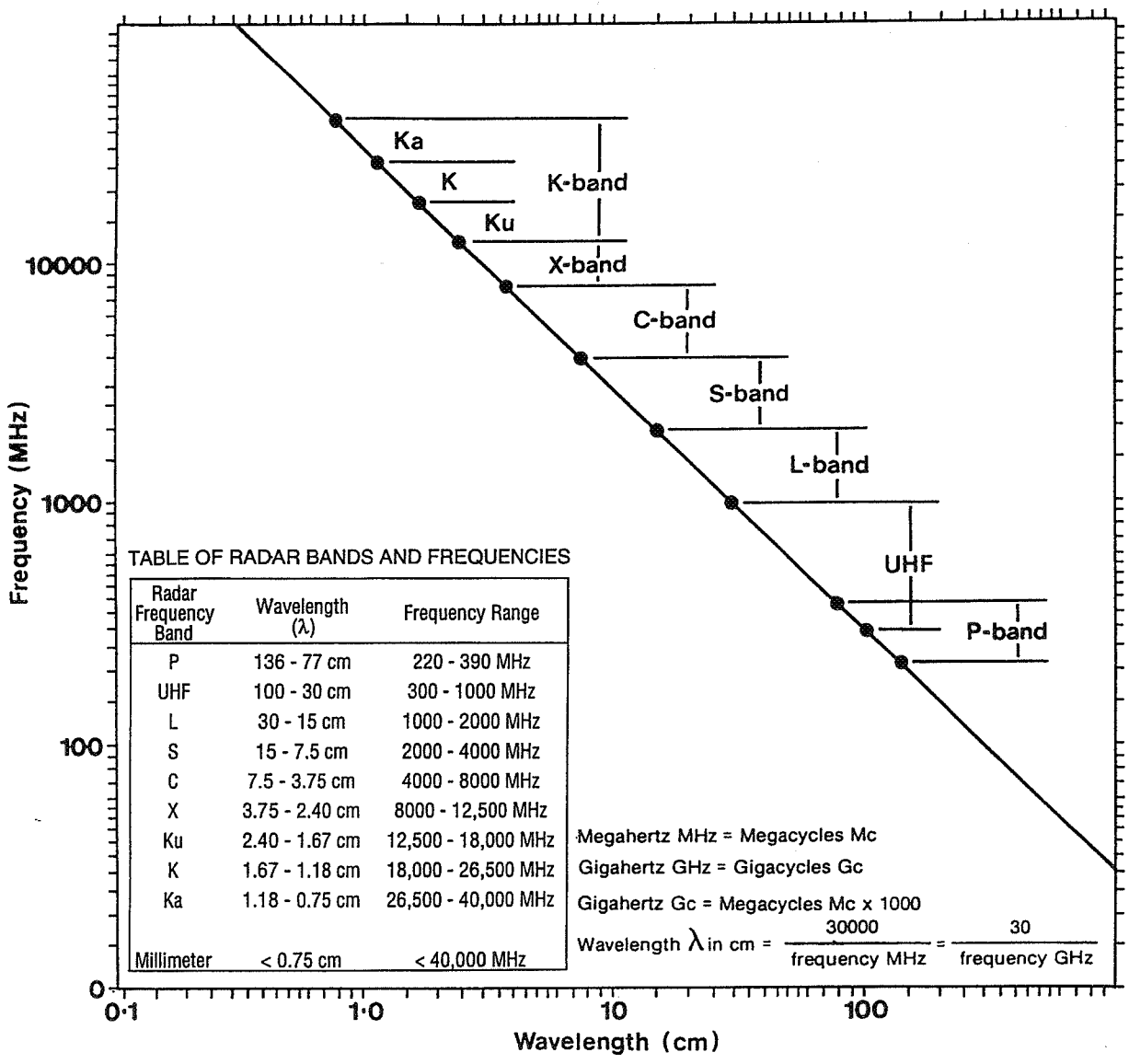


Figure 3-7. Nomogram for frequency/wavelength. Conventional radars utilize the frequency range from 230 to 40,000 HMz (Megahertz, HMz = Megacycles, Mc) although neither end of this range is truly definitive of the frequency limitation for radar operation. A letter code of frequency-wavelength bands (K-, X-, L- etc.) was arbitrarily selected to ensure military security in the early stages of radar use. This convention has continued to be used for convenience. (From Waite, 1976).

Explanation of the theory behind the operation of a SAR is beyond the discussion in this Chapter. The reader is referred to Sections 2-1 and 2-4 in this volume for an excellent description of SAR from the system perspective. Several publications by Ulaby *et al.,* (1981, 1982 and 1986) and Curlander and McDonough (1991) serve as additional sources on the theory and operation of SAR systems. As the name indicates, SAR systems set up a synthetic aperture through the forward motion of the aircraft or spacecraft and are especially suited for spacecraft operation where real aperture antenna lengths are too long to be practical. In the case of SAR systems, each target is sampled as long as it is being illuminated in the lobe of transmitted energy or lies within the beam width (Figure 3-6B).

3-1.4 WAVELENGTH/FREQUENCY

Radar wavelength and frequency are inter-related as is seen in Equations 3.11 and 3.12.

$$c = f\lambda \tag{3.11}$$

or

$$\lambda = \frac{c}{f} \tag{3.12}$$

where c is the speed of light (3×10^8 ms^{-1}), f is frequency and λ is wavelength. The conversion of frequency to wavelength is accomplished by Equation 3-12. In order to calculate λ in centimeters (cm), the value of c is given in terms of cm (3×10^{10} cms^{-1}) and the value of frequency in terms of hertz (Hz). For example, a radar imaging system operating at a frequency of 1000 MHz (Megahertz) or 1×10^9 Hz will have a wavelength of 30cm. A nomogram to calculate λ or f and a list of the conventional radar bands is provided in Figure 3-7. The definition and nomenclature for these radar bands, although now adopted as convention, are arbitrary (having been established by the US military during World War II for security reasons). Other classification systems were established outside of the United States; however, the system presented above appears to be the most widely used.

Common radar bands used in aircraft and spacecraft radar imaging systems and their nominal wavelengths are given in Tables 3-2 and 3-3. More detailed tables on satellite SAR systems (Table 2-6) and airborne SAR systems (Table 2-8) are provided in Chapter 2.

Most aircraft and spacecraft radars operate in a single wavelength mode. An early multiwavelength radar system was developed and utilized by the Naval Research Laboratory. The system operated in four channels or wavelengths (X, C, L and P). Unfortunately, the system was classified and very little of the imagery was made available to the geoscience community. However, Jet Propulsion Laboratory (JPL) has distributed multiband (C-, L- and P-band) data from AIRSAR since 1988. In October 1994, the first spaceborne multiband radar (SIR-C/X-SAR) was successfully launched, and collection of simultaneously acquired multiwavelength (multifrequency) radar (X-, C- and L- bands) from spacecraft was initiated. Black and white and false color composite examples of these images are provided in this and several subsequent chapters. These data sets are unclassified, and their distribution to the scientific community has commenced, especially via NASA/JPL.

TABLE 3-2 Selected Aircraft Imaging Radar Systems, Operating Bands and Nominal Wavelength

Radar Band	*Nominal Wavelength(cm)*	*Aircraft System (Actual Wavelength in cm)*
Ka	1	Westinghouse APQ-97(0.86)
X	3	Motorola APS-94D(2.5)
X	3	Goodyear APQ-102(3.1)
X	3	Intera STAR-1(3.1)
X	3	CCRS(3.2)
C	5	(5.7)
C	5	JPL/NASA(5.7)
L	25	(24.0)
P	70	(68.2)

Although most geoscientists refer to radar systems in terms of wavelength, systems engineers prefer to use the frequency designation. Preferred utilization of frequency is logical since once electromagnetic radiation (EMR) is generated, the frequency remains unchanged. When EMR passes through a medium other than a vacuum, the velocity will decrease. In order to maintain the integrity of Equation 3.11, one of the two parameters must change. The changed parameter is always λ.

The selection of the operating λ of the radar system is primarily dependent on the application. For example, the appearance of vegetation, soil and water surfaces and even sub-canopies and sub-surfaces will vary as a function of wavelength. This attribute is discussed and illustrated more fully in the respective application chapters that follow. Other considerations, such as those related to the radar equation for determining power return (Equation 3.10), may also be relevant. In addition to the influence of λ on radar power, what other radar-related equations incorporate λ as one of the parameters? The definition of surface rough-

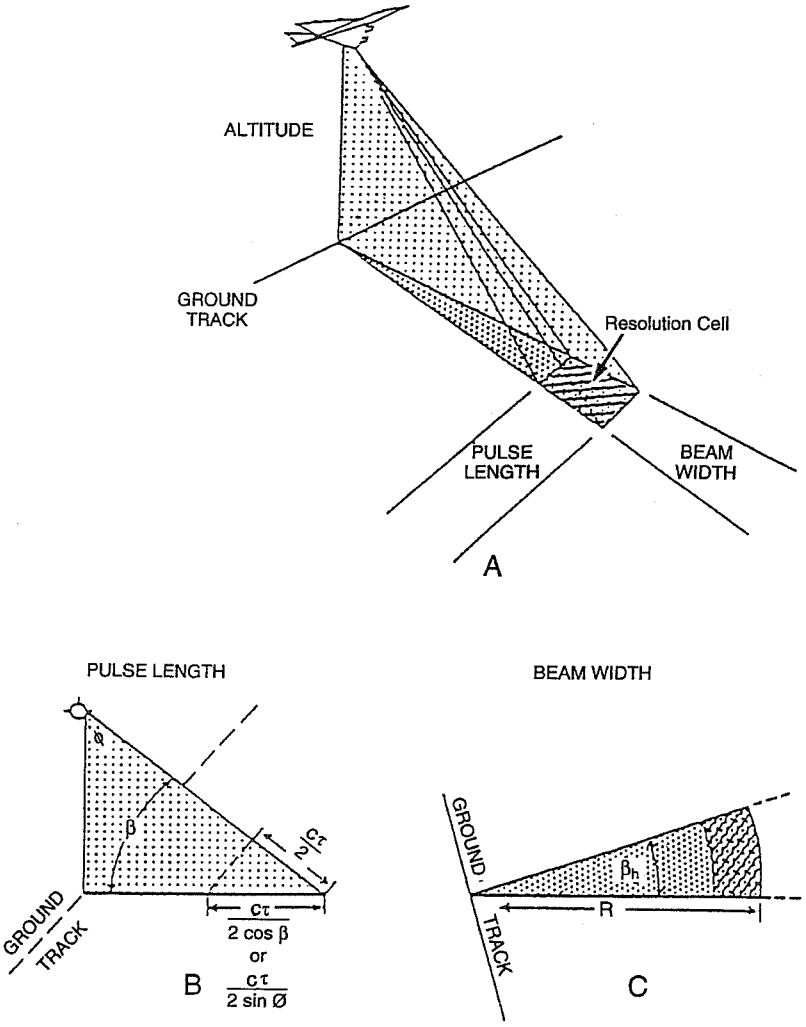
TABLE 3-3 Selected Spacecraft Imaging Radar Systems, Operating Bands and Nominal Wavelength

Radar Band	*Nominal Wavelength(cm)*	*Spacecraft System (Actual Wavelength in cm)*
X	3	SIR-C/X-SAR (3.0)
C	5	SIR-C/X-SAR (5.3)
C	5	ERS-1 (5.3)
S	10	ALMAZ-1 (9.5)
L	25	SIR-C/X-SAR (24.0)
L	25	JERS-1 (23.0)
L	25	SIR-A and B (23.5)
L	25	SEASAT-SAR (23.5)

3-1.6 RESOLUTION

An understanding of the resolution of a remote sensing system is fundamental to any interpretation. Knowledge of variables that determine resolution and the range of resolutions for the system are important considerations in determining the feasibility of using a given sensor. There is little sense in using a system with one kilometer resolution if the user requires more detailed spatial data. Since the resolution of the system equates to the Instantaneous Field Of View (IFOV), this information provides the image interpreter with information on the size of the area responsible for the radar return. Resolution also defines the necessary separation distance between two targets for them to be resolved and displayed as individual features.

Radar resolution is defined in two directions: parallel and perpendicular to the flightline (Figure 3-10). Radar resolution parallel to the flightline is termed azimuth resolution. Resolution perpendicular to the radar flightline is termed range resolution. The parameters that control radar resolution depend on the defined direction. In addition, azimuth resolution varies with the type of imaging radar system: real aperture or synthetic aperture (Figure 3-11). Chapter 2 provides a discussion of resolution and the resolution cell from the system perspective.



Range resolution at depression angle β, $c = 3 \times 10^8$ m/sec., and τ = pulse length.

Azimuth resolution = RB_h, where R is slant range and B_h is half-pwer beam width.

Figure 3-10. Radar Imaging System Resolution. (A) Pulse length and beam width in relation to ground track. (B) Range resolution. (C) Azimuth resolution for Real Aperture Radar (RAR). (From MacDonald, 1969).

3-1.6.1 Range Resolution

Range resolution in a pulse radar is directly related to the pulse length of the transmitted radar signal[1]. The shorter the pulse length, the finer the range resolution. Pulse length, the physical length of the microwave signal, is a product of c multiplied by (τ) where c is the speed of light and τ is the duration of the transmission (usually given in microseconds) and ranges from 0.4 to 1.0 micro(μ) seconds. Given the normal range of transmission bursts, the pulse length ($c\tau$) will range from approximately 8 to 210 meters. The trade off, and the reason why shorter pulse lengths are not used, is that as the pulse length diminishes so also does the total energy illuminating the target. Although a short pulse length improves range resolution, there is a point where the signal is too weak to be recorded.

Because the radar signal must travel to the target and back to the sensor, the pulse length (c) is divided by 2 to determine slant range resolution ($c\tau/2$) (see Figure 3-10A and B). Slant range resolution (S_r) is constant across the entire range (near to far) of the image or data set. However, more important to the geoscientist is the ground range resolution, which is variable across the range and more complex to calculate. To accommodate the different geometries of slant and ground range, the pulse length is divided

[1] The determination of range resolution is dependent upon the type of radar system (pulse or FM). Pulse length is important in the pulse radar whereas bandwidth is an important parameter with FM radars.

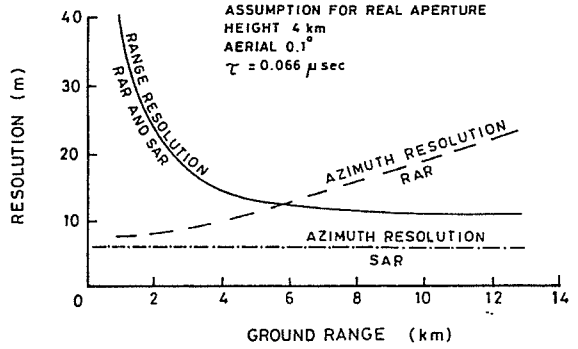


Figure 3-11. Changing radar range and azimuth resolution with distance for Real Aperture Radar (RAR) and Synthetic Aperture Radar (SAR). (Modified from Grant, 1973).

by either the cosine of the depression angle (β) or the sine of the look angle. The resultant equation for ground range resolution (G_r) is as follows:

$$G_r = \frac{c\tau}{2\sin\phi} = \frac{c\tau}{2\cos\beta} \tag{3.13}$$

It is important to understand how Equation 3.13, and therefore G_r, behaves across the swath width (range) of the image. When the look angle is zero (0°) (the ultimate near range position) and the imaging system is looking directly below the aircraft, the equation increases to infinity and G_r becomes infinitely large. When the look angle is 90° and the imaging system is effectively looking at the horizon, the range resolution is at a minimum — equal to the slant range resolution ($G_r = S_r$). Range resolution is fairly unique in that it improves from the near to the far range. Figure 3-12 illustrates the change in G_r from near to far range.

Since satellite active microwave systems generally operate with a small range of look (or depression) angles (Figure 3-13), ground range resolution will not vary much unless small look angles are used (Figure 3-12). However, this is not the case with many aircraft-mounted radar systems. The necessity of operating with a larger range of look angles in order to increase the area imaged on aircraft radars is demonstrated in Figure 3-13. In the case of an aircraft mounted imaging radar system operating with a pulse duration (τ) of 0.1 μ seconds and collecting data between look angles (ϕ) of 15° to 75°, the ground range resolution will vary from approximately 58 meters in the near range ($\phi = 15°$) to approximately 15.5 meters in the far range ($\phi = 75°$). The same system will have a slant range resolution of 15 meters across the entire range of look angles.

3-1.6.2 Azimuth Resolution – Real Aperture (RAR)

In the azimuth direction (parallel to the flightline), the resolution is determined by the type of imaging system as well as operational parameters (Figure 3-10A and C). Azimuth resolution (A_R) with a real aperture (brute force) imaging radar systems is most simply defined in the Equation 3.14 below by the beam width (β_h) and the distance in the range (R).

$$A_R = \beta_h R \tag{3.14}$$

Beam width (β_h) represents the width of the area being imaged; obviously azimuth resolution improves as the beam width decreases. The best azimuth resolution with a real aperture radar is in the near range where the beam width (β_h) is narrower. From Equation 3.14 it is apparent that range resolution degenerates in a linear fashion with increasing distance from the sensor or from near to far range. For example, an imaging radar operating at a β_h = 1.8 milliradians with an R = 6 km would have an azimuth resolution of 10.8 m as shown in Equation 3.15. If R is doubled to 12 km, then A_R is also doubled to 21.6 m.

$$A_R = (1.8\text{x}10^{-3}\text{ radians})(6.0\text{x}10^{3}\text{ meters}) \tag{3.15}$$

$$A_R = 10.8\text{ m}$$

Beam width can be further expressed in terms of the wavelength of the imaging system (λ) and the physical length of the radar antenna (l).

$$\beta_h = \frac{\lambda}{l} \tag{3.16}$$

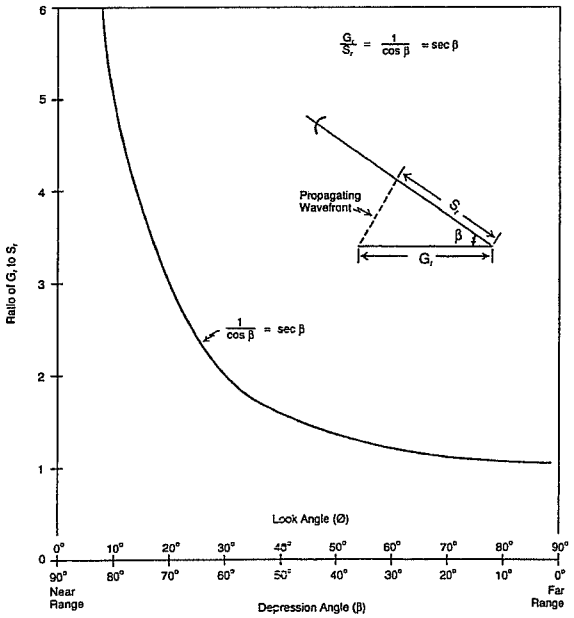


Figure 3-12. Relationship of G_r and S_r with varying look angle (ϕ) and depression angle (β). Note: The equation for G_r/S_r is an approximation and applies to short distances on the ground. The full formula over long distances is G_r/S_r = secβ + tanβ. (From Lewis, 1971).

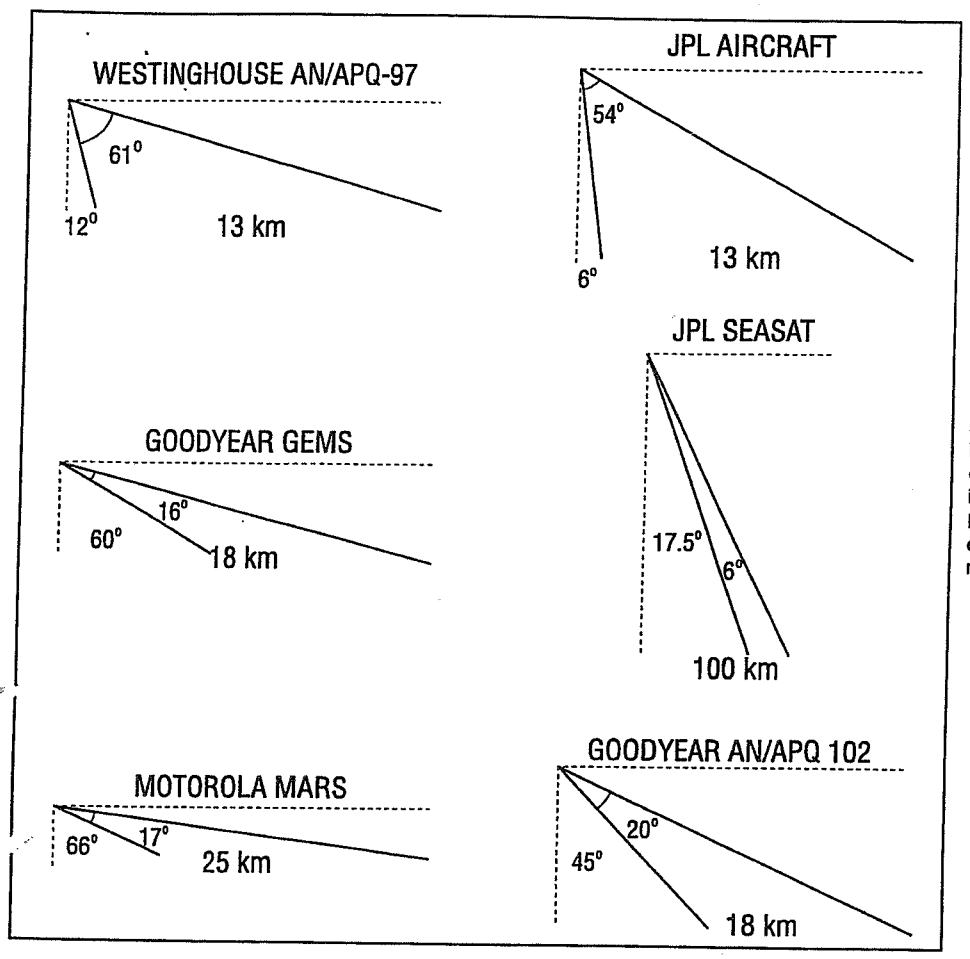


Figure 3-13. Range of swath angles indicating image coverage for selected aircraft and spacecraft imaging radar systems. Near range Look Angle (ϕ) and imaging swath in kilometers (km) are given.

Azimuth resolution for a real aperture radar system therefore becomes

$$A_R = (R)\frac{\lambda}{l} \tag{3.17}$$

From Equation 3.17 it is apparent that in addition to range (R) there are two parameters that can be altered to increase azimuth resolution of imaging radar: wavelength (λ) can be decreased and/or antenna length increased. Both of these parameters present limitations. First, the low range of the electromagnetic spectrum considered to be microwave energy is approximately 1mm, although most active microwave imaging systems operate at wavelengths above 1cm. Antenna lengths are limited by the size of the aircraft and government regulations, which consider the stability of the aircraft's design.

The equation for A_R can be further expanded to help better understand other parameters that influence azimuth resolution. From Equation 3.14 ($A_R = \beta_h R$), it can be seen that A_R increases directly with R. Since

$$R = \frac{H}{\cos\phi} = \frac{H}{\sin\beta}$$

where H is the altitude of the sensor, ϕ is look angle and β is depression angle, Equation 3.17 can be expanded to

$$A_R = \left(\frac{H}{\cos\phi}\right)\frac{\lambda}{l} = \left(\frac{H}{\sin\beta}\right)\frac{\lambda}{l} \tag{3.18}$$

Using Equation 3.18, it can be seen that the azimuth dimension of the resolution cell increases and spatial resolution decreases with increasing altitude of the sensor and increasing look angle (ϕ) or decreasing depression angle (β). Recalling Equation 3.13, it should be apparent that azimuth and range resolution are affected differently by look angle or depression angle. The resolution cell in the azimuth direction will increase from near to far range thereby reducing the azimuthal spatial resolution.

Equation 3.18 also presents the possibility of tradeoffs between the variables in the equation (H,λ,l and ϕ). Aircraft-

mounted RAR radar systems generally operate at low altitude helping to maintain reasonable resolution in the azimuth direction which is also improved by reducing λ, for example by operating a K-band (1 cm) or X-band (3 cm) system rather than an L-band (25 cm) system. All else being equal, the azimuth resolution of a K-band system would be 25x better than an L-band system. Changing wavelength changes surface roughness sensitivity and penetration capabilities. Azimuth resolution could also be improved by operating at small look angles; however, operating at small look angles would result in a decrease in range resolution. A reduction of altitude would improve spatial resolution but would also result in a reduction of the area covered or swath width at a given set of ϕ. Increasing the range of look angles (ϕ) between near and far range would help to compensate for the smaller swath width at lower altitudes; however, the effect of changing look angles (ϕ) on several other important parameters must be considered. The influence of ϕ on range and azimuth resolution has already been demonstrated (see Equations 3.13 and 3.18), but it also plays an important part on radar return since ϕ affects 1)incident angle, 2) sensitivity to surface roughness, 3) location (surface or subsurface) of radar return from the target and 4) the angle of refraction when radar penetration occurs. Trade-offs and interrelationships between and among the various radar parameters must be understood by the geoscientist in order to maximize interpretation accuracies and/or minimize incorrect assessments regarding the capabilities of RAR imaging systems.

3-1.6.3 Azimuth Resolution — Synthetic Aperture Radar (SAR)

Although a detailed explanation of how a synthetic aperture radar (SAR) system works and the derivation of the equation are covered in detail in Chapter 2, a brief discussion follows. As the name states, a SAR system sets up a synthetic aperture (antenna) and stores the signal strength and phase relationships as the target passes through the wide beam generated by the small antenna. The data from the multiple views of the target are processed in such a way that a longer antenna is synthesized. Limitations include the stability of the antenna and the duration of the time the target is in the radar beam.

Azimuth resolution in a synthetic aperture radar (SAR) system is radically different than in real aperture radar (RAR). The equation for the maximum attainable azimuth or along-track resolution for a point target in a SAR system is given simply as

$$A_S = \frac{l}{2} \tag{3.19}$$

where l is the antenna length.

This is one of the more incredible equations in the discipline of radar remote sensing, incredible in both simplicity and importance to remote sensing. A comparison of the two azimuth resolution equations (Equations 3.18 and 3.18) reveals several important differences. Perhaps the most notable is the omission of R (range distance) or its equivalent $H/\cos\phi$ from the equation for A_R (Equation 3.14 or Equation 3.17). What are the ramifications of this change, and of what importance is it to radar remote sensing? Obviously the omission of R from the equation defining azimuth resolution means that the azimuth resolution of an SAR system is independent of range distance or sensor altitude. Theoretically, the spatial resolution in the azimuth direction from a SAR imaging system will be the same from an aircraft platform as it is from a spacecraft. There is no other remote sensing system with this capability!

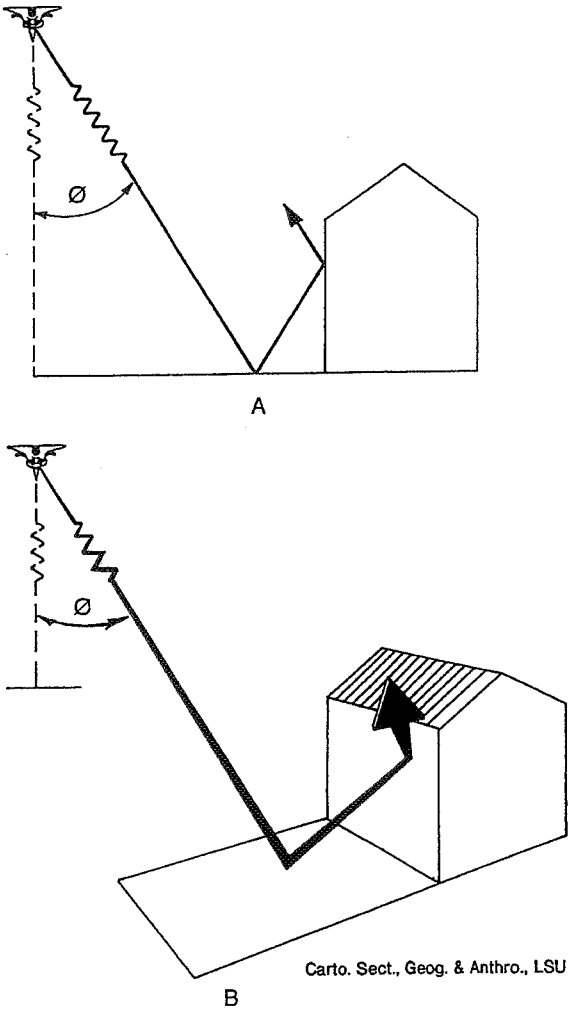


Figure 3-14. Effect of flightline and feature orientation on radar return. (A) Represents the case where flightline and feature orientation are parallel (dihedral corner reflector) or close to parallel resulting in a high radar return. (B) Represents the case where flightline and feature orientation are oriented at an angle greater than 10° and a low radar return results. Look angle (ϕ) is constant in both cases. (Lewis *et al.*, 1997).

However, there is a limitation to Equation 3.19. The coherent nature of the SAR signal results in speckle or fading that is discussed later in Sections 3-1.9 and 3-1.10. To improve image quality, speckle is minimized but at the expense of azimuth resolution. The image is divided into sub-beams. Each sub-beam provides an

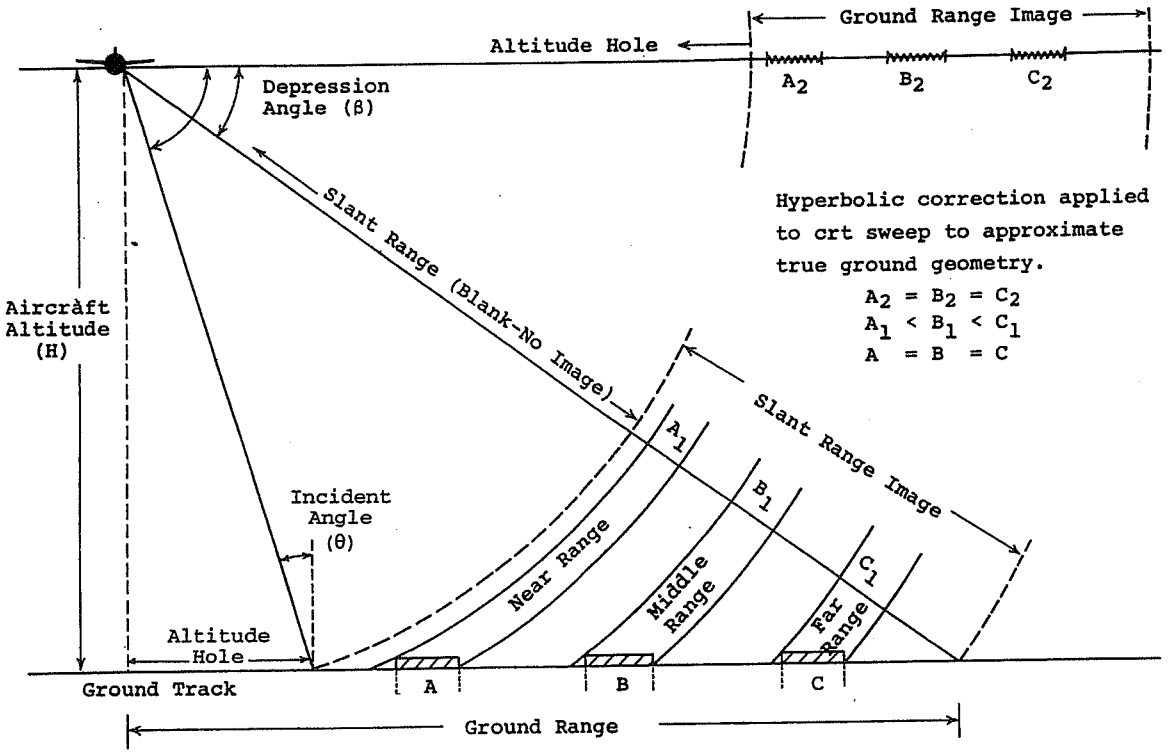


Figure 3-15. Radar Image Presentation. Geometry or scale differences in slant range and ground range images and explanation of the correction factor. Although the example used became outdated with the introduction of digital data and rectification programs, it serves to explain the differences and corrections more graphically. (Modified from Barr, 1968).

independent "look." Each look (N) is incoherently averaged producing an image with reduced speckle and reduced azimuth resolution. The approximate resolution of the incoherently averaged image is given in Equation 3.20.

$$A_S = (N)\frac{l}{2} \tag{3.20}$$

Using the equation above and the antenna length for a SIR-C SAR (12m), the maximum across-track resolution is 6 m. Speckle reduction processing using $N = 4$ looks will degrade the across-track resolution to approximately 24 m (NASA, 1986).

Another interesting characteristic of this equation is that the antenna length (l) is now in the numerator whereas, previously, in the equation for RAR (Equations 3.17 and 3.18), antenna length was part of the denominator. This reversal means that, technically, the smaller the antenna the better the resolution (the opposite of RAR). The wavelength of the system (λ) in Equations 3.17 and 3.18 has been omitted from Equation 3.20 indicating that azimuth resolution in a SAR system is also independent of λ.

3-1.7 LOOK DIRECTION/ORIENTATION

Look direction refers to the orientation of the feature to the flightline. Orientation is not significant if the feature is perfectly symmetrical or perfectly random; however, it becomes extremely important if there is a preferred orientation to the receiving area of the impinging radar beam. Orientation is especially important for urban and many other settled areas. Urban areas, because of selected orientation of street patterns and therefore buildings, provide optimum opportunities for large fluctuations of power return related to orientation. Streets and buildings are generally smooth to impinging radars; they combine to produce dihedral corner reflectors (Figure 3-14A). Orientation within 10° to 15° of the flightline results in a high radar return whereas at a higher angle of orientation away from the flightline, the urban area would produce a low radar return (Figure 3-14B). The influence of orientation has also been noted on plowed fields and cropland, the far range side of river banks, the edges of forest clearcuts, sand dunes and railroads. The preferred orientation of topographic and geologic features has also been noted. The acquisition of multilook direction radar coverage whenever feasible helps to maximize information extraction from radar imagery.

11.3 Information sensed by radar

Since radar is an **active system** and supplies its own radiation, it is able to collect information that is similar to that collected with passive systems (signal strength, angle to an object) as well as information that is unique to active systems (range, or distance to the object and the relative velocity of the object sensed). In this section we consider the system and target characteristics that affect each of these types of information.

11.3.1 Signal Strength

A radar system transmits pulses of microwave energy in a small range of direction and records the strength of the signal scattered or reflected from objects within the system's field of view. Consider an antenna that approximates a point source, radiating in all directions. If a pulse of peak power, P_t , is transmitted, the power density (irradiance) on a sphere of radius R centered on the point source will be:

$$\frac{P_t}{4\pi R^2} \tag{11.1}$$

Most real antennas are directive, i.e., they emit radiation into a fixed solid angle. This focusing of radiation may be expressed as the gain of the transmitting antenna, G_t. Thus the energy density reaching a point at a distance R from the antenna is:

$$\frac{P_t G_t}{4\pi R^2} \tag{11.2}$$

The target intercepts a portion of the radiated power, and re-radiates (reflects, scatters) power back to the radar. The efficiency with which the target returns energy in the direction of a receiving antenna is expressed as the radar cross-section, σ. Thus, the energy reflected/scattered per unit solid angle is:

$$\frac{P_t G_t}{4\pi R^2}\left|\frac{\sigma}{4\pi}\right| \tag{11.3}$$

The concept of cross-section, as its name suggests, is that of effective area of the scattering target. An object that is 1 m^2 and scatters the same amount of energy as a perfectly diffuse scattering sphere with a cross-sectional area of only 0.1 m^2 has a scattering cross-section of 0.1 m^2. The units of σ are then area units, usually square meters[1].

Assuming that the transmitting and receiving antennae are collocated, the power density at the radar is then

[1] The scattering cross-section is sometimes expressed as the ratio of the received backscatter to that from an equivalent isotropic scatterer of the same size. It is then called σ_0, a dimensionless quantity that is typically expressed in decibels (db).

$$\frac{P_t G_t}{4\pi R^2}\left|\frac{\sigma}{4\pi}\right|\left|\frac{1}{R^2}\right| = \frac{P_t G_t \sigma}{(4\pi R^2)^2} \tag{11.4}$$

The receiving antenna focuses the incoming radiation so that the power received at the antenna, the **Radar Equation,** is:

$$P_r = \frac{P_t G_t \sigma}{(4\pi R^2)^2} A_r \tag{11.5}$$

where A_r = effective receiving aperture.

Antenna theory gives a relationship between antenna gain and effective area, A, of the antenna. Let A_t be the effective area of the transmitting antenna and A_r the effective area of the receiving antenna. Then, the gain of the antennas is given by:

transmitting	receiving	
$G_t = \frac{4\pi A_t}{\lambda^2}$	$G_r = \frac{4\pi A_r}{\lambda^2}$	(11.6a,b)

If a single antenna is used for both transmitting and receiving:

$$G_t = G_r = G \qquad A_t = A_r = \lambda^2 G / 4\pi \tag{11.7a,b}$$

The radar equation may then be expressed either in terms of the antenna area of the antenna:

$$P_r = \frac{P_t A^2 \sigma}{4\pi\lambda^2 R^4} \tag{11.8}$$

or, alternatively, in terms of the area of the antenna:

$$P_r = \frac{P_t G^2 \lambda^2}{(4\pi)^3}\frac{\sigma}{R^4} \tag{11.9}$$

Unfortunately, there is also substantial noise contaminating the signal. Some of the noise is external (atmosphere, earth, power facilities, other radars, etc.), and some of the noise is generated in the transmission of the radar pulse and reception of the return signal. The maximum radar range, R_{max}, is the distance beyond which the target cannot be detected. It is reached when the received signal, P_r is just equal to the minimum detectable signal, S_{min}, above the noise level.

From Equation 14.8 it is clear that the power will drop off as the 4^{th} power of the distance, and that one may increase the received power by increasing the size of the antenna, A, or using shorter wavelengths, λ. The atmosphere limits the usable wavelengths and large antennas are more than merely inconvenient on aircraft and satellites.

The radar cross-section is the factor that relates to the scattering surface. This is roughly analogous to reflectance in the optical range. The radar cross-section incorporates the slope, roughness, area, dielectric properties, and angular response of the target into a single factor. Issues of specular, diffuse and volume reflectance/scattering apply here much as they do in the

visible domain. The difference is that the scale of roughness is much different and instead of the index of refraction, we are concerned with the dielectric of the scattering surface.

11.3.2 Surface Roughness

Radar system use wavelengths that are on the order of centimeters to meters. Distinguishing between surfaces that will reflect/scatter specularly versus surfaces that will reflect/scatter diffusely requires that we understand the nature of roughness at these wavelengths. Surfaces that are smooth at these wavelengths are vastly different from surfaces that are smooth at optical wavelengths. The criterion for a smooth surface is that a surface may be considered smooth (specular) if:

$$\Delta h < \frac{\lambda}{A \cos \theta_0} \tag{11.10}$$

where Δh is the root-mean-square variation in the surface height, λ is the wavelength of the radar pulse and θ_0 is the incidence angle. There are various ways of defining the value for the constant A. For the Rayleigh Criterion, A = 8. For the Fraunhoffer criterion, A = 32. One may also use Equation 14.10 to define rough or moderately rough surfaces. One commonly used states that:

$$\text{If } \Delta h < \frac{\lambda}{25 \cos \theta_0} \text{ the surface is smooth and}$$

$$\text{if } \Delta h < \frac{\lambda}{4.4 \cos \theta_0} \text{ the surface is rough.}$$

The categories of rough, smooth an intermediate surfaces for various satellite radar systems is shown in Table 11.2. A schematic drawing of the relationship between the radar cross-section and the incidence angle for surfaces in the different roughness ranges is shown in Figure 11.7.

		frequency	wavelength	dep. angle	Smooth	Intermediate	Rough
		GHz	λ (cm)	γ (deg)			
Ka	**0.75 - 1.10 cm**						
		35.0 GHz	0.86 cm	40.0 deg	< 0.05 cm	0.05 cm - 0.3 cm	> 0.3 cm
K	**1.10 - 1.67 cm**						
		20.0 GHz	1.5 cm	40.0 deg	< 0.09 cm	0.09 cm - 0.53 cm	> 0.53 cm
Ku	**1.67 - 2.40 cm**						
		15.0 GHz	2.0 cm	40.0 deg	< 0.12 cm	0.12 cm - 0.71 cm	> 0.71 cm
X	**2.40 - 3.75 cm**						
	SIR-C	10.0 GHz	3.0 cm	50.0 deg	< 0.16 cm	0.16 cm - 0.89 cm	> 0.89 cm
C	**3.75 - 7.5 cm**						
	Radarsat-1,2	5.36 GHz	5.6 cm	20.0 deg	< 0.65 cm	0.65 cm - 3.72 cm	> 3.72 cm
	Radarsat-1,2	5.36 GHz	5.6 cm	50.0 deg	< 0.29 cm	0.29 cm - 1.66 cm	> 1.66 cm
	ERS-1,2	5.30 GHz	5.7 cm	68.0 deg	< 0.24 cm	0.24 cm - 1.39 cm	> 1.39 cm
	SIR-C	5.20 GHz	5.8 cm	50.0 deg	< 0.30 cm	0.3 cm - 1.71 cm	> 1.71 cm
S	**7.5 - 15 cm**						
		3.00 GHz	10.0 cm	68.0 deg	< 0.43 cm	0.43 cm - 2.45 cm	> 2.45 cm
L	**15 - 30 cm**						
	SIR-A,C	1.28 GHz	23.5 cm	68.0 deg	< 1.02 cm	1.02 cm - 5.77 cm	> 5.77 cm
	Seasat	1.28 GHz	23.5 cm	68.0 deg	< 1.02 cm	1.02 cm - 5.77 cm	> 5.77 cm
	JERS-1	1.28 GHz	23.5 cm	68.0 deg	< 1.02 cm	1.02 cm - 5.77 cm	> 5.77 cm
P	**30 - 100 cm**						
		0.50 GHz	60.0 cm	68.0 deg	< 2.59 cm	2.59 cm - 14.7 cm	> 14.7 cm

Table 11.3: Roughness scales for various satellite radar systems.

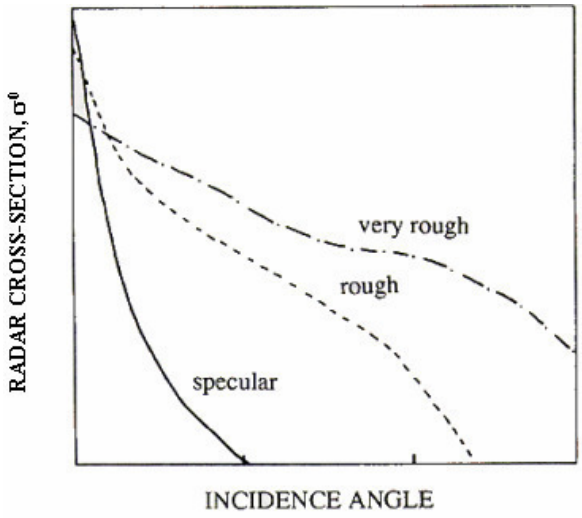


Figure 11.2: Variation in the radar cross-section with incidence angle and surface roughness.

11.3.3 Dielectric constant

The radar cross-section is also related to the dielectric constant of the target material. Recall from Chapter 2, that the dielectric constant is represented as a complex number:

$$\varepsilon_r = \varepsilon' + i\varepsilon'' \qquad (11.11)$$

where ε' and ε" are the real and imaginary part of the dielectric constant. As with the index of refraction, the full dielectric constant describes both the relative speed of light and the absorptive properties of the medium. In this case, the permittivity, ε_r' described the ability of the material to be polarized by the external electric field and ε" (the loss factor) quantifies the efficiency with which the electromagnetic energy is converted to heat (absorbed). For non-polar materials (which includes most natural and non-metalic man-made materials) the value of ε' is quite low. Furthermore, the absorption of most materials in the microwave is quite low. Water is the major exception to this rule. Water is a polar molecule and, in liquid form, also absorbs microwave radiation very effectively. The change in the two terms of the dielectric constant with respect to wavelength is shown in Figure 11.3. Note that the real part of the dielectric constant for water is quite high (~80) for all wavelengths over a meter and that the absorption peaks at about 10 cm. (Actually it's at 12 cm, the wavelength used for microwave ovens since the absorbed radiation is converted into heat.)

Since the dielectric constant for water is much higher than for most other natural materials at radar wavelengths, water tends to dominate the variation in radar return due to dielectric properties. Radar return is then strongly influenced by the moisture content in soil and vegetation (Figure 11.4). A high moisture content implies a high dielectric constant and a high radar reflectivity. A low moisture content implies a low dielectric constant and a relatively weak radar return (assuming that texture and shape factors can be ignored). The reflectivity of snow and ice are also sensitive to the liquid water content. Water content also affects the penetration depth of radar into the soil (or snow or ice). In very dry soils, radar may penetrate and record returns from several meters below the soil surface. In saturated soils the penetration depth may be little more than a centimeter.

Water itself reflects radar radiation well, however, it often appears dark in radar imagery because the water is "smooth" at radar wavelengths and the radar pulse is reflected away from the radar rather than being scattered back. When the water surface is disturbed by waves or turbulence, the radar return can be quite strong (

Figure 11.5).

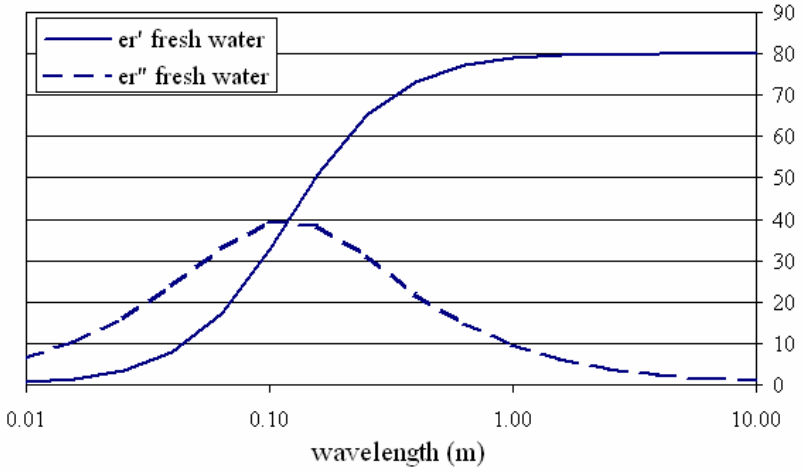


Figure 11.3: The real (polarizability), ε', and imaginary (absorbtion), ε'', of fresh water at 20°C with respect to wavelength. Note that water absorbs most strongly at ~12 cm – precisely the wavelength chosen for microwave ovens since the absorbed radiation is converted into heat.

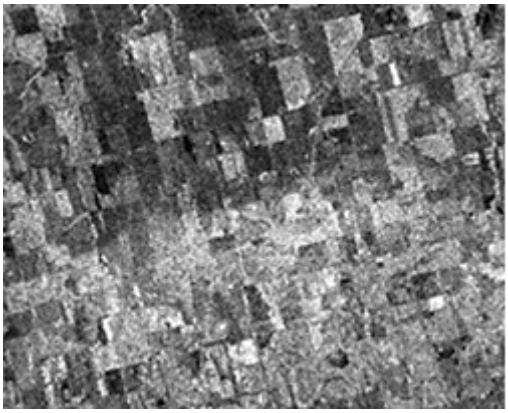


Figure 11.4: Radar image acquired near Melfort, Sasdkatchewan, 7 July 92, shows the effect of a localized precipitation event on the microwave backscatter. Areas where precipitation has recently occurred can be seen as a bright tone (bottom half) and those areas unaffected by the event generally appear darker (upper half).

http://visibleearth.nasa.gov/cgi-bin/viewrecord?1535

Figure 11.5: This is a radar image of an offshore drilling field about 150 km (93 miles) west of Bombay, India, in the Arabian Sea. The drilling platforms appear as bright white spots. The dark streaks are extensive oil slicks surrounding many of the drilling platforms. The oil flattens the water surface resulting in specular reflection of the radar radiation. There are also two forms of ocean waves shown in this image. The dominant group of large waves (upper center) are called internal waves. These waves are formed below the ocean surface at the boundary between layers of warm and cold water and they appear in the radar image because of the way they change the ocean surface. Ocean swells, which are waves generated by winds, are shown throughout the image but are most distinct in the blue area adjacent to the internal waves.

This image was acquired by the Spaceborne Imaging Radar-C/X-Band Synthetic Aperture Radar (SIR-C/X-SAR) aboard the space shuttle Endeavour on October 9, 1994. The colors are assigned to different frequencies and polarizations of the radar as follows: Red is L-band vertically transmitted, vertically received; green is the average of L-band vertically transmitted, vertically received and C-band vertically transmitted, vertically received; blue is C-band vertically transmitted, vertically received. The image is located at 19.25 degrees north latitude and 71.34 degrees east longitude and covers an area 20 km by 45 km (12.4 miles by 27.9 miles). SIR- C/X-SAR, a joint mission of the German, Italian and United States space agencies, is part of NASA's Mission to Planet Earth.

11.3.4 Angle to Target

The strength of the radar return is very dependent on the angle of incidence of the radar beam on the surface. Basically, the more nearly perpendicular the surface is to the direction of the radar beam, the stronger the return. Since the radar beam spans an array of angles, the same material at different ranges, will reflect/scatter the radiation differently. For example, consider the example illustrated in FIGURE showing a single material with surfaces of various orientations viewed from several different viewing. At angle 1, surface A is nearly perpendicular

to the viewing directions and yields the brightest return, surface B is hidden, and surface C is at a very oblique angle and is therefore dark. As the angle changes, the brightness of the return of each of the three surfaces varies and the pattern is essentially a function of the viewing angle.

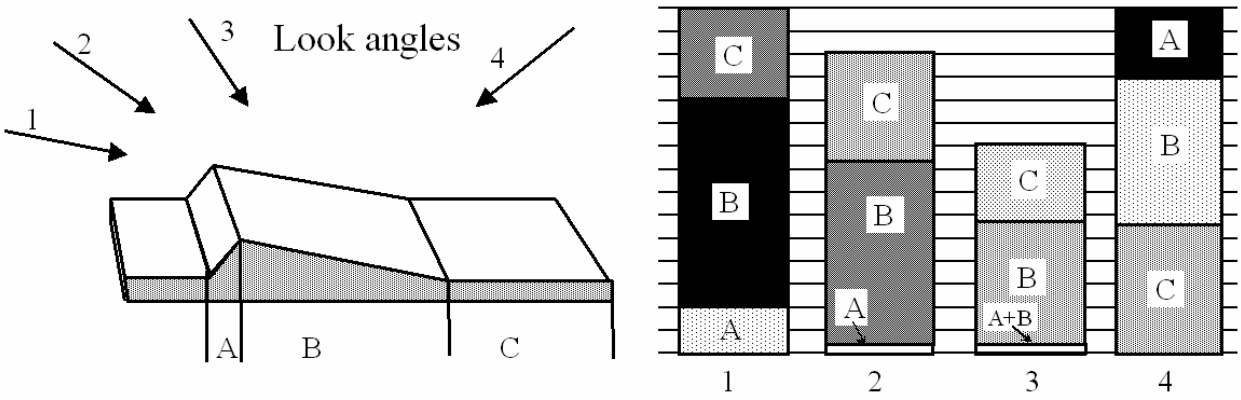


Figure 11.6: The brightness and relative size of features are dependent on the viewing angle.

11.3.5 Range to Target

Ranging – measuring the distance from the radar to a target – is at the heart of most radar systems. Range is determined by measuring the time of travel of the microwave pulse to the target and back. There are two non-imaging instruments that rely on ranging and direction finding: the Plan Position Indicator – one of the earliest forms of radar – and the radar altimeter.

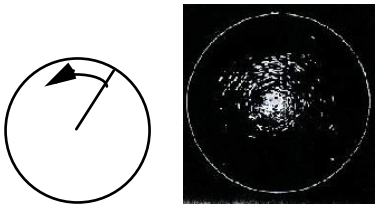


Figure 11.7: PPI display is essentially a map of the surrounding area with the antenn located at the center of the display.

The intensity of the radar return is often quite low – the intensity falls off as $1/r^4$ – and the detected signal will likely be corrupted by extraneous noise. Since each pulse typically has a duration (pulse width) of 10-50 microseconds, rather than looking for the echo directly in the received signal, the echo is sought in the correlation between a reference signal (a copy of the original transmitted signal) and the received signal. This *cross-correlation* will be large at the lag time associated with the time duration of the signal to the reflecting object and its return, and the random noise in the signal will be small when averaged over a time appropriate for the correlation.

If the radar transmits a very short pulse which is reflected (scattered) by an object at distance R from radar, total distance traveled by the pulse is 2R; the time, t, required for the"round trip" is then found from the relation: 2R = c * t, where c is the velocity of light. Thus the range of the target is given by:

$$R = c * t / 2 \qquad (11.12)$$

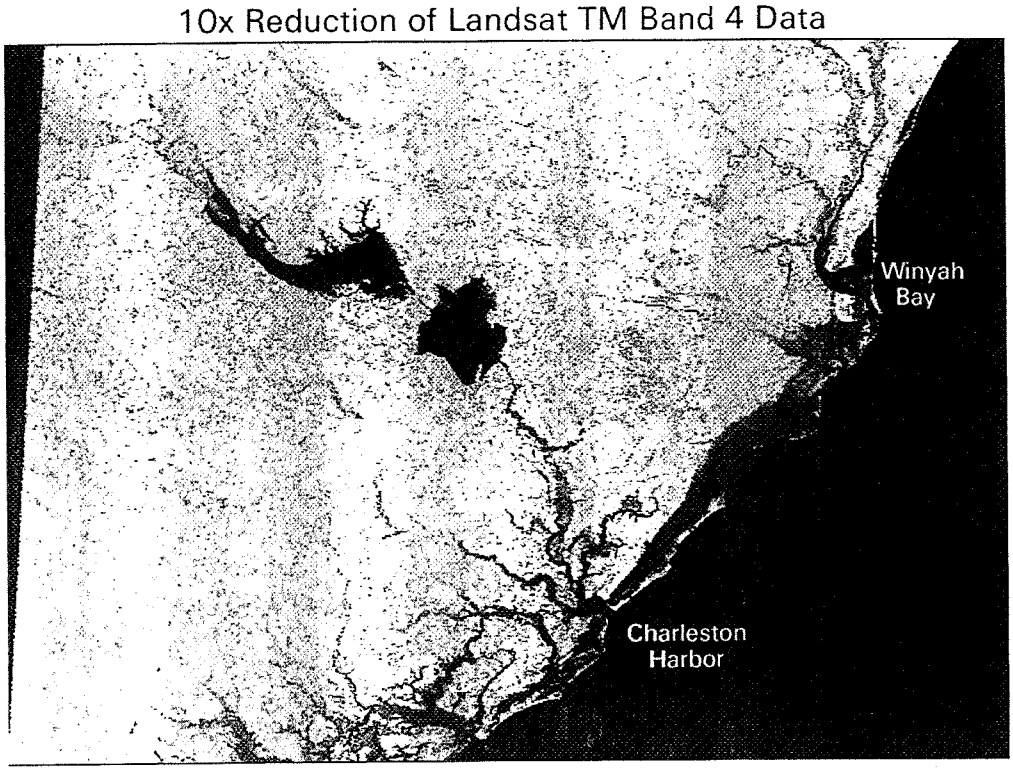


Figure 7-2 This 516-row by 696-column image represents only one one-hundredth of the data found in the 5160-row by 6960-column original Thematic Mapper image. It was created by sampling every tenth row and tenth column of the band 4 image.

Transects

The ability to extract brightness values along a user-specified *transect* between two points in an image is important in many remote sensing applications. For example, consider transects A, B, and C passed through the thermal plume in Figure 7-7b (color section). The brightness values encountered along each transect and reported in Table 7-1 were obtained only after the original image was geometrically rotated 16° clockwise so that the end points of each transect fell on the same scan line. This ensured that the number of meters in each temperature class along each transect was accurately measured. If the analyst extracts transects where the end points do not fall on the same scan line (or column), the hypotenuse of stair-stepped pixels must be considered instead of the simple horizontal pixel distance. This relationship is demonstrated in Figure 7-7c.

A histogram of transect B is shown in Figure 7-7d. The relationship between the original brightness values and the class intervals of the density-sliced map is provided. By counting the number of pixels along a transect in specific temperature class intervals within the plume and counting the total number of pixels of river (Table 7-1), it is possible to determine the proportion of the thermal plume falling within specific temperature class intervals (Jensen et al., 1983 and 1986). For example, in 1981 in South Carolina a thermal plume could not be >2.8° above river ambient temperature for more than one-third of the width of the river. Transect information extracted from thermal infrared imagery and summarized in Table 7-1 could be used to determine if the plume was in compliance.

Contrast Enhancement

Remote sensors record reflected and emitted radiant flux exiting from Earth's surface materials. Ideally, one material would reflect a tremendous amount of energy in a certain wavelength, while another material would reflect much less energy in the same wavelength. This would result in *contrast* between the two types of materials when recorded by a remote sensing system. Unfortunately, different materials

Table 7-1. Savannah River Thermal Plume Apparent Temperature Transects[a]

		Relationship of Class to Ambient River Temperature						
		Class 1 Dark blue Ambient	Class 2 Light blue +1°C	Class 3 Green 1.2°–2.8°C	Class 4 Yellow 3.0°–5.0°C	Class 5 Orange 5.2°–10°C	Class 6 Red 10.2°–20°C	Class 7 White >20°C
		Brightness Value Range for Each Class Interval (Refer to Table 5-3)						
Transect[b]	Average Width of River[c]	74–76	77–80	81–89	90–100	101–125	126–176	177–255
A	32 pixels = 89.6 m	15/42[d]	17/47.6	—	—	—	—	—
B	38 pixels = 106.4 m	25/70	1/2.8	1/2.8	2/5.6	1/2.8	5/14	3/8.4
C	34 pixels = 95.2 m	19/53.2	2/5.6	2/5.6	2/5.6	6/16.8	3/8.4	—

[a] Source: Jensen et al., 1983 and 1986.
[b] Each transect was approximately 285 m in length (66 pixels at 2.8 m/pixel). Transect measurements in the river were made only after the image was rotated so that the beginning and ending pixels of the transect fell on the same scan line.
[c] Includes one mixed pixel of land and water on each side of the river.
[d] Notation represents pixels and meters; for example, 15 pixels represent 42 m.

often reflect similar amounts of radiant flux throughout the visible, near-infrared, and mid-infrared portion of the electromagnetic spectrum, resulting in a relatively *low contrast* image. In addition, besides this obvious low contrast characteristic of biophysical materials, there are cultural factors at work. For example, people in developing countries often use natural building materials (e.g., wood and soil) in the construction of urban areas (Haack et al., 1995). This results in much lower contrast remotely sensed imagery for such areas than for urban areas in developed countries where concrete, asphalt, and fertilized green vegetation may be more prevalent. Thus, biophysical materials themselves are an important factor, and humans may further confuse the issue by bringing the materials together in diverse ways.

An additional factor in the creation of low-contrast remotely sensed imagery is the sensitivity of the detectors. For example, the detectors on most sensing systems are designed to record a relatively wide range of scene brightness values (e.g., 0 to 255) without becoming saturated. Saturation occurs if the radiometric sensitivity of a detector is insufficient to record the full range of intensities of reflected or emitted energy emanating from the scene. The Landsat TM detectors, for example, must be sensitive to reflectance from diverse biophysical materials such as dark volcanic basalt outcrops or snow (possibly represented as *BV*s of 0 and 255, respectively). However, very few scenes are composed of brightness values that utilize the full sensitivity range of the Landsat TM detectors. Therefore, this results in relatively low contrast imagery, with brightness values that usually range from 0 to 100.

To improve the contrast of digital remotely sensed data, it is desirable to utilize the entire brightness range of the display medium, which is generally a video CRT display or hard-copy output device (discussed in Chapter 5). Digital methods may be more satisfactory than photographic techniques for contrast enhancement because of the precision and wide variety of processes that can be applied to the imagery. There are linear and nonlinear digital contrast enhancement techniques.

Linear Contrast Enhancement

Contrast enhancement (referred to as a *contrast stretching*) expands the original input brightness values to make use of the total range or sensitivity of the output device. To illustrate the linear contrast stretch process, consider the Charleston, S.C., TM band 4 image produced by a sensor system whose image output levels can vary from 0 to 255 (Figure 7-8a). A histogram of this image is provided. We will assume that our output device (a high resolution black and white CRT) can display 256 shades of gray (i.e., $quant_k$ = 255). The histogram and associated statistics of this band 4 subimage reveal that the scene is composed of brightness val-

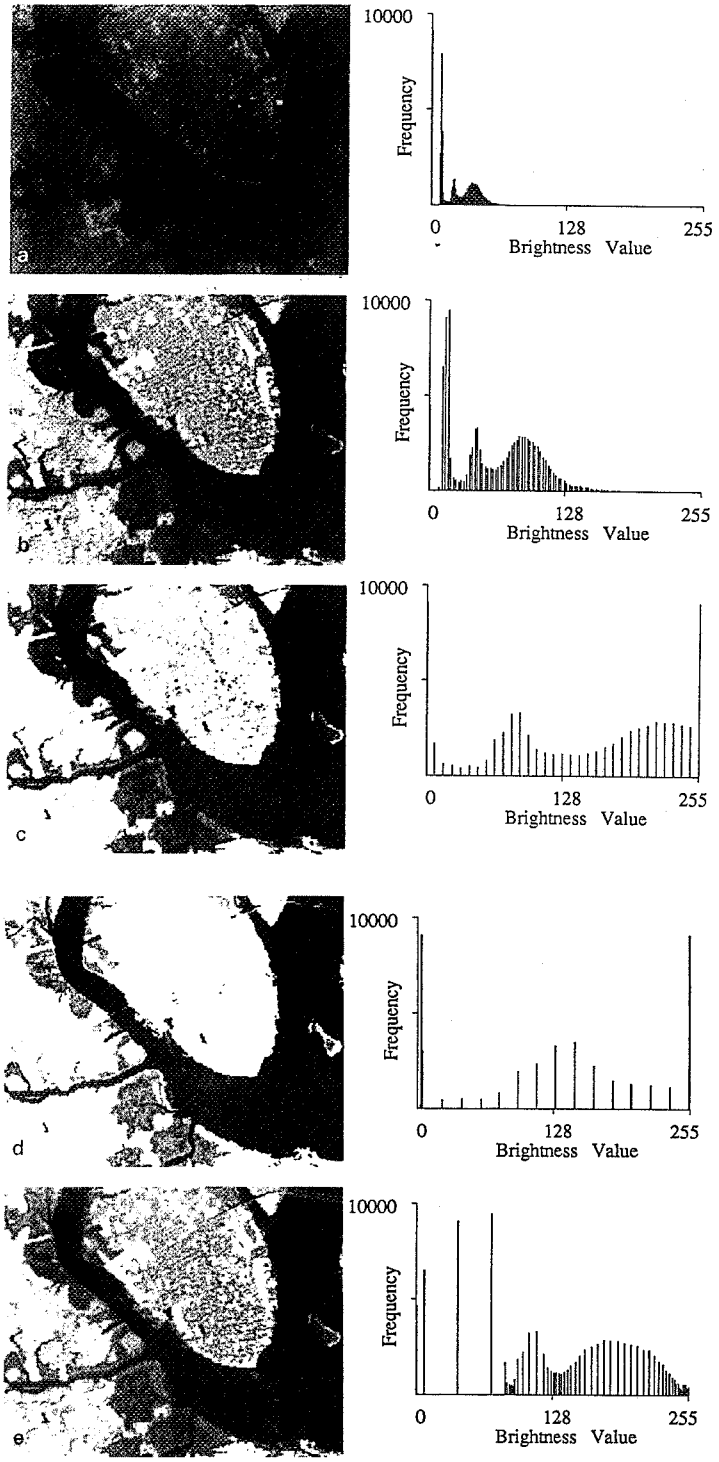


Figure 7-8 (a) Original Thematic Mapper band 4 data of Charleston, S.C., not contrast stretched, and its histogram. (b) Minimum–maximum contrast stretch applied to the data and the resultant histogram. (c) One standard deviation ($\pm 1\sigma$) percentage linear contrast stretch applied to the data and resultant histogram. (d) Specific percentage linear contrast stretch designed to highlight wetland and the resultant histogram. (e) Application of histogram equalization and resultant histogram.

ues ranging from a minimum of 4 (i.e., $\min_4 = 4$) to a maximum value of 105 (i.e., $\max_4 = 105$), with a mean of 27.3 and a standard deviation of 15.76 (Table 4-7). When these data are displayed on the CRT without any contrast enhancement, we use less than one-half of the full range of brightness values that could be displayed (i.e., brightness values between 0 and 3 and between 106 and 255 are not used). The image is rather dark, low in contrast, with no distinctive bright areas (Figure 7-8a). It is difficult to visually interpret such an image. A more useful display can be produced if we expand the range of original brightness values to use the full dynamic range of the video display.

Linear contrast enhancement is best applied to remotely sensed images with Gaussian or near-Gaussian histograms, that is, when all the brightness values fall generally within a single, relatively narrow range of the histogram and only one mode is apparent. Unfortunately, this is rarely the case especially for scenes that contain both land and water bodies. To perform a linear contrast enhancement, the analyst examines the image statistics and determines the minimum and maximum brightness values in the band, $\min_k$ and $\max_k$, respectively. The output brightness value BV_{out}, is computed according to the equation

$$BV_{out} = \left(\frac{BV_{in} - \min_k}{\max_k - \min_k}\right)\text{quant}_k \qquad (7\text{-}1)$$

where BV_{in} is the original input brightness value and quant_k is the range of the brightness values that can be displayed on the CRT (e.g., 255). In the Charleston, S.C., example, any pixel with a BV_{in} of 4 would now have a BV_{out} of 0, while any pixel with a BV_{in} of 105 would have a BV_{out} of 255. All original brightness values between 5 and 104 would be linearly distributed between 0 and 255, respectively. The application of this enhancement to the Charleston band 4 TM data is shown in Figure 7-8b. This is commonly referred to as a *min–max contrast stretch*. Most image processing systems provide for the display of a before and after histogram, as well as a graph of the relationship between the input brightness value (BV_{in}) and the output brightness value (BV_{out}). For example, the histogram of the min–max contrast stretch discussed is shown in Figure 7-8b. The logic of such a linear stretch is shown diagrammatically in Figure 7-9. Note the linear relationship between the brightness values of the input and output pixels and how the slope of the line becomes steeper as the minimum is increased or the maximum is decreased.

Image analysts often specify $\min_k$ and $\max_k$ that lie a certain percentage of pixels from the mean of the histogram. This is called a *percentage linear contrast stretch*. For example, consider setting the minimum and maximum ± 1 standard deviation ($\pm 1\sigma$) from the mean. For the Charleston TM band 4 data, the minimum becomes 12 and the maximum 43. All values between 12 and 43 are linearly contrast stretched to lie within the range 0 to 255. All values between 0 and 11 are now 0, and those between 44 and 255 are set equal to 255. This results in more pure black-and-white pixels in the Charleston scene, dramatically increasing the contrast of the image (Figure 7-8c). The histogram associated with this percentage contrast stretch is shown. The information content of the pixels that saturated at 0 and 255 is lost. The slope of a percentage linear contrast stretch is much greater than for a simple min–max stretch (refer to Figure 7-9).

The results of applying a min–max and $\pm 1\sigma$ percentage linear contrast stretch of the thermal data are shown in Figures 7-10 b and c along with representative histograms. The $\pm 1\sigma$ contrast stretch effectively "burns out" the thermal plume, yet provides more detail about the temperature characteristics of vegetation present on each side of the river.

It is not necessary that the same percentage be applied to each tail of the distribution. For example, what linear stretch would be appropriate if we were interested only in the wetland around Charleston Harbor? An analysis of the original histogram and quick assessment of some of the wetland band 4 brightness values reveal that most of the wetlands have brightness values from 13 to 27. In Figure 7-8d, the values from 0 to 12 are converted to 0 (black), 13 to 27 are linearly contrast stretched from 0 to 255, and values from 28 to 255 are sent to 255 (white). This enhancement yields additional information on the smooth cordgrass (*Spartina alterniflora*) at the expense of all the water and upland land cover. Similarly, the thermal plume is composed primarily of values from 81 to 170. A contrast stretch to highlight just the plume is found in Figure 7-10d, where values from 0 to 80 are sent to 0 (black), values from 81 to 170 are linearly contrast stretched to have values from 0 to 255, and all values from 171 to 255 now have a value of 255 (white). Similar logic may be applied to enhance specific features of interest in a scene.

When the histogram of an image is not Gaussian in nature (i.e., it is bimodal, trimodal, etc.), it is possible to perform a piecewise linear contrast stretch to the imagery of the type shown in Figure 7-11. Here the analyst identifies a number of linear enhancement steps that expand the brightness ranges in the modes of the histogram. In effect, this corresponds to setting up a series of $\min_k$ and $\max_k$ and using Equation (7-1) within user-selected regions of the histogram. This powerful contrast enhancement method should be used by per-

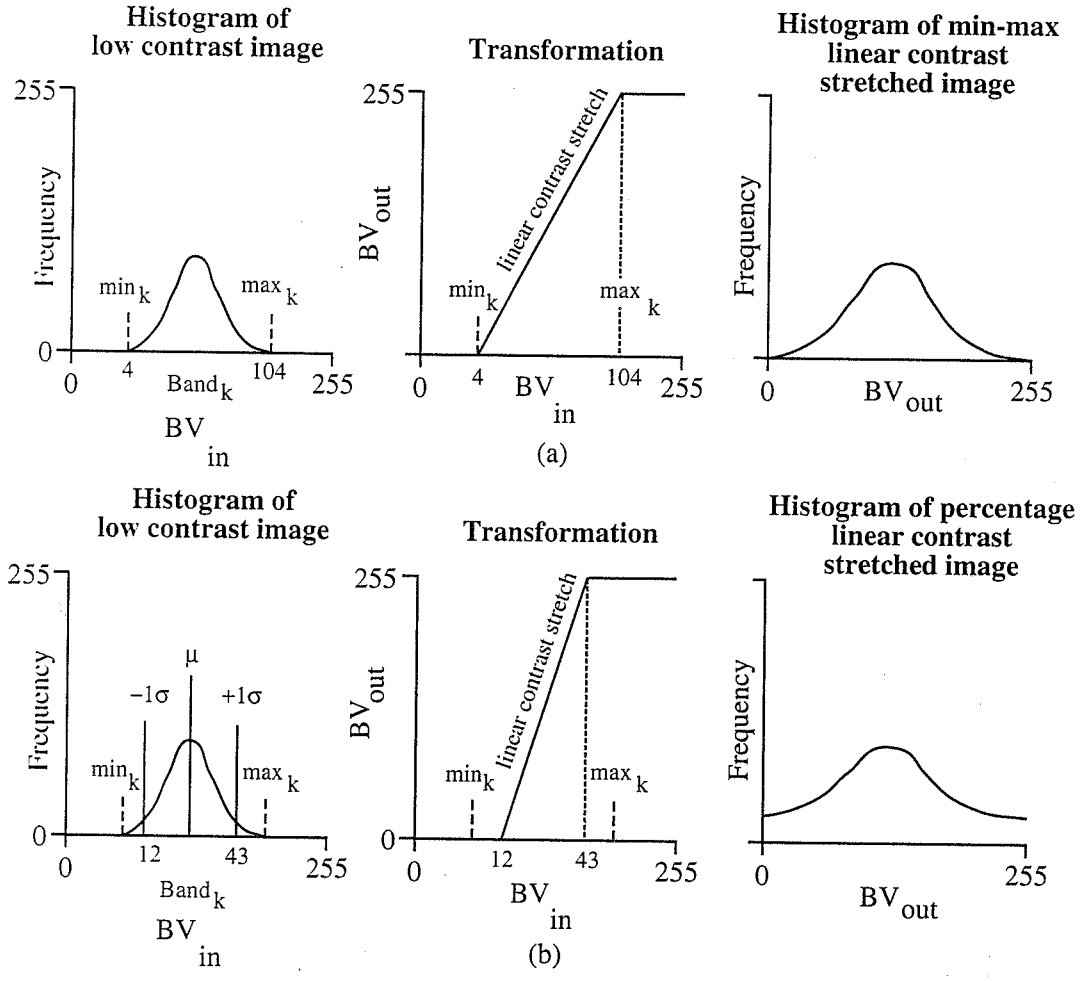


Figure 7-9 (a) Theoretical result of applying a *minimum–maximum* contrast stretch to normally distributed remotely sensed data. The histograms before and after the transformation are shown. The minimum and maximum brightness values encountered in band k are min_k and max_k, respectively. (b) Theoretical result of applying a ±1 standard deviation *percentage linear* contrast stretch. This moves the min_k and max_k values ±34% from the mean into the tails of the distribution.

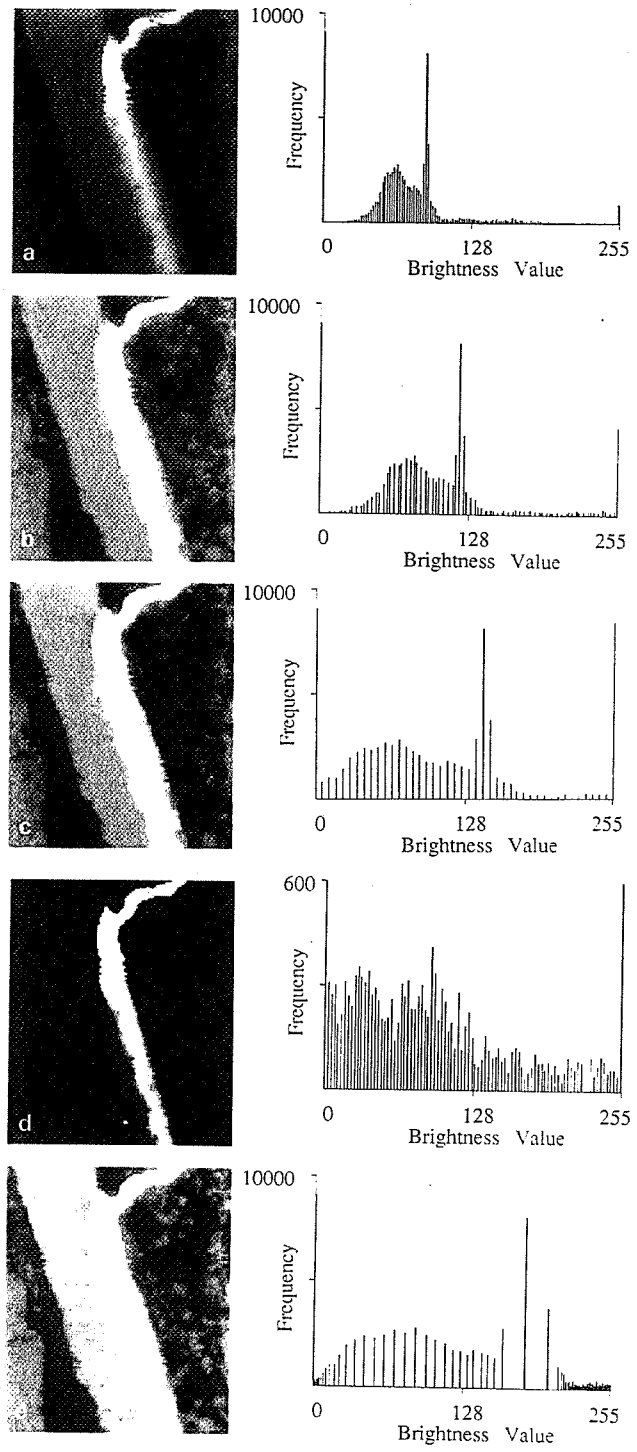


Figure 7-10 (a) Original predawn thermal-infrared data of the Savannah River, not contrast stretched, and its histogram. (b) Minimum–maximum contrast stretch applied to the data and the resultant histogram. (c) One standard deviation ($\pm 1\sigma$) percentage linear contrast stretch applied to the data and resultant histogram. (d) Specific percentage linear contrast stretch designed to highlight the thermal plume and the resultant histogram. (e) Application of histogram equalization and resultant histogram.

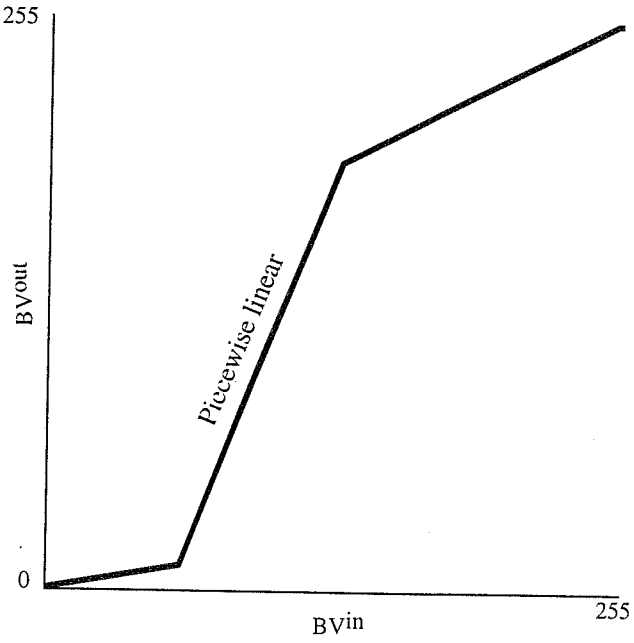


Figure 7-11 Logic of a piecewise linear contrast stretch for which selective pieces of the histogram are linearly contrast stretched. Notice that the slope of the linear contrast enhancement changes.

sons intimately familiar with the modes of the histogram and what they represent in the real world. Such contrast-stretched data are rarely used in subsequent image classification.

Nonlinear Contrast Enhancement

Nonlinear contrast enhancements may also be applied. One of the most useful is *histogram equalization*. The algorithm passes through the individual bands of the dataset and assigns approximately an equal number of pixels to each of the user-specified output gray-scale classes (e.g., 32, 64, and 256). Histogram equalization applies the greatest contrast enhancement to the most populated range of brightness values in the image. It automatically reduces the contrast in the very light or dark parts of the image associated with the tails of a normally distributed histogram.

Histogram equalization is found on many image processing systems because it requires very little information from the analyst to implement (usually just the number of output brightness value classes and the bands to be equalized), yet it is often very effective. Because of its wide availability, it is instructive to review how the equalization takes place using a hypothetical dataset (modified from Gonzalez and Wintz, 1977). For example, consider an image that is composed of just 64 rows and 64 columns (4096 pixels) with the range of brightness values that each pixel can assume, $quant_k$, limited to just 0 to 7 (Table 7-2). A histogram of this hypothetical image is shown in Figure 7-12a, and the frequency of occurrence of the individual brightness values, $f(BV_i)$, is summarized in Table 7-2. For example, there are 790 pixels in the scene with a brightness value of 0 [i.e., $f(BV_0) = 790$] and 1023 pixels with a brightness value of 1 [i.e. $f(BV_1) = 1023$]. We can compute the probability of the ith brightness value, p_i, by dividing each of the frequencies, $f(BV_i)$, by the total number of pixels in the scene (i.e., $n = 4096$). Thus, the probability of encountering a pixel with a brightness value of 0 in the scene is approximately 19% [i.e., $p_0 = f(BV_i)/n = 790/4096 = 0.19$]. A plot of the probability of occurrence of each of the eight brightness values for the hypothetical scene is shown in Figure 7-12b. This particular histogram has a large number of pixels with low brightness values (0 and 1), making it a relatively low contrast scene.

Table 7-2. Statistics for a 64 × 64 Hypothetical Image with Brightness Values from 0 to 7[a]

Brightness Value, BV_i	L_i	Frequency $f(BV_i)$	Probability[b] $p_i = f(BV_i)/n$
BV_0	0/7 = 0.00	790	0.19
BV_1	1/7 = 0.14	1023	0.25
BV_2	2/7 = 0.28	850	0.21
BV_3	3/7 = 0.42	656	0.16
BV_4	4/7 = 0.57	329	0.08
BV_5	5/7 = 0.71	245	0.06
BV_6	6/7 = 0.85	122	0.03
BV_7	7/7 = 1.00	81	0.02

[a] Source: modified from Gonzalez and Wintz, 1977.
[b] $n = 4096$ pixels.

The next step is to compute a transformation function k_i for each individual brightness value. One way of conceptualizing the histogram equalization process is to use the notation shown in Table 7-3. For each brightness value level BV_i in the $quant_k$ range of 0 to 7 of the original histogram, a new cumulative frequency value k_i is calculated:

$$k_i = \sum_{i=0}^{quant_k} \frac{f(BV_i)}{n} \qquad (7\text{-}2)$$

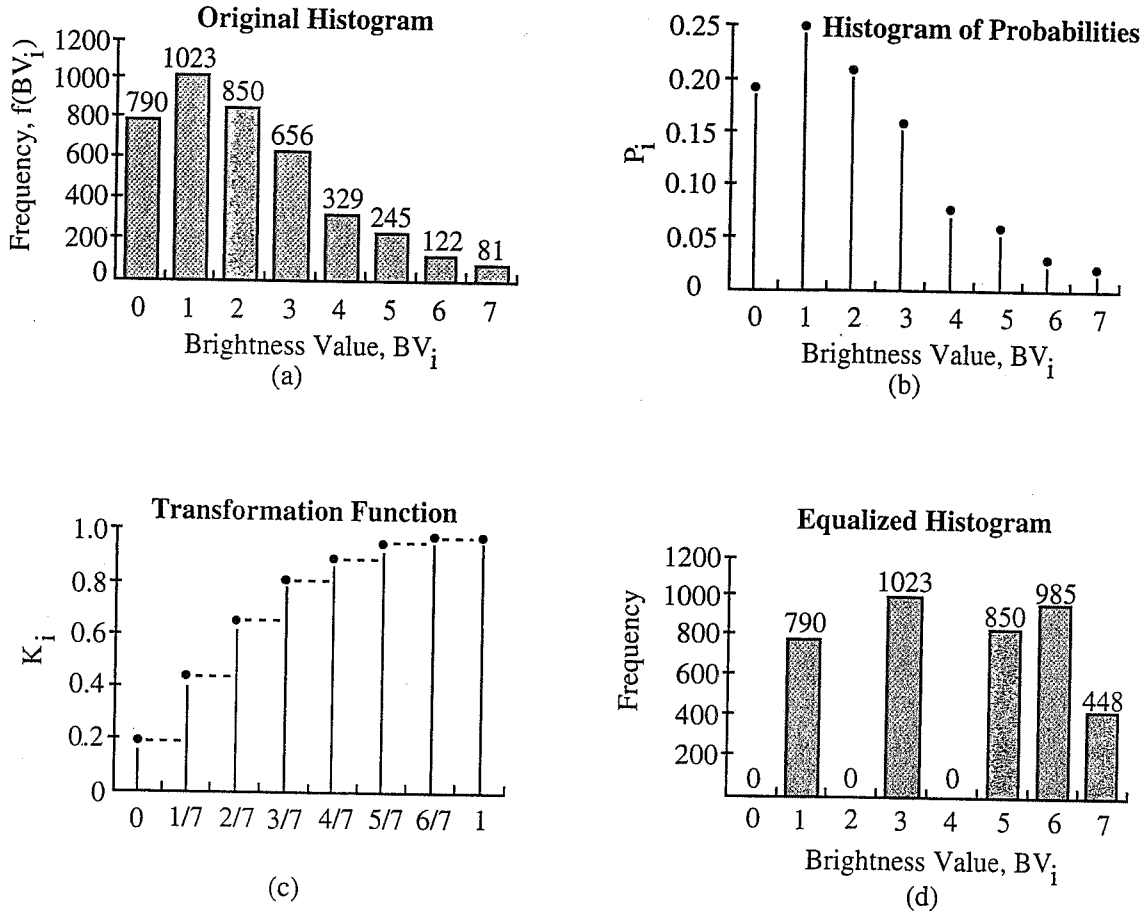


Figure 7-12 Histogram equalization process applied to hypothetical data (adapted from Gonzalez and Wintz, 1977). (a) Original histogram showing the frequency of pixels in each brightness value. (b) Original histogram expressed in probabilities. (c) The transformation function. (d) The equalized histogram showing the frequency of pixels in each brightness value.

where the summation counts the frequency of pixels in the image with brightness values equal to or less than BV_i, and n is the total number of pixels in the entire scene (4096 in this example). The histogram equalization process iteratively compares the transformation function k_i with the original values of L_i, to determine which are closest in value. The closest match is reassigned to the appropriate brightness value. For example, in Table 7-3 we see that $k_0 = 0.19$ is closest to $L_1 = 0.14$. Therefore, all pixels in BV_0 (790 of them) will be assigned to BV_1. Similarly, the 1023 pixels in BV_1 will be assigned to BV_3, the 850 pixels in BV_2 will be assigned to BV_5, the 656 pixels in BV_3 will be assigned to BV_6, the 329 pixels in BV_4 will also be assigned to BV_6, and all 448 brightness values in BV_{5-7} will be assigned to BV_7. The new image will not have any pixels with brightness values of 0, 2, or 4. This is evident when evaluating the new histogram (Figure 7-12d). When analysts see such gaps in image histograms, it is usually a good indication that histogram equalization or some other operation has been applied.

Histogram-equalized versions of the Charleston TM band 4 data and the thermal plume data are found in Figure 7-8e and 7-10e, respectively. Histogram equalization is dramatically different from any other contrast enhancement because the data have been redistributed according to the cumulative frequency histogram of the data, as described. Note that after histogram equalization, some pixels that originally had different values are now assigned the same value (perhaps a loss of information), while other values that were once very close together are now spread out, increasing the contrast between them. Therefore, while this enhancement may improve the visibility of detail in an image, it also alters the relationship between brightness values and image structure (Russ, 1992). For these reasons, it is not wise to extract texture or biophys-

Table 7-3. Example of How a Hypothetical 64 × 64 Image with Brightness Values from 0 to 7 is Histogram Equalized

Frequency, $f(BV_i)$	790	1023	850	656	329	245	122	81
Original brightness value, BV_i	0	1	2	3	4	5	6	7
$L_i = \frac{\text{brightness value}}{n}$	0	0.14	0.28	0.42	0.57	0.71	0.85	1.0
Cumulative frequency transformation: $k_i = \sum_{i=0}^{\text{quant}_k} \frac{f(BV_i)}{n}$	$\frac{790}{4096}$ =0.19	$\frac{1813}{4096}$ = 0.44	$\frac{2663}{4096}$ = 0.65	$\frac{3319}{4096}$ = 0.81	$\frac{3648}{4096}$ = 0.89	$\frac{3893}{4096}$ = 0.95	$\frac{4015}{4096}$ = 0.98	$\frac{4096}{4096}$ = 1.0
Assign original BV_i class to the new class it is closest to in value	1	3	5	6	6	7	7	7

ical information from imagery that has been histogram equalized.

Another type of nonlinear contrast stretch involves scaling the input data *logarithmically* as diagrammed in Figure 7-13. This enhancement has the greatest impact on the brightness values found in the darker part of the histogram. It could be reversed to enhance values in the brighter part of the histogram by scaling the input data using an inverse log function as shown.

The selection of a contrast enhancement algorithm depends on the nature of the original histogram and the elements of the scene that are of greatest interest to the user. An experienced image analyst can usually identify an appropriate contrast enhancement algorithm by examining the image histogram and then experimenting until satisfactory results are obtained. Most contrast enhancements cause some useful information to be lost. However, that which remains should be of value. Contrast enhancement is applied primarily to improve visual image analysis. It is *not* good practice to contrast stretch the original imagery and then use the enhanced imagery for computer-assisted classification, change detection, and the like. Contrast stretching can distort the original pixel values, often in a nonlinear fashion.

Band Ratioing

Sometimes differences in brightness values from identical surface materials are caused by topographic slope and aspect, shadows, or seasonal changes in sunlight illumination angle and intensity. These conditions may hamper the ability of an interpreter or classification algorithm to identify correctly surface materials or land use in a remotely sensed image. Fortunately, *ratio* transformations of the remotely sensed data can, in certain instances, be applied to reduce the effects of such environmental conditions (Avery and Berlin, 1992). In addition to minimizing the effects of environmental factors, ratios may also provide unique information not available in any single band that is useful for discriminating between soils and vegetation (Satterwhite, 1984).

The mathematical expression of the ratio function is

$$BV_{i,j,r} = \frac{BV_{i,j,k}}{BV_{i,j,l}} \tag{7-3}$$

where $BV_{i,j,r}$ is the output ratio value for the pixel at row i, column j; $BV_{i,j,k}$ is the brightness value at the same location in band k, and $BV_{i,j,l}$ is the brightness value in band l. Unfortunately, the computation is not always simple since $BV_{i,j} = 0$ is possible. However, there are alternatives. For example, the mathematical domain of the function is 1/255 to 255 (i.e., the range of the ratio function includes all values beginning at 1/255, passing through 0 and ending at 255). The way to overcome this problem is simply to give any $BV_{i,j}$ with a value of 0 the value of 1. Alternatively, some like to add a small value (e.g., 0.1) to the denominator if it equals zero.

To represent the range of the function in a linear fashion and to encode the ratio values in a standard 8-bit format (values from 0 to 255), normalizing functions are applied. Using this

a vector of brightness values, which might be the blue, green and red components of the pixel in a colour scene or, for a remote sensing multispectral image, may be the various spectral response components for the pixel. Most image enhancement techniques relate to scalar images and also to the scalar components of vector imagery. Such is the case with all techniques given in this chapter. Enhancement methods that relate particularly to vector imagery tend to be transformation oriented. Those are treated in Chap. 6.

4.2 The Image Histogram

Consider a spatially quantised scalar image such as that corresponding to one of the Landsat thematic mapper bands; in this case the brightness values are also quantised. If each pixel in the image is examined and its brightness value noted, a graph of number of pixels with a given brightness versus brightness value can be constructed. This is referred to as the histogram of the image. The tonal or radiometric quality of an image can be assessed from its histogram as illustrated in Fig. 4.1. An image which makes good use of the available range of brightness values has a histogram with occupied bins (or bars) over its full range, but without significantly large bars at black or white.

An image has a unique histogram but the reverse is not true in general since a histogram contains only radiometric and no spatial information. A point of some importance is that the histogram can be viewed as a discrete probability distribution since the relative height of a particular bar, normalised by the total number of pixels in the image segment, indicates the chance of finding a pixel with that particular brightness value somewhere in the image.

4.3 Contrast Modification in Image Data

4.3.1 Histogram Modification Rule

Suppose one has available a digital image with poor contrast, such as that in Fig. 4.1a, and it is desired to improve its contrast to obtain an image with a histogram that has a good spread of bars over the available brightness range, resembling that in Fig. 4.1c. In other words, a so-called contrast stretching of the image data is required. Often the degree of stretching desired is apparent. For example the original histogram may occupy brightness values between 40 and 75 and we might wish to expand this range to the maximum possible, say 0 to 255. Even though the modification is somewhat obvious it is necessary to express it in mathematical terms in order to relegate it to a computer. Contrast modification is a mapping of brightness values, in that the

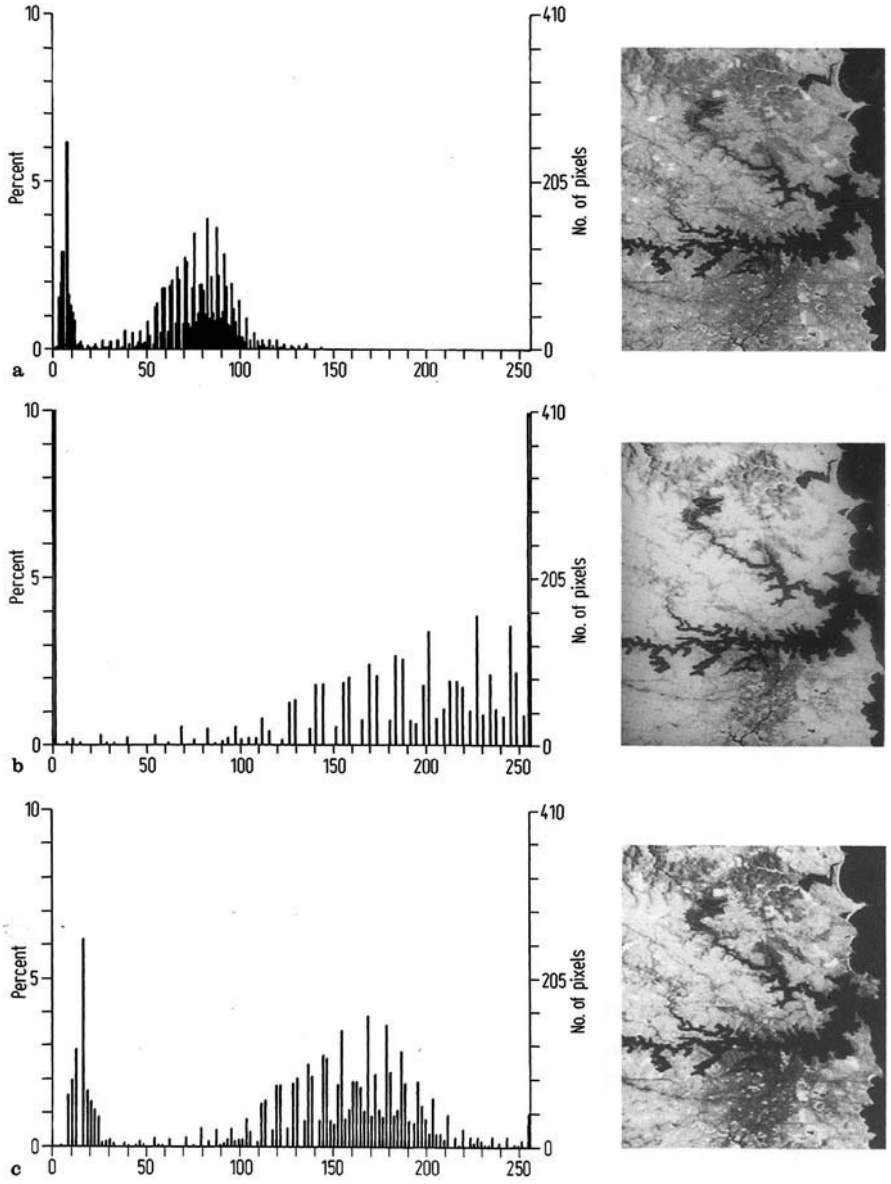


Fig. 4.1. Examples of image histograms. The image in **a** shows poor contrast since its histogram utilizes a restricted range of brightness value. The image in **b** is very contrasty with saturation in the black and white regions resulting in some loss of discrimination of bright and dull features. The image in **c** makes optimum use of the available brightness levels and shows good contrast. Its histogram shows a good spread of bars but without the large bars at black and white indicative of the saturation in image **b**

brightness value of a particular histogram bar is respecified more favourably. The bars themselves though are not altered in size, although in some cases some bars may be mapped to the same new brightness value and will be superimposed. In general, however, the new histogram will have the same number of bars as the old. They will simply be at different locations.

The mapping of brightness values associated with contrast modification can be described as

$$y = f(x) \tag{4.1}$$

where x is the old brightness value of a particular bar in the histogram and y is the corresponding new brightness value.

In principle, what we want to do in contrast modification is find the form of $f(x)$ that will implement the desired changes in pixel brightness and thus in the perceived contrast of the image. Sometimes that is quite simple; on other occasions $f(x)$ might be quite a complicated function. In the following sections we look at simple contrast changes first.

4.3.2 Linear Contrast Modification

The most common contrast modification operation is that in which the new (y) and old (x) brightness values of the pixels in an image are related in a linear fashion, i.e. so that (4.1) can be expressed

$$y = f(x) = ax + b\,.$$

A simple numerical example of linear contrast modification is shown in Fig. 4.2, whereas a poorly contrasting image that has been radiometrically enhanced by linear contrast stretching is shown in Fig. 4.3.

The look-up table for the particular linear stretch in Fig. 4.2 has been included in the figure. In practice this would be used by a computer routine to produce the new image. This is done by reading the brightness values of the original version, pixel by pixel, substituting these into the left hand side of the table and then reading the new brightness value for a pixel from the corresponding entry on the right hand side of the table. It is important to note in digital image handling that the new brightness values, just as the old, most be discrete, and cover usually the same range of brightnesses. Generally this will require some rounding to integer form of the new brightness values calculated from the mapping function $y = f(x)$. A further point to note in the example of Fig. 4.2 is that the look-up table is undefined outside the range 2 to 4 of inputs. To do so would generate output brightness values that are outside the range valid for this example. In practice, linear contrast stretching is generally implemented as the saturating linear contrast enhancement technique in Sect. 4.3.3 following.

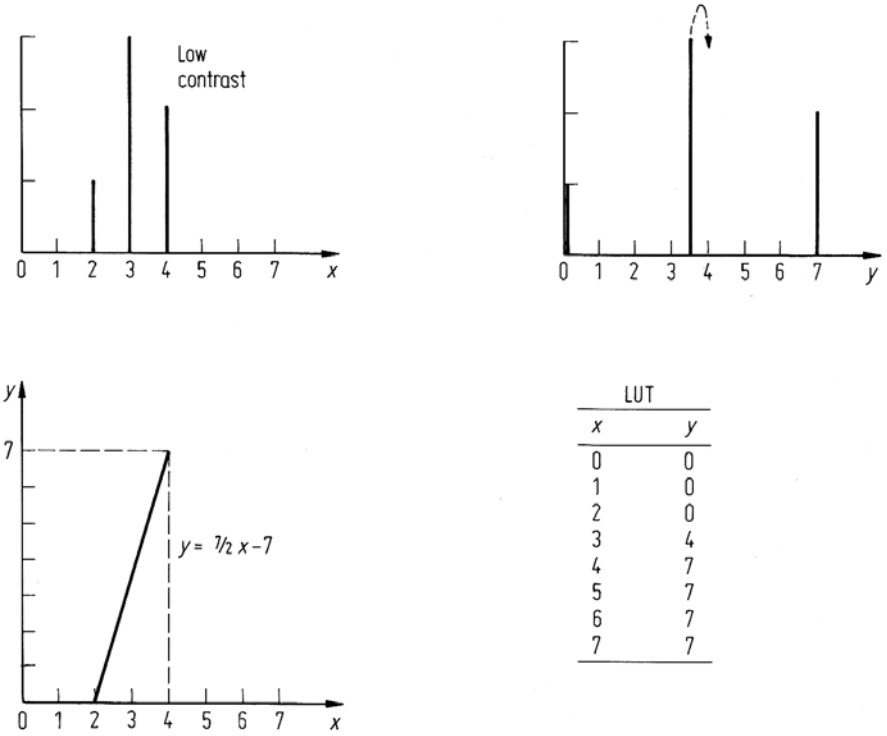


Fig. 4.2. Simple numerical example of linear contrast modification. The available range of discrete brightness values is 0 to 7. Note that a non-integral output brightness value might be indicated. In practice this is rounded to the nearest integer

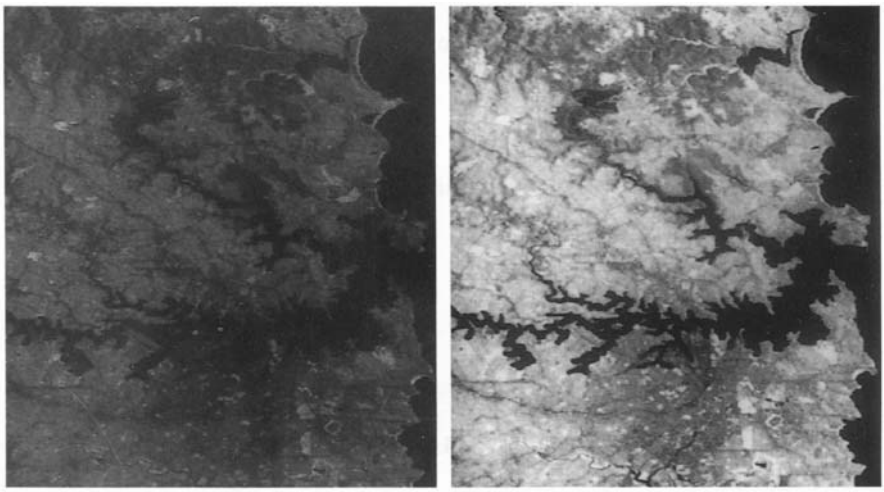


Fig. 4.3. Linear contrast modification of the image in **a** to produce the visually better product in **b**

4.3.3
Saturating Linear Contrast Enhancement

Frequently a better image product is given when linear contrast enhancement is used to give some degree of saturation at the black and white ends of the histogram. Such is the case, for example, if the darker regions in an image correspond to the same ground cover type within which small radiometric variations are of no interest. Similarly, a particular region of interest in an image may occupy a restricted brightness value range; saturating linear contrast enhancement is then employed to expand that range to the maximum possible dynamic range of the display device with all other regions being mapped to either black or white. The brightness value mapping function $y = f(x)$ for saturating linear contrast enhancement is shown in Fig. 4.4, in which B_{max} and B_{min} are the user-determined maximum and minimum brightness values that are to be expanded to the lowest and highest brightness levels supported by the display device.

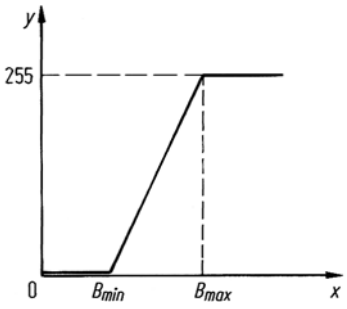


Fig. 4.4. Saturating linear contrast mapping

4.3.4
Automatic Contrast Enhancement

Most remote sensing image data is too low in brightness and poor in contrast to give an acceptable image product if displayed directly in raw form. This is a result of the need to have the dynamic range of satellite and aircraft sensors so adjusted that a variety of cover types over many images can be detected without leading to saturation of the detectors or without useful signals being lost in noise. As a consequence a single typical image will contain a restricted set of brightnesses.

Image display systems frequently implement an automatic contrast stretch on the raw data in order to give a product with good contrast.

Typically the automatic enhancement procedure is a saturating linear stretch. The cut-off and saturation limits B_{min} and B_{max} are chosen by determining the mean brightness of the raw data and its standard deviation and then making B_{min} equal to the mean less three standard deviations and B_{max} equal to the mean plus three standard deviations.

4.3.5
Logarithmic and Exponential Contrast Enhancement

Logarithmic and exponential mappings of brightness values between original and modified images are useful for enhancing dark and light features respectively. The mapping functions are depicted in Fig. 4.5, along with their mathematical expressions. It is particularly important with these that the output values be scaled to lie within the range of the device used to display the product (or the range appropriate to files used for storage in a computer memory) and that the output values be rounded to allowed, discrete values.

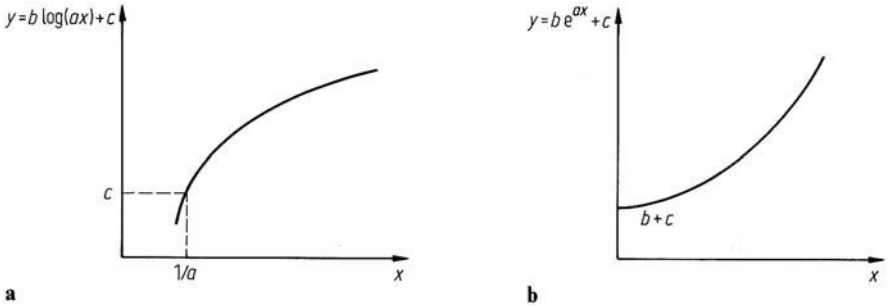


Fig. 4.5. Logarithmic **a** and exponential **b** brightness mapping functions. The parameters a, b and c are usually included to adjust the overall brightness and contrast of the output product

4.3.6
Piecewise Linear Contrast Modification

A particularly useful and flexible contrast modification procedure is the piecewise linear mapping function shown in Fig. 4.6. This is characterised by a set of user specified break points as shown. Generally the user can also specify the number of

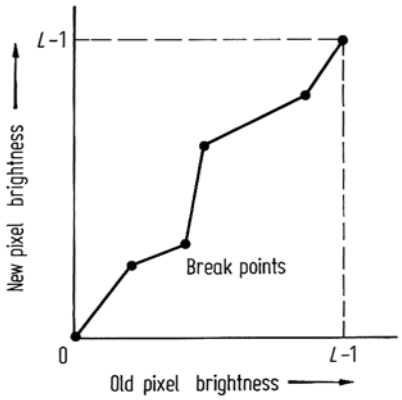


Fig. 4.6. Piecewise linear contrast modification function, characterised by the break points shown. These are user specified (as new, old pairs). It is clearly important that the function commence at 0,0 and finish at $L-1$, $L-1$ as shown, where L is the total number of brightness levels

break points. This method has particular value in implementing some of the contrast matching procedures in Sects. 4.4 and 4.5 following.

It should be noted that this is a more general version of the saturating linear contrast stretch of Sect. 4.3.3.

4.4 Histogram Equalization

4.4.1 Use of the Cumulative Histogram

The foregoing sections have addressed the task of simple expansion (or contraction) of the histogram of an image. In many situations however it is desirable to modify the contrast of an image so that its histogram matches a preconceived shape, other than a simple closed form mathematical modification of the original version. A particular and important modified shape is the uniform histogram in which, in principle, each bar has the same height. Such a histogram has associated with it an image that utilises the available brightness levels equally and thus should give a display in which there is good representation of detail at all brightness values. In practice a perfectly uniform histogram cannot be achieved for digital image data; the procedure following however produces a histogram that is quasi-uniform on the average. The method of producing a uniform histogram is known generally as histogram equilization.

It is useful, in developing the actual methods to be used for histogram equalisation, if we regard the histograms as continuous curves as depicted in Fig. 4.7, adapted from Castleman (1996). In this $h_i(x)$ represents the original image histogram (the "input" to the modification process) and $h_o(y)$ represents the histogram of the image after it has had its contrast modified (the "output" from the modification process).

In Fig. 4.7 the number of pixels represented by the range y to $y + \delta y$ in the modified histogram must, by definition in the diagram, be equal to the number of pixels represented in the range x to $x + \delta x$ in the original histogram. Given that $h_i(x)$ and $h_o(y)$ are strictly density functions, this implies

$$h_i(x)\delta x = h_o(y)\delta y$$

so that in the limit as $\delta x,\ \delta y \to 0$, using simple calculus

$$h_o(y) = h_i(x)\frac{\mathrm{d}x}{\mathrm{d}y} \tag{4.2}$$

We can use this last expression in two ways. First, if we know the original (input) histogram – which is usually always the case – and the function $y = f(x)$, we can determine the resulting (output) histogram. Alternatively, if we know the original histogram, and the shape of the output histogram we want – e.g. "flat" in the case of contrast equalisation – then we can use (4.2) to help us find the $y = f(x)$ that will generate that result. Our interest here is in the second approach.

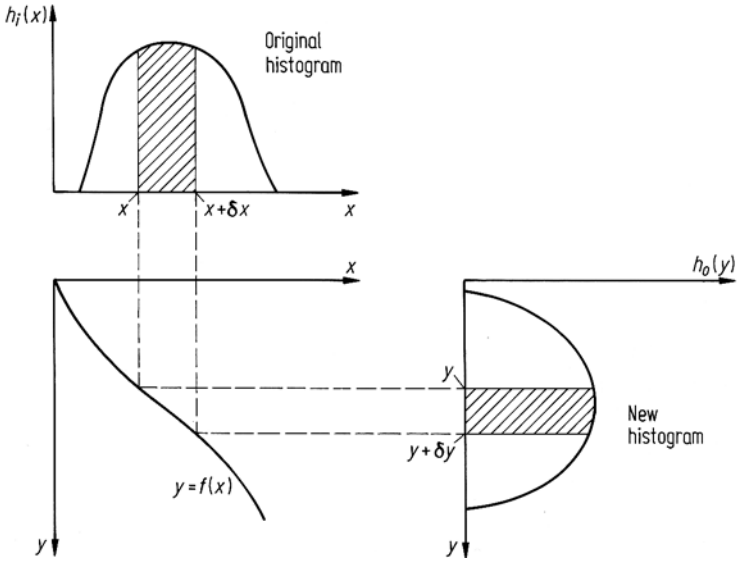


Fig. 4.7. Diagrammatic representation of contrast modification by the brightness value mapping function $y = f(x)$

Note that if $y = f(x)$, and thus $x = f^{-1}(y)$, (4.2) can be expressed

$$h_o(y) = h_i(f^{-1}(y))\frac{df^{-1}(y)}{dy}$$

which is a mathematical expression for the modified histogram[1,2].

To develop the brightness value modification procedure for contrast equalisation it is convenient to re-express (4.2) as

$$\frac{dy}{dx} = \frac{h_i(x)}{h_o(y)}$$

For a uniform histogram $h_o(y)$ and thus $1/h_o(y)$ should be constant – i.e. independent of y. This is a mathematical idealisation for real data, and rarely will we achieve a totally flat modified histogram, as the examples in the following will show. However,

[1] This requires the inverse $x = f^{-1}(y)$ to exist. For the contrast modification procedures used in remote sensing that is generally the case. Should an inverse not exist – for example if $y = f(x)$ is not monotonic – Castleman (1996) recommends treating the original brightness value range x as a set of contiguous sub-ranges within each of which $y = f(x)$ is monotonic.

[2] If we apply this expression to the brightness value modification function for linear contrast enhancement we have $y = ax + b$, giving $x = \frac{y-b}{a}$ so that $h_o(y) = \frac{1}{a}h\left(\frac{y-b}{a}\right)$. Relative to the original histogram, the modified version is shifted because of the effect of b, is spread or compressed depending on whether a is greater or less than 1 and is modified in amplitude. The last effect only relates to the continuous function and cannot happen with discrete brightness value data.

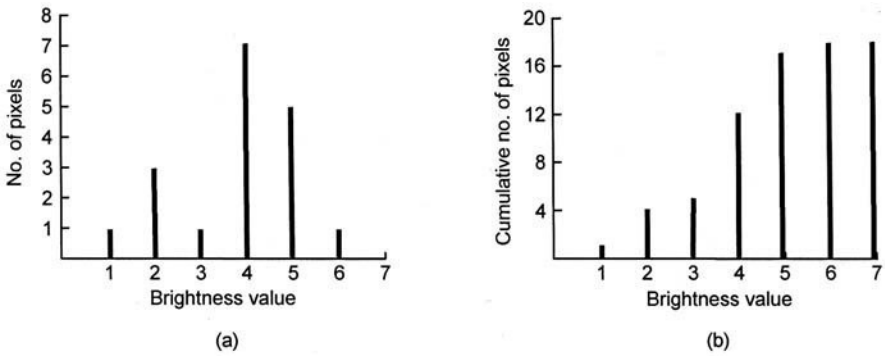


Fig. 4.8. a simple histogram and **b** the corresponding cumulative histogram

making this assumption mathematically will generate for us the process we need to adopt to equalise image histograms. With this we can write the last expression as

$$\frac{dy}{dx} = \text{constant}\, h_i(x)$$

so that

$$dy = \text{constant}\, h_i(x)dx$$

giving by integration

$$y = \text{constant} \int h_i(x)dx \ .$$

How should we interpret the integral on the right hand side of this last expression? In effect it is the continuous version of a cumulative histogram which, in discrete form, is a graph of the number of pixels below a given brightness value as a function of brightness value as illustrated in Fig. 4.8. The cumulative histogram is computed by summing the bars of the ordinary histogram from left to right.

If we call the cumulative histogram $C(x)$, then

$$y = \text{constant}\, C(x)$$

is the brightness value modification formula for histogram (contrast) equalisation. How do we find the value of the "constant"? We note first that the range of values of y is required to be 0 to $L-1$ to match the L brightness values available in the image. Secondly, note that the maximum value of $C(x)$ is N, the total number of pixels in the image, as seen in Fig. 4.8. Thus the constant needs to be $(L-1)/N$ in order to generate the correct range for y. In summary, the brightness value mapping function that gives contrast equalisation is

$$y = \frac{L-1}{N} C(x) \ . \tag{4.3}$$

where $C(x)$ is the discrete cumulative histogram.

Equation (4.3) is, in effect, a look-up table that can be used to move histogram bars to new brightness value locations. To illustrate the concept, consider the need to

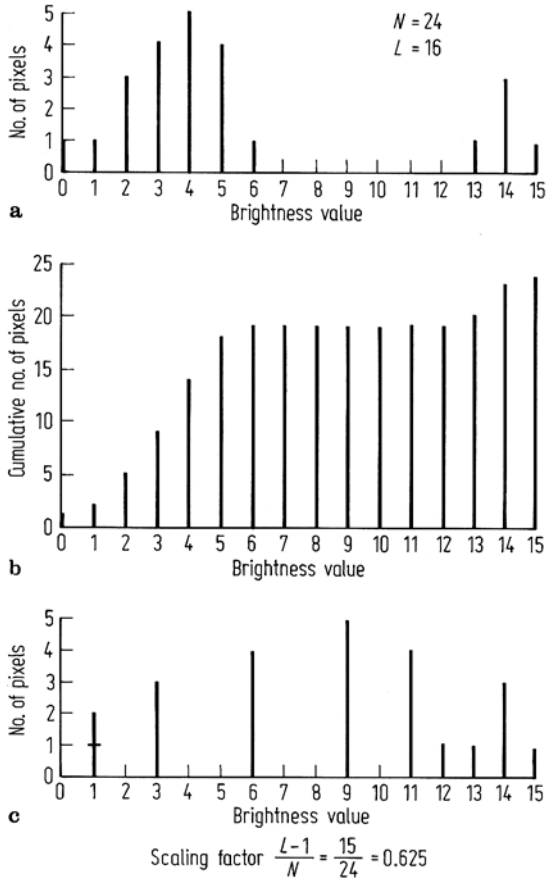
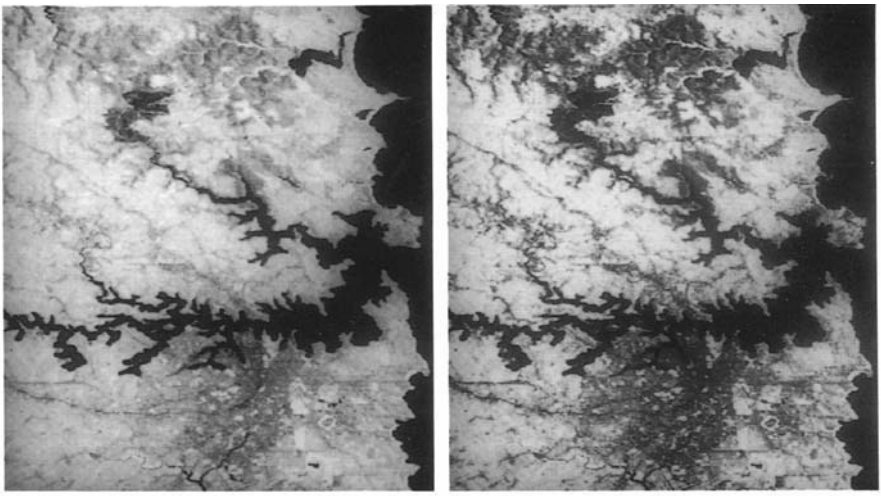


Fig. 4.9. Example of histogram equalisation. **a** Original histogram; **b** Cumulative histogram used to produce the look up table in Table 4.1; **c** The resulting quasi-uniform histogram

"flatten" the simple histogram shown in Fig. 4.9a. This corresponds to a hypothetical image with 24 pixels, each of which can take on one of 16 possible brightness values. The corresponding cumulative histogram is shown in Fig. 4.9b, and the scaling factor in (4.3) is $(L-1)/N = 15/24 = 0.625$. Using (4.3) the new brightness value location of a histogram bar is given by finding its original location on the abcissa of the cumulative histogram (x) and then reading its unscaled new location (y) from the ordinate. Multiplication by the scaling factor then produces the required new value. It is likely, however, that this may not be one of the discrete brightness values available (for the output display device) in which case the associated bar is moved to the nearest available brightness value. This procedure is summarised, for the example at hand, in Table 4.1, and the new, quasi-uniform histogram is given in Fig. 4.9c. It is important to emphasise that additional brightness values cannot be created nor can pixels from a single brightness value in an original histogram be distributed over several brightness values in the modified version. All that can be done is to re-map the brightness values to give a histogram that is as uniform as possible. Sometimes

Table 4.1. Look up table generation for histogram equalization example

Original brightness value	Unscaled new value	Scaled new value	Nearest available brightness value
0	1	0.63	1
1	2	1.25	1
2	5	3.13	3
3	9	5.63	6
4	14	8.75	9
5	18	11.25	11
6	19	11.88	12
7	19	11.88	12
8	19	11.88	12
9	19	11.88	12
10	19	11.88	12
11	19	11.88	12
12	19	11.88	12
13	20	12.50	13
14	23	14.40	14
15	24	15.00	15



a b

Fig. 4.10. Image with linear contrast stretch **a** compared with the same image enhanced with a stretch from histogram equalization **b**

this entails some bars from the original histogram being moved to the same new location and thereby being superimposed, as is observed in the example.

In practice, the look up table created in Table 4.1 would be applied to every pixel in the image by feeding into the table the original brightness value for the pixel and reading from the table the new brightness value.

Figure 4.10 shows an example of an image with a simple linear contrast modification compared to the same image but in which contrast modification by histogram

Table 4.2. Look up table for histogram equalization using 8 output brightnesses from 16 input brightnesses

Original brightness value	Unscaled new value	Scaled new value	Nearest available brightness value
0	1	0.29	0
1	2	0.58	1
2	5	1.46	1
3	9	2.63	3
4	14	4.08	4
5	18	5.25	5
6	19	5.54	6
7	19	5.54	6
8	19	5.54	6
9	19	5.54	6
10	19	5.54	6
11	19	5.54	6
12	19	5.54	6
13	20	5.83	6
14	23	6.70	7
15	24	7.00	7

equalization has been implemented. Many of these subtle contrast changing techniques only give perceived improvement of detail on some image types and sometimes require all components of a colour composite image to be so processed before an "improvement" is noticeable.

It is not necessary to retain the same number of distinct brightness values in an equalized histogram as in the original. Sometimes it is desirable to have a smaller output set and thereby produce a histogram with (fewer) bars that are closer in height than would otherwise be the case. This is implemented by redefining L in (4.3) to be the new total number of bars. Repeating the example of Table 4.1 and Fig. 4.9 for the case of $L = 8$ (rather than 16) gives the look up table of Table 4.2. Such a strategy would be an appropriate one to adopt when using an output device with a small number of brightness values (grey levels).

4.4.2 Anomalies in Histogram Equalization

Images with extensive homogeneous regions will give rise to histograms with large bars at the corresponding brightness values. A particular example is a Landsat multispectral scanner infrared image with a large expanse of water. Because histogram equalization creates a histogram that is uniform on the average by grouping smaller bars together, the equalized version of an image such as that just described will have poor contrast and little detail – quite the opposite to what is intended. The reason for this can be seen in the simple illustration of Fig. 4.11. The cumulative histogram used as the look-up table for the enhancement is dominated by the large bar at brightness value 0. The resulting image would be mostly grey and white with little grey level discrimination.

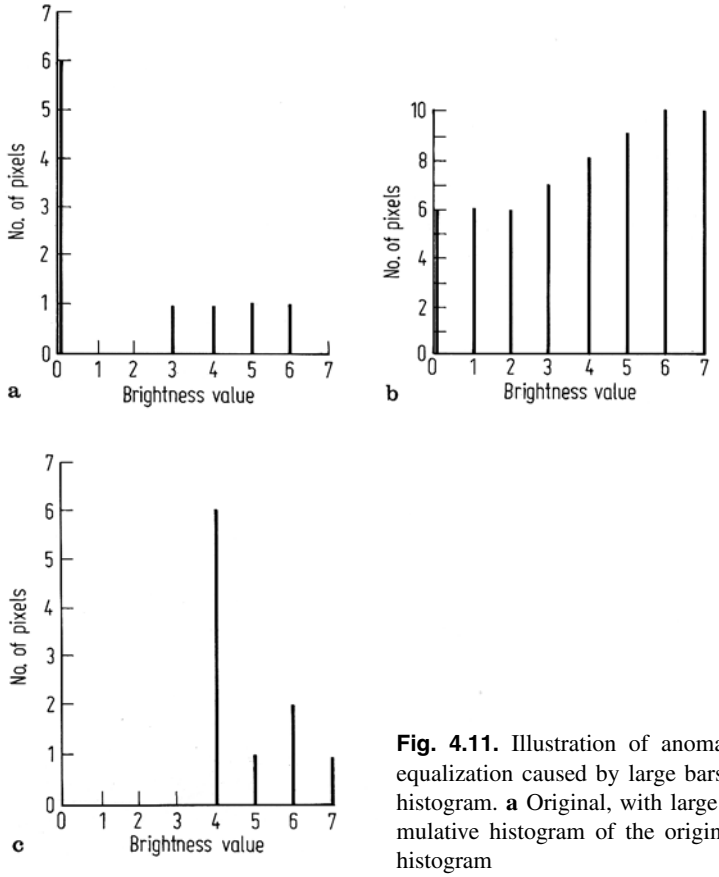


Fig. 4.11. Illustration of anomalous histogram equalization caused by large bars in the original histogram. **a** Original, with large bar at 0; **b** Cumulative histogram of the original; **c** Equalized histogram

A similar situation happens when the automatic contrast enhancement procedure of Sect. 4.3.4 is applied to images with large regions of constant brightness. This can give highly contrasting images on colour display systems; an acceptable display may require some manual adjustment of contrast taking due regard of the abnormally large histogram bars.

To avoid the anomaly in histogram equalization caused by the types of image discussed it is necessary to reduce the significance of the dominating bars in the image histograms. This can be done simply by arbitarily reducing their size when constructing the look up table, remembering to take account of this in the scale factor of (4.3). Another approach is to produce the cumulative histogram and thus look-up table on a subset of the image that does not include any, or any substantial portion, of the dominating region. Hogan (1981) has also provided an alternative procedure, based upon accumulating the histogram over "buckets" of brightness value. Once a bucket is full to a prespecified level, a new bucket is started.

4.5 Histogram Matching

4.5.1 Principle of Histogram Matching

Frequently it is desirable to match the histogram of one image to that of another image and in so doing make the apparent distribution of brightness values in the two images as close as possible. This would be necessary for example when a pair of contiguous images are to be joined to form a mosaic. Matching their histograms will minimise the brightness value variations across the join. In another case, it might be desirable to match the histogram of an image to a pre-specified shape, other than the uniform distribution treated in the previous section. For example, it is often found of value in photointerpretation to have an image whose histogram is a Gaussian function of brightness, in which most pixels have mid-range brightness values with only a few in the extreme white and black regions. The histogram matching technique, to be derived now, allows both of these procedures to be implemented.

The process of histogram matching is best looked at as having two stages, as depicted in Fig. 4.12. Suppose it is desired to match the histogram of a given image, $h_i(x)$, to the histogram $h_o(y)$; $h_o(y)$ could be a pre-specified mathematical expression or the histogram of the second image. Then the steps in the process are to equalize the histogram $h_i(x)$ by the methods of the previous section to obtain an intermediate histogram $h^*(z)$, which is then modified to the desired shape $h_o(y)$.

If $z = f(x)$ is the transformation that flattens $h_i(x)$ to produce $h^*(z)$ and $z = g(y)$ is the operation that would flatten the reference histogram $h_o(y)$ then the overall mapping of brightness values required to produce $h_o(y)$ from $h_i(x)$ is

$$y = g^{-1}(z),\ z = f(x) \quad \text{or} \quad y = g^{-1}\{f(x)\}. \tag{4.4}$$

If, as is often the case, the number of pixels and brightness values in $h_i(x)$ and $h_o(y)$ are the same, then the $(L-1)/N$ scaling factor in (4.3) will cancel in (4.4) and can therefore be ignored in establishing the look up table which implements the contrast matching process. Should the number of pixels be different, say N_1 in the image to be modified and N_2 in the reference image then a scaling factor of

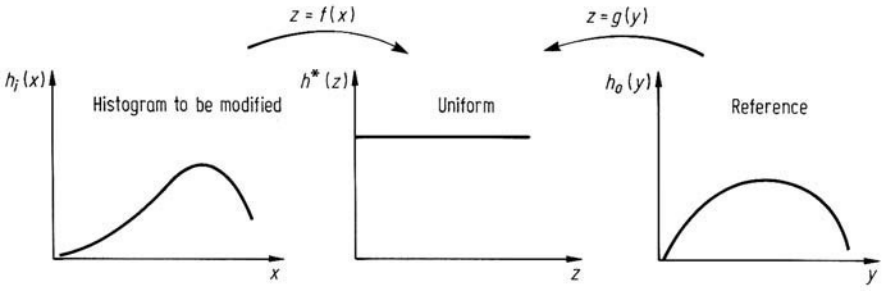


Fig. 4.12. The stages in histogram matching

N_2/N_1 will occur in (4.4). All scaling considerations can be bypassed however if the cumulative histograms are always scaled to some normalised value such as unity, or 100% (of the total number of pixels in an image).

4.5.2 Image to Image Contrast Matching

Figure 4.13 illustrates the steps implicit in (4.4) in matching source and reference histograms. In this case the reference histogram is that of a second image. Note that the procedure is to use the cumulative histogram of the source image to obtain new brightness values in the manner of the previous section by reading ordinate values corresponding to original brightness values entered on the abcissa. The new values are then entered into the *ordinate* of the cumulative reference histogram and the final brightness values (for the bars of the source histogram) are read from the *abcissa*; i.e. the cumulative reference histogram is used in reverse as indicated by the g^{-1} operation in (4.4). The look up table for this example is shown in Table 4.3. Again,

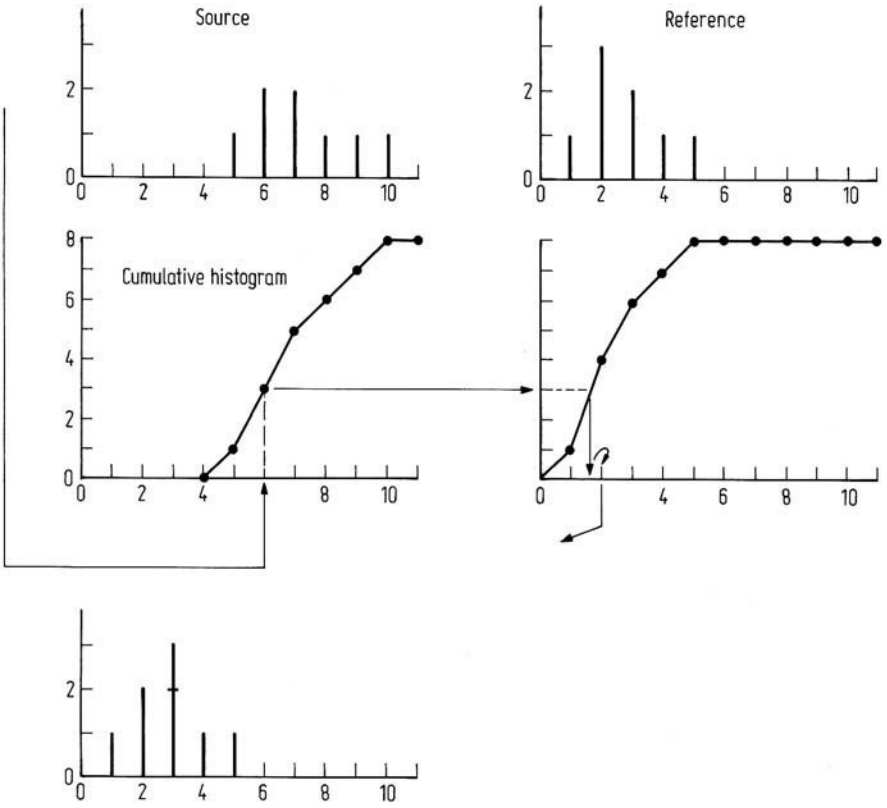


Fig. 4.13. An illustration of the steps in histogram matching

Table 4.3. Look up table generation for contrast matching

Source histogram brightness values x	Intermediate (equalized) values z	Modified values y	Nearest available brightness values
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	1	1	1
6	3	1.8	2
7	5	2.6	3
8	6	3	3
9	7	4	4
10	8	8	5
11	8	8	5

note that some of the new brightness values produced may not be in the available range; as before, they are adjusted to the nearest acceptable value.

An example using a pair of contiguous image segments is shown in Fig. 4.14. Because of seasonal differences the contrasts are quite different. Using the cumulative histograms an acceptable matching is achieved. Such a process, as noted earlier, is an essential step in producing a mosaic of separate contiguous images. Another step is to ensure geometric integrity of the join. This is done using the geometric registration procedures of Sect. 2.5.

4.5.3 Matching to a Mathematical Reference

In some applications it is of value to pre-specify the desired shape of an image histogram to give a modified image with a particular distribution of brightness values. To implement this it is necessary to take an existing image histogram and modify it according to the procedures of Sect. 4.5.1. The reference is a mathematical function that describes the desired shape. A particular example is to match an image histogram to a Gaussian or normal shape. Often this is referred to as applying a "gaussian stretch" to an image; it yields a modified version with few black and white regions and in which most detail is contained in the mid-grey range. This requires a reference histogram in the form of a normal distribution. However since a cumulative version of the reference is to be used, it is really a cumulative normal distribution that is required. Fortunately cumulative normal tables and curves are readily available. To use such a table in the contrast matching situation requires its ordinate to be adjusted to the total number of pixels in the image to be modified and its abcissa to be chosen to match the maximum allowable brightness range in the image. The latter requires consideration to be given to the number of standard deviations of the Gaussian distribution to be contained in the total brightness value range, having in mind that the Gaussian function is continuous to $\pm\infty$. The mean of the distribution is placed

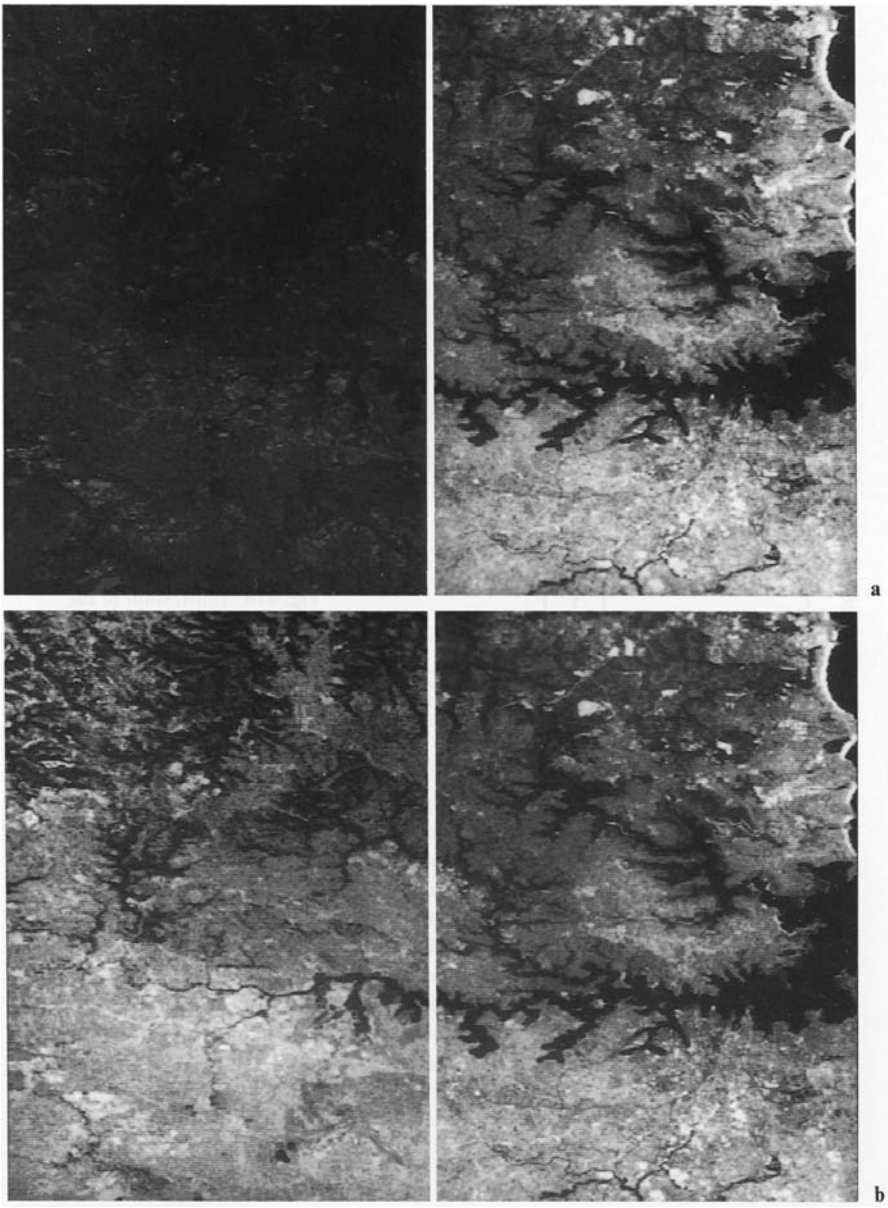


Fig. 4.14. a Contiguous Landsat multispectral scanner images showing contrast and brightness differences resulting from seasonal effects. The left hand image is an autumn scene and that on the right a summer scene, both of the northern suburbs of Sydney, Australia. **b** The same image pair but in which the histogram of the autumn scene has been matched to that of the summer scene

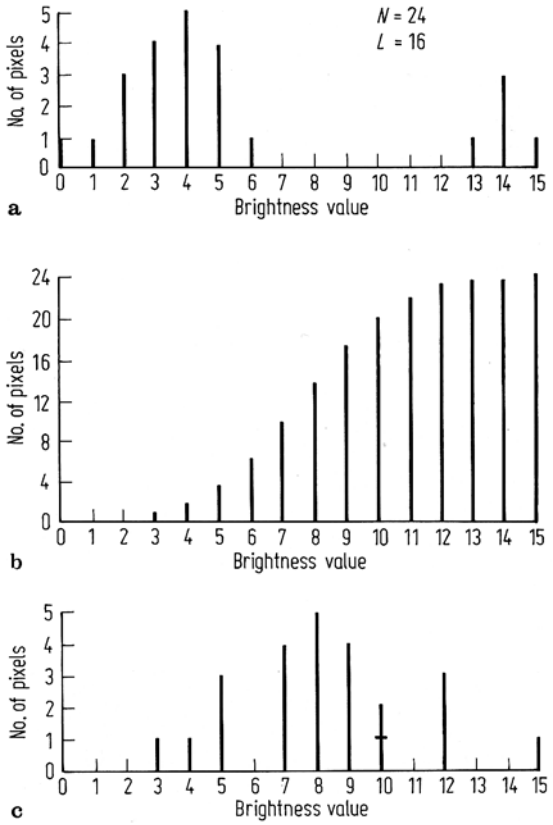


Fig. 4.15. Illustration of the modification of an image histogram to a pseudo-Gaussian shape. **a** Original histogram; **b** Cumulative normal histogram; **c** Histogram matched to Gaussian reference

usually at the mid-point of the brightness scale and commonly the standard deviation is chosen such that the extreme black and white regions are three standard deviations from the mean. A simple illustration is shown in Fig. 4.15.

or emitted by the terrain. It is instructive to review how such error may be removed from the remotely sensed data before they are analyzed further.

Correction for Sensor System Detector Error

Sometimes the remote sensing system simply does not function properly, resulting in radiometric error in the remotely sensed data. Several of the more common radiometric errors include line drop-outs, striping or banding, and line-start problems. For example, if one of the six detectors in the Landsat MSS fails to function during a scan, this can result in a brightness value of zero for every pixel, *j*, in a particular line, *i*. This is called a *line drop-out* and may appear as a completely black line in the band, *k*, of imagery. This is a serious condition because there is no way to restore data that were never acquired. However, it is possible to improve the visual interpretability of the data by introducing estimated brightness values for each bad scan line. The first problem is to locate each bad line. A simple thresholding algorithm can flag any scan line having a mean brightness value at or near zero. Once identified, it is then possible to evaluate the output for a pixel in the preceding line ($BV_{i-1,j,k}$) and succeeding line ($BV_{i+1,j,k}$) and assign the output pixel ($BV_{i,j,k}$) in the drop-out line the average of these two brightness values:

$$BV_{ijk} = \text{Int}\left(\frac{BV_{i-1,j,k} + BV_{i+1,j,k}}{2}\right) \quad (6\text{-}1)$$

This is performed for every pixel in a bad scan line. The result is an image consisting of interpolated data every *n*th line that is more visually interpretable than one with horizontal black lines running systematically throughout the entire image. The restoration method is also applicable to drop-out lines that occur randomly.

Sometimes, a detector does not fail completely, but simply goes out of adjustment; for example, it provides readings perhaps twice as great as the other detectors for the same band. This is often referred to as *n*-line *striping*. The data are valid but should be corrected (or restored) to have the same general contrast as the other detectors per scan. First, the bad scan lines must be identified in the scene. This is usually accomplished by computing a histogram of the values for each of the *n* detectors over a homogeneous area, such as a body of water. If one detector's mean or median is significantly different from the others, it is probable that this detector is out of adjustment. It may require a bias (additive or subtractive) correction or a more severe gain (multiplicative) correction. Then the errors arising from the maladjusted detector are not as noticeable.

Some scanning systems such as the Landsat Thematic Mapper generate a unique kind of scan-line noise, which may be a function of (1) relative gain and/or offset differences among the 16 detectors within a band (causing striping) and/or (2) variations between neighboring forward (west to east) and reverse (east to west) scans of all 16 detectors, causing *banding*. Crippen (1989) and Helder et al. (1992) provide filtering methods for the cosmetic removal of scan-line noise from Landsat TM-P imagery. However, Crippen (1989) cautions that the procedure may not be suitable for data that are going to be used to extract quantitative biophysical information. It is also possible to compute the Fourier transform (a special type of image enhancement algorithm to be discussed) of the various bands of remotely sensed data and apply a correction to remove striping and/or banding (an example is provided in Chapter 7).

Occasionally, scanning systems fail to collect data at the beginning of a scan line. This is called a *line-start* problem. Also, a detector may abruptly stop collecting data somewhere along a scan and produce results similar to the line drop-out previously discussed. Ideally, when data are not collected, the sensor system would be programmed to remember what was not collected and place any good data in their proper location within the scan. Unfortunately, this is not always the case. For example, the first pixel (column 1) in band *k* on line *i* (i.e., $BV_{i,1,k}$) might be improperly located at column 50 (i.e., $BV_{i,50,k}$). If the line-start problem is always associated with a horizontal bias of 50 columns, it can be corrected using a simple horizontal adjustment. However, if the amount of the line start displacement is random, it is difficult to restore the data without extensive human interaction on a line-by-line basis. A considerable amount of MSS data collected by Landsats 2 and 3 have this problem. Data not recorded by the detectors can never be restored.

Correction for Environmental Attenuation Error

Even when the remote sensing system is functioning properly, radiometric error may be introduced into the remote sensor data. The two most important sources of environmental attenuation include: (1) atmosphere attenuation caused by scattering and absorption in the atmosphere and (2) topographic attenuation. We will first address the correction of atmospheric attenuation.

Removal of Atmospheric Effects in Remotely Sensed Data

Atmospheric attenuation may be removed from remotely sensed data using one of several approaches (Cracknell and

The radiometric correction of topographically induced effects is still in its infancy. Civco (1989) identified some important considerations:

- The digital elevation model of the study area should have a spatial resolution comparable to the spatial resolution of the digital remote sensor data.
- When correcting for topographic effect using a Lambertian surface assumption, remotely sensed data are often overcorrected. Slopes facing away from the sun (i.e., north-facing slopes in the northern hemisphere) appear brighter than sun-facing slopes (south-facing slopes) of similar composition. This suggests that the ideal correction is less than the traditional cosine of the incidence angle.
- Most techniques consider only the direct solar beam contribution to irradiance and forget that the diffuse component also illuminates the local topography.
- There is often strong anisotropy of apparent reflectance, which is wavelength dependent, and solar, surface, and sensor system geometry should be considered in the modeling process (Leprieur et al., 1988).
- The amount of correction required is a function of wavelength. Particular attention must be given to middle-infrared bands, which are severely affected by the topographic effect (Kawata et al., 1988).
- It is difficult to remove the topographic effect completely from severely shadowed areas such as deep valleys (Kawata et al., 1988).

Geometric Correction of Remote Sensor Data

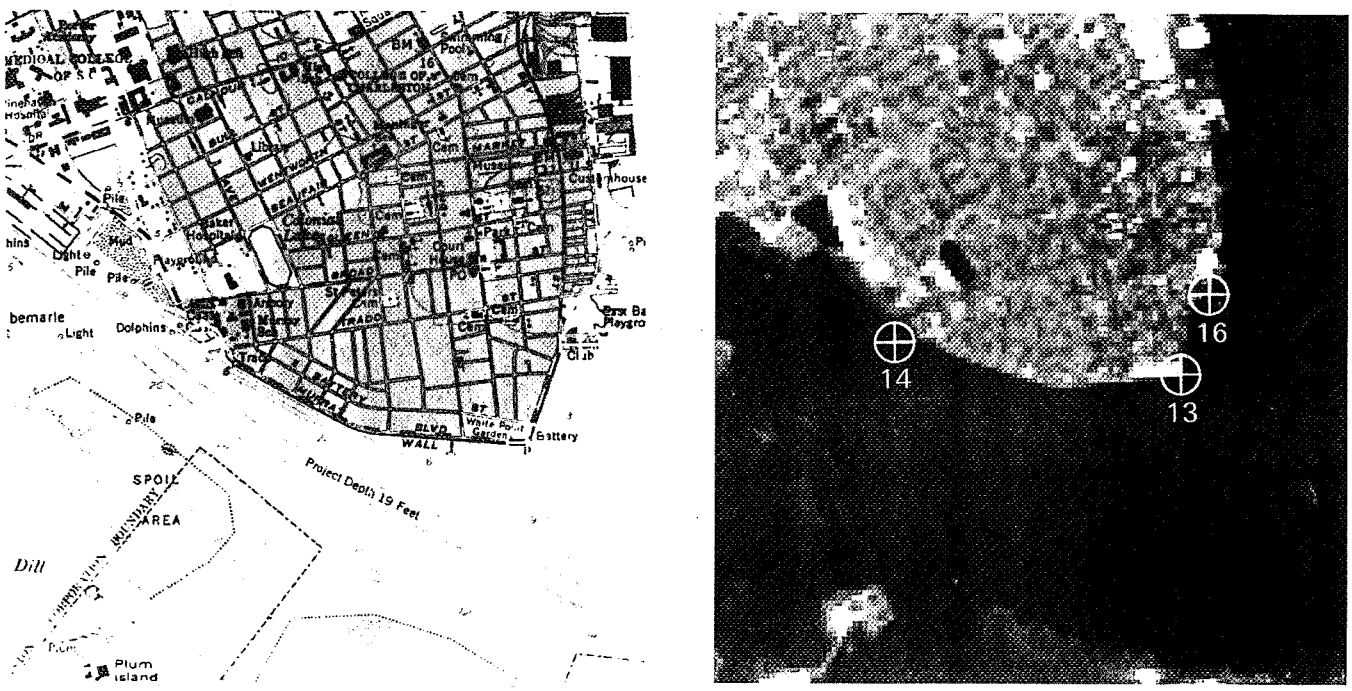
Remotely sensed data usually contain both systematic and unsystematic geometric errors. Some of the more important sources are summarized in Table 6-4. These errors may be divided into two classes: (1) those that can be corrected using data from platform ephemeris and knowledge of internal sensor distortion and (2) those that cannot be corrected with acceptable accuracy without a sufficient number of ground control points. A *ground control point* (GCP) is a point on the surface of Earth where both image coordinates (measured in rows and columns) and map coordinates (measured in degrees of latitude and longitude, feet, or meters) can be identified. Those geometric distortions that can be corrected through analysis of sensor characteristics and ephemeris include scan skew, mirror-scan velocity nonlinearities, panoramic distortion, spacecraft velocity, and perspective geometry (including Earth's curvature). Those that can only be corrected through the use of GCPs are sensor system attitude (roll, pitch, and yaw) and/or altitude (Bernstein, 1983).

Table 6-4. Sources of Geometric Error in Remote Sensing Scanning Systems[a]

Systematic Distortions
Scan skew: Caused by the forward motion of the platform during the time required for each mirror sweep. The ground swath is not normal to the ground track but is slightly skewed, producing cross-scan geometric distortion.
Mirror-scan velocity: The mirror scanning rate is usually not constant across a given scan, producing along-scan geometric distortion.
Panoramic distortion: The ground area imaged is proportional to the tangent of the scan angle rather than to the angle itself. Because data are sampled at regular intervals, this produces along-scan distortion.
Platform velocity: If the speed of the platform changes, the ground track covered by successive mirror scans changes, producing along-track scale distortion.
Earth rotation: Earth rotates as the sensor scans the terrain. This results in a shift of the ground swath being scanned, causing along-scan distortion.
Perspective: For some applications it is desirable to have the MSS images represent the projection of points on Earth on a plane tangent to Earth with all projection lines normal to the plane. This introduces along-scan distortion.
Nonsystematic Distortions
Altitude: If the sensor platform departs from its normal altitude or the terrain increases in elevation, this produces changes in scale.
Attitude: One sensor system axis is usually maintained normal to Earth's surface and the other parallel to the spacecraft's direction of travel. If the sensor departs from this attitude, geometric distortion results.

[a] Several of the systematic distortions (scan skew, mirror-scan velocity, and panoramic) may not be found in remote sensor data collected by charge-coupled-device (CCD) sensor systems, such as the SPOT High Resolution Visible (HRV).

Most commercially available remote sensor data (e.g., from SPOT Image Corporation or EOSAT Corporation) already have much of the systematic error removed. Unless otherwise processed, however, the unsystematic error remains in



(a) (b)

Figure 6-11 Example of image-to-map rectification. (a) USGS 7.5′ quadrangle of Charleston, S.C. with ground control points in UTM coordinates identified. (b) Unrectified 11/09/82 Landsat TM band 4 data with ground control points in row and column coordinates identified.

the image, making it nonplanimetric (i.e., not in its proper x, y planimetric position). This section focuses on two common geometric correction procedures often used by Earth scientists to make the digital remote sensor data of value: image-to-map rectification and image-to-image registration.

Image-to-map rectification is the process by which the geometry of an image is made planimetric. Whenever accurate area, direction, and distance measurements are required, image-to-map geometric rectification should be performed. It may not, however, remove all distortion caused by topographic relief displacement in images. The image-to-map rectification process normally involves selecting GCP image pixel coordinates (row and column) with their map coordinate counterparts (e.g., meters in northing and easting in a universal transverse mercator map projection). For example, Figure 6-11 displays three GCPs (points 13, 14, and 16) easily identifiable by an image analyst in both a USGS 7.5-minute quadrangle and an unrectified Landsat TM band 4 image of Charleston, S.C. It will be demonstrated how the mathematical relationship between the image coordinates and map coordinates of selected GCPs is computed and the image is made to fit the geometry of the map. GCP map coordinate information does not always have to come from a planimetric map. Instead, global positioning system (GPS) instruments may be taken into the field to obtain the coordinates of objects to within ±5 m when differentially corrected (Clavet et al., 1993). GPS collection of map coordinate information to be used for image rectification is especially effective in poorly mapped regions of the world or where rapid change has made existing maps obsolete (Welch et al., 1992).

Image-to-image registration is the translation and rotation alignment process by which two images of like geometry and of the same geographic area are positioned coincident with respect to one another so that corresponding elements of the same ground area appear in the same place on the registered images (Chen and Lee, 1992). This type of geometric correction is used when it is not necessary to have each pixel assigned a unique x, y coordinate in a map projection. For example, we might want simply to compare two images obtained on different dates to see if any change has taken place in them. While it is possible to rectify both of the images to a standard map projection and then evaluate them (and this is often done), this may not be necessary to simply identify the change between the two images.

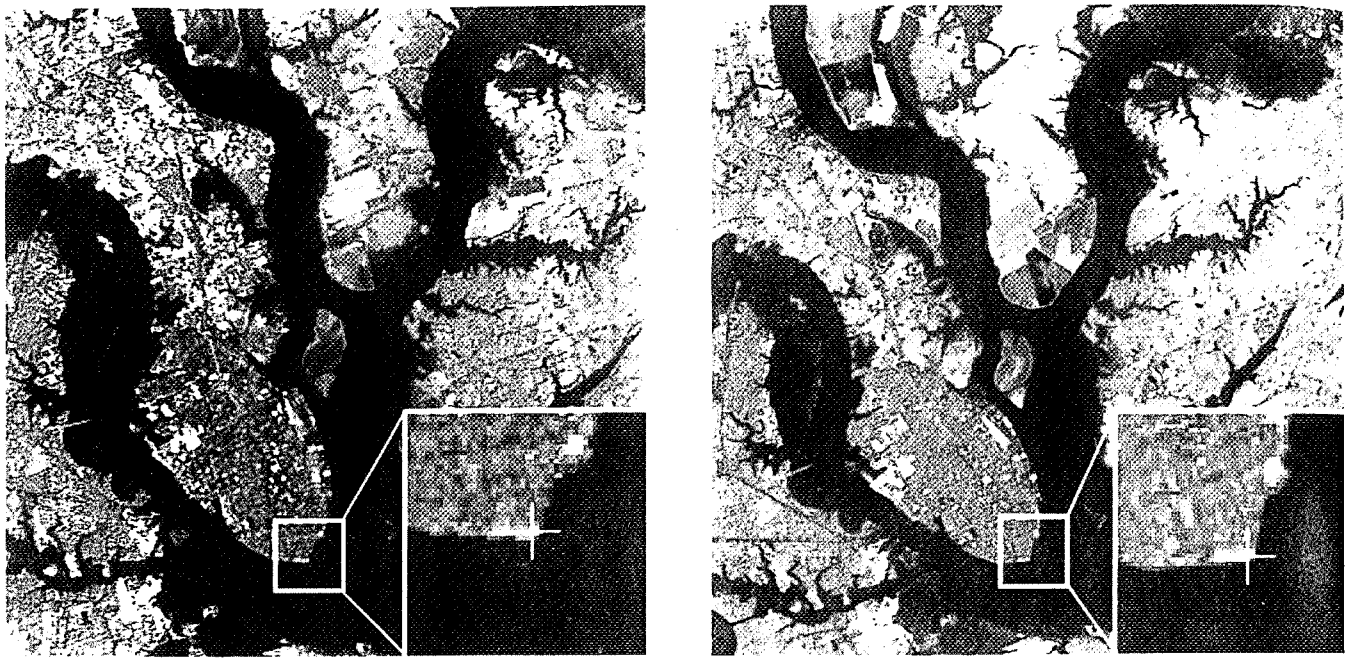


Figure 6-12 Example of image-to-image registration. (a) Previously rectified 11/09/82 Landsat TM band 4 data resampled to 30 × 30 m pixels using nearest-neighbor resampling logic and a UTM projection. (b) Unrectified 10/14/87 Landsat TM data to be registered to the rectified 1992 Landsat scene.

It is interesting that the same general image processing principles are used in both image rectification and image registration. The difference is that in image-to-map rectification the reference is a map in a standard map projection, while in image-to-image registration the reference is another image. It should be obvious that if an image is used as the reference base (rather than a map) any other image registered to it will inherit the geometric errors existing in the reference image. Because of this characteristic, most serious Earth science remote sensing research is based on analysis of data that have been rectified to a map base. However, when conducting rigorous change detection between two or more dates of remotely sensed data, it may be useful to select a *hybrid* approach involving both image-to-map rectification and image-to-image registration (Jensen et al., 1993).

An example of the hybrid approach is demonstrated in Figure 6-12 where a 10/14/87 Landsat TM image is being registered to a rectified 11/09/82 Landsat TM scene. In this case, the 1982 base year image was previously rectified to a universal transverse mercator map projection with 30 × 30 m pixels. Ground control points are being selected to register the 1987 image to the rectified 1982 base year image. It is often very difficult to locate good ground control points in remotely sensed data, especially in rural areas (e.g., forests, wetland, and water bodies). The use of the *rectified* base year image as the map allows many more common GCPs to be located in the unrectified 1987 imagery. For example, edges of water bodies and fields or the intersection of small stream segments are not usually found on a map but may be easy to identify in the rectified and unrectified imagery. The optimum method of selecting such points is to have both the rectified base year image and the image to be rectified on the screen at the same time (Figure 6-12). This dual-display greatly simplifies GCP selection. Some image processing systems even allow the GCP selected to be reprojected onto the image to be corrected (with the appropriate transformation coefficients to be discussed) to determine the quality of the GCP point. Also, some systems allow the analyst to extract floating point row and column coordinates of GCPs (instead of just integer values) through the use of a chip extraction algorithm that zooms in and does subpixel sampling, as demonstrated in Figure 6-12. GCP subpixel row and column coordinates often improve the precision of the image-to-map rectification or image-to-image registration. Some scientists have developed methods of automatically extracting GCPs common to two images, which can be used during image-to-image registration (Ton and Jain, 1989; Chen and Lee, 1992). However, most image-to-map rectification still relies heavily on human interaction.

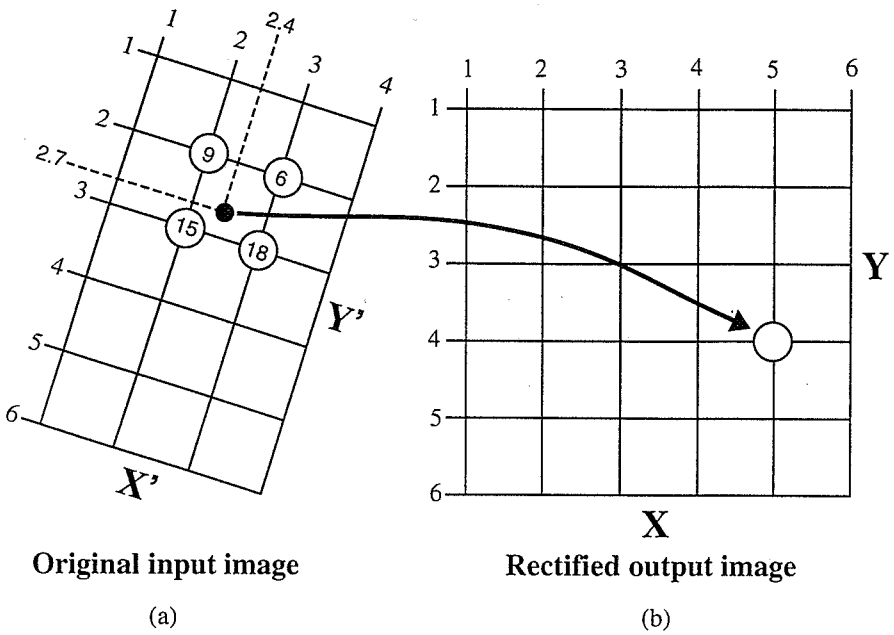


Figure 6-13 This diagram illustrates how the rectified output grid (X, Y) is filled with brightness values from the geometrically distorted input grid (X', Y'). In particular, we are trying to fill pixel 5, 4 in the output grid with a value from the appropriate location in the input grid. The appropriate location is computed using Equation 6-18 with the six necessary coefficients and a value of 5 for the X value and a value of 4 for the Y value in the output image. The equation then computes the X', Y' coordinates of the location in the original image to obtain the appropriate brightness value. The location in this example is 2.4, 2.7. Because these coordinates are not integers, it is necessary to use an intensity interpolation technique to compute the exact brightness value to be placed in X, Y location 5, 4 of the rectified image.

The following example focuses on image-to-map geometric rectification because it is the most frequently used method of removing geometric distortion from remotely sensed data.

Image-to-Map Geometric Rectification

Two basic operations must be performed to geometrically rectify a remotely sensed image to a map coordinate system:

1. The geometric relationship between the input pixel location (row and column) and the associated map coordinate of this same point (x, y) must be identified (Figure 6-13). This will establish the nature of the geometric coordinate transformation that must be applied to rectify or relocate every pixel in the original input image (x', y') to its proper position in the rectified output image (x, y). This process is called *spatial interpolation.*

2. Pixel brightness values must be determined. Unfortunately, there is not a direct one-to-one relationship between the movement of input pixel values to output pixel locations. It will be shown that a pixel in the rectified output image often requires a value from the input pixel grid that does not fall neatly on a row-and-column coordinate. When this occurs, there must be some mechanism for determining the brightness value (BV) to be assigned to the new rectified pixel. This process is called *intensity interpolation.*

Spatial Interpolation Using Coordinate Transformations

As discussed earlier, some distortions in remotely sensed data may be removed or mitigated using techniques that model systematic orbital and sensor characteristics. Unfortunately, this does not remove error produced by changes in attitude (roll, pitch, and yaw) or altitude. Such errors are generally unsystematic and are best removed by identifying GCPs in the original imagery and on the reference map and then mathematically modeling the geometric distortion present. Image-to-map rectification requires that polynomial equations be fit to the GCP data using least-squares criteria to model the corrections directly in the image domain without explicitly identifying the source of the distortion

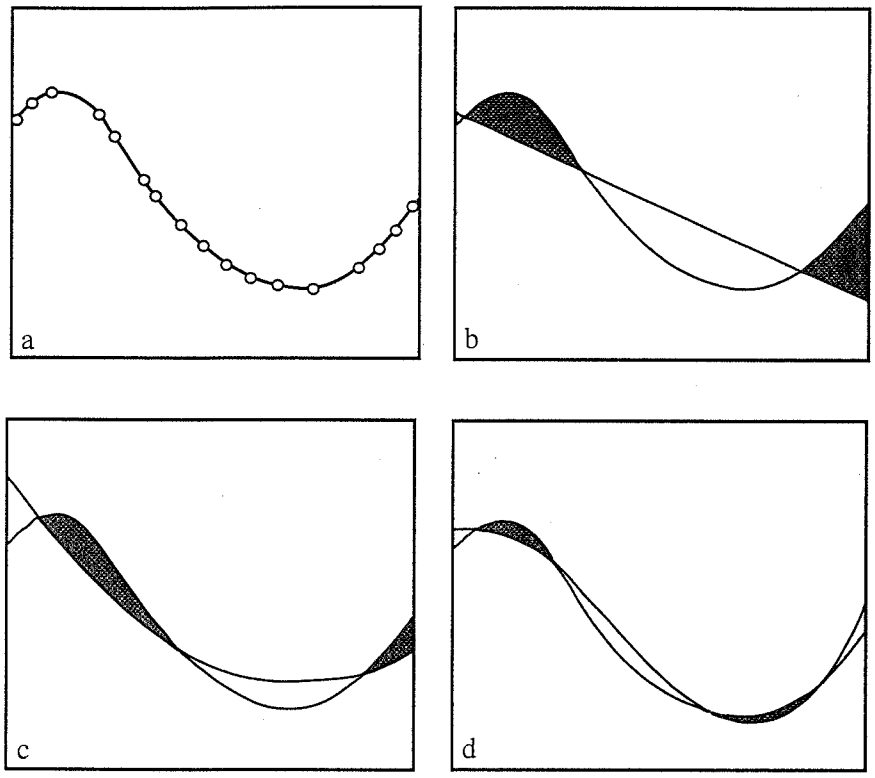


Figure 6-14 Concept of how different-order affine transformations fit a hypothetical surface illustrated in cross section. (a) Original surface. (b) A first-order linear transformation fits a plane to the data. (c) Second-order quadratic fit. (d) Third-order cubic fit (Jensen et al., 1988).

(Novak, 1992). Depending on the distortion in the imagery, the number of GCPs used, and the degree of topographic relief displacement in the area, higher-order polynomial equations may be required to geometrically correct the data. The *order* of the rectification is simply the highest exponent used in the polynomial. For example, Figure 6-14 demonstrates how different-order affine transformations fit a hypothetical surface (Jensen et al., 1988). Generally, for moderate distortions in a relatively small area of an image (e.g., a quarter of a Landsat TM scene), a first-order, six-parameter, affine transformation is sufficient to rectify the imagery to a geographic frame of reference.

This type of transformation can model six kinds of distortion in the remote sensor data, including translation in x and y, scale changes in x and y, skew, and rotation (Novak, 1992). When all six operations are combined into a single expression it, becomes

$$x' = a_0 + a_1x + a_2y \qquad (6\text{-}18)$$

$$y' = b_0 + b_1x + b_2y$$

where x and y are positions in the output-rectified image or map, while the x' and y' represent corresponding positions in the original input image.

Using these six coordinate transform coefficients that model distortions in the original scene, it is possible to transfer (relocate) pixel values from the original distorted image x', y' to the grid of the rectified output image, x, y. However, before applying the rectification to the entire set of data, it is important to determine how well the six coefficients derived from the least-squares regression of the initial GCPs account for the geometric distortion in the input image. The method used most often involves the computation of the root-mean-square error (RMS_{error}) for each of the ground control points (Ton and Jain, 1989).

Let us consider for a moment the nature of the GCP data. We first identify a point in the image such as a road intersection. Its column and row coordinates in the original image we will call x_{orig} and y_{orig}. The x and y position of the same road intersection is then measured from the reference map in degrees, feet, or meters. These two sets of GCP coordinates

and many others selected by the analyst are used to compute the six coefficients discussed in Equation 6-18. Now, if we were to input the map x and y values for the first GCP back into Equation 6-18, with all the coefficients in place, we would get computed x' and y' values that are supposed to be the location of this point in the image space. Ideally, x' would equal x_{orig} and y' would equal y_{orig}. Unfortunately, this is rarely the case. Any discrepancy between the values represents image geometric distortion not corrected for by the six-coefficient coordinate transformation.

A simple way to measure such distortion is by computing the RMS_{error} for each control point using the following equation:

$$RMS_{error} = \sqrt{(x' - x_{orig})^2 + (y' - y_{orig})^2} \qquad (6\text{-}19)$$

where x_{orig} and y_{orig} are the original row and column coordinates of the GCP in the image and x' and y' are the computed or estimated coordinates in the original image. The square root of the squared deviations represents a measure of the accuracy of this GCP in the image. By computing RMS_{error} for all GCPs, it is possible to (1) see which GCPs exhibit the greatest error and (2) sum all the RMS_{error}.

Normally, the user specifies a certain amount (a threshold) of acceptable total RMS error. If an evaluation of the total RMS_{error} reveals that a given set of control points exceeds this threshold, it is common practice to (1) delete the GCP from the analysis that has the greatest amount of individual error, (2) recompute the six coefficients, and (3) recompute the RMS_{error} for all points. This process continues until one of the following occurs: the total RMS_{error} is less than the threshold specified or too few points remain to perform a least-squares regression to compute the coefficients. Once the acceptable RMS_{error} is reached, we can proceed to the intensity interpolation phase of geometric rectification, which attempts to fill an output grid (x, y) with brightness values found within the original input grid (x', y').

Intensity Interpolation

This process involves the extraction of a brightness value from an x', y' location in the original (distorted) input image and its relocation to the appropriate x, y coordinate location in the rectified output image. This pixel filling logic is used to produce the output image line by line, column by column. Most of the time the x' and y' coordinates to be sampled in the input image are real numbers (i.e. they are not integers). For example, in Figure 6-13 we see that pixel 5, 4 (x, y) in the output image is to be filled with the value from coordinates 2.4, 2.7 (x', y') in the input image. When this occurs, there are several methods of brightness value (BV) interpolation that can be applied, including nearest neighbor, bilinear interpolation, and cubic convolution. The practice is commonly referred to as *resampling*.

In zero-order or *nearest-neighbor* interpolation, the brightness value closest to the x', y' coordinate specified is assigned to the output x, y coordinate. For example, in Figure 6-13, the output pixel 5, 4 (x, y) requests the brightness value in the original input image at location 2.4, 2.7 (x', y'). There is no value at this location. However, there are nearby values at the integer grid intersections. A nearest-neighbor rule would assign the output pixel (x, y) the value of 15, which is the value found at the nearest input pixel.

This is a computationally efficient procedure. It is especially liked by Earth scientists because it does not alter the pixel brightness values during resampling (Duggin and Robinove, 1990). It is often the very subtle changes in brightness value that make all the difference when discriminating one type of vegetation from another, an edge associated with a geologic lineament, or different levels of turbidity or temperature in a lake. Other interpolation techniques to be discussed use averages to compute the output intensity value, often removing valuable spectral information.

First-order or *bilinear interpolation* assigns output pixel values by interpolating brightness values in two orthogonal directions in the input image. It basically fits a plane to the four pixel values nearest to the desired position (x', y') in the input image and then computes a new brightness value based on the weighted distances to these points. For example, the distances from the requested x', y' position at 2.4, 2.7 in the input image in Figure 6-13 to the closest four input pixel coordinates (2, 2; 3, 2; 2, 3; 3, 3) are computed in Table 6-5. The closer a pixel is to the desired x', y' location, the more weight it will have in the final computation of the average. The weighted average of the new brightness value (BV_{wt}) is computed according to the equation

$$BV_{wt} = \frac{\sum_{k=1}^{4} \frac{Z_k}{D_k^2}}{\sum_{k=1}^{4} \frac{1}{D_k^2}} \qquad (6\text{-}20)$$

where Z_k are the surrounding four data point values, and D_k^2 are the distances squared from the point in question (x', y') to these data points. In our example, the weighted average of BV_{wt} is 13.53 (truncated to 13), as shown in Table 6-5. The average without weighting is 12. In many respects this method acts as a spatial moving filter that subdues extremes in brightness value throughout the output image. The

Table 6-5. Bilinear Interpolation of a Weighted Brightness Value (BV_{wt}) at Location x',y' Based on the Analysis of Four Sample Points in Figure 6-13.

Sample Point Location (column, row)	Value at Sample Point, Z	Distance from x', y' to the Sample Point, D	D_k^2	$\frac{Z}{D_k^2}$	$\frac{1}{D_k^2}$
2, 2	9	0.806	0.65	13.85	1.539
3, 2	6	0.922	0.85	7.06	1.176
2, 3	15	0.500	0.25	60.00	4.000
3, 3	18	0.670	0.45	40.00	2.222
				Σ 120.91	Σ 8.937
				BV_{wt} = 120.91/8.937 =13.53	

method is also more computationally demanding than the nearest-neighbor method.

Cubic convolution resampling assigns values to output pixels in much the same manner as bilinear interpolation, except that the weighted values of 16 input pixels surrounding the location of the desired x', y' pixel are used to determine the value of the output pixel.

Implementation

To appreciate digital image rectification, it is useful to demonstrate the logic by applying it to a real data set such as the Charleston, S.C., thematic mapper image. In doing so we are concerned with the rectification of the remotely sensed data to a map (i.e., an image-to-map rectification). The type of map and its associated projection are important parameters and must be carefully selected.

There are several types of map projections to be considered in this context. On an *equivalent* (equal-area) map projection, a circle of diameter n drawn at any location on the map will encompass exactly the same geographic area. This characteristic is useful if the scientist is interested in comparing land-use area, density, and so on. Unfortunately, to maintain the equal-area attribute, the shapes, angles, and scale in parts of the map may be distorted.

Conversely, a *conformal* map projection maintains correct shape and distance around a given point on the map. Because angles at each point are correct on conformal maps, the scale in every direction around any point is constant. This allows the analyst to measure distance and direction between relatively near points with good precision. For our purposes, this means that, for image areas covering a few contiguous 7.5-minute quadrangle sheets, accurate spatial measurement is possible if the data are rectified to a conformal map projection.

One of the most often used projections for rectifying remotely sensed data is the transverse mercator projection. It is made from a normal mercator projection by rotating the cylinder (the developable surface) so that it lies tangent along a meridian. The central meridians, the equator, and each line 90° from the central meridian are straight lines (Figure 6-15). The central meridian normally has a constant scale. Any lines parallel to the central meridian are lines of constant scale. The universal transverse mercator projection often used by the USGS in their topographic mapping program has a central scale factor of 0.9996 and is composed of 60 zones, each 6° of longitude wide, with a central meridian placed every sixth meridian beginning with 177° west.

In our example, the Charleston, S.C., 7.5 minute quadrangle (photorevised 1979) was selected as the appropriate base map with which to rectify the TM data (Figure 6-11). It lies in UTM zone 17 and has a 1000-m UTM grid overlay. Twenty ground control points were located on the map and the UTM easting and northing of each point were identified (Table 6-6). The same 20 GCPs were then identified in the TM data according to their row and column coordinates (also in Table 6-6). The location of points 13, 14, and 16 are shown in Figure 6-11. The GCPs should be located uniformly throughout the region to be rectified and not congested into one small area simply because there are more easily identifiable points in that area.

The 20 GCPs were input to the least-squares regression procedure previously discussed to identify (1) the coefficients of

CHAPTER 2

Optical Radiation Models

2.1 Introduction

Passive remote sensing in the optical regime (visible through thermal) depends on two sources of radiation. In the visible to shortwave infrared, the radiation collected by a remote sensing system originates with the sun. Part of the radiation received by a sensor has been reflected at the earth's surface and part has been scattered by the atmosphere, without ever reaching the earth. In the thermal infrared, thermal radiation is emitted directly by materials on the earth and combines with self-emitted thermal radiation in the atmosphere as it propagates upward. In this chapter, we outline the basic models appropriate to the optical region from the visible through the thermal infrared. The science of radiometry is applied here as a means to an end, i.e., to define the major processes involved in remote sensing in this spectral region. For thorough treatments of optical radiometry in remote sensing, the books by Slater (1980) and Schott (1996) are recommended.

2.2 Visible to Shortwave Infrared Region

All materials on the earth's surface passively absorb and reflect solar radiation in the 0.4 to 3μm spectral range. Some materials also transmit solar radiation; for example, water bodies and plant canopies. At longer wavelengths, materials at normal temperatures begin to actively emit thermal radiation, which is discussed later in this chapter. A description of how solar radiation propagates and is modified prior to sensing by an optical system follows.

2.2.1 Solar Radiation

The source of energy for remote sensing in the *solar-reflective* spectral region (visible to shortwave IR) is the sun. The sun is a near-perfect *blackbody* radiator; that is, it emits radiation at nearly the maximum efficiency possible for a body at its effective temperature. The *spectral radiant exitance*, M_λ, from the sun can therefore be modeled by *Planck's blackbody equation*,

$$M_\lambda = \frac{C_1}{\lambda^5 [e^{C_2/(\lambda T)} - 1]} \tag{2-1}$$

where

T is the blackbody's temperature in Kelvin (K),

$C_1 = 3.74151 \times 10^8$ W-m^{-2}-μm^4, and

$C_2 = 1.43879 \times 10^4$ μm-K.

Using these values and specifying the wavelength, λ, in μm and the temperature in Kelvin results in units for spectral radiant exitance of W-m^{-2}-μm^{-1} (Slater, 1980), i.e., power (or flux) per unit area of the sun's surface, per unit wavelength interval.

The blackbody function peaks at a wavelength given by *Wien's Law*,

$$\lambda|_{max} = 2898 / T . \tag{2-2}$$

where the blackbody's temperature is in Kelvin and the wavelength at which the radiation is a maximum is in μm. Thus, as the temperature of a blackbody increases, the wavelength of maximum radiant exitance decreases.

We, of course, are interested in the radiation that reaches the earth. To calculate that, the spectral radiant exitance of Eq. (2-1) must first be propagated to the top of the earth's atmosphere and then converted to *spectral irradiance*, E_λ^0. This transformation is accomplished with the following equation,

$$\textit{top-of-the atmosphere:} \quad E_\lambda^0 = \frac{M_\lambda}{\pi} \times \frac{\text{area solar disk}}{(\text{distance-to-earth})^2} . \tag{2-3}$$

The units of spectral irradiance are the same as those of spectral radiant exitance, namely those of *spectral flux density*, W-m^{-2}-μm^{-1}. The spectral content of the radiation does not change in the transit through space, but the magnitude of solar irradiance at the earth changes by a few percent as the sun-earth distance changes throughout the year. The exo-atmospheric flux density is plotted in Fig. 2-1 using a blackbody model and a more detailed, empirical function stored in the atmospheric modeling program MODTRAN. As can be seen from the figure, the blackbody model with a temperature of 5900K is a good approximation to measured solar radiation.

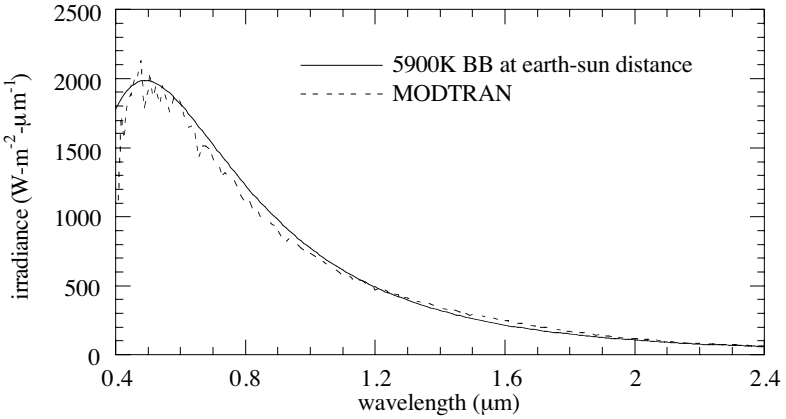


FIGURE 2-1. Comparison of the exo-atmospheric (top-of-the-atmosphere) solar spectral irradiance as used in the atmospheric modeling code MODTRAN (Berk et al., 1989) to that produced by a blackbody object at a temperature of 5900K and at the same distance as the sun. Deviations from the blackbody model are mainly in narrow absorption lines present in the actual solar spectrum.

As the wavelength increases to the shortwave IR, less radiation is available from the sun for signal detection by remote sensing. If we ignore atmospheric effects, the solar radiation arriving at the earth is matched by self-emitted thermal energy from the earth at about 4.5 μm (Fig. 2-2). The wavelength at which they are equal contributors to the *at-sensor radiance above the atmosphere* depends on earth surface reflectance and emissivity (see Sect. 2.3) and the atmosphere, and can range between 2.5 and 6 μm (Slater, 1980).

2.2.2 Radiation Components

The major radiation transfers of concern in the visible through SWIR spectral regions are shown in Fig. 2-3. There are generally three significant components in the upwelling at-sensor radiation:

- the unscattered, surface-reflected radiation, L_λ^{su}

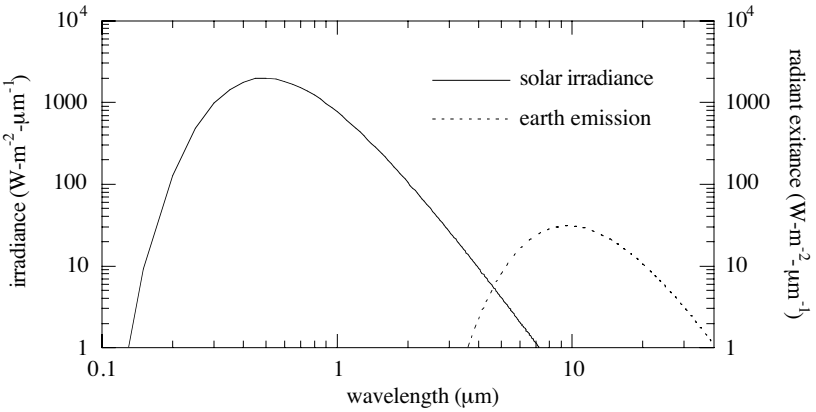


FIGURE 2-2. Spectral distributions at the top-of-the-atmosphere for the two radiation sources in the visible through thermal infrared spectral regions. The earth is assumed to be a blackbody at T = 300K and the sun a blackbody at T = 5900K, and atmospheric effects are ignored. Blackbody curves for different temperatures cannot cross each other; it only appears to occur here because the solar radiant exitance has been scaled as described in Eq. (2-3).

- the down-scattered, surface-reflected skylight, L_λ^{sd}
- the up-scattered path radiance, L_λ^{sp}

We can therefore write for the total upwelling radiance at a high altitude or satellite sensor,[1]

$$L_\lambda^s = L_\lambda^{su} + L_\lambda^{sd} + L_\lambda^{sp} \ . \tag{2-4}$$

In the following sections, we will look at appropriate models for each component.

Surface-reflected, unscattered component (L_λ^{su})

The atmosphere is an unavoidable influence in satellite and high altitude aerial remote sensing in the visible through shortwave IR. It scatters and absorbs radiation between the sun and the earth along the solar path, and again between the earth and the sensor along the view path. The fraction of radiation that initially arrives at the earth's surface is called the *solar path transmittance*, $\tau_s(\lambda)$, and is, by definition, between zero and one, and unitless. A typical spectral transmittance curve for the solar path, as generated from atmospheric models, is shown in Fig. 2-4.

1. We use the superscript *s* to denote "solar," thereby distinguishing these terms from thermal emitted radiation, described later.

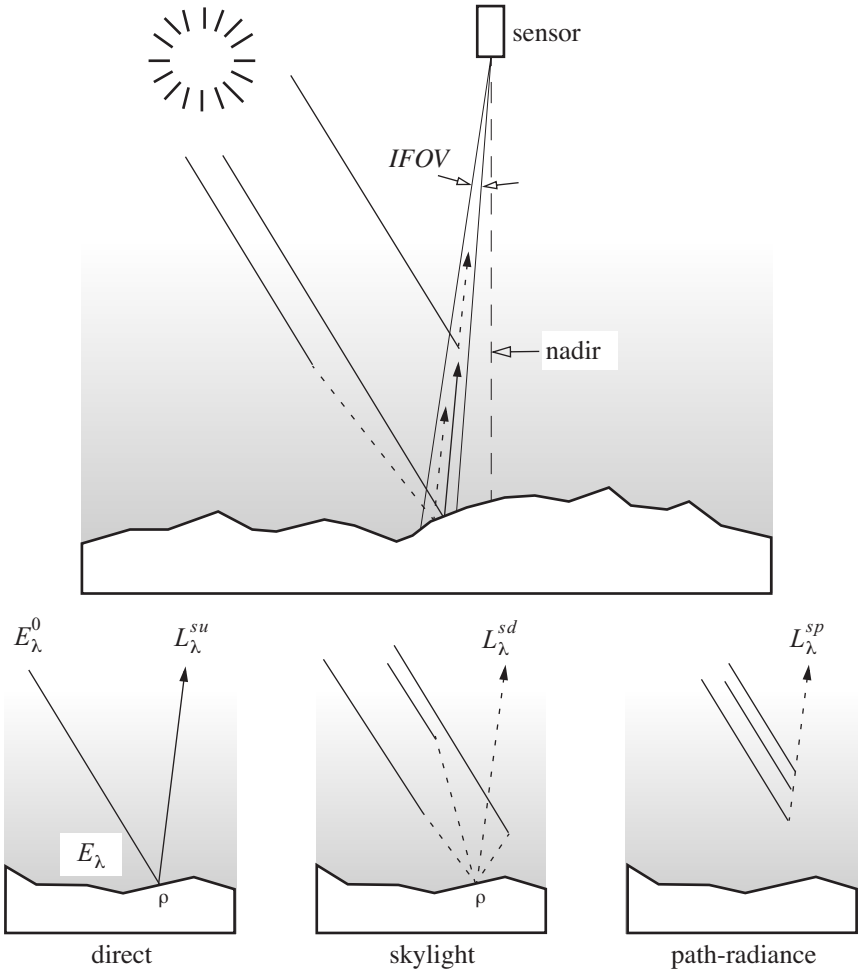


FIGURE 2-3. The most significant radiation components seen by the sensor in solar reflective remote sensing are the "direct" component, the "skylight" component, and the "path radiance" component (commonly called "haze"). The shading in the diagram portrays the decreasing atmospheric density at higher altitudes. Other radiation can reach the sensor, such as from the "adjacency" component, consisting of a direct reflection from a nearby GIFOV, followed by either an up-scatter directly to the sensor, or a down-scatter into the GIFOV of interest, followed by a second reflection toward the sensor. The adjacency phenomenon increases local spatial correlation among pixels and reduces the contrast of dark-light boundaries, such as a shoreline. Multiple surface reflections and atmospheric scatterings generally are of less significance, because the radiation magnitude is reduced by each reflection or scattering event. Note that, because of the great distance to the sun, its rays are effectively parallel when entering the earth's atmosphere.

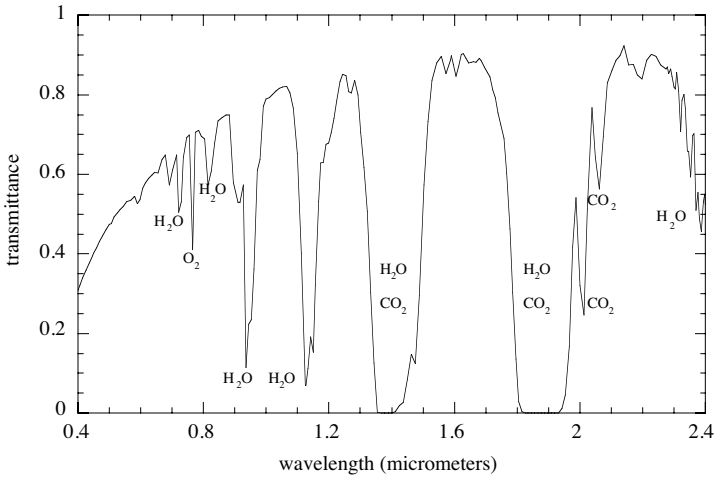


FIGURE 2-4. Transmittance of the atmosphere as calculated by the program MODTRAN. This is the transmittance along the solar path, i.e., between the sun and the earth's surface, for a solar elevation angle of 45°. The absorption bands are primarily associated with water vapor and carbon dioxide. This figure, and many of the others in this chapter, are produced with specific atmospheric parameters in the modeling program MODTRAN (Berk et al., 1989). The atmosphere model is "mid-latitude summer" with a "visibility" of 23 km (a relatively high value, implying that Mie scattering effects are small relative to Rayleigh scattering effects). The solar elevation angle is 45° and the azimuth angle from north is 135°. Special thanks to Dr. Kurt Thome, who performed the MODTRAN runs and provided assistance in data interpretation. The reader should realize that these curves are intended to be only examples; the specifics of any particular atmospheric condition will be different, and other atmospheric modeling programs, e.g., 6S (Vermote et al., 1997), may yield somewhat, but not vastly, different curves.

The molecular absorption bands of water and carbon dioxide cause "deep" absorption features that, in two bands near 1.4 μm and 1.9 μm, completely block transmission of radiation. These spectral regions, therefore, are avoided for remote sensing of the earth's surface. They can be useful, however, for the detection of cirrus clouds, which are not easily distinguished from lower altitude clouds or surface features, at other wavelengths (Gao *et al.*, 1993). Since cirrus clouds are above most of the atmosphere's water vapor, they return a signal to a high-altitude sensor, while the signals from the earth's surface and lower altitude clouds are absorbed in the 1.4 μm band. This application is the motivation for the MODIS band 26 (Table 1-1). The water absorption bands near 0.9 μm and 1.1 μm are much narrower, but can still block the surface signal significantly in narrowband remote sensors and can reduce the signal in broadband sensors when the water vapor content of the atmosphere is high.

The transmittance generally decreases towards the blue spectral region. This is largely the result of light being scattered out of the solar path by air molecules, whose diameter is less than λ/π (Slater, 1980). The magnitude of scattered light is approximated by pure *Rayleigh scattering* with a λ^{-4} dependence on wavelength. The short wavelength blue light is scattered out of the direct transmitting path more than the longer wavelength red light, which explains the red sky at sunrise and sunset, when the solar path through the atmosphere is greatest. In an atmosphere containing aerosols and particulates (smoke, smog, dust, haze, fog), *Mie scattering* is also present. For particle sizes larger than about $2\lambda/\pi$, the magnitude of Mie scattering does not depend on wavelength. Thick clouds are white from above because of Mie scattering by water droplets. For particle sizes in the range from about λ/π to about $2\lambda/\pi$, there is a wavelength dependence that is less than that of Rayleigh scattering. Real atmospheres with a full range of molecular, aerosol, and particulate sizes exhibit a combination of Rayleigh and Mie scattering.

The spectral effect of atmospheric transmission on the solar radiation is shown in Fig. 2-5. The atmosphere significantly alters the spectral irradiance before it arrives at the earth. Mathematically, the irradiance E_λ on a plane perpendicular to the solar path and at the earth's surface is given by,

$$\textit{earth's surface:} \quad E_\lambda = \tau_s(\lambda) E_\lambda^0 \tag{2-5}$$

where τ_s is the solar path atmospheric transmittance. Note that, by the definition of transmittance, E_λ must be less than or equal to E_λ^0.

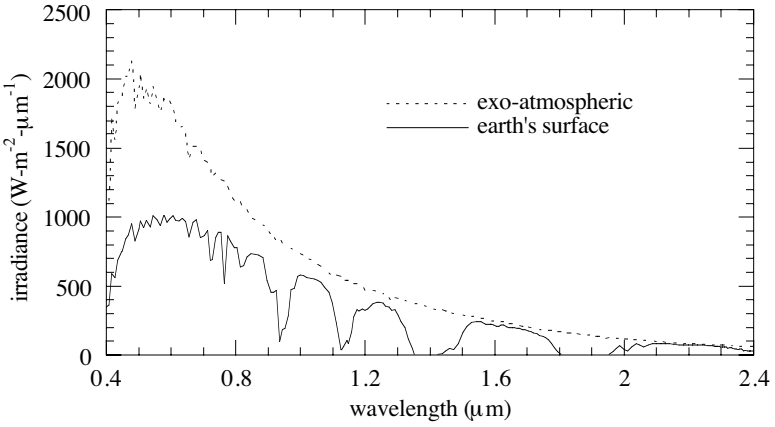


FIGURE 2-5. Solar irradiance in the visible and shortwave IR regions (for a solar elevation angle of 45°), above the atmosphere and at the earth's surface. The ratio of these two curves is the path transmittance depicted in Fig. 2-4.

With the exceptions of cast shadows or clouds, E_λ^0 can be assumed to be constant across the *GFOV* of a sensor such as ETM+. The irradiance at the surface depends on the incident angle, being a maximum if the surface is perpendicular to the incident angle, and less as the angle decreases. The decrease varies as the cosine of the angle, which can be calculated by a dot product of two vectors (Fig. 2-6) (Horn, 1981). The incident irradiance E_λ in Eq. (2-5) must then be modified to account for terrain shape as follows,

$$\textit{earth's surface:} \quad \begin{aligned} E_\lambda(x,y) &= \tau_s(\lambda) E_\lambda^0 \boldsymbol{n}(x,y) \bullet \boldsymbol{s} \\ &= \tau_s(\lambda) E_\lambda^0 \cos[\theta(x,y)] \end{aligned} \tag{2-6}$$

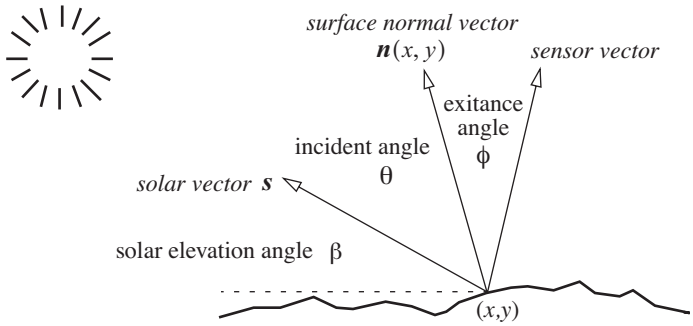


FIGURE 2-6. The geometry of solar direct irradiance on the earth's surface. The unit length vector $\boldsymbol{s}$ *points to the sun and the unit length vector* $\boldsymbol{n}(x,y)$ *is normal to the surface. The solar elevation angle is* β *and the solar zenith angle is* $90° - \beta$*. The solar incidence angle to the surface is* θ *and the exitance angle from the surface normal toward the sensor is* ϕ*. The cosine of the angle* θ *is given by the vector dot product* $\boldsymbol{n}(x,y) \bullet \boldsymbol{s}$ *. For simplicity in this diagram, the surface normal is assumed to lie in the vertical plane through the solar vector, but the dot product calculation is valid for any two vectors. Note this terrain-related effect does not involve the view angle of the sensor.*

The next energy transfer occurs upon reflectance at the earth's surface. The irradiance downward onto a *Lambertian* surface is converted to the *surface radiance* L_λ leaving the surface with the aid of a geometric factor π and a *diffuse spectral reflectance* ρ,

$$\textit{earth's surface:} \quad \begin{aligned} L_\lambda(x,y) &= \rho(x,y,\lambda) \frac{E_\lambda(x,y)}{\pi} \\ &= \rho(x,y,\lambda) \frac{\tau_s(\lambda) E_\lambda^0}{\pi} \cos[\theta(x,y)] \, . \end{aligned} \tag{2-7}$$

Like transmittance, reflectance is, by definition, unitless and between zero and one. Example spectral reflectance curves for some natural earth surface materials were shown in Chapter 1. Reflectance varies with wavelength and spatial location but does not depend on the view (sensor)

direction for a truly Lambertian surface.[2] Deviation from this simple model by real materials is expressed in a *Bi-directional Reflectance Distribution Function (BRDF)*, which can be measured for a particular material by finding the ratio of outgoing radiance to incoming irradiance, as a function of incident and view angles. The quantity $\rho(x, y, \lambda)/\pi$ in Eq. (2-7) is then replaced by the BRDF (Schott, 1996).

The outgoing radiance from the earth's surface traverses the atmosphere to the sensor. The transmittances of the atmosphere along a nadir view path and along a view path at 40° from nadir are shown in Fig. 2-7. These curves differ from the one in Fig. 2-4 only by virtue of different path lengths through the atmosphere. Sensors such as Landsat TM with a small *FOV* show little within-scene atmospheric variation (other than localized clouds, smoke, and haze) over areas with moderate terrain relief, since the path through the atmosphere is nearly constant for all pixels. In areas of high terrain relief, even Landsat images can show pixel-to-pixel atmospheric variation due to pixel-to-pixel altitude differences (Proy *et al*., 1989). Wide *FOV* sensors, such as the AVHRR, can show considerable scan-angle effects due to changes in the atmospheric path in the cross-scan direction. Note that the differences among the curves in Fig. 2-4 and Fig. 2-7 are relatively greater at shorter wavelengths. Visible and NIR imagery will therefore be particularly sensitive to changes in view angle across the *FOV* (Plate 2-1). Atmospheric view path and vegetation canopy radiance variation with view angle are the rationale for the MISR sensor that records images from nine cameras in four spectral bands at different in-track view angles (Plate 2-2).

We now must modify Eq. (2-7) according to the *view path* transmittance, $\tau_v(\lambda)$, to obtain the *at-sensor radiance*,

$$\textit{at-sensor:} \quad \begin{aligned} L_\lambda^{su} &= \tau_v(\lambda) L_\lambda \\ &= \rho(x, y, \lambda) \frac{\tau_v(\lambda)\tau_s(\lambda)E_\lambda^0}{\pi} \cos[\theta(x, y)] \, . \end{aligned} \tag{2-8}$$

This component carries the signal of interest, namely the spatial-spectral reflectance distribution, $\rho(x, y, \lambda)$.

Surface-reflected, atmosphere-scattered component (L_λ^{sd})

The sensor also sees radiance arising from radiation that is scattered *downward* by the atmosphere ("skylight") and then reflected at the earth upward into the *IFOV* of the pixel of interest. This term, $L_\lambda^{sd}(x, y)$, is responsible for the commonly-observed fact that shadows are not totally dark. The

2. A Lambertian surface exhibits equal radiance in all directions. Visually, we say it appears equally bright at any view angle. Such a surface is also termed *perfectly diffuse*, with no mirror-like specular reflection. Many natural surfaces are approximately Lambertian within a limited range of view angles, typically 20°–40°; as the view angle increases beyond that, most materials become non-Lambertian and show unequal reflectance in different directions. This property can be measured by sensors such as the Muli-angle Imaging SpectroRadiometer (MISR) to better characterize the surface radiative characteristics (Diner *et al*., 1989).

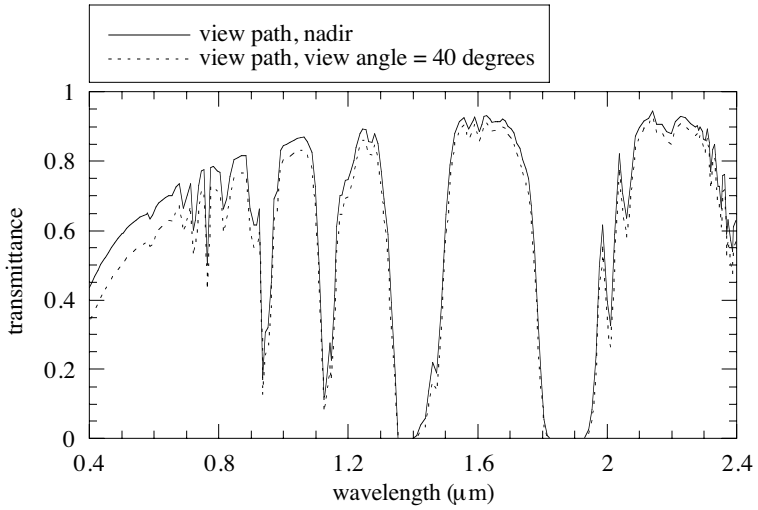


FIGURE 2-7. Atmospheric path transmittance as viewed by a satellite sensor. View angle is defined as the angle from nadir. Note the similarity in shape, but differences in magnitude, compared to Fig. 2-4. The transmittance for a view angle of 40° is less because the path through the atmosphere is longer, leading to more scattering and absorption.

reflected-skylight term is proportional to the diffuse surface reflectance, ρ, and the irradiance at the surface due to skylight, E_λ^d. This quantity is used because it is directly measurable with ground-based instruments. We accommodate the possibility that the sky may not be entirely visible from the pixel of interest due to intervening topography with a factor, $F(x, y)$,[3]

$$\textit{at-sensor: } L_\lambda^{sd} = F(x,y)\rho(x,y,\lambda)\frac{\tau_v(\lambda)E_\lambda^d}{\pi}. \tag{2-9}$$

Path-scattered component (L_λ^{sp})

The path radiance term is a combination of molecular *Rayleigh scattering*, which varies with wavelength as λ^{-4} , and aerosol and particulate *Mie scattering,* which depends less strongly, or not at all, on wavelength. The combined effect of Rayleigh and Mie scattering in a clear atmosphere results in a net wavelength dependence of between λ^{-2} and $\lambda^{-0.7}$ (Curcio, 1961).

3. *F* is the *fraction* of the sky hemisphere that is visible from the pixel of interest. For completely flat terrain, *F* equals one. A detailed description and calculated examples are given in Schott (1996).

Path radiance can vary within a scene, for example, between a rural and an urban area, or in the presence of smoke plumes from fires. It also will vary with view angle, which is particularly important for a wide *FOV* sensor (such as AVHRR) or a sensor pointed off-nadir (such as SPOT). For scenes of homogeneous landscapes and for nadir-viewing sensors with relatively small *FOV*s (such as TM or ETM+), the path radiance is reasonably assumed to be a constant over the entire scene, and we simply write this term as L_λ^{sp} .

Total at-sensor, solar radiance (L_λ^s)

The total at-sensor, solar radiation is the sum of the three components described previously,

at-sensor:

$$\begin{aligned} L_\lambda^s(x,y) &= L_\lambda^{su}(x,y) + L_\lambda^{sd}(x,y) + L_\lambda^{sp} \\ &= \rho(x,y,\lambda)\frac{\tau_v(\lambda)\tau_s(\lambda)E_\lambda^0}{\pi}\cos[\theta(x,y)] + F(x,y)\rho(x,y,\lambda)\frac{\tau_v(\lambda)E_\lambda^d}{\pi} + L_\lambda^{sp} \\ &= \rho(x,y,\lambda)\frac{\tau_v(\lambda)}{\pi}\{\tau_s(\lambda)E_\lambda^0\cos[\theta(x,y)] + F(x,y)E_\lambda^d\} + L_\lambda^{sp} \ . \end{aligned} \tag{2-10}$$

The essence of Eq. (2-10) is that:

- the total spectral radiance received by the sensor is linearly proportional to the surface diffuse reflectance, modified by
- a multiplicative, spatially- and spectrally-variant factor that depends on the terrain shape, and
- an additive, spatially-invariant, spectrally-variant term due to view path scattering.

As an example, Fig. 2-8 shows how the at-sensor radiance components vary with wavelength for a reflecting surface described by Kentucky Bluegrass (Fig. 1-7). The sum of these two components is the total radiance seen by a satellite or high altitude aircraft sensor and is compared to the spectral profile of an AVIRIS pixel containing grass and trees in Fig. 2-9. Even though there has been no attempt to emulate the atmospheric conditions, surface reflectance magnitude, or solar angles of this particular AVIRIS scene, the data are remarkably similar in shape to the results of the MODTRAN simulation. The MODTRAN atmospheric models are evidently realistic, and if *actual* atmosphere values are used in predicting the radiances of a particular image with a program such as MODTRAN, good agreement is possible. The problem, of course, is that such calibration data are almost never available for a given scene. Measurements of the atmosphere and ground reflectance in conjunction with aircraft or satellite overpasses require substantial planning and effort, and are subject to the vagaries of weather and equipment failure. Therefore, there has been a great amount of interest in *image-based* atmospheric calibration techniques (Teillet and Fedosejevs, 1995).

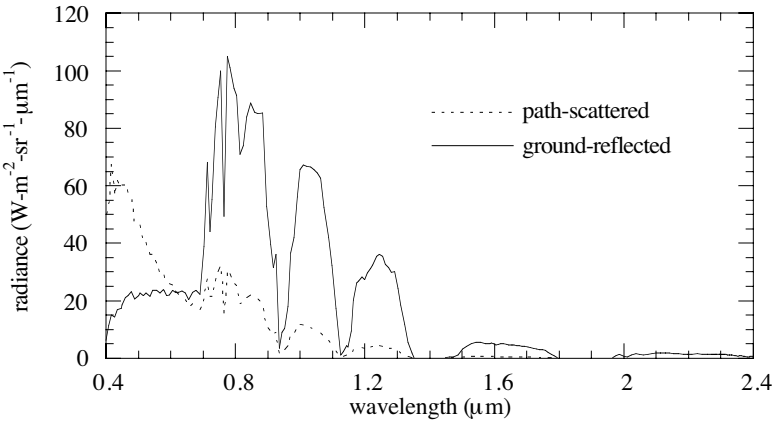


FIGURE 2-8. The path-scattered and ground-reflected components of the total upwelling radiance seen by a satellite sensor for a surface reflectance of Kentucky Bluegrass. These components, as defined in MODTRAN, are related to the terms of Eq. (2-10) as follows. The path-scattered component is L_λ^{sp}, plus radiation that is reflected by the surface in a direction other than toward the sensor (remember, we assume the surface is perfectly diffuse and reflects equally in all directions), and is then scattered into the IFOV (we have not included this term in our discussion). The strong increase in the path-scattered component below 0.7 μm is due to molecular scattering and is primarily the L_λ^{sp} term, since the surface reflectance here is relatively low. Above 0.7 μm, the influence of the reflected and then scattered component is apparent. The ground-reflected component is the sum of L_λ^{su} and L_λ^{sd}. In the ground-reflected component, little information about the grass signature is seen until above 0.7 μm, where the reflectance becomes relatively high. The ground-reflected component only exceeds the path-scattered component above 0.7 μm, but both contain information about the signal (grass reflectance). Note the atmospheric water vapor absorption bands near 0.9, 1.1, 1.4, and 1.9 μm (compare to Fig. 2-4).

The at-sensor spectral radiance and the spectral reflectance of Kentucky bluegrass are dramatically different (Fig. 2-10)! How then is it possible to even hope to use remote sensing to recognize different terrestrial materials? The major reason why this drastic alteration of spectral "signatures" is not, in fact, a disaster, is that many image processing algorithms rely only on *relative spectral differences* among pixels. It is clear, however, that *remote sensing data must be corrected for atmospheric, topographic, and solar effects if they are to be compared to a library of spectral reflectance curves*. Furthermore, *relative* atmospheric correction is needed if data signatures from one image date are to be compared to those from another date. These aspects of data calibration are discussed in detail in Chapter 7.

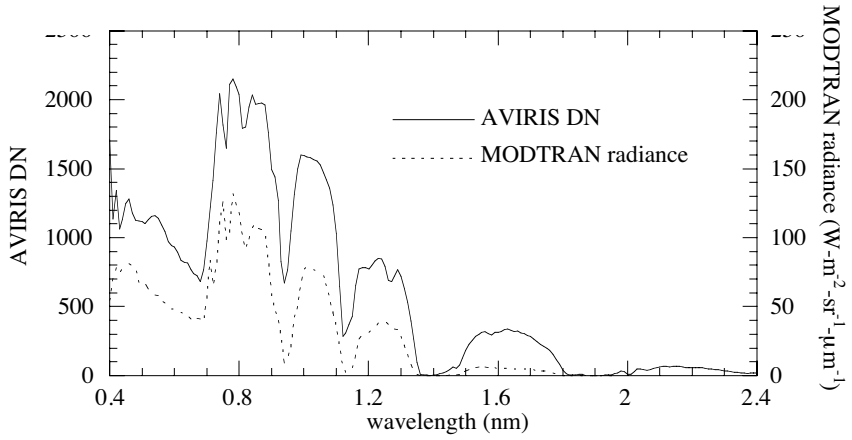


FIGURE 2-9. The spectral response of Kentucky Bluegrass as predicted by the MODTRAN model and a plot of a mixed grass and trees response from the AVIRIS image of Palo Alto (Plate 1-3). The shapes of the curves are quite similar, even though the MODTRAN model parameters have no direct relation to this particular AVIRIS data.

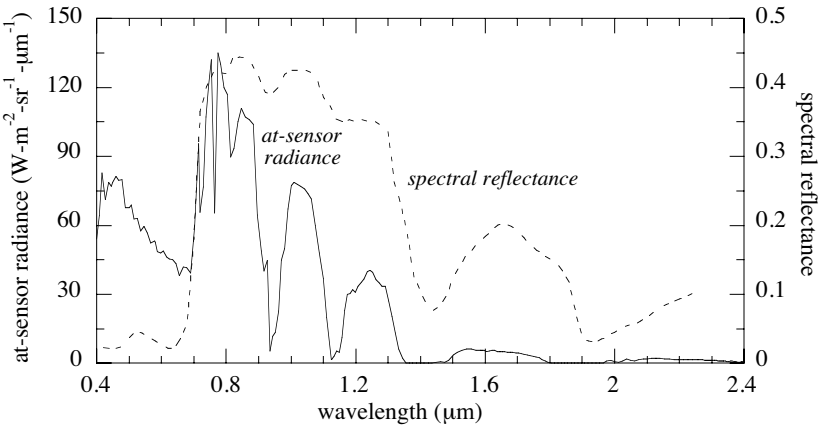


FIGURE 2-10. Comparison of the reflectance and remotely-sensed radiance spectral signals for grass. The characteristic features in the reflectance curve are masked and modified by all the factors involved in optical remote sensing. The most significant feature however, the "red edge" near 0.71 μm, is preserved.

2.2.3 Image Examples in the Solar Region

Some image examples are used in this section to illustrate the theory and models of optical radiation transfer in the solar reflective portion of the spectrum.

Terrain shading

The cosine factor arising from terrain relief (Eq. (2-10)) can be important to the spatial content of images as demonstrated in Fig. 2-11. The *Digital Elevation Model (DEM)*[4] has a *GSI* of 30m, matching the TM *GSI*. A *shaded relief* image is created from this DEM by calculating the $\cos[\theta(x,y)]$ term at each pixel and assuming that the surface reflectance is diffuse (i.e., the surface is Lambertian) and constant everywhere. In this case, we have set the solar elevation and azimuth angles to correspond to those for the particular TM image in Fig. 2-11. The shading caused by the terrain relief is clearly predictable with the DEM data.

The differences that remain between the contrast-stretched TM image and the shaded relief image of Fig. 2-11 are partially due to the reflectance spatial variations that are not modeled in the relief image. The modeling of vegetation reflectance from plant canopies is quite complex and is a matter of considerable research interest. The reader is referred to an extensive review (Goel, 1988) as a starting point for this subject and to Liang (2004) for more details.

The influence of terrain geometry on remote sensing measurements has been discussed in detail in Sjoberg and Horn (1983) and Proy *et al*. (1989), including the possibility of reflectance by neighboring ground elements in mountainous terrain into the *IFOV* at a pixel of interest. Looking at Eq. (2-10), we see that the image spatial content results from the product of two spatially-varying terms, the reflectance and the terrain-dependent cosine factor (assuming the down-scattered term is small). Either may be considered to *spatially modulate* the other. This relationship is useful in certain algorithms for image fusion, as described in Chapter 8.

Shadowing

Analysis of the DEM in Fig. 2-11 provides information on shadowing in the TM image (Dubayah and Dozier, 1986; Giles *et al*., 1994). For example, the *self-shadowed* points can be found directly from the shaded relief image (see Exercise 2-1); these are terrain segments that face away from the solar irradiance direction. With a "line-of-sight" algorithm, one can also find the pixels that lie within *projected shadows* (Fig. 2-12). Because of the mild terrain relief and high solar elevation for the TM image, there are relatively few projected shadows. Both types of shadow regions may be used to estimate atmospheric path radiance if a reasonable estimate is made of the reflectance within the shadowed area.

4. A DEM is a regular spatial grid of elevation values. It can be created in a variety of ways, for example, indirectly via analysis of stereo imagery (see Chapter 8) or satellite Interferometric SAR (InSAR) data, or directly from airborne laser LIght Detection And Ranging (LIDAR) measurements.

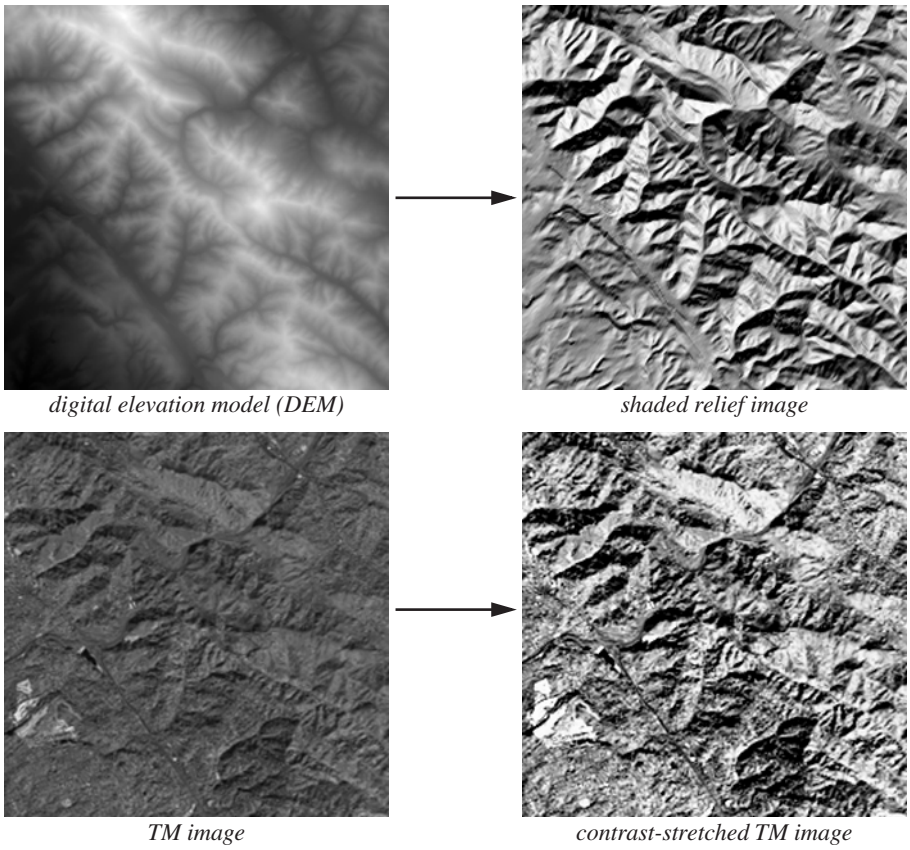
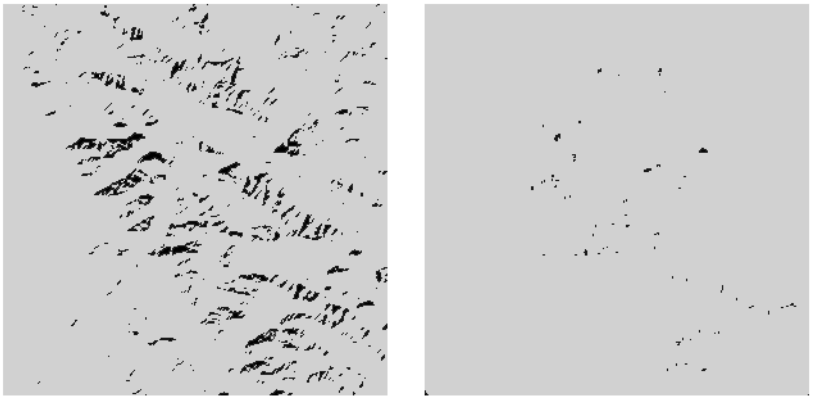


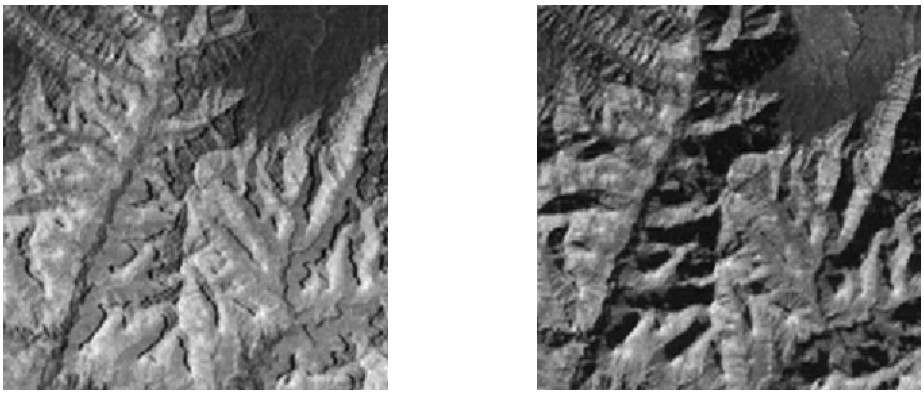
FIGURE 2-11. The influence of terrain relief on image structure is depicted with a co-registered DEM and TM band 4 image near Berkeley, California. Elevation is coded as brightness in the DEM image. The solar elevation angle of 35° and azimuth angle of 151° for the October 25, 1984, TM image are used to create a shaded relief image from the DEM. The contrast of the TM image is significantly lower than that of the shaded relief image because of atmospheric contrast reduction and the fact that the displayed image contrast is controlled by the high reflectance man-made feature in the lower left. If the contrast of the TM image is adjusted to match that of the shaded relief image, the similarity of the two is evident. The residual differences are due to surface cover reflectance variation in the TM image and reflected skylight within shadows, neither of which are modeled in the shaded relief image. (The DEM is courtesy of William Acevedo and Len Gaydos of the U.S. Geological Survey.)



self-shadowed pixels *projected shadows*

FIGURE 2-12. Maps of the self-shadowed pixels and projected shadows for the solar angles and DEM of Fig. 2-11. (The projected shadows map was produced by Justin Paola, Oasis Research Center.)

An image's appearance can vary dramatically with different solar angles, as illustrated in Fig. 2-13. The difference in acquisition dates of four months creates a dramatically different image because the changes in solar elevation and, to a lesser extent, azimuth cause changes in the $\cos[\theta(x,y)]$ term of Eq. (2-10) and in shadowing.



June 11, 1981 *October 20, 1980*

FIGURE 2-13. Landsat MSS images of the Grand Canyon, Arizona, acquired on two dates. The lower sun elevation of 38° for the October image dramatically increases the shadowing in the Canyon, compared to the June image with a sun elevation of 65°.

Atmospheric correction

A portion of a Landsat TM scene near Oakland, California, is shown for bands 1 through 4 in Fig. 2-14. These data, although not calibrated for sensor gain and offset (Chapters 3 and 7), nevertheless indicate the general properties of atmospheric influence as a function of wavelength. A common technique for atmospheric correction of multispectral imagery uses a "dark object" as a calibration target (Chavez, 1988). The dark object may be a region of cast shadow or, as in this case, a deep body of water. It is *assumed* that the dark object has uniformly zero radiance for all bands, and that any non-zero measured radiance must be due to atmospheric scattering into the object's pixels.[5] For the images in Fig. 2-14, the Briones Reservoir is a good choice for a dark object (it is significantly darker than the San Pablo Reservoir at shorter wavelengths). Averaging the darkest pixels in the Reservoir yields *DN*s of 53, 20, 11, and 14 for bands 1 through 4. It is tempting at this point to remove the atmospheric path scattering bias by simply subtracting these *DN* values from every pixel in the corresponding band, given that the at-sensor radiance-to-*DN* calibration is linear. However, that would not correct the data for the sensor gain and would therefore not be a complete calibration. Calibration of the raw image *DN*s to at-sensor radiance and correction to reflectance are discussed in Chapter 7.

2.3 Midwave to Thermal Infrared Region

At longer wavelengths, beyond the SWIR and into the MWIR spectral region, the importance of solar radiation declines and that of emitted thermal radiation increases for objects that are Lambertian reflectors (Fig. 2-2). At the longer wavelengths of the TIR, direct solar radiation is not a factor compared to self-emitted thermal radiation, other than solar-induced heating of the surface. The only exception to these statements is if the objects of interest are specular reflectors; in that case the solar reflected component may exceed the emitted component, even in the TIR (Slater, 1996). We will ignore this relatively rare circumstance in the remainder of this section.

2.3.1 Thermal Radiation

Every object at a temperature above absolute zero (0K) emits thermal radiation due to kinetic energy of molecules within the object. The radiation obeys Planck's equation (Eq. (2-1)) *if* the object is a perfect emitter and absorber, i.e., a blackbody. Real objects are not perfect emitters or absorbers, and Eq. (2-1) is modified by an emission efficiency factor, the *emissivity*, which is generally a function of wavelength. Real materials are sometimes referred to as *greybodies*, but their spectral radiant exitance may not follow the blackbody curve because of the wavelength dependence of emissivity.

5. If any data are available on the actual reflectance of the "dark object," then it should be used to obtain a better estimate of the at-sensor path radiance for this correction (Teillet and Fedosejevs, 1995).