Chapter 2

Physics

Wan Bakx
Klaus Tempfli
Valentyn Tolpekin
Tsehaie Woldai

Introduction

Geospatial Data Acquisition (GDA) challenges us to make choices: on which one of the many sensors available should the agronomist rely for accurate yield predictions? If he or she chooses a sensor producing several images, such as a multispectral scanner, which image or which combination of images to use? How to properly process sensor recordings to increase the chances of a correct interpretation? When interpreting a colour image, what causes the sensation *red*? Instead of writing a thick book of recipes to answer such questions for every application, we can better review the physics of RS. Understanding the basics of electromagnetic (EM) radiation will help you in making more profound choices and enable you to deal with sensors of the future.

A standard photograph is an image of an object or scene that very closely resembles direct sensing with our eyes. The sensation of colour is caused by EM radiation. Red, green and blue relate to forms of radiation that we commonly refer to as light. *Light* is EM radiation that is visible to the human eye. As we are interested in Earth Observation, our light source is the Sun. The Sun emits light, the Earth's surface features reflect light, and the photosensitive cells (cones and rods) in our eyes detect light. When we look at a photograph, it is the light reflected from the photograph that allows us to interpret the photograph. Light is not the only form of radiation from the Sun and other bodies. The sensation *warm*, for example, is is the result of thermal emissions. Another type of emissions, ultraviolet (UV) radiation, triggers our body to generate vitamin D and also produces a suntan.

This chapter explains the basic characteristics of EM radiation, its sources and what we call the EM spectrum, the influence of the atmosphere on EM radiation, interactions of EM radiation with the Earth's surface, and the basic principles of sensing EM radiation and generic properties of sensors.

2.1 Waves and photons

EM wave

EM radiation can be modelled in two ways: by waves, or by radiant particles called photons. The first publications on the wave theory date back to the 17th century. According to the wave theory, light travels in a straight line (unless there are external influences) with its physical properties changing in a wave-like fashion. Light waves have two oscillating components: an electric field and a magnetic field. We refer, therefore, in this context to electromagnetic waves. The two components interact—an instance of a positive electric field coincides with a moment of negative magnetic field (Figure 2.1). The wave behaviour of light is common to all forms of EM radiation. All EM waves travels at the speed of light, which is approximately equal to 2.998×10^8 m s^{-1}. This is fast, but the distances in space are literally astronomical: it takes eight minutes for the sunlight to reach the Earth, thus when we see, a sunrise, for example, the light particles actually left the Sun that much earlier. Because they travel in a straight line, we use the notion of light rays in optics.

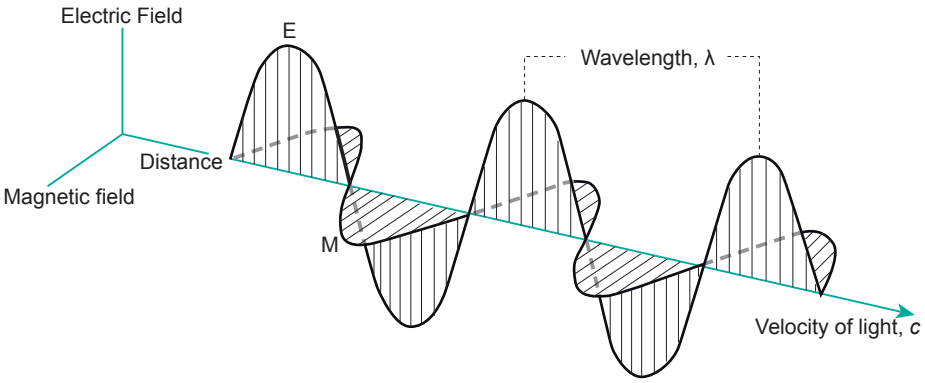


Figure 2.1
The two oscillating components of EM radiation: an electric and a magnetic field.

A sine wave can be described as:

$$e = \alpha sin\left(\frac{2\pi}{\lambda}x + \varphi\right). \tag{2.1}$$

where α is the amplitude of the wave, φ is the phase (it depends on time) and λ is the *wavelength*. The wavelength is a differentiating property of the various types of EM radiation and is usually measured in micrometres (1 μm $= 10^{-6}$ m). Blue light is EM radiation with a wavelength of around 0.45 μm. Red light, at the other end of the colour spectrum of a rainbow, has a wavelength of around 0.65 μm (Figure 2.2). Electromagnetic radiation outside the range 0.38–0.76 μm is not visible to the human eye.

wavelength

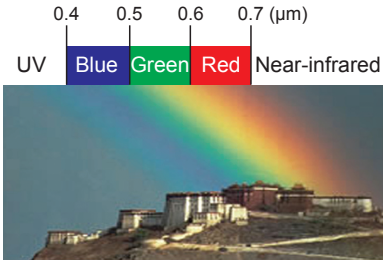


Figure 2.2
The spectrum of light.

We call the amount of time needed by an EM wave to complete one cycle the *period* of

the wave. The reciprocal of the period is called the *frequency* of the wave. Thus, the frequency ν is the number of cycles of the wave that occur in one second. We usually measure frequency in hertz (1 Hz = 1 cycle s^{-1}). Since the speed of light c is constant, the relationship between wavelength and frequency is:

period and frequency

$$c = \lambda \times \nu. \tag{2.2}$$

Obviously, a short wavelength implies a high frequency, while long wavelengths are equivalent to low frequencies. Blue light has a higher frequency than red light (Figure 2.3).

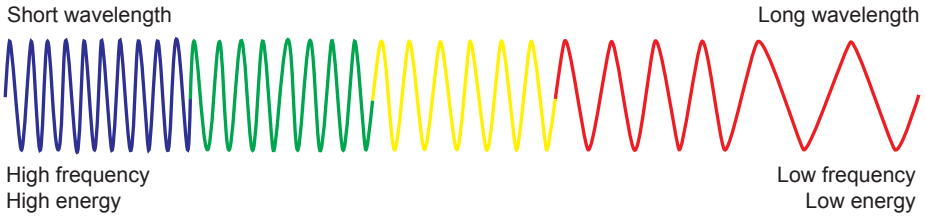


Figure 2.3
Relationship between wavelength, frequency and energy.

Although wave theory provides a good explanation for many EM radiation phenomena, for some purposes we can better rely on particle theory, which explains EM radiation in terms of photons. We take this approach when quantifying the radiation detected by a multispectral sensor (see Section 2.6). The amount of energy carried by a photon of a specific wavelength is:

energy of a photon

$$Q = h \times \nu = h \times \frac{c}{\lambda}, \tag{2.3}$$

where Q is the energy of a photon measured in joules (J) and h is Planck's constant ($h \approx 6.626 \times 10^{-34}$ J s).

The energy carried by a single photon of light is just sufficient to excite a single molecule of a photosensitive cell of the human eye, thus contributing to vision. It follows from Equation 2.3 that long-wavelength radiation has a low level of energy while short-wavelength radiation has a high level. Blue light has more energy than red light (Figure 2.3). EM radiation beyond violet light is progressively more dangerous to our body as its frequency increases. UV radiation can already be harmful to our eyes, so we wear sunglasses to protect them. An important consequence of Formula 2.3 for RS is that it is more difficult to detect radiation of longer wavelengths than radiation of shorter wavelengths.

2.2 Sources of EM radiation

All matter with a temperature above absolute zero emits EM radiation because of molecular agitation. *Planck's law of radiation* describes the amount of emitted radiation per unit of solid angle in terms of the wavelength and the object's temperature:

Planck

$$L(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k T}} - 1}, \tag{2.4}$$

where h is the Planck's constant, $k \approx 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant, λ is the wavelength (m), c is the speed of light and T is the absolute temperature (K). $L(\lambda, T)$ is called the spectral radiance.

radiometric units

We can use different measures to quantify radiation. The amount of radiative energy is commonly expressed in joules (J). We may, however, be interested in the radiative energy per unit of time, called the *radiant power*. We measure the power in watts (W = J s^{-1}). *Radiant emittance* is the power emitted from a surface; it is measured in watts per square metre (W m^{-2}). *Spectral radiant emittance* characterizes the radiant emittance per wavelength; it is measured in W m^{-2} μm^{-1} (this is the unit used in Figure 2.4). *Radiance* is another quantity frequently used in RS. It is the radiometric quantity that describes the amount of radiative energy being emitted or reflected in a specific direction per unit of projected area per unit of solid angle and per unit of time. Radiance is usually expressed in W sr^{-1} m^{-2} (sr is steradian, unit of solid angle). *Spectral radiance* used in Equation 2.4 is radiance per wavelength and is measured in W sr^{-1} m^{-2} μm^{-1}. *Irradiance* is the amount of incident radiation on a surface per unit of area and per unit of time. Irradiance is usually expressed in W m^{-2}.

Planck's law of radiation is only applicable to black bodies. A black body is an idealized object with assumed extreme properties that helps us when explaining EM radiation. A black body absorbs 100% of incident EM radiation; it does not reflect anything and thus appears perfectly black. Because of its perfect absorptivity, a black body emits EM radiation at every wavelength (Figure 2.4). The radiation emitted by a black body is called black-body radiation. Real objects can re-emit some 80 to 98% of the radiation received. The emitting ability of real objects is expressed as a dimensionless ratio called emissivity $\epsilon(\lambda)$ (with values between 0 and 1). The *emissivity* of a material depends on the wavelength; it specifies how well a real body made of that material emits radiation as compared to a black body.

The Sun behaves similarly to a black body. It is a prime source of the EM radiation that plays a role in Earth Observation, but it is not the only source. The global mean temperature of the Earth's surface is 288 K and over a finite period the temperatures of objects on the Earth rarely deviate much from this mean. The surface features of the Earth therefore emit EM radiation. Solar radiation constantly replenishes the energy that the Earth radiates into space. The Sun's temperature is about 6000 K. Planck's law of radiation is illustrated in Figure 2.4 for the approximate temperature of the Sun (about 6000 K) and the ambient temperature of the Earth's surface (288 K). The figure shows that for very hot surfaces (e.g. the Sun), spectral emittance of a black body peaks at short wavelengths. For colder surfaces, such as the Earth, spectral emittance peaks at longer wavelengths. This behaviour is described by *Wien's displacement law*:

Wien's displacement law

$$\lambda_{max} = \frac{b}{T}, \tag{2.5}$$

where λ_{max} is the wavelength of the radiation maximum (μm), T is the temperature (K) and $b \approx$2898 μm K is a physical constant.

We can use Wien's law to predict the position of the peak of the black-body curve if we know the temperature of the emitting object. The temperature of the black body determines the most prominent wavelength of black-body radiation. At room temperature, black bodies emit predominantly infrared radiation. When a black body is heated beyond 4450 K (approximately 4700 °C) emission of light becomes dominant, from red, through orange, yellow, and cyan, (at 6000 K) to blue, beyond which the emitted energy includes increasing amounts of ultraviolet radiation. At 6000 K a black body emits radiation of all visible wavelengths in approximately equal amounts, creating the sensation of white to us. Higher temperatures correspond to a greater contribution of radiation of shorter wavelengths.

black body

The following description illustrates the physics of what we see when a blacksmith heats a piece of iron or what we observe when looking at a candle. The flame appears

light-blue at the outer edge of its core; there the flame is hottest, with a temperature of 1670 K. The centre, with a temperature of 1070 K, appears orange. More generally, flames may burn with different colours (depending on the material being burnt, the surrounding temperature and the amount of oxygen present) and accordingly have different temperatures (in the range of 600 °C to 1400 °C). Colour tells us something about temperature. We can use colour, for example, to estimate the temperature of a lava flow from a safe distance. More generally, if we can build sensors that allow us to detect and quantify EM radiation of different wavelengths (also outside the visible range), we can use RS recordings to estimate the temperature of objects. You may also notice from the black-body radiation curves (Figure 2.4) that the intensity of EM radiation increases with increasing temperature; the total radiant emittance at a certain temperature is the area under the spectral emittance curve.

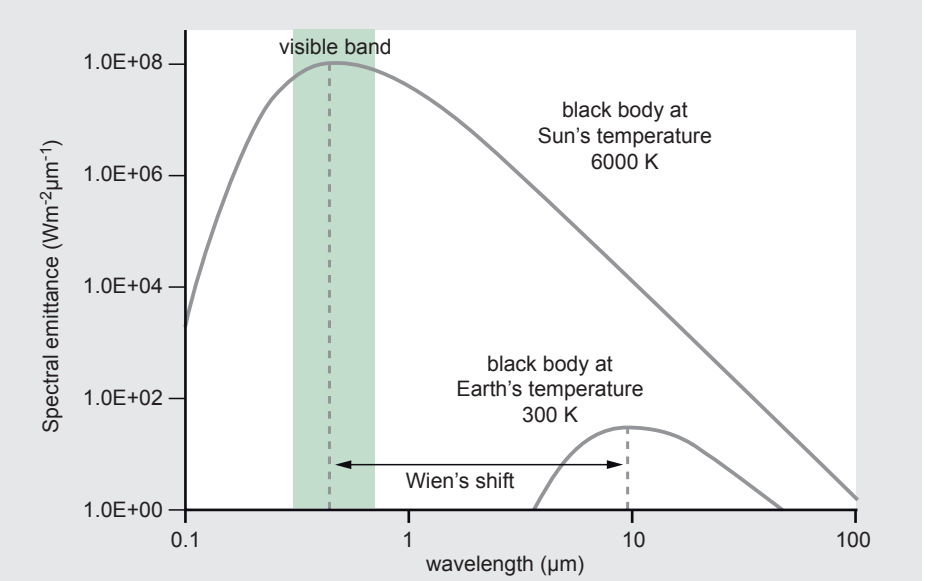


Figure 2.4
Illustration of Planck's law of radiation for the Sun (6000 K) and for the average surface temperature (300 K) the Earth. Note the logarithmic scale for both x- and y-axes. The broken lines mark the wavelength of the emission maxima for the two temperatures.

If you were interested in monitoring forest fires, which typically burn at 1000 K, you could immediately turn to wavelength bands around 2.9 µm, where the radiation maximum for those fires is to be expected. For ordinary land surface temperatures of around 300 K, wavelengths from 8 to 14 µm are most useful.

You can probably now understand why reflectance remote sensing (i.e. based on reflected sunlight) uses short wavelengths in the visible and short-wave infrared, and thermal remote sensing (based on emitted Earth radiation) uses the longer wavelengths in the range 3–14 µm. Figure 2.4 also shows that the total energy (integrated area under the curve) is considerably higher for the Sun than for the cooler Earth's surface. This relationship between surface temperature and total amount of radiation is described by the *Stefan-Boltzmann law*.

Stefan-Boltzmann law

$$M = \sigma T^4, \tag{2.6}$$

where M is the total radiant emittance (W m^{-2}), σ is the *Stefan-Boltzmann constant* ($\sigma \approx 5.6697 \times 10^{-8}$ (W m^{-2} K^{-4}), and T is the temperature in K.

The Stefan-Boltzmann law states that colder objects emit only small amounts of EM radiation. Wien's displacement law predicts that the peak of the radiation distribution will shift to longer wavelengths as the object gets colder. In Section 2.1 you will have

learnt that photons at long wavelengths have less energy than those at short wavelengths. Hence, in thermal RS we are dealing with a small amount of low energy photons, which makes their detection difficult. As a consequence of that, we often have to reduce spatial or spectral resolution when acquiring thermal data, to guarantee an acceptable signal-to-noise ratio.

2.3 Electromagnetic spectrum

We call the total range of wavelengths of EM radiation the *EM spectrum*. Figure 2.2 illustrates the spectrum of visible light; Figure 2.5 illustrates the wider range of EM spectrum. We refer to the different portions of the spectrum by name: gamma rays, X-rays, UV radiation, visible radiation (light), infrared radiation, microwaves, and radio waves. Each of these named portions represents a range of wavelengths, not one specific wavelength. The EM spectrum is continuous and does not have any clear-cut class boundaries.

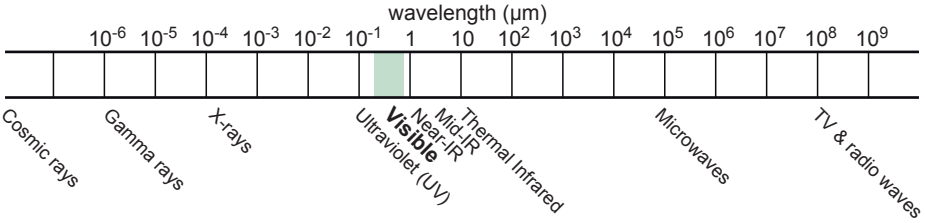


Figure 2.5
The EM spectrum.

Different portions of the spectrum have differing relevance for Earth Observation, both in the type of information that we can gather and the volume of geospatial data acquisition (GDA). The majority of GDA is accomplished by sensing in the visible and infrared range. The UV portion covers the shortest wavelengths that are of practical use for Earth Observation. UV radiation can reveal some properties of minerals and the atmosphere. Microwaves are at the other end of the useful range for Earth Observation; they can, among other things, provide information about surface roughness and the moisture content of soils.

light and colour

The "visible portion" of the spectrum, with wavelengths producing colour, is only a very small fraction of the entire EM wavelength range. We call objects "green" when they reflect predominately EM radiation of wavelengths around 0.54 μm. The intensity of solar radiation has its maximum around this wavelength (see Figure 2.8) and the sensitivity of our eyes is peaked at green-yellow. We know that colour effects our emotions and we usually experience green sceneries as pleasant. We use colour to distinguish between objects and we can use it to estimate temperature. We also use colour to visualize EM radiation we cannot see directly. Section 5.1 elaborates how we can "produce colour" by adequately "mixing" the three primary colours red, green and blue.

near-infrared, short-wave infrared

Radiation beyond red light, with larger wavelengths in the spectrum, is referred to as infrared (IR). We can distinguish vegetation types and the stress state of plants by analysing *near-infrared* (and *mid-infrared*) radiation—this works much better than trying to do so by colour. For example, deciduous trees reflect more near-infrared (NIR) radiation than conifers do, so they show up brighter on photographic film that is sensitive to infrared. Dense green vegetation has a high reflectance in the NIR range, which decreases with increasing damage caused by plant disease (see also Section 2.5.1). Mid-IR is also referred to as short-wave infrared (SWIR). SWIR sensors are used to monitor surface features at night.

Infrared radiation with a wavelength longer than 3 μm is termed thermal infrared (TIR) because it produces the sensation of "heat". Near-IR and mid-IR do not produce a sensation of something being hot. Thermal emissions of the Earth's surface (288 K) have a peak wavelength of 10 μm (see Figure 2.4). A human body also emits "heat" radiation, with a maximum at $\lambda \approx 10$ μm. Thermal detectors for humans are, therefore, designed such that they are sensitive to radiation in the wavelength range 7–14 μm. NOAA's thermal scanner, with its interest in heat issuing from the Earth's surface, detects thermal IR radiation in the range 3.5– 12.5 μm. Object temperature is a kind of quantity often needed for studying a variety of environmental problems, as well as being useful for analysing the mineral composition of rocks and the evapotranspiration of vegetation.

thermal infrared

2.4 Interaction of atmosphere and EM radiation

Before the Sun's radiation reaches the Earth's surface, three RS-relevant interactions in the atmosphere have occurred: absorption, transmission, and scattering. The transmitted radiation is then either absorbed by the surface material or reflected. Before reaching a remote sensor, the reflected radiation is also subject to scattering and absorption in the atmosphere (Figure 2.6).

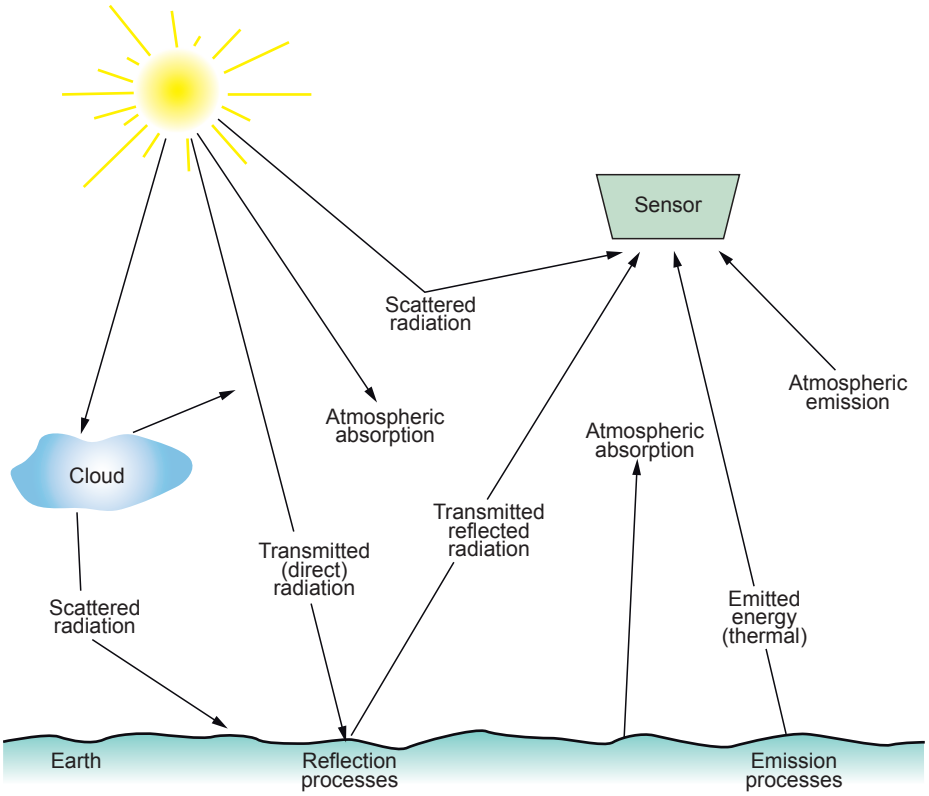


Figure 2.6
Interactions of EM radiation with the atmosphere and the Earth's surface.

2.4.1 Absorption and transmission

As it moves through the atmosphere, EM radiation is partly absorbed by various molecules. The most efficient absorbers of solar radiation in the atmosphere are ozone (O_3), water vapour (H_2O) and carbon dioxide (CO_2).

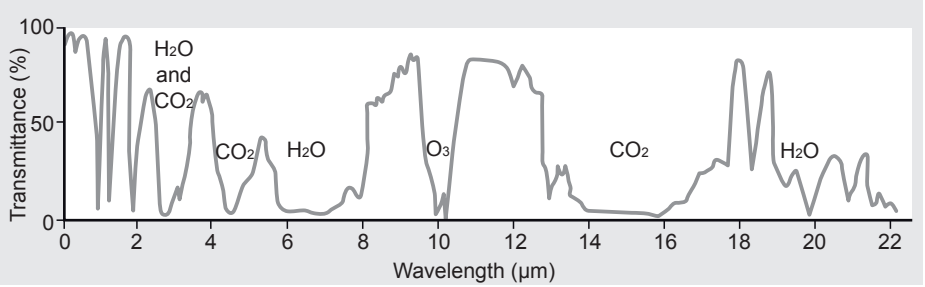


Figure 2.7
Atmospheric transmittance.

atmospheric transmission

Figure 2.7 shows a schematic representation of atmospheric transmission in the wavelength range 0–22 µm. From this figure it can be seen that many of the wavelengths are not useful for remote sensing of the Earth's surface, simply because the corresponding radiation cannot penetrate the atmosphere. Only those wavelengths outside the main absorption ranges of atmospheric gases can be used for remote sensing. The useful ranges are referred to as *atmospheric transmission windows* and include:

- the window from 0.4 to 2 µm. The radiation in this range (visible, NIR, SWIR) is mainly reflected radiation. Because this type of radiation follows the laws of optics, remote sensors operating in this range are often referred to as optical sensors.
- three windows in the TIR range, namely two narrow windows around 3 and 5 µm, and a third, relatively broad window extending from approximately 8 µm to 14 µm.

Because of the presence of atmospheric moisture, strong absorption occurs at longer wavelengths. There is hardly any transmission of radiation in the range from 22 µm to 1 mm. The more or less "transparent" range beyond 1 mm is the microwave range.

Solar radiation observed both with and without the influence of the Earth's atmosphere is shown in Figure 2.8. Solar radiation measured outside the atmosphere resembles black-body radiation at 6000 K. Measuring solar radiation at the Earth's surface shows that there the spectral distribution of the solar radiation is very ragged. The relative dips in this curve indicate the absorption by different gases in the atmosphere. We also see from Figure 2.8 that the total intensity in this range (i.e. the area under the curve) has decreased by the time the solar energy reaches the Earth's surface, after having passed through the atmosphere.

2.4.2 Atmospheric scattering

Atmospheric scattering occurs when particles or gaseous molecules present in the atmosphere cause EM radiation to be redirected from its original path. The amount of scattering depends on several factors, including the wavelength of the radiation in relation to the size of particles and gas molecules, the amount of particles and gases, and the distance the radiation travels through the atmosphere. On a clear day the colours are bright and crisp, and approximately 95% of the sunlight detected by our eyes, or a comparable remote sensor, is radiation reflected from objects; 5% is light scattered in the atmosphere. On a cloudy or hazy day, colours are faint and most of the radiation received by our eyes is scattered light. We may distinguish three types of scattering according to the size of particles in the atmosphere causing it. Each has a different relevance to RS.

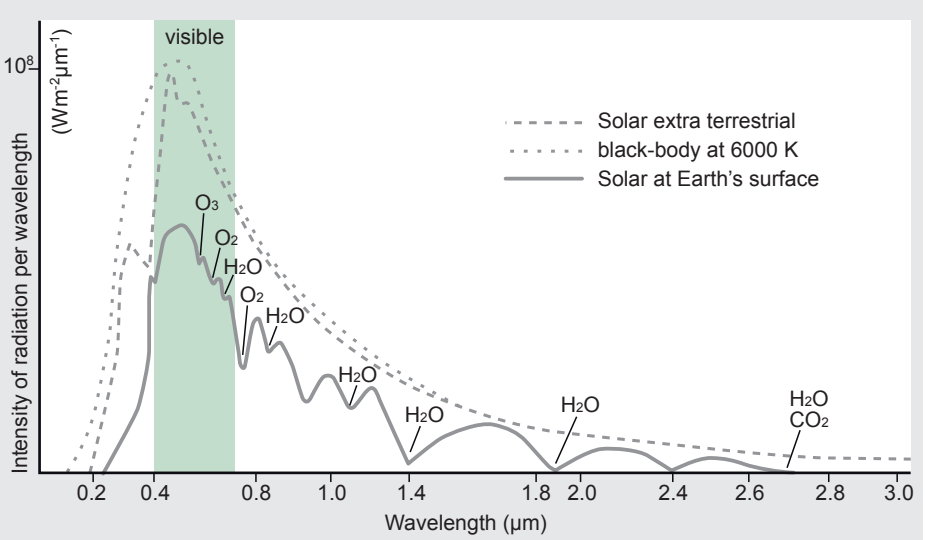


Figure 2.8
Radiation curves of the Sun and a black body at the Sun's temperature.

Rayleigh scattering dominates where electromagnetic radiation interacts with particles that are smaller than the wavelengths of light. Examples of such particles are tiny specks of dust and molecules of nitrogen (NO_2) and oxygen (O_2). Light of shorter wavelengths (e.g. blue) is scattered more than light of longer wavelengths (e.g. red); see Figure 2.9.

Rayleigh scattering

Figure 2.9
Rayleigh scattering is caused by particles smaller than the wavelengths of light and is greater for small wavelengths.

In the absence of particles and scattering, the sky would appear black. During the day, solar radiation travels the shortest distance through the atmosphere; Rayleigh scattering causes a clear sky to be observed as blue. At sunrise and sunset, the sunlight travels a longer distance through the Earth's atmosphere before reaching the surface. All the radiation of shorter wavelengths is scattered after some distance and only the longer wavelengths reach the Earth's surface. As a result we do not see a blue but an orange or red sky (Figure 2.10).

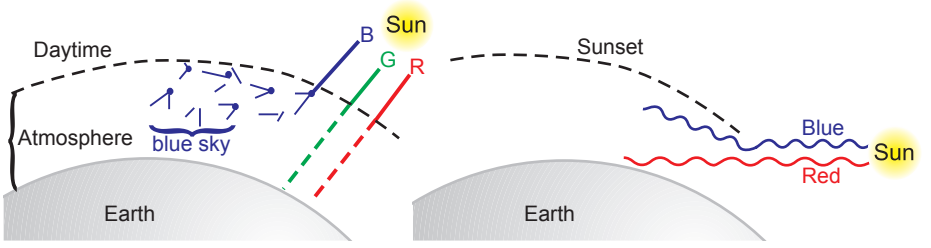


Figure 2.10
Rayleigh scattering causes us to see a blue sky during the day and a red sky at sunset.

Rayleigh scattering disturbs RS in the visible spectral range from high altitudes. It causes a distortion of the spectral characteristics of the reflected light as compared to measurements taken on the ground: due to Rayleigh scattering, the shorter wavelengths are overestimated. This accounts for the blueness of colour photos taken from

high altitudes. In general, Rayleigh scattering diminishes the "crispness" of photos and thus reduces their interpretability. Similarly, Rayleigh scattering has a negative effect on digital classification using data from multispectral sensors.

Mie scattering

Mie scattering occurs when the wavelength of EM radiation is similar in size to particles in the atmosphere. The most important cause of Mie scattering is the presence of aerosols: a mixture of gases, water vapour and dust. Mie scattering is generally restricted to the lower atmosphere, where larger particles are more abundant, and it dominates under overcast, cloudy conditions. Mie scattering influences the spectral range from the near-UV up to mid-IR range and has a greater effect on radiation of longer wavelengths than Rayleigh scattering.

non-selective scattering

Non-selective scattering occurs when particle sizes are much larger than the radiation wavelength. Typical particles responsible for this effect are water droplets and larger dust particles. Non-selective scattering is independent of the wavelength within the optical range. The most prominent example of non-selective scattering is that we see clouds as white bodies. A cloud consists of water droplets; since they scatter light of every wavelength equally, a cloud appears white. A remote sensor like our eye cannot "see through" clouds. Moreover, clouds have a further limiting effect on optical RS: clouds cast shadows (Figure 2.11).

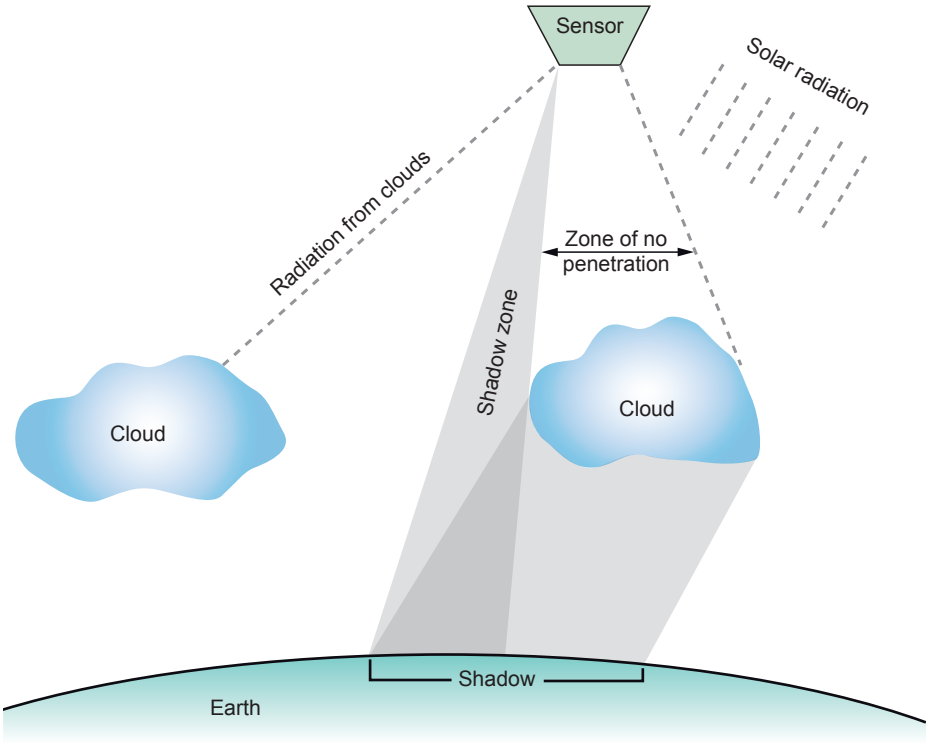


Figure 2.11
Direct and indirect effects of clouds for optical remote sensing.

2.5 Interactions of EM radiation with the Earth's surface

The EM radiation that reaches an object interacts with it. As a result of this interaction, EM radiation is absorbed, transmitted or reflected by the object. The energy conservation law, applied to interaction of EM radiation with the object, states that *all* incident EM radiation (I) is absorbed (A), reflected (R), or transmitted (T):

$$A(\lambda) + R(\lambda) + T(\lambda) = I(\lambda) \tag{2.7}$$

It is important to note that Equation 2.7 applies for each wavelength. Dividing both sides of Equation 2.7 by I we get.

$$\frac{A(\lambda)}{I(\lambda)} + \frac{R(\lambda)}{I(\lambda)} + \frac{T(\lambda)}{I(\lambda)} = \alpha(\lambda) + \rho(\lambda) + \tau(\lambda) = 1 \tag{2.8}$$

where $\alpha(\lambda)$ is absorptance, $\rho(\lambda)$ is reflectance and $\tau(\lambda)$ is transmittance of the object, all depend on wavelength λ and range from 0 to 1. For opaque objects $\tau(\lambda) = 0$ and Equation 2.8 reduces to

$$\alpha(\lambda) + \rho(\lambda) = 1 \tag{2.9}$$

Absorption of EM radiation leads to an increase in the object's temperature, while emission of EM radiation leads to a decrease in the object's temperature. The amount of emitted EM radiation is determined by the object's temperature (see Planck's law) and emissivity $\epsilon(\lambda)$. In equilibrium the total amounts of absorbed and emitted radiation at all wavelength are equal and the object's temperature is constant.

equilibrium

Kirchhoff's law of thermal radiation states that in equilibrium absorptance and emissivity at each wavelength are equal:

Kirchhoff's law

$$\alpha(\lambda) = \epsilon(\lambda) \tag{2.10}$$

The reflectance, transmittance and absorptance will vary with wavelength and type of target material. Here and further in the book we define a *target* as an object on the Earth surface that is being detected or sensed. Also the surface of target influences interaction of EM radiation and the target. Two types of reflection that represent the two extremes of the way in which radiation is reflected by a target are "specular reflection" and "diffuse reflection" (Figure 2.12). In the real world, usually a combination of both types is found.

reflection

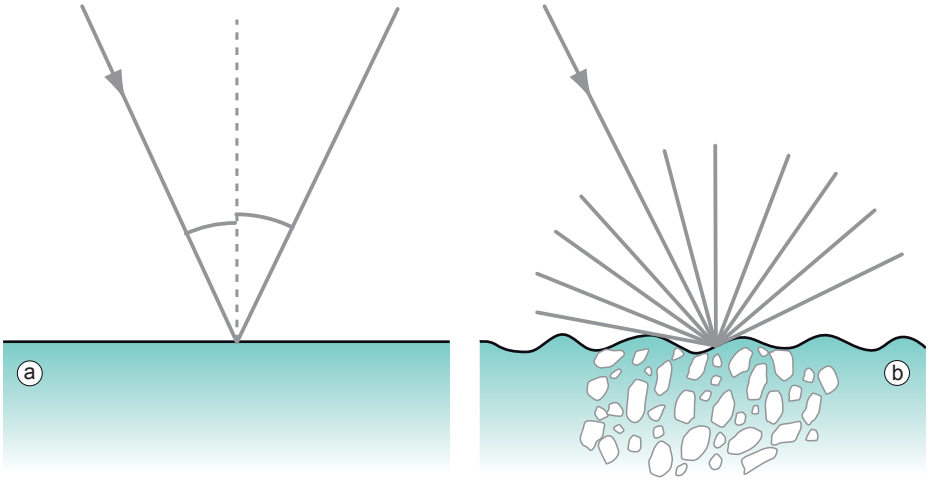


Figure 2.12
Schematic diagrams showing (a) specular and (b) diffuse reflection.

- *Specular reflection*, or mirror-like reflection, typically occurs when a surface is smooth and (almost) all of the radiation is directed away from the surface in a

single direction. Specular reflection can occur, for example, for a water surface or a glasshouse roof. It results in a very bright spot (also called "hot spot") in the sensed image.

- *Diffuse reflection* occurs in situations where the surface is rough and the radiation is reflected almost uniformly in all directions.

Whether a particular target reflects specularly, diffusely, or both, depends on the surface roughness relative to the wavelength of the incident radiation.

2.5.1 Spectral reflectance curves

reflectance measurements

We can establish for each type of material of interest a *reflectance curve*. Such a curve shows the portion of the incident radiation ρ that is reflected as a function of wavelength λ (expressed as percentage; see Figure 2.13). Remote sensors are sensitive to ranges, albeit narrow, of wavelengths, not just to one particular λ, for example the "spectral band" from $\lambda = 0.4$ µm to $\lambda = 0.5$ µm. The spectral reflectance curve can be used to estimate the overall reflectance in such bands by calculating the mean of reflectance measurements in the respective ranges. Reflectance measurements can be carried out in a laboratory or in the field, in the latter case using a field spectrometer. Reflectance curves are typically collected for the optical part of the electromagnetic spectrum and large efforts are made to store collections of typical curves in "spectral libraries". The reflectance characteristics of some common land cover types are discussed in the following subsections.

Vegetation

The reflectance characteristics of vegetation depend on the properties of the leaves, including the orientation and structure of the leaf canopy. The amount of radiation reflected for a particular wavelength depends on leaf pigmentation, thickness and composition (cell structure), and on the amount of water in the leaf tissue. Figure 2.13 shows an ideal reflectance curve of healthy vegetation. In the visible portion of the spectrum, the reflection of the blue and red components of incident light is comparatively low, because these portions are absorbed by the plant (mainly by chlorophyll) for photosynthesis; the vegetation reflects relatively more green light. The reflectance in the NIR range is highest, but the amount depends on leaf development and cell structure. In the SWIR range, reflectance is mainly determined by the free water in the leaf tissue; more free water results in less reflectance. Wavelengths around 1.45 µm and 1.95 µm are, therefore, called water absorption bands. The plant may change colour when its leaves dry out, for instance at harvest time for a crop (e.g. to yellow). At this stage there is no photosynthesis, which causes reflectance in the red portion of the spectrum to become higher. Also, the leaves will dry out, resulting in a higher reflectance of SWIR radiation, whereas reflectance in the NIR range may decrease. As a result, optical remote sensing can provide information about the type of plant and also about its health.

Bare soil

Reflectance from bare soil depends on so many factors that it is difficult to give one typical soil reflectance curve. The main factors influencing reflectance are soil colour, moisture content, the presence of carbonates, and iron oxide content. Figure 2.14 gives the reflectance curves for the five main types of soil occurring in the U.S.A. Note the typical shapes of most of the curves, which are convex shape in the range 0.5–1.3 µm and dip at 1.45 µm and 1.95 µm. These dips correspond to water absorption bands and are caused by the presence of soil moisture. Iron-dominated soil (e) has quite a different reflectance curve since iron absorption dominates at longer wavelengths.

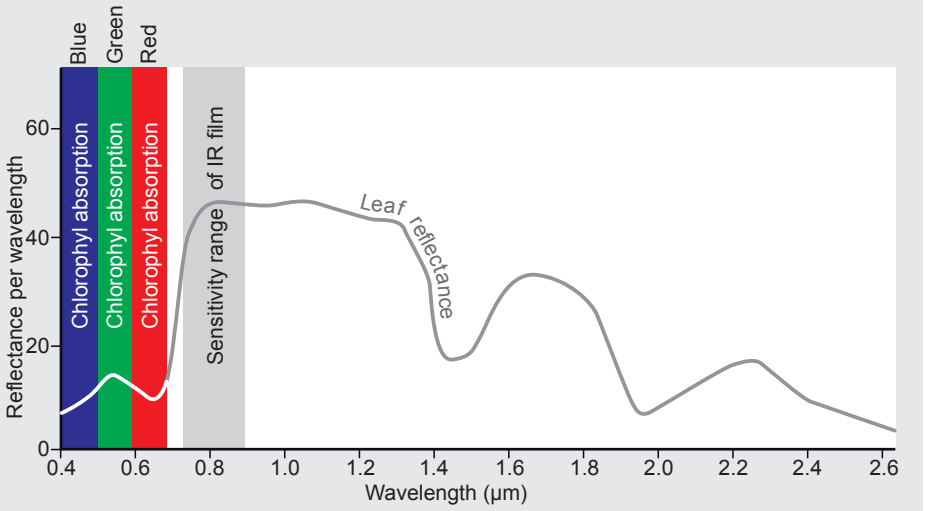


Figure 2.13
An idealized spectral reflectance curve of healthy vegetation.

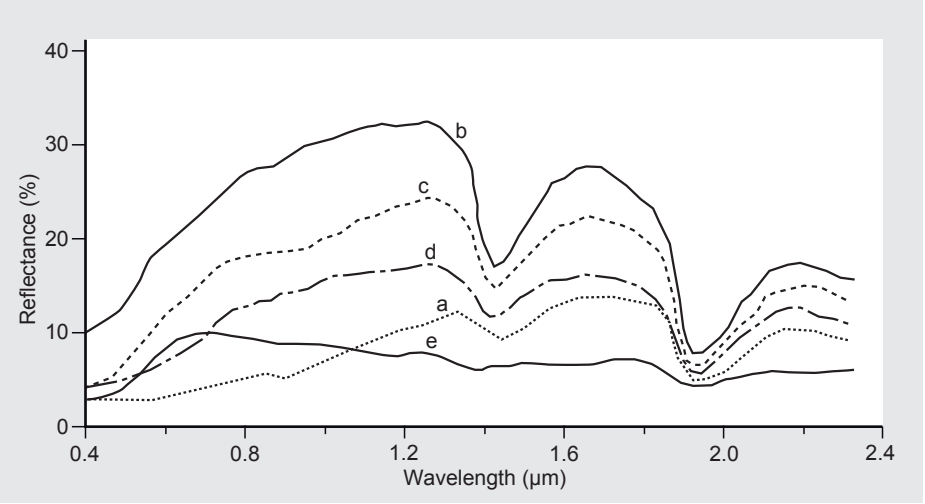


Figure 2.14
Spectral reflectance of five mineral soils: (a) organic dominated, (b) minimally altered, (c) iron altered, (d) organic affected and (e) iron dominated (from [71]).

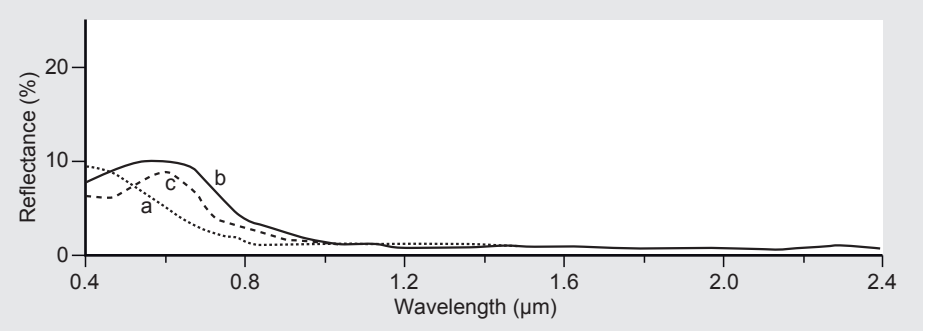


Figure 2.15
Typical effects of chlorophyll and sediments on water reflectance: (a) ocean water, (b) turbid water, (c) water with chlorophyll (from [71]).

Water

Compared to vegetation and soils, water has a lower reflectance. Vegetation may reflect up to 50% and soils up to 30–40%, while water reflects at most 10% of the incident

radiation. Water reflects EM radiation in the visible range and a little in the NIR range. Beyond 1.2 μm, all radiation is absorbed. Spectral reflection curves for water of different compositions are given in Figure 2.15. Turbid (silt loaded) water has the highest reflectance. Water containing plants or algae has a pronounced reflectance peak for green light because of the chlorophyll present.

2.6 Sensing of EM radiation

The review of properties of EM radiation shows that different forms of radiation can provide us with different information about terrain-surface features and that different applications of Earth Observation are likely to benefit from sensing in different ranges of the EM spectrum. A geoinformatics engineer who wants to discriminate objects for topographic mapping will prefer to use an optical sensor operating in the visible range. An environmentalist who needs to monitor heat losses of a nuclear power plant will use a sensor that detects thermal emission. A geologist interested in surface roughness, because it indicates to him rock type, will rely on microwave sensing. Different demands combined with different technical solutions have resulted in a multitude of sensors. In this section we will classify various remote sensors and discuss their common features. Peculiarities will then be treated later in appropriate sections.

2.6.1 Sensing properties

A *remote sensor* is a device that detects EM radiation, quantifies it and, usually, records it in an analogue or digital form. A remote sensor may also transmit recorded data (to a receiving station on the ground). Many sensors used in Earth Observation detect reflected solar radiation. Others detect the radiation emitted by the Earth itself. There are, however, some obstacles to be overcome. The Sun does not always shine brightly and there are regions on the globe almost permanently under cloud cover. There are also regions that have seasons with very low Sun elevation, so that objects cast long shadows over long periods. Furthermore, at night there are only emissions—and perhaps moonlight. Sensors detecting reflected solar radiation are useless at night and face problems when dealing with unfavourable seasonal and weather conditions. Sensors detecting emitted terrestrial radiation do not directly depend on the Sun as a source of illumination; they can be operated any time. The Earth's emissions, we have learned, occurs only at longer wavelengths because of the relatively low surface temperature and because long EM waves do not hold much energy, which makes them more difficult to sense.

obstacles to sensing

Luckily we do not have to rely only on solar and terrestrial radiation. We can build instruments that emit EM radiation and then detect the radiation returning from the target object or surface. Such instruments are called *active sensors*, as opposed to passive ones, which measure reflected solar or terrestrial radiation (Figure 2.16). An example of an active sensor is a laser rangefinder, a device that can be bought for a few euros in any DIY store. Another very common active sensor is a camera with a flash unit (which will operate below certain levels of light). The same camera without the flash unit is a passive sensor. The main advantages of active sensors are that they can be operated day and night and have a controlled illuminating signal. They are often designed to work in an EM spectrum range that is less affected by the atmosphere and weather conditions. Laser and radar instruments are the most prominent active sensors for GDA.

active versus passive RS

Most remote sensors measure either the intensity or the phase of EM radiation. Some—like a simple laser rangefinder—only measure the elapsed time between sending a radiation signal and receiving it back. Radar sensors may measure both intensity and

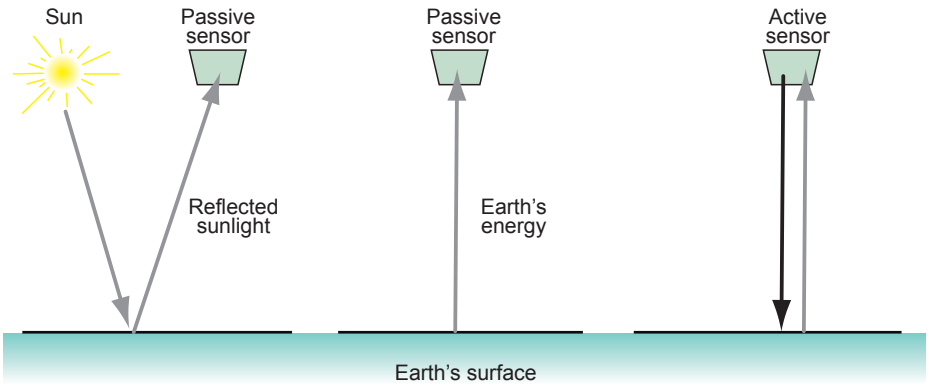


Figure 2.16
A remote sensor measures reflected or emitted radiation. An active sensor has its own source of radiation.

phase. Phase measuring sensors are used for precise ranging (distance measurement), e.g. by GPS "phase receivers" or continuous-wave laser scanners. The intensity of radiation can be measured from the photon energy striking the sensor's radiation-sensitive surface.

intensity or phase

By considering the following equation, you can relate the intensity measure of reflected radiation to Figure 2.6 and link the Figures 2.13 to 2.17. When sensing reflected light, radiance at the sensor is equal to the radiance at the Earth's surface attenuated by atmospheric absorption, plus the radiance of scattered light:

measuring radiance

$$L = \frac{\rho E \tau}{\pi} + \text{sky radiance} \tag{2.11}$$

where L is the total radiance at the sensor, E is the irradiance (the intensity of the incident solar radiation, attenuated by the atmosphere) at the Earth's surface, ρ is the terrain reflectance, and τ is the atmospheric transmittance. The radiance at the Earth's surface depends on the irradiance and the terrain surface reflectance. The irradiance, in turn, stems from direct sunlight and diffuse light, the latter caused by atmospheric scattering, particularly on a hazy days (see Figure 2.17). This indicates why you should study radiometric correction (Subsection 5.1.3 and Subsection 5.2.2), to enable you to make better inferences about surface features.

The radiance is observed for a *spectral band*, not for a single wavelength. A *spectral band* or wavelength band is an interval of the EM spectrum in which the average radiance is measured. Sensors such as a panchromatic camera, a radar sensor and a laser scanner only measure in one specific band, while a multispectral scanner or a digital camera measures in several spectral bands at the same time. Multispectral sensors have several *channels*, one for each spectral band. Figure 2.18 shows spectral reflectance curves, together with the spectral bands, of some popular satellite-based sensors. Sensing in several spectral bands simultaneously allows us to relate properties that show up well in specific spectral bands. For example, reflection characteristics in the spectral band 2 to 2.4 μm (as recorded by Landsat-5 TM channel 7) tell us something about the mineral composition of soil. The combined reflection characteristics in the red and NIR bands (from Landsat-5 TM channels 3 and 4) can tell us something about biomass and plant health.

spectral band

Landsat MSS (MultiSpectral Scanner), the first civil space-borne Earth Observation sensor, had sensing elements (detectors) for three rather broad spectral bands in the visible range of the spectrum, each with a width of 100 nm, and one broader band in the NIR range. A hyperspectral scanner uses detectors for many more, but narrower, bands, which may be as narrow as 20 nm, or even less. We say a hyperspectral sensor

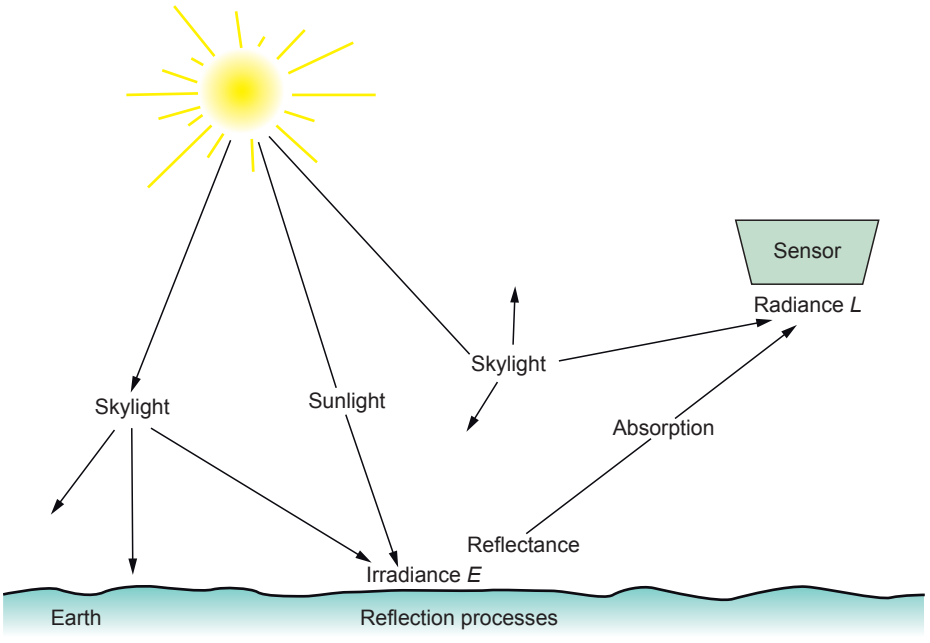


Figure 2.17
Radiance at the sensor, adapted from [67].

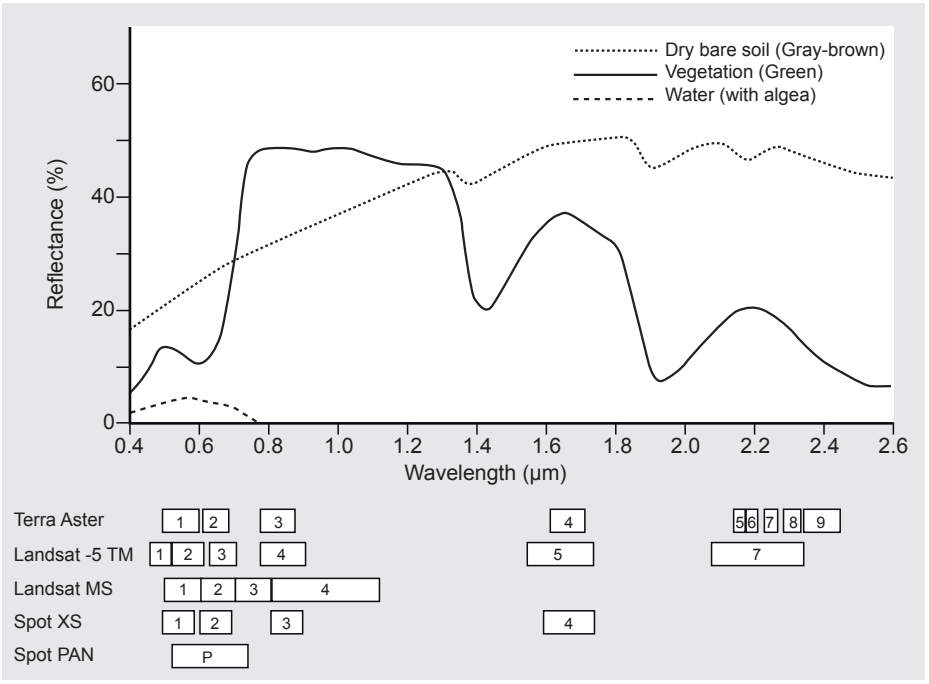


Figure 2.18
Spectral reflectance curves and spectral bands of some multispectral sensors.

spectral resolution

has a higher 'spectral resolution' than a multispectral one. A laser instrument can emit (and detect) almost monochrome radiation, with a wavelength band no wider than 10 nm. A camera loaded with panchromatic film or a space-borne electronic sensor with a panchromatic channel (such as SPOT PAN or WorldView-1) records the intensity of radiation of a broad spectral band covering the entire visible range of the EM spectrum. Panchromatic—which stands for "across all colours"—recording is compa-

rable with the function of the 120 million rods of a human eye. They are brightness sensors and cannot sense colour.

In a camera loaded with panchromatic film (*black & white film*), the silver halide crystals of the light-sensitive emulsion detect radiation. The silver halide grains turn to silver metal when exposed to light, the more so the higher the intensity of the incident light. Each light ray from an object/scene triggers a chemical reaction of some particular grain. This way, variations in radiance within a scene are detected and an image of the scene is created at the time of exposure. The record obtained is only a latent image; the film has to be developed to turn it into a photograph.

photographic detector

Digital cameras and multispectral scanners are examples of sensors that use electronic detectors instead of photographic ones. An electronic detector (CCD, CMOS, photodiode, solid state detector, etc.) is made of semiconductor material. The detector accumulates a charge by converting the photons incident upon its surface to electrons. (It was Einstein who won the Nobel prize for discovering and explaining that there is an emission of electrons when a negatively charged plate of light-sensitive (semiconductor) material is subject to a stream of photons.) The electrons can then be made to flow as a current from the plate. So the charge can be converted to a voltage (electrical signal). The charge collected is proportional to the radiance at the detector (the amount of radiation "deposited" in the detector). In a process called A/D conversion, the electrical signal is *sampled* and quantified. The output is a digital number (DN), which is recorded. A DN is an integer within a fixed range. Older remote sensors used 8 bits for recording, which allows a differentiation of radiance into $2^8 = 256$ levels (i.e. DNs in the range 0 to 255). The recently launched (in 2007) WorldView-1 sensor records with a *radiometric resolution* of 11 bits $\left(2^{11} = 2048\right)$. ASTER records the visible spectral band using 8 bits and the thermal infrared band using 12 bits. A higher radiometric resolution requires more storage capacity but has the advantage of offering data with greater information content (see Figure 2.19).

AD conversion

radiometric resolution

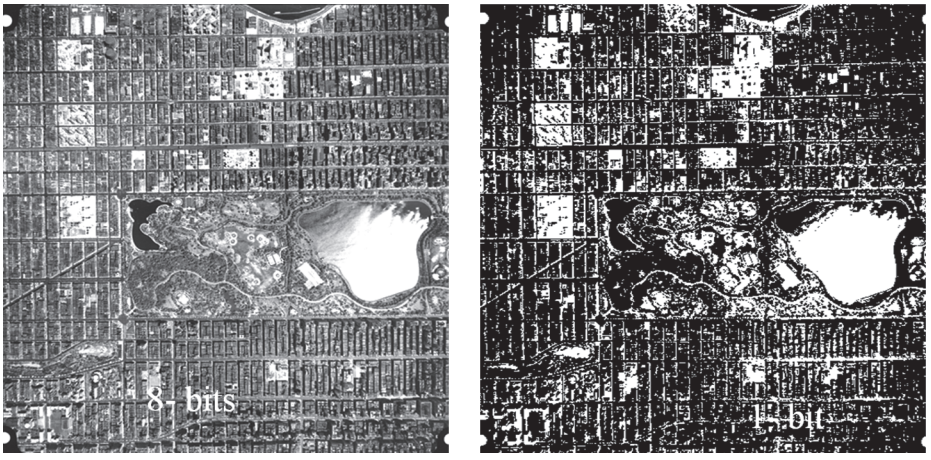


Figure 2.19
8-bit versus 11-bit radiometric resolution.

A digital panchromatic camera has an array of detectors instead of silver halide crystals suspended in gelatine on a polyester base of photographic film. Each detector (e.g. a CCD, which stands for charge-coupled device) is very small, in the order of 9 µm × 9 µm. Space-borne cameras use larger detectors than aerial cameras to ensure that enough photons are collected despite the great distances at which they operate from the Earth. At the moment of exposure, each detector yields one DN, so in total we obtain a data set that represents an image similar to the one created by "exciting" photographic material in a film camera.

imaging,
spatial resolution

When arranging the DNs in a two-dimensional array, we can readily visualize them as grey values. We refer to the obtained "image" as a *digital image* and to a sensor producing digital images as an *imaging sensor*. The array of DNs represents an image in terms of discrete picture elements, called *pixels*. The value of a pixel—its DN—corresponds to the radiance of the light reflected from the small ground area viewed by the relevant detector. The smaller the detector, the smaller will be the area on the ground that corresponds to one pixel. The size of the "ground resolution cell" is often referred to as "pixel size on the ground". Early digital cameras for consumers had 2×10^6 CCDs per spectral band (named 2 megapixel cameras); today we can get for the same price a 10 megapixel camera. The latter has much smaller CCDs so that they can fit on the same board, with the consequence that an image can reveal much more detail; we would say the *spatial resolution* of the image is higher.

A digital camera for the consumer market does not record intensity values for a single (panchromatic) spectral band, but for three bands simultaneously, namely for red, green, and blue light, in order to obtain colour images. This is comparable with our eyes: we have three types of cones, one for each primary colour. The data set obtained for one shot taken with the camera (the *image file*) therefore contains three separate digital images (Figure 2.20). Multispectral sensors record in as many as 14 bands simultaneously (e.g. ASTER). For convenience, a single digital image is then often referred to as "band" and the total image file as a *multi-band image*.

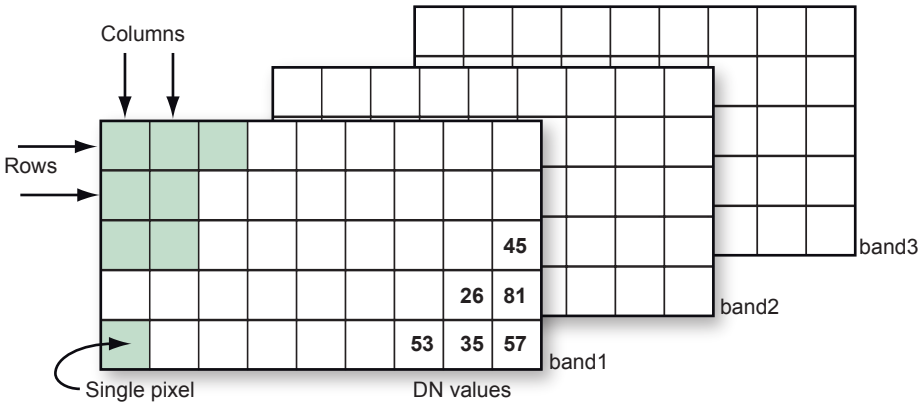


Figure 2.20
An image file comprises a digital image for each of the spectral bands of the sensor. The DN values for each band are stored in a row-column arrangement.

storage media

Various storage media are used for recording the huge amount of data produced by electronic detectors: solid state media (such as memory cards as used in consumer cameras), magnetic media (disk or tape) and optical discs (some video cameras); satellites usually have several recorders on board.

Light sensor systems often transmit data to ground receiving stations at night. Data can also be transmitted directly to a receiving station using satellite communication technology. Airborne sensors often use the hard disk of a laptop computer as a recording device. The huge amounts of data collected demand efficient data management systems. This issue will be examined in Section 8.4.

2.6.2 Classification of sensors

Remote sensors can be classified and labelled in different ways. According to whatever our prime interest in Earth Observation may be—geometric properties, spectral differences, or an intensity distribution of an object or scene—we can distinguish three salient types of sensors: altimeters, spectrometers, and radiometers.

Laser and radar altimeters are non-imaging sensors that provide information about

the elevation of water and land surfaces.

altimeter

Thermal sensors, such as the channels 3 to 5 of NOAA's AVHRR or the channels 10 to 14 of Terra's ASTER, are called (imaging) radiometers. Radiometers measure radiance and typically sense in one broad spectral band or in only a few bands, but with high radiometric resolution. Panchromatic cameras and passive microwave sensors are other examples of radiometers. The spatial resolution depends on the wavelength band of the sensor. Panchromatic radiometers can have a very high spatial resolution, whereas microwave radiometers have a low spatial resolution because of the low levels of energy inherent in this spectral range. Scatterometers are non-imaging radiometers. Radiometers are used for a wide range of applications: for example, detecting forest/bush/coal fires; determining soil moisture and plant response; monitoring ecosystem dynamics; and analysing energy balance across land and sea surfaces.

radiometer

Spectrometers measure radiance in many (usually about 100 or 200) narrow, contiguous spectral bands and therefore have a high spectral resolution. Their spatial resolution is moderate to low. The prime use of imaging spectrometers is to identify surface materials—from the mineral composition of soils, to concentrations of suspended matter in surface water and chlorophyll content. There are also androgynous sensors: spectro-radiometers, imaging laser scanners, and Fourier spectrometers, for example.

spectrometers

We can also group the multitude of remote sensors used for GDA according to the spectral domains in which they operate (Figure 2.21). The following list gives a short description of each group and refers to the section in which they are treated in more detail.

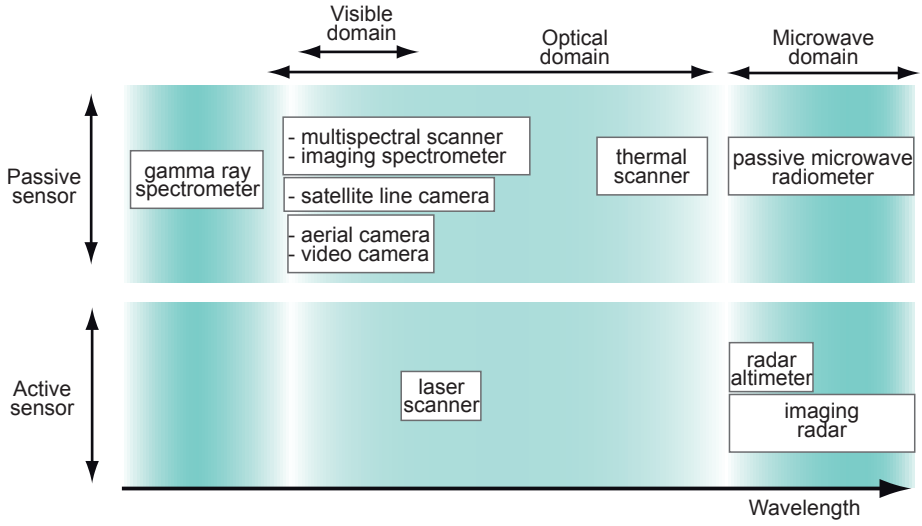


Figure 2.21
Overview of the sensors that are described in this book.

- gamma ray spectrometers are mainly used in mineral exploration.

gamma ray sensors

- aerial film cameras have been the remote sensing workhorse for decades. Today, they are used primarily for large-scale topographic mapping, cadastral mapping, and orthophoto production for urban planning, to mention a few examples; they are discussed in Section 4.6.

film cameras

- digital aerial cameras are not conquering the market as quickly as digital cameras did on the consumer market. These cameras use CCD arrays instead of film; they are treated together with optical scanners in Section 4.1. Line cameras operated from satellites have very similar properties.

digital cameras

video cameras

- digital video cameras are not only used to record movies. They are also used in aerial Earth Observation to provide low cost (and low resolution) images for mainly qualitative purposes, for instance to provide visual information about an area covered by "blind" airborne laser scanner data. Handling images from video cameras is similar to dealing with images from digital "still" cameras; this is not explicitly discussed any further in this book.

multispectral scanners

- multispectral scanners are mostly operated from satellites and other space vehicles. The essential difference between multispectral scanners and satellite line cameras is the imaging/optical system employed: multispectral scanners use a moving mirror to "scan" a line (i.e. a narrow strip on the ground) and a single detector instead of recording intensity values of an entire line at one instant by an array of detectors as for line cameras. Multispectral scanners are treated in Section 4.1. Figure 2.22 shows an image obtained by combining the images of Landsat TM channels 4, 5 and 7, which are displayed in red, green and blue, respectively. Section 5.1 explains how to produce such a "false colour" image.

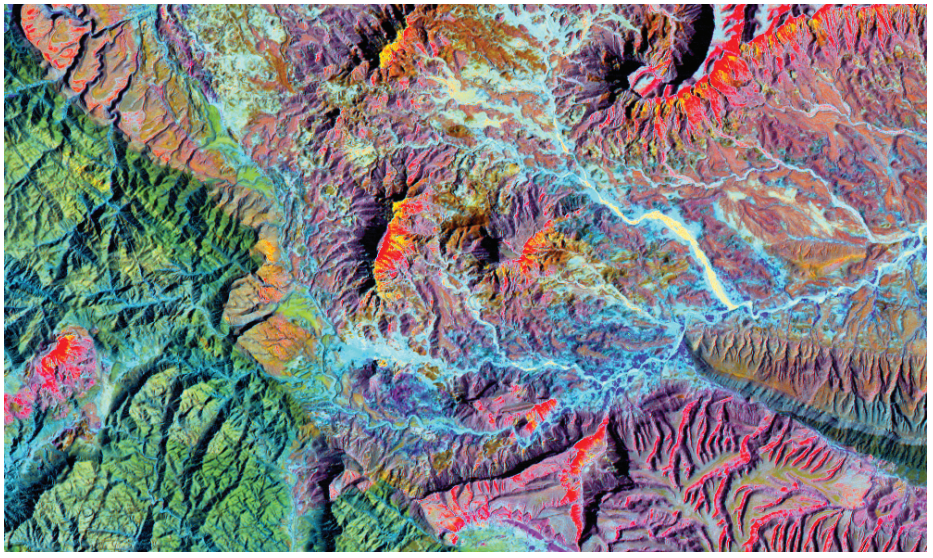


Figure 2.22
Landsat-5 TM false colour composite of an area of 30 km $\times$ 17 km.

imaging spectrometers

- hyperspectral scanners are imaging spectrometers with a scanning mirror; they are treated in detail in Section 4.3.

thermal scanners

- thermal scanners are placed here in the optical domain purely for the sake of convenience. They exist as special instruments and as a component of multispectral radiometers; they are included in Section 4.2. Thermal scanners provide us with data that can be directly related to object temperature. Figure 2.23 is an example of a thermal image acquired by an airborne thermal scanner at night.

microwave radiometers

- passive microwave radiometers detect emitted radiation of the Earth's surface in the 10 to 1000 mm wavelength range. These radiometers are mainly used in mineral exploration, for monitoring soil-moisture changes, and for snow and ice detection. Microwave radiometers are not discussed further in this book.

- laser scanners are the scanning variant of laser rangefinders and altimeters (as on ICESat). Laser scanners measure the distance from the laser instrument to many points of the target in "no time" (e.g. 150,000 points in one second). Laser

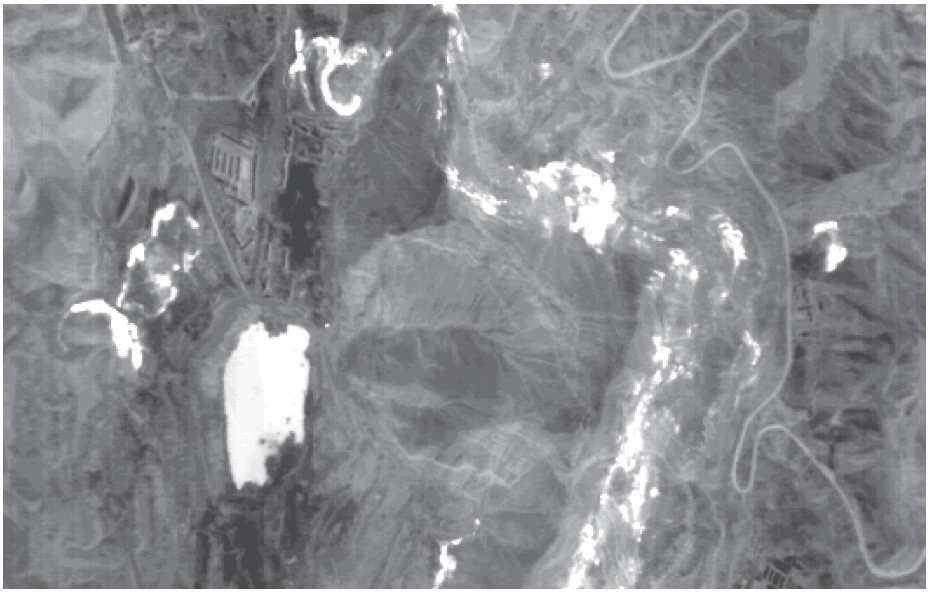


Figure 2.23
'Thermal image' at night of a coal mining area affected by underground coal fires. Darker tones represent colder surfaces, while lighter tones represent warmer areas. Most of the warm spots are due to coal fires, except for the large white patch, which is a lake; at that time of the night, apparently the temperature of the water was higher than the temperature of the land. On the ground, the area depicted is approximately 4 km across.

ranging is often referred to as LIDAR (LIght Detection And Ranging). The prime application of airborne laser scanning (ALS) is for creating high resolution digital surface models and digital terrain models (see Section 5.3). We can also create a digital terrain model (DTM) from photographs and similar panchromatic images. However, because of the properties of laser radiation, ALS has important advantages in areas of dense vegetation and for sandy deserts and coastal areas. Surface modelling is of interest for many applications, such as, for example, biomass estimation of forests, volume calculations for open-pit mining (see Figure 2.24), flood plain mapping, and 3D modelling of cities. Laser scanning is dealt with in more detail in Section 4.5.

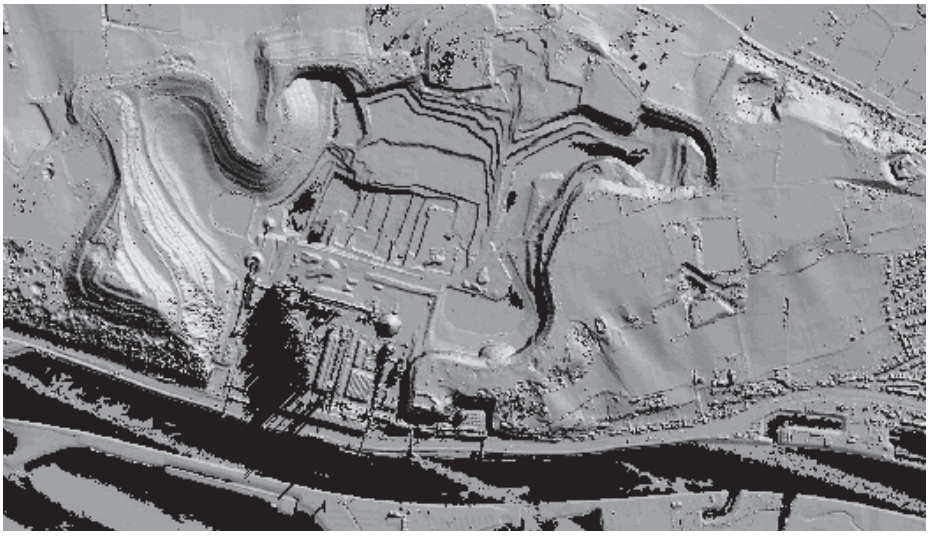


Figure 2.24
Pictorial representation of a digital surface model of the Sint Pietersberg open-pit mine in the Netherlands. The size of the pit is roughly 2 km $\times$ 1 km. The terraced rim of the pit is clearly visible. The black strip near the bottom of the image is the River Meuse. Courtesy Survey Department, Rijkswaterstaat.

- imaging radar (RAdio Detection And Ranging) operates in the spectral range

imaging radar

10–1000 mm. Radar instruments are active sensors and because of the range of wavelengths used they can provide data day and night, under all weather conditions. Radar waves can penetrate clouds; only heavy rainfall affects imaging to some degree. One of its applications is, therefore, the mapping of areas that are subject to permanent cloud cover. Figure 2.25 shows an example of a SAR (Synthetic Aperture Radar) image from the ERS-1 satellite. Radar data from the air or space can also be used to create surface models. Radar imaging has a peculiar geometry and processing raw radar data is not simple. Radar is treated in Section 4.4.

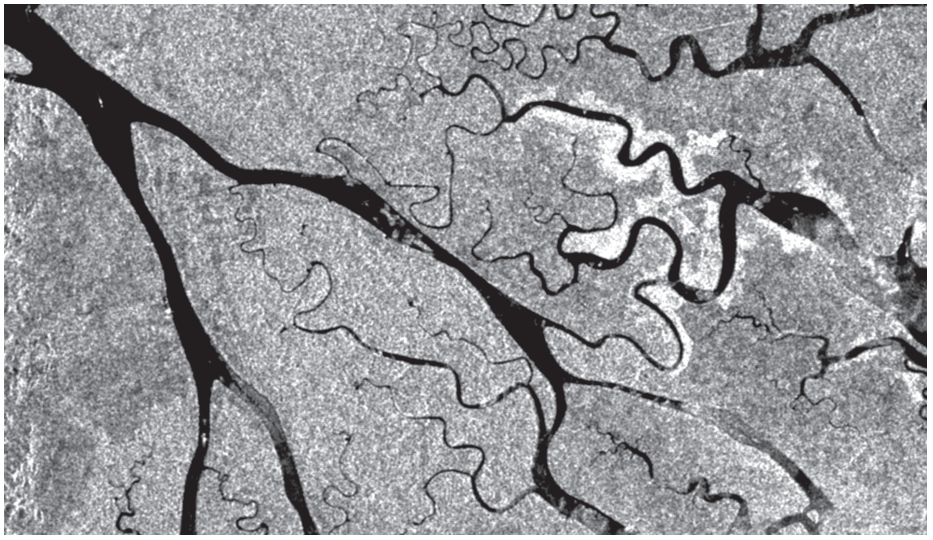


Figure 2.25
An ERS-1 SAR image of the Mahakam Delta, Kalimantan. The image shows different types of land cover. The river is black. The darker patch of land on the the left is inland tropical rainforest. The rest is a mixed forest of Nipa palm and mangrove on the delta. The right half of the image shows light patches, where the forest has been partly cleared. The image covers an area on the ground of 30 km x 15 km.

radar altimeters

- radar altimeters are used to measure elevation profiles of the Earth's surface that is parallel to the receiving satellite's orbit. Radar altimeters operate in the 10–60 mm range and allow us to calculate elevation with an accuracy of 20–50 mm. Radar altimeters are useful for measuring relatively smooth surfaces.

sonar

- for the sake of completeness, sonar, another active sensor, is included here. Sonar, which stands for SOund NAvigation Ranging, is used, for example, for mapping river beds and sea floors, and for detecting obstacles underwater. Sonar works by emitting a small burst of sound from a ship. The sound is reflected off the bottom of the body of water. The time taken for the reflected pulse to be received corresponds to the depth of the water. More advanced systems also record the intensity of the return signal, thus giving information about the material on the sea floor. In its simplest form, sonar "looks" vertically and is operated very much like a radar altimeter. The body of water will be traversed in paths resembling a grid; not every point below the water surface will be monitored. The distance between data points depends on the ship's speed, the frequency of the measurements, and the distance between the adjacent paths.

  One of the most accurate systems for imaging large areas of the ocean floor is side-scan sonar. It is an imaging system that works in a way that is somewhat similar to side-looking airborne radar (see Section 4.4). The images produced by side-scan sonar systems are highly accurate and can be used to delineate even very small ($<$ 1 cm) objects. From sonar data, we can produce contour maps of sea floors and other water bodies, which can be used, for example, for navigation and water-discharge studies.

Chapter 3

Spatial referencing and satellite-based positioning

Richard Knippers
Klaus Tempfli

Introduction

In the early days of geoinformation science, spatially referenced data usually originated within national boundaries, i.e. these data were derived from printed maps published by national mapping organizations. Nowadays, users of geoinformation are combining spatial data from a given country with global spatial data sets, reconciling spatial data from published maps with coordinates established by satellite positioning techniques, and integrating their spatial data with that from neighbouring countries.

To perform these kinds of tasks successfully, we need to understand basic spatial referencing concepts. Section 3.1 discusses the relevance and actual use of reference surfaces, coordinate systems and coordinate transformations. We will explain the principles of spatial referencing as applied to mapping, the traditional application of geoinformation science. These principles are generally applicable to all types of geospatial data.

Section 3.2 discusses satellite-based systems of spatial positioning. The development of these global positioning systems has made it possible to unambiguously determine any position in space. This and related developments have laid the foundations for the integration of all spatial data within a single, global 3D spatial-reference system, which we may expect to emerge in the near future.

3.1 Spatial referencing

One of the defining features of geoinformation science is its ability to combine spatially referenced data. A frequently occurring issue is the need to combine spatial data from different sources that use different spatial reference systems. This section provides

a broad background of relevant concepts relating to the nature of spatial reference systems and the translation of data from one spatial referencing system into another.

3.1.1 Reference surfaces

The surface of the Earth is far from uniform. Its oceans can be treated as reasonably uniform, but the surface or topography of its land masses exhibits large vertical variations between mountains and valleys. These variations make it impossible to approximate the shape of the Earth with any reasonably simple mathematical model. Consequently, two main reference surfaces have been established to approximate the shape of the Earth : one is called the *Geoid*, the other the *ellipsoid*; see Figure 3.1.

geoid and ellipsoid

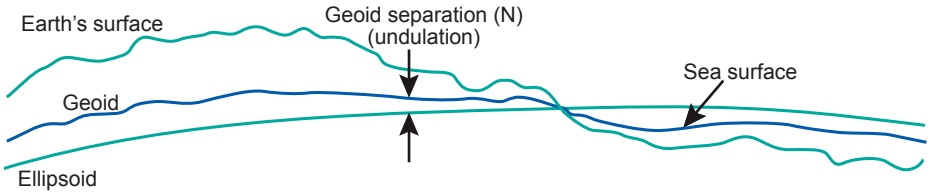


Figure 3.1
The Earth's surface and two reference surfaces used to approximate it: the Geoid; and a reference ellipsoid. The Geoid separation (N) is the deviation between the Geoid and the reference ellipsoid.

The Geoid and the vertical datum

We can simplify matters by imagining that the entire Earth's surface is covered by water. If we ignore effects of tides and currents on this *global ocean*, the resultant water surface is affected only by gravity. This has an effect on the shape of this surface because the direction of gravity—more commonly known as the plumb line—is dependent on the distribution of mass inside the Earth. Owing to irregularities or mass anomalies in this distribution, the surface of the *global ocean* would be undulating. The resulting surface is called the Geoid (Figure 3.2). A plumb line through any surface point would always be perpendicular to the surface.

plumb line

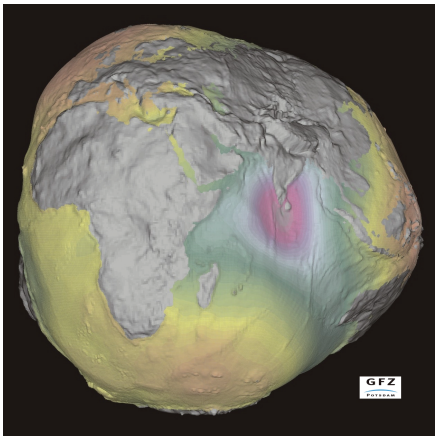


Figure 3.2
The Geoid, exaggerated to illustrate the complexity of its surface. Image: GFZ German Research Centre for Geosciences.

The Geoid is used to describe *heights*. In order to establish the Geoid as a reference for heights, the ocean's water level is registered at coastal locations over several years using tide gauges (mareographs). Averaging the registrations largely eliminates variations in sea level over time. The resultant water level represents an approximation to the Geoid and is termed mean sea level.

mean sea level

For the Netherlands and Germany, local mean sea level is related to the Amsterdam Tide Gauge (zero height). We can determine the height of a point in Enschede with

respect to the Amsterdam Tide Gauge using a technique known as geodetic levelling (Figure 3.3). The result of this process will be the height of the point in Enschede above local mean sea level. Height determined with respect to a tide gauge station is known as *orthometric height* (height H above the Geoid) .

Several definitions of local mean sea levels (also called local vertical datums) appear throughout the world. They are parallel to the Geoid but offset by up to a couple of metres to allow for local phenomena such as ocean currents, tides, coastal winds, water temperature and salinity at the location of the tide gauge. Care must be taken when using heights from another local vertical datum . This might be the case, for example, in areas along the border of adjacent nations.

local vertical datums

Even within a country, heights may differ depending on the location of the tide gauge to which mean sea level is related. As an example, the mean sea level from the Atlantic to the Pacific coast of the U.S.A. differs by 0.6–0.7 m. The tide gauge (zero height) of the Netherlands differs -2.34 m from the tide gauge (zero height) of neighbouring Belgium.

The local vertical datum is implemented through a levelling network (Figure 3.3a), which consists of benchmarks whose height above mean sea level has been determined through geodetic levelling . The implementation of the datum enables easy user access. Surveyors, for example, do not need to start from scratch (i.e. from the Amsterdam tide gauge) each time they need to determine the height of a new point. They use the benchmark of the levelling network that is closest to the new point (Figure 3.3b).

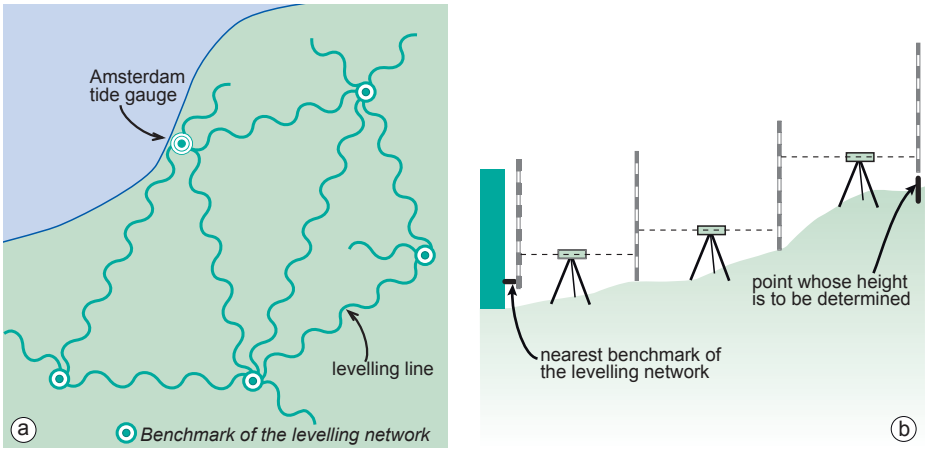


Figure 3.3
A levelling network implements a local vertical datum: (a) network of levelling lines starting from the Amsterdam Tide Gauge, showing some of the benchmarks; (b) how the orthometric height (H) is determined for some point by working from the nearest benchmark.

As a result of satellite gravity missions, it is currently possible to determine height (H) above the Geoid to centimetre levels of accuracy. It is foreseeable that a global vertical datum may become ubiquitous in the next 10–15 years. If all geodata, for example maps, were to use such a global vertical datum, heights would become globally comparable, effectively making local vertical datums redundant for users of geoinformation.

The ellipsoid

We have defined a physical surface, the Geoid, as a reference surface for heights. We also need, however, a reference surface for the description of the *horizontal coordinates* of points of interest. Since we will later want to project these horizontal coordinates onto a mapping plane, the reference surface for horizontal coordinates requires a mathematical definition and description. The most convenient geometric reference

horizontal coordinates

is the *oblate ellipsoid* (Figure 3.4). It provides a relatively simple figure that fits the Geoid to a first-order approximation (for small-scale mapping purposes we may use the *sphere*). An ellipsoid is formed when an ellipse is rotated around its minor axis. This ellipse, which defines an ellipsoid or *spheroid*, is called a meridian ellipse (notice that ellipsoid and spheroid are used here to refer to the same).

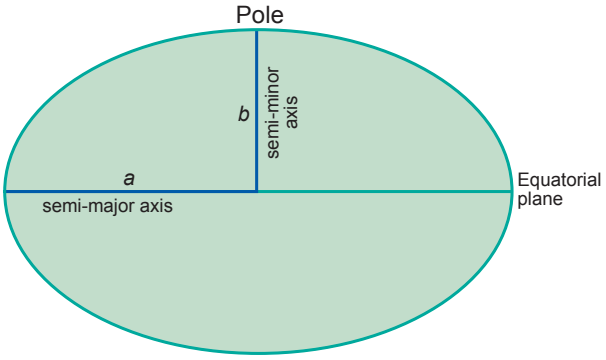


Figure 3.4
An oblate ellipsoid, defined by its semi-major axis a and semi-minor axis b.

The shape of an ellipsoid may be defined in a number of ways, but in geodetic practice it is is usually defined by its semi-major axis and flattening (Figure 3.4). Flattening f is dependent on both the semi-major axis a and the semi-minor axis b:

$$f = \frac{a-b}{a}.$$

The ellipsoid may also be defined by its semi-major axis a and its eccentricity e, which can be expressed as:

$$e^2 = 1 - \frac{b^2}{a^2} = \frac{a^2-b^2}{a^2} = 2f - f^2.$$

Given one axis and any one of the other three parameters, the other two can be derived. Typical values of the parameters for an ellipsoid are:

$$a = 6378135.00\ m,\ b = 6356750.52\ m,\ f = \frac{1}{298.26},\ e = 0.08181881066$$

local ellipsoids

Many different sorts of ellipsoids have been defined. Local ellipsoids have been established to fit the Geoid (mean sea level) well over an area of local interest, which in the past was never larger than a continent. This meant that the differences between the Geoid and the reference ellipsoid could effectively be ignored, allowing accurate maps to be drawn in the vicinity of the datum (Figure 3.5).

global ellipsoids

With increasing demands for global surveying, work is underway to develop global reference ellipsoids. In contrast to local ellipsoids, which apply only to a specific country or localized area of the Earth's surface, global ellipsoids approximate the Geoid as a mean Earth ellipsoid. The International Union for Geodesy and Geophysics (IUGG) plays a central role in establishing these reference figures.

In 1924, the general assembly of the IUGG in Madrid introduced the ellipsoid determined by Hayford in 1909 as the international ellipsoid. According to subsequently acquired knowledge, however, the values for this ellipsoid give an insufficiently accurate approximation. At the 1967 general assembly of the IUGG in Luzern, the 1924 reference system was replaced by the Geodetic Reference System 1967 (GRS 1967) el-

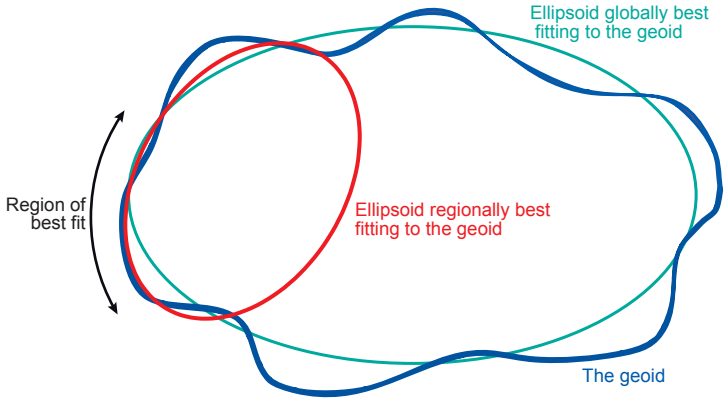


Figure 3.5
The Geoid, its global best-fit ellipsoid, and a regional best-fit ellipsoid for a chosen region. Adapted from: Ordnance Survey of Great Britain. *A Guide to Coordinate Systems in Great Britain.*

lipsoid. Later, in 1980, this was in turn replaced by the Geodetic Reference System 1980 (GRS80) ellipsoid.

Name	*a* (m)	*b* (m)	*f*
International (1924)	6378388.	6356912.	1:297.000
GRS 1967	6378160.	6356775.	1:298.247
GRS 1980 & WGS84	6378137.	6356752.	1:298.257

Table 3.1
Three global ellipsoids defined by a semi-major axis a, semi-minor axis b, and flattening f. For all practical purposes, the GRS80 and WGS84 can be considered to be identical.

The local horizontal datum

Ellipsoids have varying positions and orientations. An ellipsoid is positioned and oriented with respect to the local mean sea level by adopting a latitude (ϕ) and longitude (λ) and ellipsoidal height (h) of what is called a fundamental point and an azimuth to an additional point. We say that this defines a *local horizontal datum*. Note that the term horizontal datum and geodetic datum are treated as equivalent and interchangeable terms.

Several hundred local horizontal datums exist in the world. The reason for this is obvious: different local ellipsoids of varying position and orientation had to be adopted to provide a best fit of the local mean sea level in different countries or regions. The Potsdam Datum, the local horizontal datum used in Germany is an example of a local horizontal datum. The fundamental point is located in Rauenberg and the underlying ellipsoid is the Bessel ellipsoid (a = 6,377,397.156 m, b = 6,356,079.175 m). We can determine the latitude and longitude (ϕ, λ) of any other point in Germany with respect to this local horizontal datum using geodetic positioning techniques, such as triangulation and trilateration. The result of this process will be the geographic (or horizontal) coordinates (ϕ, λ) of a new point in the Potsdam Datum.

triangulation networks

A local horizontal datum is determined through a triangulation network. Such a network consists of monumented points that form a network of triangular mesh elements (Figure 3.6). The angles in each triangle are measured, in addition to at least one side of the triangle; the fundamental point is also a point in the triangulation network. The angle measurements and the adopted coordinates of the fundamental point are then used to derive geographic coordinates (ϕ, λ) for all monumented points of the triangulation network.

Within this framework, users do not need to start from scratch (i.e. from the fundamental point) in order to determine the geographic coordinates of a new point. They

can use the monument of the triangulation network that is closest to the new point. The extension and re-measurement of the network is nowadays done through satellite measurements.

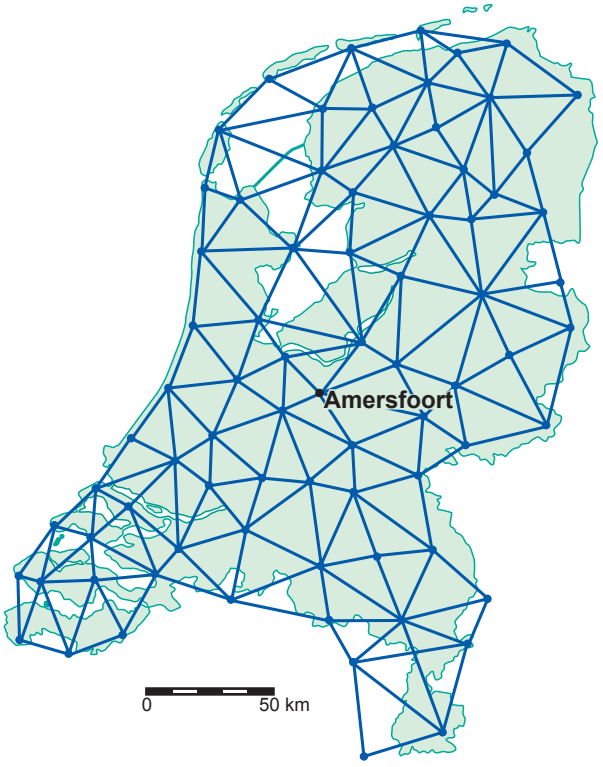


Figure 3.6
The old primary triangulation network in the Netherlands was made up of 77 points (mostly church towers). The extension and re-measurement of the network is done nowadays through satellite measurements. Adapted from original figure by "Dutch Cadastre and Land Registers", now called *het Kadaster*.

The global horizontal datum

With increasing demands for global surveying, activities are underway to establish global reference surfaces. The motivation in this is to make geodetic results mutually compatible and to be able to provide coherent results to other disciplines, e.g. astronomy and geophysics.

The most important global (geocentric) spatial reference system for the geoinformation community is the International Terrestrial Reference System (ITRS) . This is a three-dimensional coordinate system with a well-defined origin (the centre of mass of the Earth) and three orthogonal coordinate axes (X, Y, Z). The Z-axis points towards a mean North Pole. The X-axis is oriented towards the mean Greenwich meridian and is orthogonal to the Z-axis. The Y-axis completes the right-handed reference coordinate system (Figure 3.7a).

ITRS

The ITRS is realized through the International Terrestrial Reference Frame (ITRF), a distributed set of ground control stations that measure their position continuously using GPS (Figure 3.7b). Constant re-measuring is needed because of the addition of new control stations and ongoing geophysical processes (mainly tectonic plate motion) that deform the Earth's crust at measurable global, regional and local scales. These deformations cause positional differences over time and have resulted in more than one realization of the ITRS. Examples are the ITRF96 and the ITRF2000. The ITRF96 was established on 1 January 1997, which means that the measurements use data acquired up to 1996 to fix the geocentric coordinates (X, Y and Z in metres) and velocities (posi-

ITRF

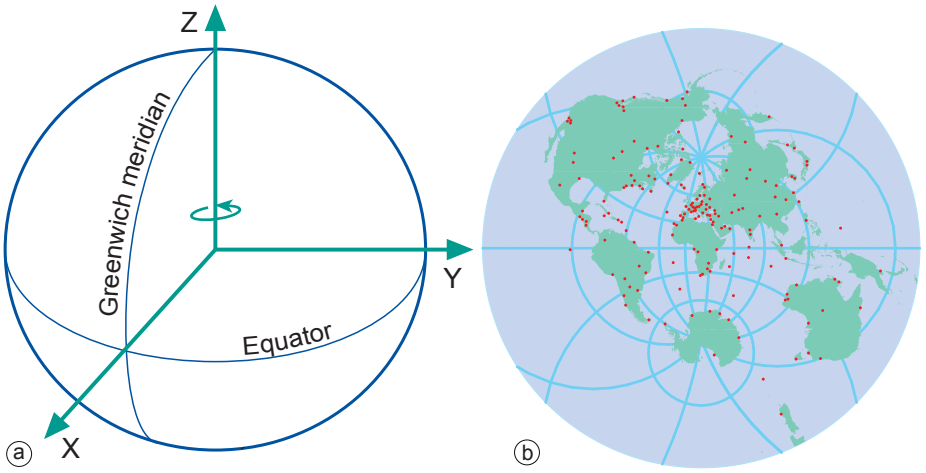


Figure 3.7
(a) The International Terrestrial Reference System (ITRS) and; (b) the International Terrestrial Reference Frame (ITRF) visualized as a distributed set of ground control stations (represented by red dots).

tional change in X, Y and Z in metres per year) of the different stations. The velocities are used to propagate measurements to other epochs (times). The trend is to use the ITRF everywhere in the world for reasons of global compatibility.

geocentric datums

GPS uses the World Geodetic System 1984 (WGS84) as its reference system. It has been refined on several occasions and is now aligned with the ITRF to within a few centimetres worldwide. Global horizontal datums, such as ITRF2000 or WGS84, are also called geocentric datums because they are geocentrically positioned with respect to the centre of mass of the Earth. They became available only recently (roughly, since the 1960s), as a result of advances in extra-terrestrial positioning techniques. [1]

Since the range and shape of satellite orbits are directly related to the centre of mass of the Earth, observations of natural or artificial satellites can be used to pinpoint the centre of mass of the Earth, and hence the origin of the ITRS[2]. This technique can also be used for the realization of global ellipsoids and datums at levels of accuracy required for large-scale mapping.

To implement the ITRF in a particular region, a densification of control stations is needed to ensure that there are enough coordinated reference points available in that region. These control stations are equipped with permanently operating satellite positioning equipment (i.e. GPS receivers and auxiliary equipment) and communication links. Examples of (networks consisting of) such permanent tracking stations are the Actief GNSS Referentie Systeem Nederland (AGRS) in the Netherlands and the Satellitenpositionierungsdienst der deutschen Landesvermessung (SAPOS) in Germany.

We can transform ITRF coordinates (X, Y and Z in metres) into geographic coordinates (ϕ, λ, h) with respect to the GRS80 ellipsoid without the loss of accuracy. However the ellipsoidal height h obtained through this straightforward transformation has no physical meaning and is contrary to our intuitive human perception of height. Therefore, we use instead the height, H, above the Geoid (see Figure 3.8). It is foreseeable that global 3D spatial referencing in terms of (ϕ, λ, H) could become ubiquitous in the next 10–15 years. If, by then, all published maps are also globally referenced the underlying spatial referencing concepts will become transparent and, hence, irrelevant to users of geoinformation.

3D spatial referencing

[1] Extra-terrestrial positioning techniques include, for example, Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), Global Positioning System (GPS), and Very Long Baseline Interferometry (VLBI).

[2] In the case of an idealized spherical Earth, it is one of the focal points of the elliptical orbits.

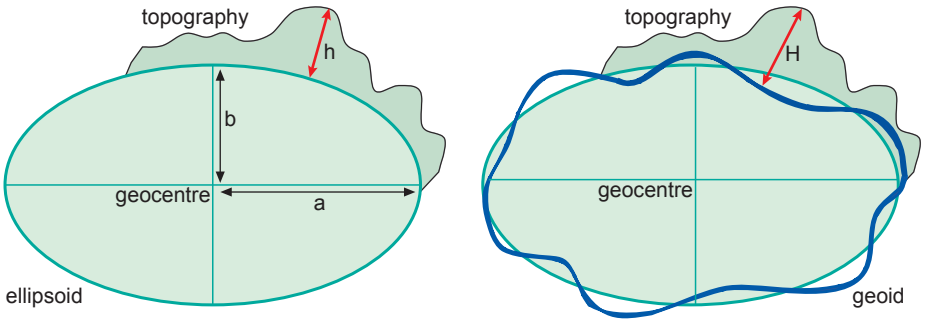


Figure 3.8
Height h above the geocentric ellipsoid, and height H above the Geoid. h is measured orthogonal to the ellipsoid, H orthogonal to the Geoid.

Hundreds of existing local horizontal and vertical datums are still relevant because they form the basis of map products all over the world. For the next few years we still have to deal with both local and global datums, until the former are eventually phased out. During the transition period, we will need tools to transform coordinates from local horizontal datums to a global horizontal datum and vice versa (see Subsection 3.1.4). The organizations that usually develop transformation tools and make them available to the user community are provincial or national mapping organizations (NMOs) and cadastral authorities.

3.1.2 Coordinate systems

spatial coordinate systems

planar coordinate systems

Spatial data are special, because they are spatially referenced. Different kinds of coordinate systems are used to position data in space. Here we distinguish between *spatial* and *planar* coordinate systems. *Spatial* (or global) coordinate systems locate data either on the Earth's surface in a 3D space or on the Earth's reference surface (ellipsoid or sphere) in a 2D space. *Planar* coordinate systems, on the other hand, locate data on the flat surface of a map in a 2D space. Initially the 2D Cartesian coordinate system and the 2D polar coordinate system will be examined. This will be followed by a discussion of the geographic coordinate system in a 2D and 3D space and the geocentric coordinate system, also known as the 3D Cartesian coordinate system.

2D geographic coordinates (ϕ, λ)

The most widely used global coordinate system consists of lines of geographic *latitude* (phi or ϕ or φ) and *longitude* (lambda or λ). Lines of equal latitude are called parallels. They form circles on the surface of the ellipsoid.[3]. Lines of equal longitude are called meridians and form ellipses (meridian ellipses) on the ellipsoid (Figure 3.9)

The latitude (ϕ) of a point P (Figure 3.10) is the angle between the ellipsoidal normal through P' and the equatorial plane. Latitude is zero on the Equator ($\phi = 0°$), and increases towards the two poles to maximum values of $\phi = +90°$ (N 90°) at the North Pole and $\phi = -90°$ (S 90°) at the South Pole.

The longitude (λ) of the point is the angle between the meridian ellipse that passes through Greenwich and the meridian ellipse containing the point in question. It is measured on the equatorial plane from the meridian of Greenwich ($\lambda = 0°$), either eastwards through $\lambda = +180°$ (E 180°) or westwards through $\lambda = -180°$ (W 180°).

Latitude and longitude represent the geographic coordinates (ϕ, λ) of a point P' (Figure 3.10) with respect to the selected reference surface. They are always given in angular units. For example, the coordinates for the City Hall in Enschede are:[4]

[3] The concept of geographic coordinates can also be applied to a sphere.
[4] This latitude and longitude refers to the Amersfoort datum. The use of a different reference surface will

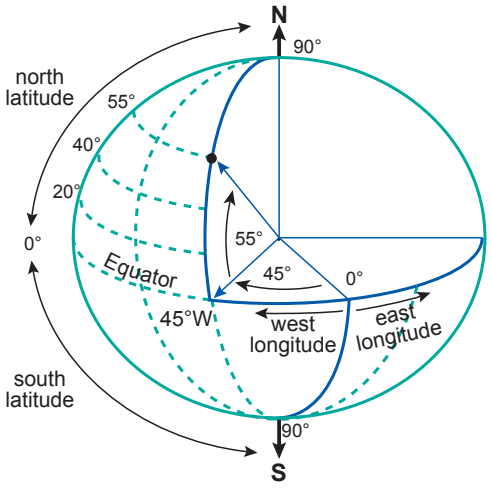


Figure 3.9
Latitude (ϕ) and longitude (λ) angles express the position of points in the 2D geographic coordinate system.

$$\phi = 52^\circ 13' 26.2'' N,\ \lambda = 6^\circ 53' 32.1'' E$$

The graticule on a map represents the projected position of the geographic coordinates (ϕ, λ) at constant intervals or, in other words, the projected position of selected meridians and parallels (Figure 3.13). The shape of the graticule depends largely on the characteristics of the map projection and the scale of the map.

3D geographic coordinates (ϕ, λ, h)

3D geographic coordinates (ϕ, λ, h) are obtained by introducing ellipsoidal height (h) into the system. The ellipsoidal height (h) of a point is the vertical distance of the point in question above the ellipsoid. It is measured in distance units along the ellipsoidal normal from the point to the ellipsoid surface. 3D geographic coordinates can be used to define a position on the surface of the Earth (point P in Figure 3.10).

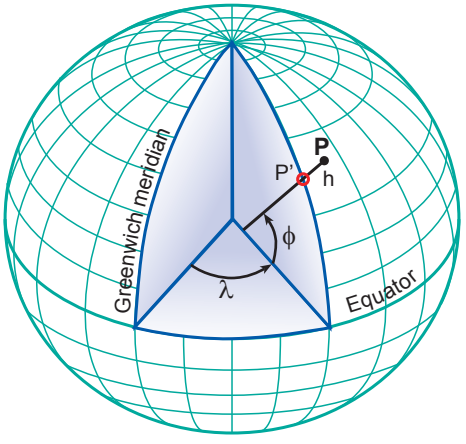


Figure 3.10
The angles of latitude (ϕ) and longitude (λ) and the ellipsoidal height (h) represent the 3D geographic coordinate system.

result in different angles of latitude and longitude.

3D geocentric coordinates (X, Y, Z)

An alternative method for defining a 3D position on the surface of the Earth is to use geocentric coordinates (X, Y, Z), also known as *3D Cartesian coordinates*. The system's origin lies at the Earth's centre of mass, with the X and Y axes on the plane of the Equator. The X-axis passes through the meridian of Greenwich and the Z-axis coincides with the Earth's axis of rotation. The three axes are mutually orthogonal and form a right-handed system. Geocentric coordinates can be used to define a position on the surface of the Earth (point P in Figure 3.11).

polar motion

The rotational axis of the Earth, however, changes position over time (referred to as *polar motion*). To compensate for this, the mean position of the pole in the year 1903 (based on observations between 1900 and 1905) is used to define what is referred to as the "Conventional International Origin" (CIO).

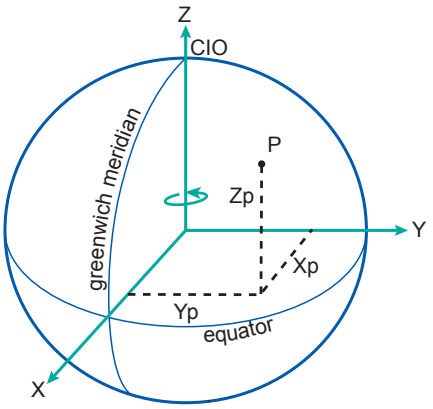


Figure 3.11
An illustration of the 3D geocentric coordinate system (see text for further explanation).

2D Cartesian coordinates (X, Y)

A flat map has only two dimensions: width (left to right) and length (bottom to top). Transforming the three dimensional Earth onto a two-dimensional map is the subject matter of map projections and coordinate transformations (Subsection 3.1.3 and Subsection 3.1.4). Here, as for several other cartographic applications, *two-dimensional Cartesian coordinates* (x, y), also known as *planar rectangular coordinates*, describe the location of any point unambiguously.

The 2D Cartesian coordinate system is one of intersecting perpendicular lines with the X-axis and the Y-axis as principal axes. The X-axis (the *Easting*) is the horizontal axis and the Y-axis (the *Northing*) is the vertical axis with an intersection at the *origin*. The plane is marked at intervals by equally-spaced coordinate lines that together form the *map grid* . Given two numerical coordinates x and y for point P, one can unambiguously specify any location P on the map (Figure 3.12).

false origin

Usually, the origin is assigned the coordinates $x = 0$ and $y = 0$. Sometimes, however, large positive values are added to the origin coordinates. This is to avoid negative values for the x and y coordinates in cases where the origin of the coordinate system is located inside the specific area of interest. The point that then has the coordinates $x = 0$ and $y = 0$ is called the *false origin*. The Rijksdriehoekstelsel (RD) in the Netherlands is an example of a system with a false origin. The system is based on the azimuthal double stereographic projection (see Section 3.1.3), with the Bessel ellipsoid used as reference surface. The origin was shifted from the projection centre (Amersfoort) towards the southwest(false origin)to avoid negative coordinates inside the country (see Figure 3.13).

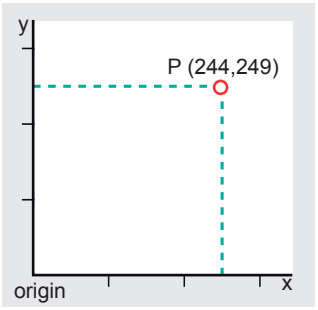


Figure 3.12
An illustration of the 2D Cartesian coordinate system (see text for further explanation).

The grid on a map represents lines having constant 2D Cartesian coordinates (Figure 3.13). It is almost always a rectangular system and is used on large- and medium-scale maps to enable detailed calculations and positioning. The map grid is usually not used on small-scale maps (about 1:1,000,000 or smaller). Scale distortions that result from transforming the Earth's curved surface to the mapping plane are so great on small-scale maps that detailed calculations and positioning become difficult.

map grid

2D Polar coordinates (α, d)

Another way of defining a point in a plane is by using polar coordinates. This is the distance d from the origin to the point concerned and the angle α between a fixed (or zero) direction and the direction to the point. The angle α is called *azimuth* or *bearing*

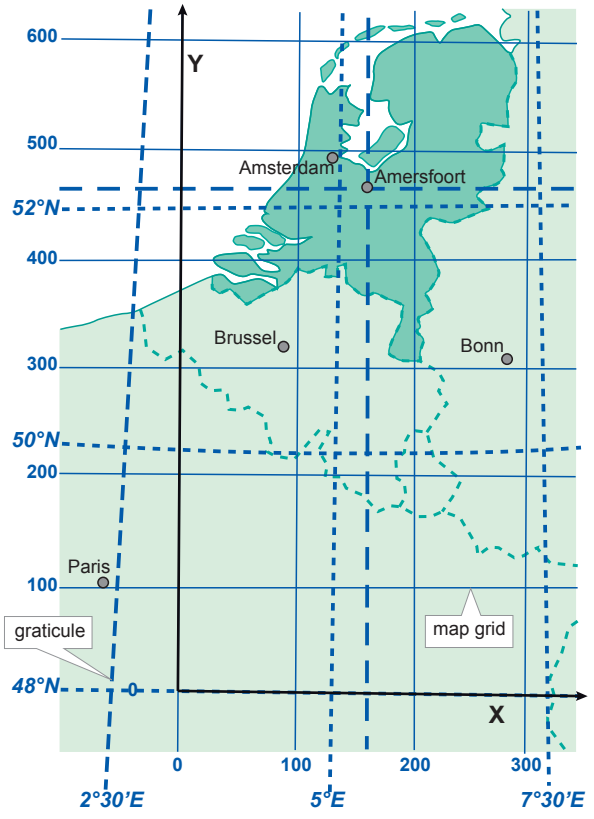


Figure 3.13
The coordinate system of the Netherlands represented by the map grid and the graticule. The origin of the coordinate system has been shifted (the false origin) from the projection centre (Amersfoort) towards the southwest.

and is measured in a clockwise direction. It is given in angular units while the distance d is expressed in length units.

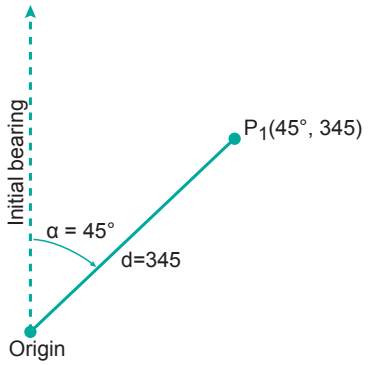


Figure 3.14
An illustration of the 2D Polar coordinate system (see text for further explanation).

Bearings are always related to a fixed direction (initial bearing) or a datum line. In principle, this reference line can be chosen freely. Three different, widely used fixed directions are: *True North*, *Grid North* and *Magnetic North*. The corresponding bearings are true (or geodetic) bearings, grid bearings and magnetic (or compass) bearings, respectively.

polar coordinates

Polar coordinates are often used in land surveying. For some types of surveying instruments, it is advantageous to make use of this coordinate system. The development of precise, remote-distance measurement techniques has led to a virtually universal preference for the polar coordinate method for detailed surveys.

3.1.3 Map projections

Maps are one of the world's oldest types of document. In the days that our planet was thought to be *flat*, a map was simply a miniature representation of a part of the world. To represent the specifically curved Earth's surface, a map needs to be a flattened representation of a part of the planet. Map projection concerns itself with ways of translating the curved surface of the Earth into a flat, 2D map.

> *Map projection* is a mathematically described technique for representing the Earth's curved surface on a flat map.

To represent parts of the surface of the Earth on a flat, printed map or a computer screen, the curved horizontal reference surface must be mapped onto a 2D mapping plane. The reference surface for large-scale mapping is usually an oblate ellipsoid; for small-scale mapping it is a sphere.[5] Mapping onto a 2D mapping plane means transforming each point on the reference surface with geographic coordinates (ϕ, λ) to a set of Cartesian coordinates (x, y) that represent positions on the map plane (Figure 3.15).

mapping equations

The actual mapping cannot usually be visualized as a true geometric projection, directly onto the mapping plane. Rather, it is achieved through mapping equations. A *forward mapping equation* transforms the geographic coordinates (ϕ, λ) of a point on the curved reference surface to a set of planar Cartesian coordinates (x, y), representing the position of the same point on the map plane:

$$(x, y) = f(\phi, \lambda)$$

[5] In practice, maps at scales of 1:1,000,000 or smaller can use the mathematically simpler sphere without the risk of large distortions. At larger scales, the more complicated mathematics of ellipsoids is needed to prevent large distortions occurring on the map.

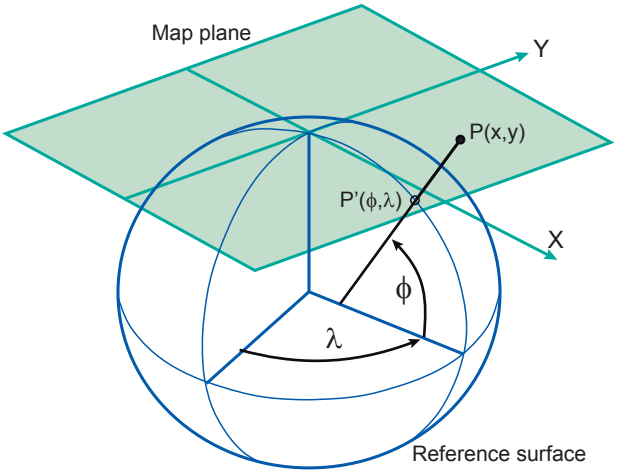


Figure 3.15
Example of a map projection in which the reference surface with geographic coordinates (ϕ, λ) is projected onto the 2D mapping plane with 2D Cartesian coordinates (x, y).

The corresponding *inverse mapping equation* transforms mathematically the planar Cartesian coordinates (x, y) of a point on the map plane to a set of geographic coordinates (ϕ, λ) on the curved reference surface:

$$(\phi, \lambda) = f(x, y)$$

The Mercator projection (spherical assumption) [106], a commonly used mapping projection, can be used to illustrate the use of mapping equations. The *forward mapping equation* for the Mercator projection is:[6]

$$x = R(\lambda - \lambda_0)$$
$$y = R \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right)$$

The *inverse mapping equation* for the Mercator projection is:

$$\phi = \frac{\pi}{2} - 2 \arctan \left(e^{-\frac{y}{R}} \right)$$
$$\lambda = \frac{x}{R} + \lambda_0$$

Classification of map projections

Many map projections have been developed, each with its own specific qualities. It is these qualities that make the resulting maps useful for certain purposes. By definition, any map projection is associated with scale distortions. There is simply no way to flatten an ellipsoidal or spherical surface without stretching some parts of the surface more than others. The amount and kind of distortions a map has depends on the type of map projection.

scale distortions

Some map projections can be visualized as true geometric projections directly onto the mapping plane—known as an azimuthal projections—or onto an intermediate surface,

[6] When an ellipsoid is used as a reference surface, the equations are considerably more complicated than those introduced here. R is the radius of the spherical reference surface at the scale of the map; ϕ and λ are given in radians; λ_0 is the central meridian of the projection; $e = 2.7182818$, the base of the natural logarithms, not the eccentricity.

which is then rolled out onto the mapping plane. Typical choices for such intermediate surfaces are cones and cylinders. These map projections are called conical or cylindrical projections, respectively. Figure 3.16 shows the surfaces involved in these three classes of projection.

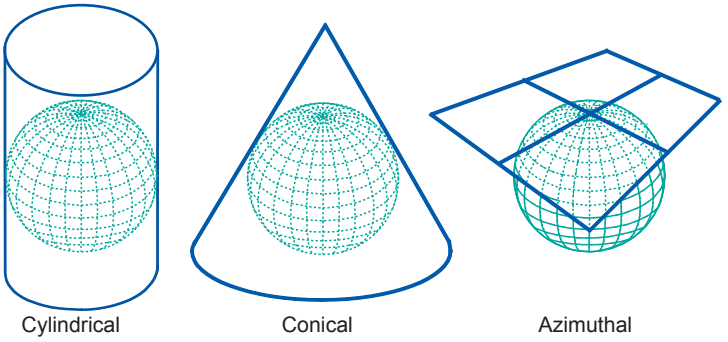


Figure 3.16
Classes of map projections

The azimuthal, conical, and cylindrical surfaces in Figure 3.16 are all *tangent* surfaces, i.e. they touch the horizontal reference surface at one point (azimuthal), or along a closed line (cone and cylinder), only. Another class of projections is obtained if the surfaces are chosen to be *secant* to (to intersect with) the horizontal reference surface; see Figure 3.17. Then, the reference surface is intersected along one closed line (azimuthal) or two closed lines (cone and cylinder). Secant map surfaces are used to reduce or average out scale errors since the line(s) of intersection are not distorted on the map.

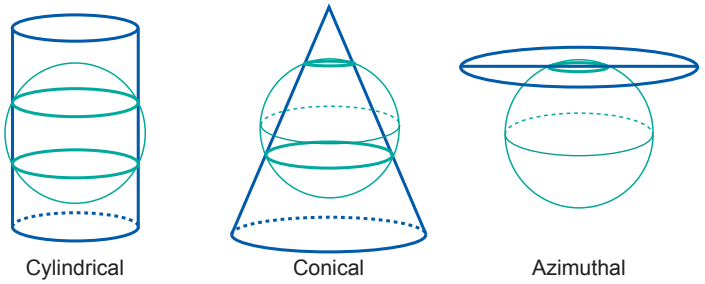


Figure 3.17
Three classes of secant projection

normal, transverse, and oblique projections

In the geometric depiction of map projections in Figures 3.16 and 3.17, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the ellipsoid or sphere, i.e. a line through the North and South poles. In this case, the projection is said to be a *normal projection*. The other cases are *transverse projections* (symmetry axis in the equatorial plane) and *oblique projections* (symmetry axis is somewhere between the rotation axis and the equator of the ellipsoid or sphere); see Figure 3.18.

The Universal Transverse Mercator (UTM) is a system of map projection that is used worldwide. It is derived from the Transverse Mercator projection (also known as Gauss-Kruger or Gauss conformal projection). UTM uses a transverse cylinder secant to the horizontal reference surface. It divides the world into 60 narrow longitudinal zones of 6 degrees, numbered from 1 to 60. The narrow zones of 6 degrees (and the secant map surface) make the distortions small enough for large-scale topographic mapping.

Normal cylindrical projections are typically used to map the world in its entirety. Conical projections are often used to map individual continents, whereas the normal az-

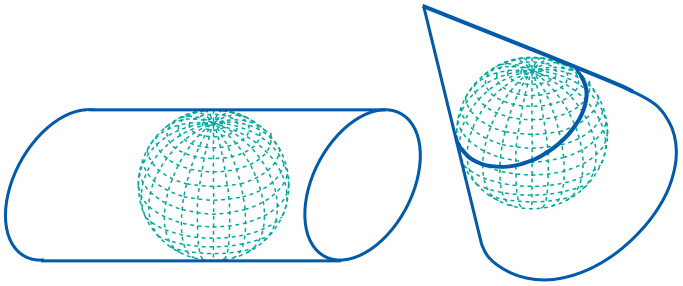


Figure 3.18
A transverse cylindrical and an oblique conical projection

imuthal projection may be used to map polar areas. Transverse and oblique aspects of many projections can be used for most parts of the world.

It is also important to consider the shape of the area to be mapped. Ideally, the general shape of the mapping area should be well-match with the distortion pattern of a specific projection. If an area is approximately circular, it is possible to create a map that minimizes distortion for that area on the basis of an azimuthal projection. Cylindrical projection is best for a rectangular area and conic projection for a triangular area.

So far, we have not specified *how* the curved horizontal reference surface is projected onto a plane, cone or cylinder. *How* this is done determines what kind of *distortions* the map will have compared to the original curved reference surface. The distortion properties of a map are typically classified according to what is *not* distorted on the map:

distortion properties

- With a *conformal* map projection, the angles between lines in the map are identical to the angles between the original lines on the curved reference surface. This means that angles (with short sides) and shapes (of small areas) are shown correctly on the map.
- With an *equal-area* (equivalent) map projection, the areas in the map are identical to the areas on the curved reference surface (taking into account the map scale), which means that areas are represented correctly on the map.
- With an *equidistant* map projection, the length of particular lines in the map are the same as the length of the original lines on the curved reference surface (taking into account the map scale).

A particular map projection can exhibit only one of these three properties. No map projection can be both conformal and equal-area, for example.

The most appropriate type of distortion for a map depends largely on the purposes for which the map will be used. Conformal map projections represent angles correctly, but as the region becomes larger they show considerable area distortions (Figure 3.19). Maps used for the measurement of angles (e.g. aeronautical charts, topographic maps) often make use of a conformal map projection such as the UTM projection.

Equal-area projections, on the other hand, represent areas correctly, but as the region becomes larger, considerable distortions of angles and, consequently, shapes occur (Figure 3.20). Maps that are to be used for measuring area (e.g. distribution maps) are often made using an equal-area map projection.

The equidistant property is achievable only to a limited degree. That is, true distances can be shown only from one or two points to any other point on the map, or in certain directions. If a map is true to scale along the meridians (i.e. no distortion in the

Figure 3.19
The Mercator projection, a cylindrical map projection with conformal properties. The area distortions are significant towards the polar regions.

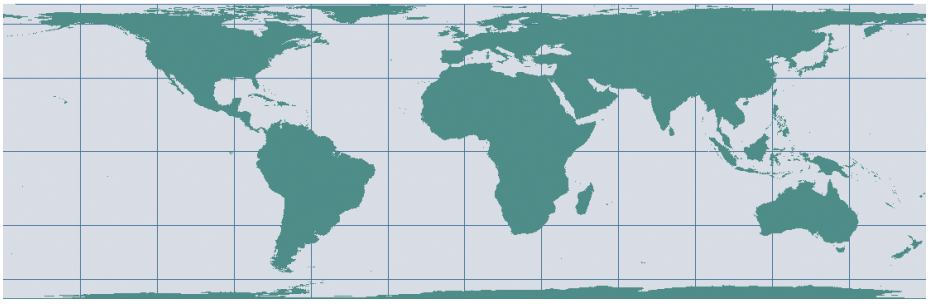


Figure 3.20
The cylindrical equal-area projection, i.e. a cylindrical map projection with equal-area properties. Distortions of shapes are significant towards the poles.

North–South direction), we say that the map is *equidistant along the meridians* (e.g. an equidistant cylindrical projection) (Figure 3.21). If a map is true to scale along all parallels we say the map is *equidistant along the parallels* (i.e. no distortion in the East–West direction). Maps for which the area and angle distortions need to be reasonably acceptable (several thematic maps) often make use of an equidistant map projection.

As these discussions indicate, a particular map projection can be classified. An example would be the classification "conformal conic projection with two standard parallels", which means that the projection is a conformal map projection, that the intermediate surface is a cone, and that the cone intersects the ellipsoid (or sphere) along two parallels. In other words, the cone is secant and the cone's symmetry axis is parallel to the rotation axis. (This would amount to the middle projection displayed in Figure 3.17). This projection is also referred to as "Lambert's conical projection" [47].

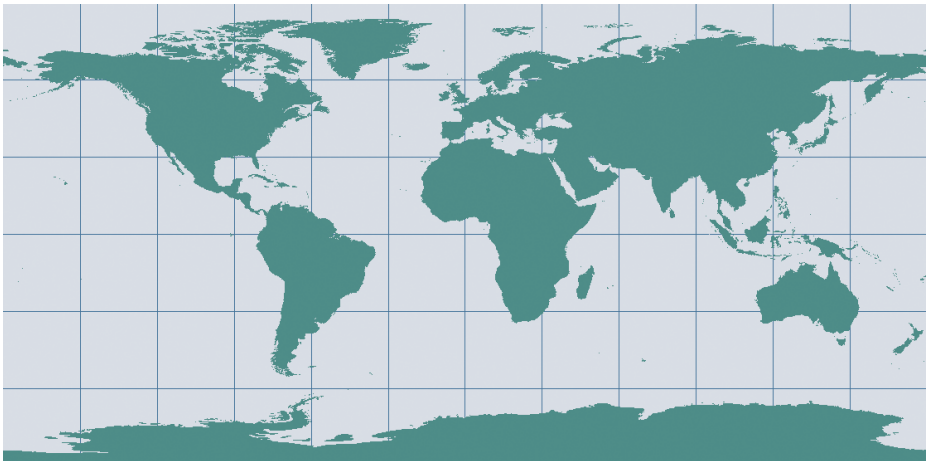


Figure 3.21
The equidistant cylindrical projection (also called Plate Carrée projection), a cylindrical map projection with equidistant properties. The map is equidistant (true to scale) along the meridians. Both shape and area are reasonably well preserved.

3.1.4 Coordinate transformations

Users of geoinformation often need transformations from a particular 2D coordinate system to another system. This includes the transformation of polar coordinates into Cartesian map coordinates, or the transformation from one 2D Cartesian (x, y) system of a specific map projection into another 2D Cartesian (x', y') system of a defined map projection. This transformation is based on relating the two systems on the basis of a set of selected points whose coordinates are known in both systems, such as ground control points or common points such as corners of houses or road intersections. Image and scanned data are usually transformed by this method. The transformations may be conformal, affine, polynomial or of another type, depending on the geometric errors in the data set.

2D Polar to 2D Cartesian transformations

The transformation of polar coordinates (α, d), into Cartesian map coordinates (x, y) is done when field measurements, i.e. angular and distance measurements, are transformed into map coordinates. The equation for this transformation is:

$$x = d \sin \alpha$$
$$y = d \cos \alpha$$

The inverse equation is:

$$\alpha = \arctan\left(\frac{x}{y}\right)$$
$$d^2 = x^2 + y^2$$

Changing map projection

Forward and inverse mapping equations are normally used to transform data from one map projection into another. The inverse equation of the source projection is used first to transform source projection coordinates (x, y) to geographic coordinates (ϕ, λ). Next, the forward equation of the target projection is used to transform the geographic coordinates (ϕ, λ) into target projection coordinates (x', y'). The first equation takes us from a projection A into geographic coordinates. The second takes us from geographic

coordinates (ϕ, λ) to another map projection B. These principles are illustrated in Figure 3.22.

Historically, GI Science has dealt with data referenced spatially with respect to the (x, y) coordinates of a specific map projection. For application domains requiring 3D spatial referencing, a height coordinate may be added to the (x, y) coordinates of the point. The additional height coordinate can be a height H above mean sea level, which is a height with a physical meaning. The (x, y, H) coordinates then represent the location of objects in a 3D space.

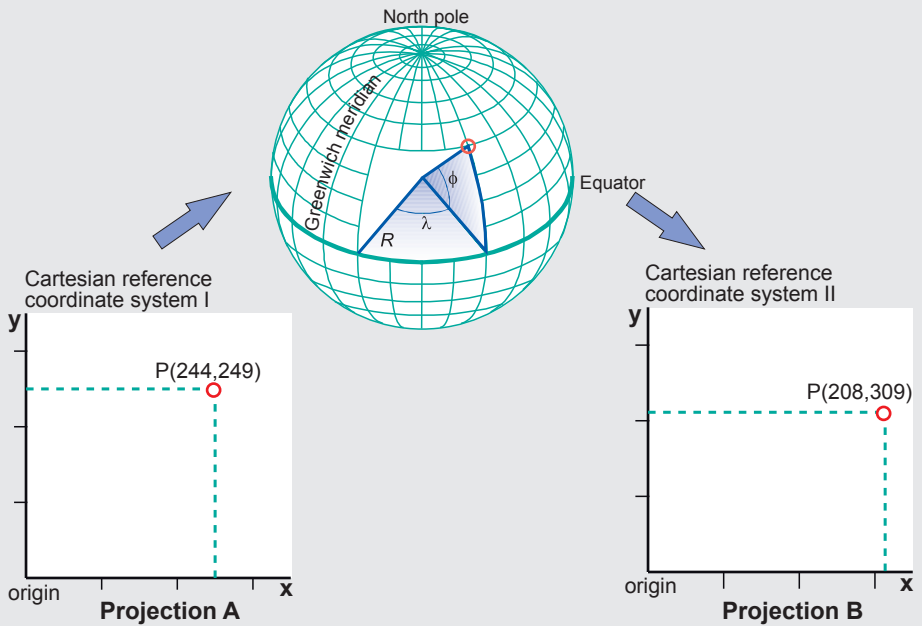


Figure 3.22
The principle of changing from one map projection to another.

Datum transformations

A change of map projection may also include a change of the horizontal datum. This is the case when the source projection is based upon a different horizontal datum than the target projection. If the difference in horizontal datums is ignored, there will not be a perfect match between adjacent maps of neighbouring countries or between overlaid maps originating from different projections. It may lead to differences of several hundreds of metres in the resulting coordinates. Therefore, spatial data with different underlying horizontal datums may require *datum transformation*.

Suppose we wish to transform spatial data from the UTM projection to the Dutch RD system, and suppose that the data in the UTM system are related to the European Datum 1950 (ED50), while the Dutch RD system is based on the Amersfoort datum. To achieve a perfect match, in this example the change of map projection should be combined with a datum transformation step; see Figure 3.23.

The inverse equation of projection A is used first to take us from the map coordinates (x, y) of projection A to the geographic coordinates (ϕ, λ, h) for datum A. A height coordinate (h or H) may be added to the (x, y) map coordinates. Next, the datum transformation takes us from these coordinates to the geographic coordinates (ϕ, λ, h) for datum B. Finally, the forward equation of projection B takes us from the geographic coordinates (ϕ, λ, h) for datum B to the map coordinates (x', y') of projection B.

Mathematically, a datum transformation is feasible via the geocentric coordinates (x, y, z)

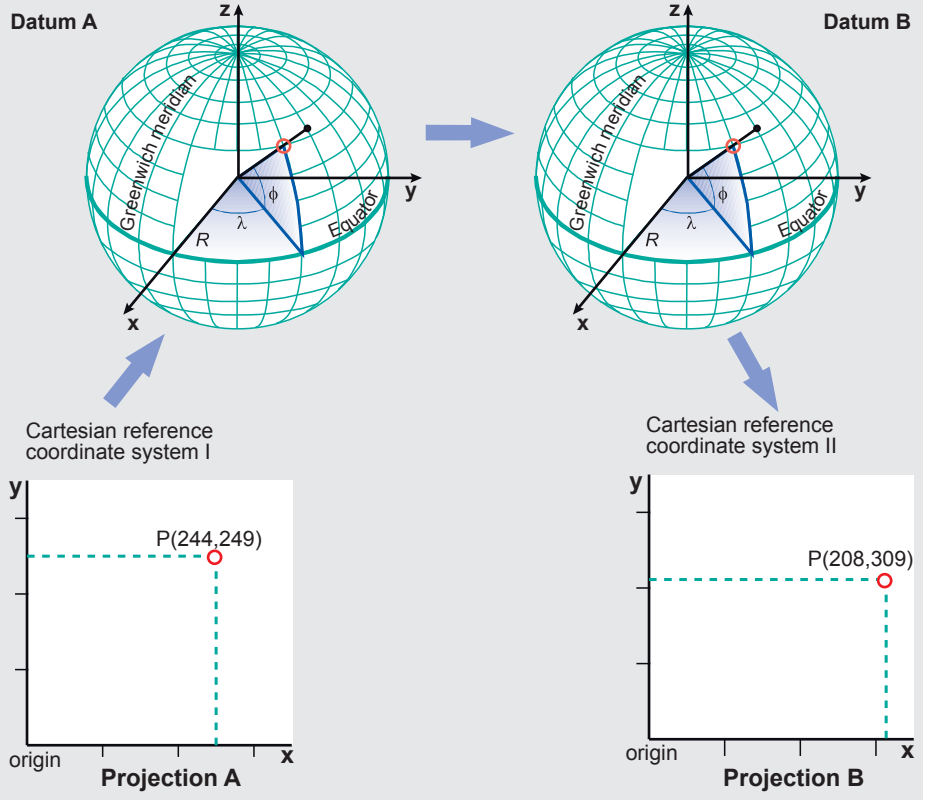


Figure 3.23
The principle of changing from one projection into another, combined with a datum transformation from datum A to datum B.

or directly by relating the geographic coordinates of both datum systems. The latter relates the ellipsoidal latitude (ϕ) and longitude (λ), and possibly also the ellipsoidal height (h), of both datum systems [59].

Geographic coordinates $(\phi,\ \lambda,\ h)$ can be transformed into geocentric coordinates $(x,\ y,\ z)$, and vice versa. The datum transformation via the geocentric coordinates implies a 3D similarity transformation. This is essentially a transformation between two orthogonal 3D Cartesian spatial reference frames together with some elementary tools from adjustment theory. The transformation is usually expressed with seven parameters: three rotation angles $(\alpha,\ \beta,\ \gamma)$, three origin shifts $(X_0,\ Y_0,\ Z_0)$ and a scale factor (s). The inputs are the coordinates of points in datum A and coordinates of the same points in datum B. The output are estimates of the seven transformation parameters and a measure of the likely error of the estimate.

datum transformation parameters

Datum transformation parameters have to be estimated on the basis of a set of selected points whose coordinates are known in both datum systems. If the coordinates of these points are not correct—often the case for points measured on a local datum system—the estimated parameters may be inaccurate and hence the datum transformation will be inaccurate.

Inaccuracies often occur when we transform coordinates from a local horizontal datum to a global geocentric datum. The coordinates in the local horizontal datum may be distorted by several tens of metres because of the inherent inaccuracies of the measurements used in the triangulation network. These inherent inaccuracies are also responsible for another complication: the transformation parameters are not unique. Their estimation depends on the particular choice of common points and whether all

seven transformation parameters, or only some of them, are estimated.

The example in Table 3.2 illustrates the transformation of the Cartesian coordinates of a point in the state of Baden-Württemberg, Germany, from ITRF to Cartesian coordinates in the Potsdam Datum. Sets of numerical values for the transformation parameters are available from three organizations:

Table 3.2
Three different sets of datum transformation parameters from three different organizations for transforming a point from ITRF to the Potsdam datum.

	Parameter	National set	Provincial set	NIMA set
scale	s	$1 - 8.3 \cdot 10^{-6}$	$1 - 9.2 \cdot 10^{-6}$	1
angles	α	$+1.04''$	$+0.32''$	
	β	$+0.35''$	$+3.18''$	
	γ	$-3.08''$	$-0.91''$	
shifts (m)	X_0	-581.99	-518.19	-635
	Y_0	-105.01	-43.58	-27
	Z_0	-414.00	-466.14	-450

1. The federal mapping organization of Germany (labelled "National set" in Table 3.2) provided a set calculated using common points distributed throughout Germany. This set contains all seven parameters and is valid for whole Germany.
2. The mapping organization of Baden-Württemberg (labelled "Provincial set" in Table 3.2) provided a set calculated using common points distributed throughout the state of Baden-Württemberg. This set contains all seven parameters and is valid only within the state borders.
3. The National Imagery and Mapping Agency (NIMA) of the U.S.A. (labelled "NIMA set" in Table 3.2) provided a set calculated using common points distributed throughout Germany and based on the ITRF. This set contains a coordinate shift only (no rotations, and scale equals unity). This set is valid for whole Germany.

The three sets of transformation parameters vary by several tens of metres, for reasons already mentioned. The sets of transformation parameters were used to transform the ITRF cartesian coordinates of a point in the state of Baden-Württemberg. Its ITRF (X, Y, Z) coordinates are:

$$(4,156,939.96\ \text{m},\ 671,428.74\ \text{m},\ 4,774,958.21\ \text{m}).$$

The three sets of transformed coordinates for the Potsdam datum are given in Table 3.3.

Table 3.3
Three sets of transformed coordinates for a point in the state of Baden-Württemberg, Germany.

Potsdam coordinates	National set (m)	Provincial set (m)	NIMA set (m)
X	$4,156,305.32$	$4,156,306.94$	$4,156,304.96$
Y	$671,404.31$	$671,404.64$	$671,401.74$
Z	$4,774,508.25$	$4,774,511.10$	$4,774,508.21$

The three sets of transformed coordinates differ by only a few metres from each other. In a different country, the level of agreement could be a within centimetres, but it can be up to tens of metres of each other, depending upon the quality of implementation of the local horizontal datum.

3.2 Satellite-based positioning

The importance of satellites in spatial referencing has already been mentioned before. Satellites have allowed us to create geocentric reference systems and to increase the level of spatial accuracy substantially. Satellite-based systems are critical tools in geodetic engineering for the maintenance of the ITRF. They also play a key role in mapping and surveying in the field, as well as in a growing number of applications requiring positioning techniques. The setting up a satellite-based positioning system requires the implementation of three hardware segments:

1. the *space segment*, i.e. the satellites that orbit the Earth and the radio signals that they emit;
2. the *control segment*, i.e. the ground stations that monitor and maintain the components of the space segment;
3. the *user segment*, i.e. the users, along with the hardware and software they use for positioning.

In satellite positioning, the central problem is to determine the values (X, Y, Z) of a receiver of satellite signals, i.e. to determine the position of the receiver with the accuracy and precision required. The degree of accuracy and precision needed depends on the application, as does timeliness, i.e. are the position values required in real time or can they be determined later during post-processing. Finally, some applications, such as navigation, require kinematic approaches, which take into account the fact that the receiver is not stationary, but moving.

Some fundamental aspects of satellite-based positioning and a brief review of currently available technologies follows.

3.2.1 Absolute positioning

The working principles of absolute, satellite-based positioning are fairly simple:

1. A satellite, equipped with a clock, sends a radio message at a specific moment that includes
   (a) the *satellite identifier*,
   (b) its *position in orbit*, and
   (c) its *clock reading*.
2. A receiver on or above the planet, also equipped with a clock, receives the message slightly later and reads its own clock.
3. From the time delay observed between the two clock readings, and knowing the speed of radio transmission through the medium between (satellite) sender and receiver, the receiver can compute the distance to the sender, also known as the satellite's *pseudorange*. This *pseudorange* is the apparent distance from satellite to receiver, computed from the time delay with which its radio signal is received.

Such a computation places the position of the receiver on a sphere of radius equal to the computed pseudorange (see Figure 3.24a). If, instantaneously, the receiver were to do the same with a message from another satellite positioned elsewhere, the position of the receiver would be placed on another sphere. The intersection of the two spheres,

trilateration

which have different centres, describes a circle as being the set of possible positions of the receiver (see Figure 3.24b). If a third satellite message is taken into consideration, the three spheres intersect at, at most, two positions, one of which is the actual position of the receiver. In most, if not all practical situations where two positions result, one of them is a highly unlikely position for a signal receiver, thus narrowing down the true position of the receiver. The overall procedure is known as *trilateration*: the determination of a position based on three distances.

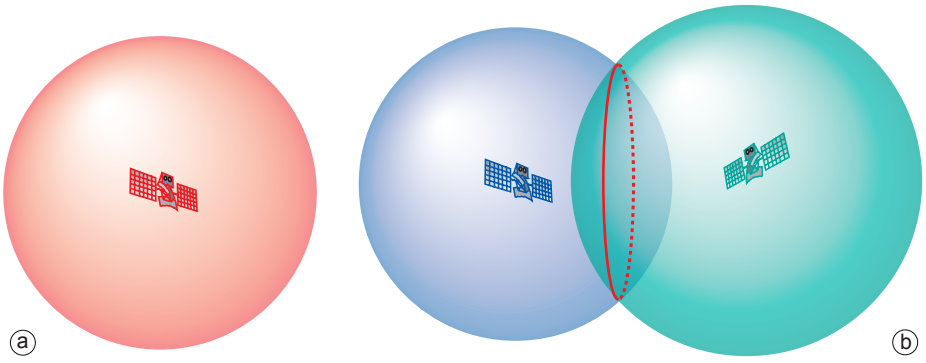


Figure 3.24
Pseudorange positioning: (a) With just one satellite, the receiver position is somewhere on a sphere, (b) With two satellites, the position is located where the two spheres intersect, i.e. in a circle. Not shown: with three satellites, its position is where the three spheres intersect.

clock bias

It would appear, therefore, that the signals of three satellites would be sufficient to determine a *positional fix* for our receiver. In theory this is true, but in practice it is not. The reason being that satellite clocks and the receiver clock are never exactly synchronized. Satellite clocks are costly, high-precision, atomic clocks that we can consider synchronized for the time being, but the receiver typically uses a far cheaper, quartz clock that is not synchronized with satellite clocks. This brings an additional unknown variable into play, namely the synchronization bias of the receiver clock, i.e. the difference in time readings between it and the satellite clocks.

3D positioning

Our set of unknown variables has now become $(X, Y, Z, \Delta t)$ representing a 3D position and a clock bias. The problem can be solved by including the information obtained from a fourth satellite message, (see Figure 3.25). This will result in the determination of the receiver's actual position (X, Y, Z), as well as its receiver clock bias Δt, and if we correct the receiver clock for this bias we effectively turn it into a high-precision atomic clock as well!

Obtaining a high-precision clock is a fortunate side-effect of using the receiver, as it allows the design of experiments distributed in geographic space that demand high levels of synchronicity. One such application is the use of wireless sensor networks for researching natural phenomena such as earthquakes or meteorological patterns, and for water management.

The positioning of mobile phone users making an emergency call is yet another application. Often callers do not know their location accurately. The telephone company can trace back the call to the receiving transmitter mast, but this may be servicing an area with a radius ranging from 300 m to 6 km. That is far too inaccurate for emergency services. If all masts in the telephony network are equipped with a satellite positioning receiver (and thus, with a very high-precision synchronized clock), however, the time of reception of the call at each mast can be recorded. The *time difference of arrival* of the call between two nearby masts describes a hyperbola on the ground of possible positions of the caller. If the call is received on three masts, two hyperbolas are described, allowing intersection and thus "hyperbolic positioning". With current technology the (horizontal) accuracy would be better than 30 m.

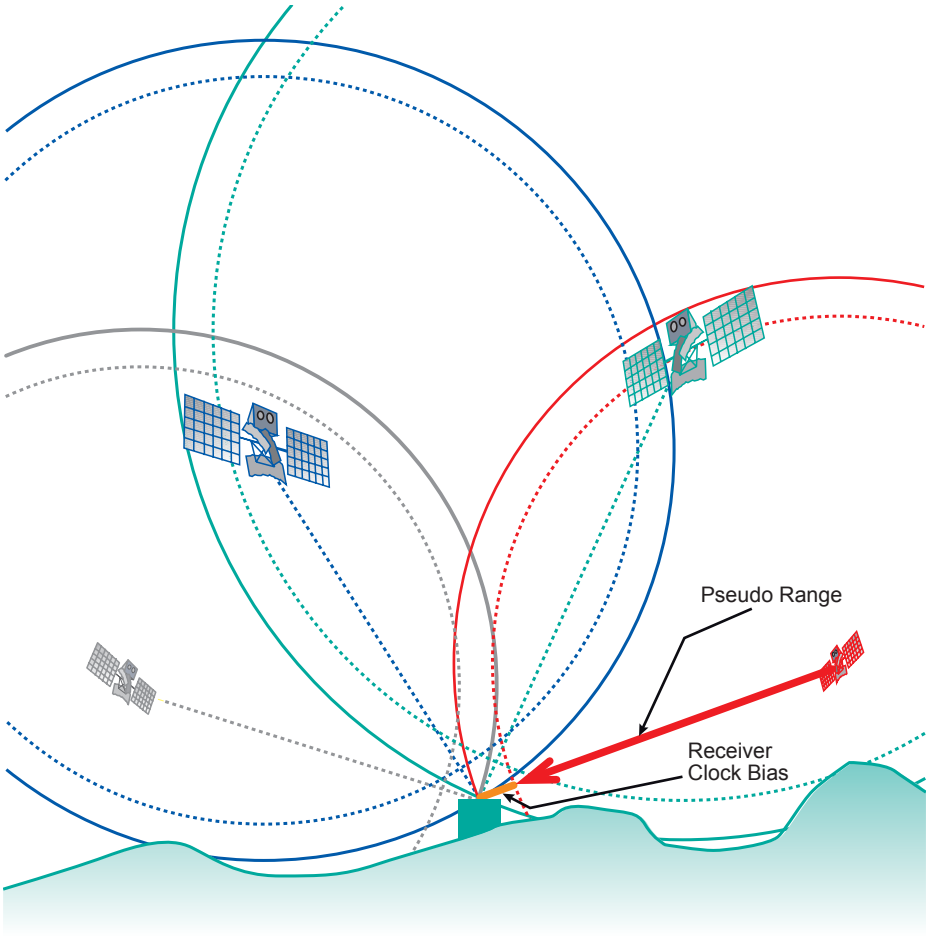


Figure 3.25
Four satellites are needed to obtain a 3D position fix. Pseudoranges are indicated for each satellite as dotted circles, representing a sphere; the actual range is represented as a solid circle, which is the pseudorange plus the range error caused by receiver clock bias.

Returning to satellite-based positioning, when only three, and not four, satellites are "in view", the receiver is capable of falling back from the above *3D positioning mode* to the inferior *2D positioning mode*. With the relative abundance of satellites in orbit around the Earth, this is a relatively rare situation, but it serves to illustrate the importance of 3D positioning.

2D positioning mode

If a 3D fix had already been obtained, the receiver simply assumes that the height above the ellipsoid has not changed since the last 3D fix. If no fix had been obtained, the receiver assumes that it is positioned at the geocentric ellipsoid adopted by the positioning system, i.e. at height $h = 0$.[7] In the receiver computations, the ellipsoid fills the slot of the missing fourth satellite sphere, and the unknown variables can therefore still be determined. Clearly, in both of these cases, the assumption upon which this computation is based is flawed and the resulting positioning in 2D mode will be unreliable—much more so if no previous fix had been obtained and one's receiver is not at all near the surface of the geocentric ellipsoid.

[7] Any receiver is capable of transforming a coordinate $(X,\ Y,\ Z)$, using a straightforward mathematical transformation, into an equivalent coordinate $(\phi,\ \lambda,\ h)$, where h is the height above the geocentric ellipsoid.

Time, clocks and world time

Before any notion of standard time existed, villages and cities simply kept track of their local time, determined from the position of the Sun in the sky. When trains became an important means of transportation, these local time systems became problematic as train scheduling required a single time system. Such a time system called for the definition of *time zones*: typically 24 geographic strips bounded by longitudes that are multiples of 15°. This and navigational demands gave rise to Greenwich Mean Time (GMT), based on the mean solar time at the meridian passing through Greenwich, United Kingdom, which is the conventional 0-meridian in geography. GMT became the world time standard of choice.

Greenwich Mean Time

GMT was later replaced by Universal Time (UT), a system still based on meridian crossings of stars, albeit distant quasars, as this approach provides more accuracy than that based on the Sun. It is still the case that the rotational velocity of our planet is not constant and the length of a solar day is increasing. So UT is not a perfect system either. It continues to be used for civilian clock time, but it has now officially been replaced by International Atomic Time (TAI). UT actually has various versions, among them UT0, UT1 and UTC. UT0 is the Earth's rotational time observed at some location. Because the Earth experiences polar motion as well, UT0 differs between locations. If we correct for polar motion, we obtain UT1, which is identical everywhere. Nevertheless, UT1 is still a somewhat erratic clock system because of the varying rotational velocity of the planet, as mentioned above. The degree of uncertainty is about 3 ms per day.

Coordinated Universal Time (UTC) is used in satellite positioning and is maintained with atomic clocks. By convention, it is always within a margin of 0.9 s of UT1, and twice annually it may be shifted to stay within that margin. This occasional shift of a *leap second* is applied at the end of 30 June or, preferably, at the end of 31 December. The last minute of such a day is then either 59 or 61 seconds long. So far, adjustments have always entailed adding a second. UTC time can only be determined to the highest precision after the fact, as atomic time is determined by the reconciliation of the observed differences between a number of atomic clocks maintained by different national time bureaus.

In recent years, we have learned to measure distance, and therefore also position, with clocks, by using satellite signals, the conversion factor being the speed of light, approximately 3×10^8 m s^{-1} in a vacuum. As a consequence, multiple seconds of clock bias could no longer be accepted, and this is where atomic clocks are at an advantage. They are very accurate time keepers, based on the exact frequencies at which specific atoms (Cesium, Rubidium and Hydrogen) make discrete energy-state jumps. Positioning satellites usually have multiple clocks on board; ground control stations have even better quality atomic clocks.

atomic clocks

Atomic clocks are not flawless, however: their timing tends to drift from true time and they, too, need to be corrected. The drift, and the change in drift over time, are monitored and included in the satellite's navigation message, so that these discrepancies can be corrected for.

3.2.2 Errors in absolute positioning

Before we continue discussing other modes of satellite-based positioning, let us take a close look at the potential for error in absolute positioning. Users of receivers are required to be sufficiently familiar with the technology in order to avoid real operating blunders such as poor receiver placement or incorrect receiver software settings, which can render positioning results virtually useless. We will skip over many of the physical and mathematical details underlying these errors; they are only mentioned

here to raise awareness and understanding among users of this technology. For background information on the calculation of positional error (specifically, the calculation of RMSE or *root mean square error*), see Subsection 5.3.2.

Errors related to the space segment

As a first source of error, operators of the control segment may, for example in times of global political tension or war, intentionally deteriorate radio signals from satellites to the general public to avoid optimal use of the system by a perceived enemy. This *selective availability*—meaning that military forces allied with the control segment *will* still have access to undisturbed signals—may cause error that has an order of magnitude larger than all other error sources combined.[8]

A second source occurs if the satellite signal contains incorrect information. Assuming that it will always know its own identifier, the satellite may make two kinds of error:

1. *Incorrect clock reading*. Even atomic clocks can be off by a small margin, and thanks to Einstein we know that moving clocks are slower than stationary clocks, due to a relativistic effect. If one understands that a clock that is off by 0.000001 s causes an computation error in the satellite's pseudorange of approximately 300 m, it becomes clear that these satellite clocks require very strict monitoring.

2. *Incorrect orbit position*. The orbit of a satellite around our planet is easy to describe mathematically if both bodies are considered point masses, but in real life they are not. For the same reasons that the Geoid is not a simply shaped surface, the gravitation pull of the Earth that a satellite experiences in orbit is not simple either. Moreover, satellite orbits are also disturbed by solar and lunar gravitation, making flight paths slightly erratic and difficult to forecast exactly.

Both types of error are strictly monitored by the ground control segment, which is responsible for correcting any errors of this nature, but it does so by applying an agreed-upon tolerance. A control station can obviously compare results of positioning computations such as those discussed above with its accurately *known* position, flagging any unacceptable errors and potentially labelling a satellite as temporarily "unhealthy" until those errors have been corrected and brought back within the agreed tolerance limits. This may be done by uploading a correction to the clock or the satellite's orbit settings.

Errors related to the medium

A third source may be due to the influence of the *medium* between sender and receiver on the satellite's radio signals. The middle atmospheric layers of the stratosphere and mesosphere are relatively harmless and of little hindrance to radio waves, but this is not true of the lower and upper layers of the atmosphere:

- *The troposphere*: the approximate 14 km-high airspace directly above the Earth's surface, which holds most of the atmosphere's oxygen and which envelops all phenomena that we call the weather. It is an obstacle that delays radio waves in a rather variable way.

- *The ionosphere*: the part of the atmosphere that is farthest from the Earth's surface. It starts at an altitude of 90 km and holds many electrically charged atoms,

[8] Selective availability was stopped at the beginning of May 2000; in late 2007 the White House decided to remove selective availability capabilities all together. However, when deemed necessary, the US government still has a range of capabilities and technology available to implement regional denial of service of civilian GPS signals in an area of conflict, effectively producing the same result.

thereby forming a protective "shield" against various forms of radiation from space, including, to some extent, radio waves. The degree of ionization shows a distinct night and day rhythm and also varies with solar activity.

The ionosphere is a more severe source of delay for satellite signals, which obviously means that pseudoranges are estimated as being larger than they actually are. When satellites emit radio signals at two or more frequencies, an estimate can be computed from differences in delay incurred for signals of different frequency, which enables correction for atmospheric delay, leading to a 10–50% improvement of accuracy. If this is not the case, or if the receiver is capable of receiving only a single frequency, a model should be applied to forecast the (especially ionospheric) delay; typically the model takes into account the time of day and current latitude of the receiver.

Errors related to the receiver's environment

Fourth in the list of sources of error is that which occurs when a radio signal is received via two or more paths between sender and receiver, typically caused by the signal bouncing off some nearby surface such as a building or rock face. The term applied to this phenomenon is *multi-path*; when it occurs the multiple receptions of the same signal may interfere with each other (see Figure 3.26). Multi-path is a source of error that is difficult to avoid.

multi-path error

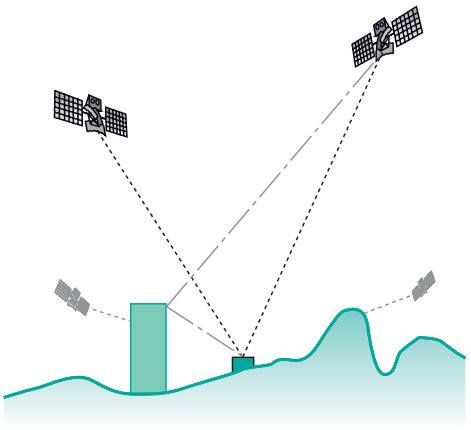


Figure 3.26
At any point in time, a number of satellites will be above the receiver's horizon. But not all of them will be "in view" (e.g. the satellites on the far left and right); and for others, multi-path signal reception may occur.

All of the above sources of error influence computation of a satellite's pseudorange. Cumulatively, they are called the *user equivalent range error* (UERE). Some error sources may affect all satellites being used by a particular receiver, e.g. selective availability and atmospheric delay, while others may be specific to one satellite, for instance, incorrect satellite information and multi-path.

range error

Errors related to the relative geometry of satellites and receiver

There is one more source of error, which is unrelated to individual radio signal characteristics: rather, this error is the result of the combination of signals from satellites used for positioning. The constellation of satellites in the sky from the receiver's perspective is the controlling factor in these cases. Referring to Figure 3.27, the sphere-intersection technique of positioning provides more precise results when the four satellites are evenly spread over the sky; the satellite constellation of Figure 3.27b is preferred over that of 3.27a. This source of error is know as geometric dilution of precision (GDOP). GDOP is lower when satellites are just above the horizon in mutually opposed compass directions. However, such satellite positions have bad atmospheric delay char-

geometric dilution of precision

acteristics, so in practice it is better if they are at least 15° above the horizon. When more than four satellites are in view, modern receivers use "least-squares" adjustment to calculate the best possible positional fix from all the signals. This gives a better solution that obtained just using the "best four", as was done previously.

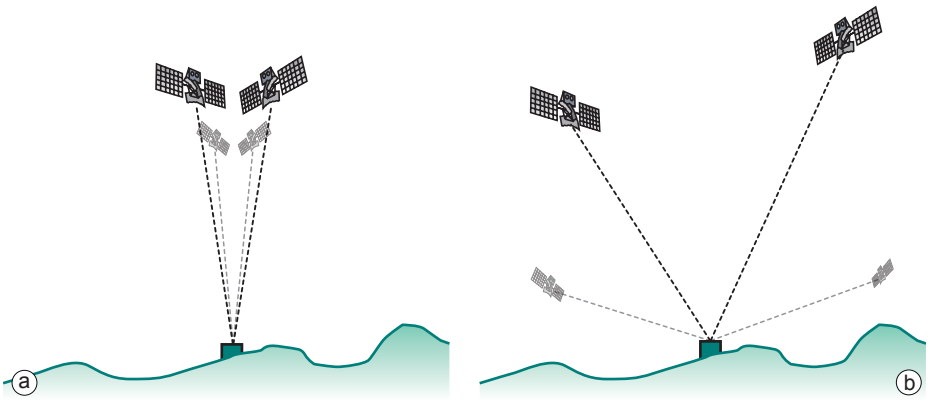


Figure 3.27
Geometric dilution of precision. The four satellites can be in a poor constellation for positioning(a) or in a better constellation (b).

Table 3.4
Indication of typical magnitudes of error in absolute satellite-based positioning

satellite clock (m)	2
satellite position (m)	2.5
ionospheric delay (m)	5
tropospheric delay (m)	0.5
receiver noise (m)	0.3
multi-path (m)	0.5
Total RMSE Range error (m): $\sqrt{2^2 + 2.5^2 + 5^2 + 0.5^2 + 0.3^2 + 0.5^2} =$	5.97

These errors are not all of similar magnitude. An overview of some typical values (without selective availability) is provided in Table 3.4. GDOP functions not so much as an independent error source but rather as a multiplying factor, decreasing the precision of position and time values obtained.

The procedure that we discussed above is known as *absolute, single-point positioning based on code measurement*. It is the fastest and simplest, yet least accurate, means of determining a position using satellites. It suffices for recreational purposes and other applications that require horizontal accuracies to within 5–10 m. Typically, when encrypted military signals can also be used, on a dual-frequency receiver the achievable horizontal accuracy is 2–5 m. Below, we discuss other satellite-based positioning techniques with better accuracies.

3.2.3 Relative positioning

One technique for trying to remove errors from positioning computations is to perform many position computations, and to determine the average over all solutions. Many receivers allow the user to do this. It should, however, be clear from the above that *averaging* may address *random* errors such as signal noise, selective availability (SA) and multi-path to some extent, but not *systematic* sources of error, such as incorrect satellite data, atmospheric delays, and GDOP effects. These sources should be removed before averaging is applied. It has been shown that averaging over 60 min in absolute, single-point positioning based on code measurements, before systematic error removal, leads to only a 10–20% improvement of accuracy. In such cases, receiver

random and systematic error

averaging is therefore of limited value and requires near-optimal conditions for long periods. Averaging is a good technique if systematic errors have been accounted for.

In relative positioning, also known as *differential positioning*, one tries to remove some of the sources of systematic error by taking into account measurements of these errors in a nearby stationary *reference receiver* that has an accurately known position. By using these systematic error findings for the reference receiver, the position of the *target receiver* of interest can be determined much more precisely.

In an optimal setting, the reference and target receiver experience identical conditions and are connected by a direct data link, allowing the target to receive correctional data from the reference. In practice, relative positioning allows reference and target receiver to be 70–200 km apart; they will experience essentially similar atmospheric signal error. Selective availability can also be addressed in this away.

For each satellite in view, the reference receiver will determine its pseudorange error. After all, its position is known to a high degree of accuracy, so it can solve any pseudorange equations to determine the error. Subsequently, the target receiver, having received the error characteristics will apply the correction for each of the satellite signals that it uses for positioning. In doing so, it can improve its accuracy to within 0.5–1 m.

The discussion above assumes we needed positioning information in real time, which called for a data link between reference and target receiver. But various uses of satellite-based positioning do not need real time data, making post-processing of the recorded positioning data suitable. If the target receiver records time and position accurately, correctional data can be used later to improve the accuracy of the originally recorded data.

Finally, mention should be made of the notion of *inverted relative positioning*. The principles are still the same as above, but with this technique the target receiver does not correct for satellite pseudorange error, but rather uses a data link to upload its positioning/timing information to a central repository, where the corrections are applied. This can be useful in cases where many target receivers are needed and budget does not allow them to be expensive.

3.2.4 Network positioning

Now that the advantages of relative positioning have been discussed, we can move on to the notion of *network positioning*: an integrated, systematic network of reference receivers covering a large area, perhaps an entire continent or even the whole globe.

The organization of such a network can take different shapes, augmenting an already existing satellite-based system. Here we discuss a general architecture, consisting of a network of *reference stations*, strategically positioned in the area to be covered, each of them constantly monitoring signals and their errors for all positioning satellites in view. One or more *control centres* receive the reference station data, verify this for correctness, and relay (uplink) this information to a *geostationary satellite*. The satellite will retransmit any correctional data to the area that it covers, so that *target receivers*, using their own approximate position, can determine how to correct for satellite signal error, and consequently obtain much more accurate position fixes.

With network positioning, accuracy in the sub-metre range can be obtained. Typically, advanced receivers are required, but the technology lends itself also for solutions with a single advanced receiver that functions in the direct neighbourhood as a reference receiver to simple ones.

3.2.5 Code versus phase measurements

Up until this point, we have assumed that the receiver determines the range of a satellite by measuring time delay of the received ranging code. There exists a more advanced range determination technique, known as *carrier phase measurement*. This typically requires more advanced receiver technology and longer observation sessions. Currently, carrier phase measurement can only be used with relative positioning, as absolute positioning using this method is not yet well developed.

The technique aims to determine the number of cycles of the (sine-shaped) radio signal between sender and receiver. Each cycle corresponds to one wavelength of the signal, which in the L-band frequencies used is 19–24 cm. Since the number of cycles of the signal cannot be measured directly, it is determined (in a long observation session) from the change in carrier phase over time. Such a change occurs because the satellite is orbiting. From its orbit parameters and the change in phase over time, the number of cycles can be derived.

With relative positioning techniques, a horizontal accuracy of 2 mm–2 cm can be achieved. This degree of accuracy makes it possible to measure tectonic plate movements, which can be as large as 10 cm per year for some locations on the planet.

3.2.6 Positioning technology

This section provides information on currently available satellite-based positioning technology. At present, two satellite-based positioning systems are operational—GPS and GLONASS—and a third is in the implementation phase—Galileo. These systems are US, Russian and European, respectively. Any of them, but especially GPS and Galileo, will be improved over time and will be augmented with new techniques.

GPS

The NAVSTAR Global Positioning System (GPS) was declared operational in 1994, providing Precise Positioning Services (PPS) to US and allied military forces, as well as US government agencies; civilians throughout the world have access to Standard Positioning Services (SPS). The GPS space segment nominally consists of 24 satellites, each of which orbits our planet in 11 h 58 min at an altitude of 20,200 km. There can be any number of satellites active, typically between 21 and 27. The satellites are organized in six orbital planes, somewhat irregularly spaced, at an angle of inclination of 55–63° to the equatorial plane; nominally four satellites orbit in each plane (see Figure 3.28). This means that a receiver on Earth will have between five and eight (rarely, even up to 12) satellites in view at any moment in time. Software packages exist to help in planning GPS surveys, identifying the expected satellite set-up for any location and time.

The NAVSTAR satellites transmit two radio signals, an L1 frequency of 1575.42 MHz and an L2 frequency of 1227.60 MHz. There is also a third and fourth signal, but these are not important for the discussion here. The role of the L2 signal is to provide a second radio signal, thereby providing a way, with (more expensive) dual-frequency receivers, of determining fairly precisely the actual ionospheric delay of the satellite signals received.

GPS uses WGS84 as its reference system,which has been refined on several occasions and is now aligned with the ITRF at the level of a few centimetres worldwide. (See also Section 3.1.1.) GPS has adopted UTC as its time system.

WGS84

For civilian applications, GPS receivers of varying quality are available, their quality depending on the embedded positioning features: supporting single- or dual-frequencies; supporting only absolute or also relative positioning; performing code

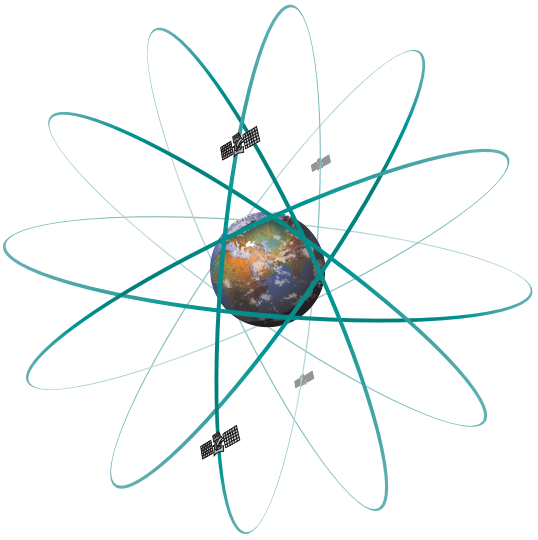


Figure 3.28
Constellation of satellites in the GPS system; here four satellites are shown in only one orbital plane.

measurements or also carrier phase measurements.

GLONASS

What GPS is to the US military, is GLONASS to the Russian military, specifically the Russian Space Forces. Both systems were primarily designed on the basis of military requirements, but GLONASS did not significantly develop civil applications as GPS did and thus it is commercially less important.

GLONASS's space segment consists nominally of 24 satellites, organized in three orbital planes, at an inclination of 64.8° to the Equator. Its orbiting altitude is 19,130 km, with a period of revolution of 11 h 16 min.

GLONASS uses the PZ-90 as its reference system and, like GPS, uses UTC as its time reference, albeit with an offset for Russian daylight.

GLONASS's radio signals are somewhat similar to those of GPS, differing only in the details: the frequency of GLONASS's L1 signal is approximately 1605 MHz (changes are underway), and its L2 signal approximately 1248 MHz; otherwise, GLONASS's system performance is rather comparable with that of GPS.

Galileo

In the 1990s, the European Union (EU) judged that it needed its own satellite-based positioning system, to become independent of the GPS monopoly and to support its own economic growth by providing services of high reliability under civilian control.The EU system is named Galileo.

The vision is that satellite-based positioning will become even bigger due to the emergence of mobile phones equipped with receivers, perhaps with some 400 million users by the year 2015. The development of the system has experienced substantial delays; currently European ministers insist that Galileo should be up and running by the end of 2013.

When completed, Galileo will have 27 satellites, with three in reserve, orbiting in one of three, equally spaced, circular orbits at an elevation of 23,222 km and inclined at 56° to the Equator. This higher inclination (when compared to that of GPS) has been chosen to provide better positioning coverage at high latitudes, such as in northern

Scandinavia, where GPS performs rather poorly.

In June 2004, the EU and the US agreed to make Galileo and GPS compatible by adopting interchangeable set-ups for their satellite signals. The effect of this agreement is that a Galileo/GPS tandem satellite system will have so many satellites in the sky (close to 60) that a receiver can almost always find an optimal constellation in view.

This will be especially useful in situations where in the past signal reception was poor, in built-up areas and forests, for instance. It will also bring the implementation of a Global Navigation Satellite System (GNSS) closer, since positional accuracy and reliability will improve. Such a system would bring the ultimate development of fully automated air and road traffic control systems much closer. Automatic aircraft landing, for instance, requires a horizontal accuracy in the order of 4 m, and a vertical accuracy of less than 1 m. Currently, these requirements cannot be reliably met.

The Galileo Terrestrial Reference Frame (GTRF) will be a realization of the ITRS and will be set up independently from that of GPS so that one system can back up the other. Positional differences between WGS84 and GTRF will be at worst only a few centimetres.

The Galileo System Time (GST) will closely follow International Atomic Time (TAI), with a time offset of less than 50 ns for 95% of the time over any period of the year. Information on the actual offset between GST and TAI, and between GST and UTC (as used in GPS), will be broadcast in the Galileo satellite signal.

Satellite-based augmentation systems

Satellite-based augmentation systems (SBAS) aim to improve the accuracy and reliability of satellite-based positioning (see Subsection 3.2.4) in support of safety-critical navigation applications, such as aircraft operations near airfields. Typically, these systems make use of an extra, now geostationary, satellite that has a large service area, for example a continent, and which sends differential data about standard positioning satellites that are currently in view in its service area. If multiple ground reference stations are used, the quality of the differential data can be quite good and reliable. Usually this satellite will use radio signals of the same frequency as those in use by the positioning satellites, so that receivers can receive the differential code without problem.

Not all advantages of satellite augmentation will be useful for all receivers. For consumer market receivers, the biggest advantage, as compared to standard relative positioning, is that SBAS provides an ionospheric correction grid for its service area, from which a correction specific for the location of the receiver can be retrieved. This is not true in relative positioning, where the reference station determines the error it experiences and simply broadcasts this information for nearby target receivers to use. With SBAS, the receiver obtains information that is best viewed as a geostatistical interpolation of errors from multiple reference stations.

More advanced receivers will be able to deploy also other differential data such as corrections on satellite position and satellite clock drift.

Currently, three systems are operational: for North America WAAS (Wide-Area Augmentation System) is in place; EGNOS (European Geostationary Navigation Overlay Service) for Europe; and MSAS (Multi-functional Satellite Augmentation System) for eastern Asia. The ground segment of WAAS consists of 24 control stations, spread over North America; that of EGNOS has 34 control stations. These three systems are compatible, guaranteeing international coverage.

Usually signals from the geostationary SBAS satellites (under various names, such as AOR, Artemis, IOR, Inmarsat, MTSAT) can be received even outside their respective

service areas. But the use of these signals there must be discouraged, as they will not help improve positional accuracy. Satellite identifiers, as shown by the receiver, have numbers above 30, setting them apart from standard positioning satellites.

Though originally intended to improve the safety of aircraft landings, SBAS, with its horizontal accuracy to within 3 m, has many other uses. At this level of accuracy, vehicle position can be determined to a specific road lane, and "automatic pilots" become a possibility.

Chapter 4

Sensors

Wim Bakker
Wan Bakx
Wietske Bijker
Karl Grabmaier
Lucas Janssen
John Horn
Gerrit Huurneman
Freek van der Meer
Christine Pohl
Klaus Tempfli
Valentyn Tolpekin
Tsehaie Woldai

4.1 Platforms and passive electro-optical sensors

Having explained the physics of sensing in Chapter 2, in this chapter we discuss sensor systems and set out to discover the logic of current electro-optical sensing technology. First, in Subsection 4.1.1, we will look at the characteristics of platforms used for geospatial data acquisition (GDA) from the air and from space: various platforms such as aircraft, space shuttles, space stations and satellites are used to carry one or more sensors for Earth Observation. Next, Subsection 4.1.2 will elaborate on frame and line cameras; the latter, which can be operated from the air or space, are also known as *pushbroom sensors*. Optical scanners (also referred to in the literature as across-track scanners or *whiskbroom scanners*) are treated in Section 4.1.3, which discusses multispectral, hyperspectral and thermal scanners in detail. Some camera systems can provide us with *stereo* images, justifying a short introduction to stereoscopy in Subsection 4.1.4.

4.1.1 Platforms and missions

Sensors used in Earth Observation can be operated at altitudes ranging from just a few centimetres above the ground—using field equipment—to those far beyond the atmosphere. Very often the sensor is mounted on a moving vehicle—which we call the *platform*—such as an aircraft or a satellite. Occasionally, static platforms are used. For example, we could mount a spectrometer on a pole to measure the changing re-

flectance of a specific crop during the day or over a whole season.

Moving platforms

To gain a wider view, we use aircraft at altitudes ranging up to approximately 20 km. Depending on the type of aerial survey and the weight of equipment and survey costs, we can choose from a variety of vehicles. Fixed-wing aircraft are used for thermal scanning and a systematic photo-coverage for topographic mapping, land titling projects, and the like. Aerial survey cameras are heavy and they are fixed to a stabilized mount set in a hole in the floor of the aircraft. Most survey airplanes fly lower than 8 km but higher than 1000 m. They can fly as slow as 150 km h^{-1}, but even at that speed image quality is already affected by motion blur unless the camera is fitted with a compensation device. Aerial survey cameras are highly sophisticated and expensive.

Airborne laser-scanner systems used to be heavy, but nowadays the total weight of the equipment can be as light as 30 kg. Laser scanners are either mounted on fixed-wing aircraft or helicopters, the latter being able to fly very slowly at low altitudes, thus allowing the acquisition of highly detailed data (at high costs per unit of area). The small-format cameras used are cheaper and lighter than large-format aerial survey cameras, making it possible to mount these systems on micro-light airplanes for urban reconnaissance, or even kites (e.g. for surveying an industrial area). Unmanned aerial vehicles (UAVs) are gaining popularity for observing dangerous areas or to reduce costs. A special type of UVA, the High Altitude Long Endurance (HALE) vehicle, can bridge the gap between manned survey aircraft and spacecraft or satellites. Typically, a HALE is a remotely operated aircraft of ultra-light weight and load that flies for months at altitudes of around 20 km.

A key advantage of aerial surveys is that they can be "targeted". The survey can be undertaken at exactly the required time and can be done with exactly the required spatial resolution by having the aircraft fly at the required altitude. Moreover, in comparison with civilian satellite RS, we can acquire images of much higher spatial resolution, enabling recognition of objects of much smaller size. With current aerial survey cameras, we can achieve a pixel size on the ground as small as 5 cm.

Satellites are launched by rocket into space, where they then circle the Earth for 5 to 12 years on a predefined orbit. The choice of orbit depends on the objectives of the sensor mission; orbit characteristics and different orbit types are explained below. A satellite must travel at high speed to orbit at a certain distance from the Earth; the closer to the Earth, the faster the speed required. A space station such as ISS has a mean orbital altitude of 400 km and travels at roughly 27,000 km h^{-1}. The Moon at a distance of 384,400 km can conveniently circle the Earth at only 3700 km h^{-1}. At altitudes of 200 km, satellites already encounter traces of the atmosphere, which causes rapid orbital and mechanical decay. The higher the altitude, the longer is the expected lifetime of the satellite.The majority of civilian Earth-observing satellites orbit at altitudes ranging from 500 to 1000 km. Here we generally find the "big boys", such as Landsat-7 (2200 kg) and Envisat (8200 kg), but the mini-satellites of the Disaster Management Constellation (DMC) also orbit in this range. DMC satellites have a weight of around 100 kg and were launched by several countries into space early in the current millennium at relatively low-cost. These satellites represent a network for disaster monitoring that provides images in three or four spectral bands with a ground pixel size of 32 m or smaller.

satellites

Satellites have the advantage over aerial survey of continuity. Meteosat-9, for example, delivers a new image of the same area every 15 minutes and it has done so every day for many years. The high temporal resolution at low cost goes together with a low

spatial resolution (pixel size on the ground of 1×1 km^2). Both the temporal and the spatial resolution of satellite remote sensors are fixed. While aerial surveys have been restricted in some countries, access to satellite RS data is commonly easier, although not every type of satellite RS image is universally available.

Aerial survey missions

Modern airborne sensor systems use a high-end GPS receiver and many also include an Inertial Measuring Unit (IMU). GPS is used for navigation and for coarse "sensor positioning". We need to know the coordinates of the exposure stations of a camera to relate points and features in the images to positions on the ground; differential GPS is applied for more precise positioning. To this end, we need a reference GPS station on the ground within some 30 km from the aircraft. Adding an IMU has two advantages: IMU readings can be used to improve the accuracy of the coordinates obtained by GPS (achieving a RMSE better than 0.1 m); and the IMU measures the attitude angles of the sensor (Figure 4.27). An IMU, an assemblage of gyros and accelerometers, is a sophisticated, heavy, and expensive instrument that was originally used only in Inertial Navigation Systems (INSs). Measuring continuously the position and attitude of the moving sensor, an IMU allows us to relate the sensor recordings to position in the terrain in near real-time. We call this *direct sensor orientation*. We need a GPS-IMU positioning and orientation system (POS) for line cameras and scanners; for frame cameras we can also solve the georeferencing problem indirectly (see Section 5.3).

direct sensor orientation

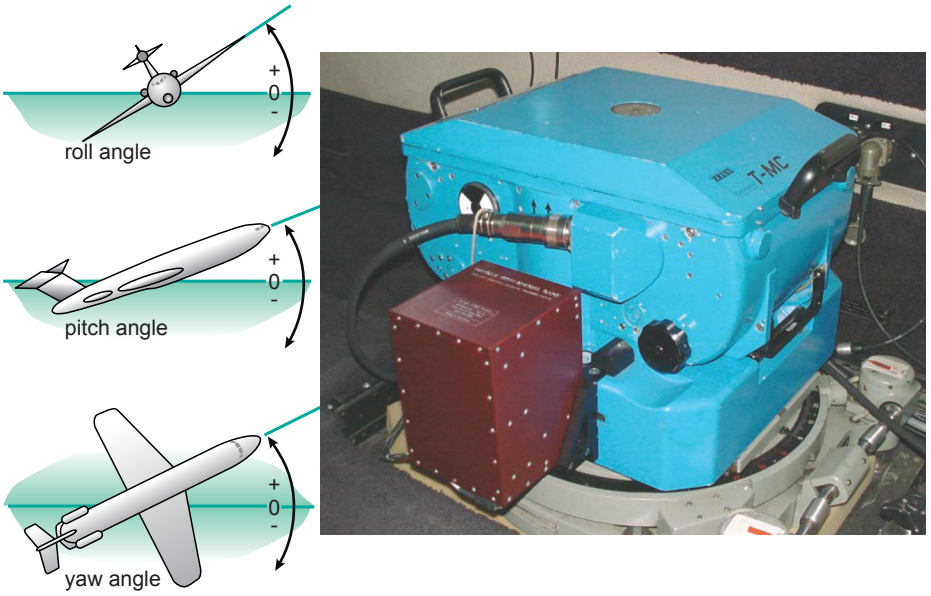


Figure 4.1
Attitude angles (left) and an IMU attached to a Zeiss RMK-TOP aerial camera (courtesy of IGI).

Mission planning and execution is usually done by commercial survey companies or, otherwise, by large national mapping agencies or the military. During missions, companies use professional software for flight planning and, most likely, one of the two integrated aircraft guidance and sensor management systems available (produced by APPLANIX or IGI). Pioneering work on computer-controlled navigation and camera management was done at ITC in the days when it still had an aerial photography and navigation department. The basics of planning aerial survey missions are explained in Section 4.6.

Satellite missions

orbit parameters

The monitoring capabilities of a satellite-borne sensor are to a large extent determined by the parameters of the satellite's orbit. An *orbit* is a circular or elliptical path described by the satellite in its movement around the Earth. Different types of orbits are required to achieve continuous monitoring (meteorology), global mapping (land cover mapping) or selective imaging (urban areas). For Earth Observation, the following orbit characteristics are relevant:

- *Orbital altitude* is the distance (in km) from the satellite to the surface of the Earth. It influences to a large extent the area that can be viewed (i.e. the *spatial coverage*) and the details that can be observed (i.e. the *spatial resolution*). In general, the higher the altitude, the larger the spatial coverage but the lower the spatial resolution.
- *Orbital inclination angle* is the angle (in degrees) between the orbital plane and the equatorial plane. The inclination angle of the orbit determines, together with the field of view (FOV) of the sensor, the latitudes up to which the Earth can be observed. If the inclination is 60 °, then the satellite orbits the Earth between the latitudes 60 ° N and 60 ° S. If the satellite is in a low-Earth orbit with an inclination of 60 °, then it cannot observe parts of the Earth at latitudes above 60 ° North and below 60 ° South, which means it cannot be used for observations of the Earth's polar regions.
- *Orbital period* is the time (in minutes) required to complete one full orbit. For instance, if a polar satellite orbits at 806 km mean altitude, then it has an orbital period of 101 minutes. The Moon has an orbital period of 27.3 days. The speed of the platform has implications for the type of images that can be acquired. A camera on a low-Earth orbit satellite would need a very short exposure time to avoid motion blur resulting from the high speed. Short exposure times, however, require high intensities of incident radiation, which is a problem in space because of atmospheric absorption. It should be obvious that the contradictory demands of high spatial resolution, no motion blur, high temporal resolution, long satellite lifetime (thus lower cost) represent a serious challenge for satellite-sensor designers.
- *Repeat cycle* is the time (in days) between two successive identical orbits. The *revisit time* (i.e. the time between two subsequent images of the same area) is determined by the repeat cycle together with the pointing capability of the sensor. *Pointing capability* refers to the possibility of the sensor–platform combination to look to the side, or forward, or backward, and not only vertically downwards. Many modern satellites have such a capability. We can make use of the pointing capability to reduce the time between successive observations of the same area, to image an area that is not covered by clouds at that moment, and to produce stereo images (see Subsection 4.1.4).

The following orbit types are most common for remote sensing missions:

- *Polar orbit* refers to orbits with an inclination angle between 80 ° and 100 °. An orbit having an inclination larger than 90 ° means that the satellite's motion is in a westward direction. Such a polar orbit enables observation of the whole globe, also near the poles. Satellites typically orbit at altitudes of 600–1000 km.
- *Sun-synchronous orbit* refers to a polar or near-polar orbit chosen in such a way that the satellite always passes overhead at the same time. Most Sun-synchronous

orbits cross the Equator mid-morning, at around 10:30 h local solar time. At that moment the Sun angle is low and the shadows that creates reveal terrain relief. In addition to day light images, a Sun-synchronous orbit also allows the satellite to record night images (thermal or radar, passive) during the ascending phase of the orbit on the night side of the Earth.

- A *Geostationary orbit* refers to orbits that position the satellite above the Equator (inclination angle: 0 °) at an altitude of approximately 36,000 km. At this distance, the orbital period of the satellite is equal to the rotational period of the Earth, exactly one sidereal day. The result is that the satellite has a fixed position relative to the Earth. Geostationary orbits are used for meteorological and telecommunication satellites.

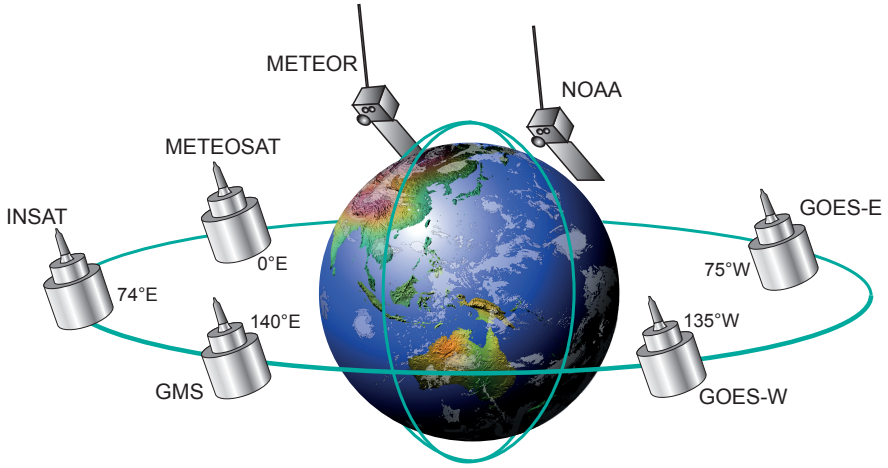


Figure 4.2
Meteorological observation by geostationary and polar satellites.

Today's meteorological weather satellite systems use a combination of geostationary satellites and polar orbiters (Figure 4.28). The geostationary satellites offer a continuous hemispherical view of almost half the Earth (45%), while the polar orbiters offer a higher spatial resolution.

RS images from satellites come with data on orbital parameters and other parameters to facilitate georeferencing of the images. High resolution sensor systems such as Ikonos or QuickBird use GPS receivers and star trackers as their POS.

The data from space-borne sensors need to be transmitted to the ground in some way. Russia's SPIN-2 satellite, with its KVR camera, used film cartridges that were dropped over a designated area on the Earth. Today's Earth Observing satellites *downlink* the data. The acquired data are sent directly to a receiving station on the ground, or via a geostationary communication satellite. One current trend is that small receiving units, consisting of a small dish with a PC, are being developed for local reception of RS data.

4.1.2 Cameras

A *digital camera* is an electro-optical remote sensor. In its simplest form, it consists of the camera body, a lens, a focal plane array of CCDs, and a storage device, but no mechanical component. The CCD array can either be an assembly of linear arrays or a matrix array (Figure 4.3). Accordingly, we talk about line cameras and frame cameras. A small-format frame camera has a single matrix chip and closely resembles a photographic camera. The chip (a) of the Figure 4.3 has three channels, one for each primary colour (red, green, blue); three elongated CCDs next to each other constitute

CCD

one square "colour pixel". Each CCD has its colour filter right on top to only transmit the required band of incident light. The linear chip (b) of the Figure 4.3 also has three channels; three lines of square CCDs are assembled next to each other. A line camera is exclusively used on a moving platform, which can be a car, an aircraft or a spacecraft. SPOT-1, launched in 1986, was the first satellite to use a line camera. Line cameras build up a digital image of an area line by line (Figure 4.4). In the older literature, therefore, it is also referred to as *pushbroom scanner*, as opposed to a *whiskbroom scanner* (see Subsection 4.1.3), which actually scans (across the track of the moving platform).

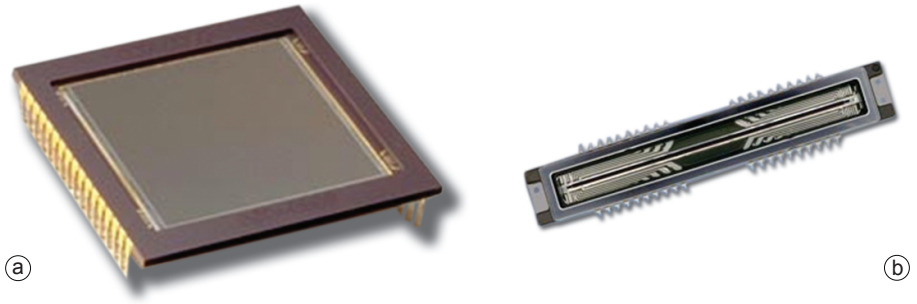


Figure 4.3
Two CCD chips: (a) matrix array Kodak KAF-16801, pixel size 9 μm; (b) linear array Kodak KLI-14403, pixel size 5 μm.

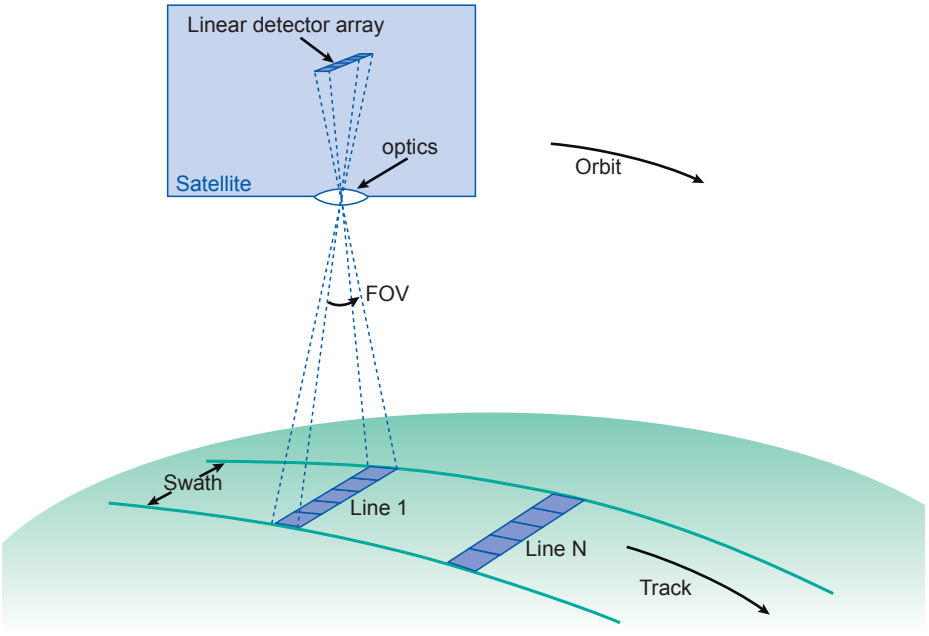


Figure 4.4
Principle of imaging by line camera on a spacecraft ('pushbrooming').

Detector arrays

Cameras are used for sensing in the visible, NIR, and SWIR portions of the spectrum. We need different types of semiconductors for sensing in this range; the semiconductors used are all solid-state detectors but are made of different material for different spectral ranges. CCDs are the most common type of semiconductor used today for sensing in the visible to very near-IR range; they are made of silicon.

The spectral sensitivity of a sensor band is commonly specified by a lower- and an upper-bound wavelength of the spectral band covered, e.g. 0.48 to 0.70 μm for the

SPOT-5 panchromatic channel. However, a detector such as a CCD is not equally sensitive to each monochromatic radiation within this band. The actual response of a CCD can be determined in the laboratory; an example of a resulting spectral response curve is shown in Figure 4.5. The lower and upper bound specification is usually chosen at the wavelengths where the 50% response is achieved. The DN produced by a detector results from averaging the spectral response of incident radiation. Figure 4.5 shows that the DNs of AVHRR channel 1 are biased towards red, whereas the brightness sensation of our eyes is dominated by yellow-green. The CCDs of a channel array do not have exactly the same sensitivity. It takes radiometric sensor calibration to determine the differences. CCDs show, moreover, varying degrees of degradation over time. Therefore, radiometric calibration needs to be done regularly. Knowing the detector's spectral sensitivity becomes relevant when we want to convert DNs to radiances (see Section 5.2).

spectral sensitivity

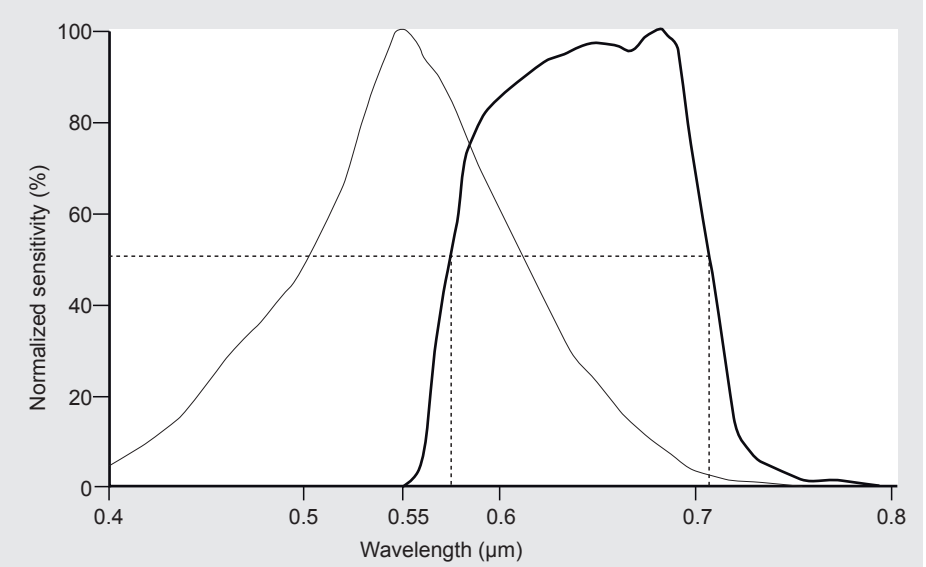


Figure 4.5
Normalized spectral response curve of (a) channel 1 of NOAA's AVHRR and (b) the spectral sensitivity of the rods of the human eye.

When compared with photographic film, most CCDs have a much higher general sensitivity and thus they need less light. The reason is that they typically respond to 70% of the incident light, whereas photographic film captures only about 2% of the incident light. They also offer a much better differentiation of intensity values in the very dark and the very bright parts of a scene.

If we were interested in a high radiometric resolution, we would like a CCD to have a wide dynamic range. *Dynamic range* is the ratio of the maximum to the minimum level of intensity that can be measured; it is also known as the signal to noise ratio of the detector. The maximum intensity is determined by the maximum charge capacity of the semiconductor cell. The minimum intensity is determined by the noise level. Noise is unwanted collected charge, for example caused by unblocked IR or UV radiation for a CCD that should be sensitive to blue light. It only makes sense to record a DN of many bits if the semiconductor cell has a wide dynamic range. It is the manufacturer's concern to ensure sufficient dynamic range to meet the radiometric resolution (expressed in bits) required by the user. We can compute the effective radiometric resolution of a sensor if the manufacturer specifies both the number of bits and the dynamic range.

dynamic range

We had line cameras in space and frame cameras in our pockets before we had any digital airborne camera offering satisfactory surveying quality. The main reason for

this is the ultra-high quality of aerial film cameras and their operational maturity, including the entire photogrammetric processing chain. Cameras on satellite platforms are exclusively line cameras, typically having a panchromatic channel and four more linear arrays (e.g. for red, green, blue, NIR). ASTER has two panchromatic channels, one linear array looking vertically down (nadir view) and the second looking backwards; the two resulting images can be used to generate stereo images. The first aerial line camera on the market was Leica's ADS40 (in 2000). It has three panchromatic detector arrays (forward, nadir, backward looking) and four multispectral ones (for RGB and NIR). One linear array consists of 12,000 CCDs.

linear arrays

Current CCD technology enables the production of very high quality linear arrays but not (yet) the very large matrix arrays that would be needed for large-format digital aerial cameras to be able to match the well-proven film-based survey camera. The two market leaders in digital aerial frame cameras, ZI and Microsoft (former Vexcel), therefore use several detector arrays for panchromatic imaging and software to compile a single large-format image from the sub-frames. ZI's DMC has, for example, 13,500 $\times$ 7,500 CCDs per sub-frame. One of the advantages of frame cameras is that the same photogrammetric software can be used as for photographs. At the moment there are about as many aerial line cameras as digital aerial frame cameras on the market.

matrix arrays

Optical system

Cameras use either lenses or telescopes to focus incident radiation onto the focal plane where the CCD surfaces are. A lens of a simple hand-held camera is a piece of glass or plastic shaped to form an image by means of refraction. The lens cone of a survey camera contains a compound lens, which is a carefully designed and manufactured assembly of glass bodies (and thus very expensive). The camera head of a digital aerial frame camera (such as Z/I's DMC and Vexcel's UltraCam) even consists of several of such lenses to focus the light rays on the respective CCD arrays. However complicated a lens may be physically, geometrically imaging through a lens is simple. The geometric model that a point of an object connects to its point in the image by a straight line and that all such lines pass through the centre of the lens (Figure 4.6) is a very close approximation of reality. We refer to the geometric imaging of a camera as "central projection".

lens, focal length, scale

An important property of a lens is its *focal length*. The focal length, f, determines, together with the length of a CCD line, the FOV of the camera. The focal length together with the *flying height* determine the size of the ground-resolution cell for a given pixel size, p. The *flying height*, H, is either the altitude of the aircraft above the ground or the orbital altitude.

field of view

$$GRC = p\frac{H}{f} \tag{4.1}$$

The focal length of Leica's ADS40 line camera is 63 mm. At a flying height of 2000 m, we would attain a ground-resolution cell size in the across-track direction of 21 cm, provided the airplane flies perfectly horizontally over flat terrain (see Figure 4.6). You would conclude correctly that the ADS40 has a CCD/pixel size of 6.5 µm. The ratio $\frac{H}{f}$ is referred to as the *scale factor* of imaging.

Space-borne cameras do not have lens cones—they have telescopes. The telescopes of Ikonos and QuickBird consist of an assembly of concave and flat mirrors, thus achieving a spatial resolution that is absolutely amazing when considering their flying height. The focal length equivalent of the Ikonos telescope is 10 m. Ikonos specifications state a ground-resolution cell size of 80 cm for a panchromatic image at nadir.

telescope

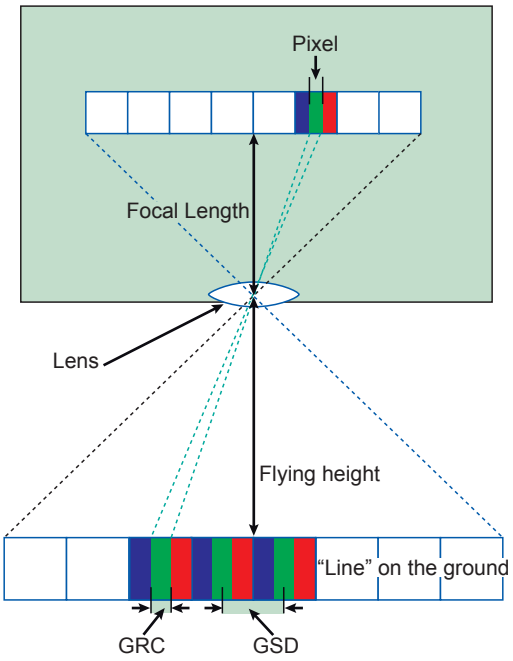


Figure 4.6
Pixel, ground resolution cell, ground sampling distance for digital cameras.

The size of a CCD determines the pixels size. A pixel projected onto the ground gives us the *ground resolution cell* (GRC) of the camera. The distance between the centres of two adjacent resolution cells of the same channel is called the *ground sampling distance* (GSD). Ideally the ground resolution cell size and the GSD are equal; the GSD then uniquely defines the spatial resolution of the sensor. This can be most easily achieved for panchromatic frame cameras. Note that the GSD is the same throughout an entire line if the terrain is flat and parallel to the focal plane (e.g. in the case of a nadir view of horizontal terrain); see Figure 4.6. If a space-borne line camera is pointed towards the left or the right of the orbit track (across-track, off-nadir viewing), we obtain an oblique image. The scale of an oblique image changes throughout the image. In the case of oblique viewing, Formula 4.1 does not apply anymore; the ground resolution cell size and the GSD increase with increasing distance from nadir. Section 5.3 explains how to deal with this.

ground resolution cell

ground sampling distance

Digital aerial cameras have several advantages over film cameras, pertaining to both the quality of images and economics. Digital aerial cameras commonly record in 5 spectral bands (panchromatic, RGB, NIR), therefore, we can obtain with one flight panchromatic stereo images, true colour images and false colour images; with a film camera we would have to fly this course three times and develop three different types of film. The radiometric quality of CCDs is better than that of photographic film. Digital cameras also allow an all-digital workflow, making processing faster and cheaper. Digital cameras can acquire images with a high likelihood of redundancy without additional costs for material and flying time; this favours automated information extraction. Finally, new geoinformation products can be generated as a result of various extended camera features. In Subsection 4.1.3 multispectral scanners are introduced. Line cameras as compared to across-track scanners have the advantage of better geometry. Airborne line cameras and scanners require gyroscopically stabilized mounts to reduce any effects of aircraft vibration and compensate for rapid movements of the aircraft. Such a stabilized mount keeps a camera in an accurate level position so that

advantages of digital cameras

it continuously points vertically downward. We want vertical images for mapping because of the better geometry. we, Therefore, also mount large-format digital frame cameras and film cameras on stabilized platforms for applications that require high-quality images.

4.1.3 Scanners

Components

An *optical scanner* is an electro-optical remote sensor with a scanning device, which is in most cases a mechanical component. In its simplest form (e.g. a thermal scanner operating in the 7 to 14 μm range), it consists of the sensor rack, a single detector with electronics, a mirror, optics for focusing, and a storage device (see Figure 4.7). A detector has a very narrow field of view (called the *instantaneous field of view* (IFOV)) of 2.5 milliradians or less. In order to image a large area, we have scan the ground across the track while the aircraft or space craft is moving. The most commonly-used scanning device is a moving mirror, which can be an oscillating mirror, a rotating mirror, or a nutating mirror. An alternative, which is used for laser scanning, is fiber optics.

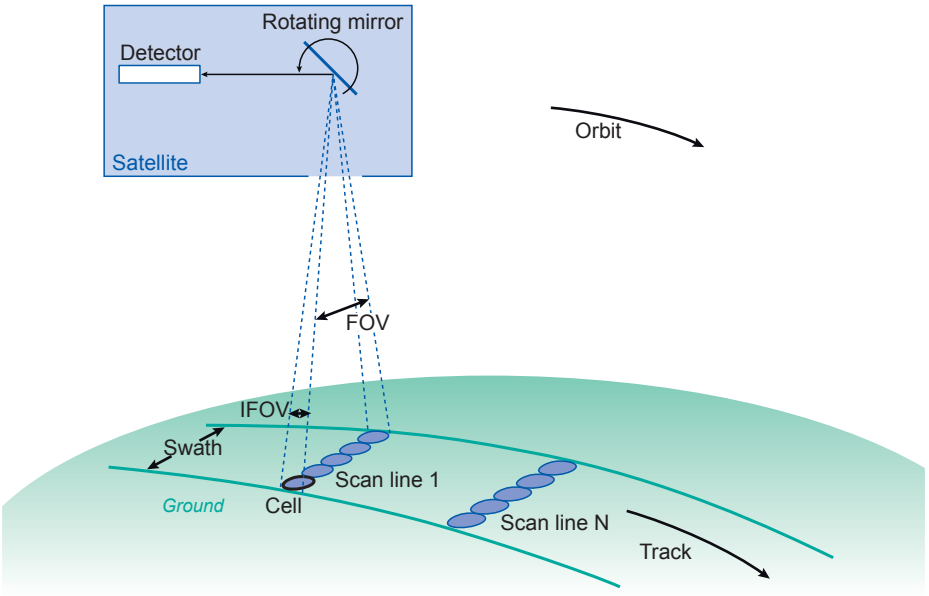


Figure 4.7
Principle of an across-track scanner.

detectors

Scanners are used for sensing in a broad spectral range, from light to TIR and beyond, to microwave radiation. Photodiodes made of silicon are used for the visible and NIR bands. Cooled photon detectors (e.g. using mercury-cadmium-telluride semiconductor material) are used for thermal scanners.

beam splitters

Most scanners are multispectral scanners, thus sensing in several bands, often including TIR (such as NOAA's AVHRR). As such, thermal scanners can be considered as being just a special type of multispectral scanner. A multispectral scanner has at least one detector per spectral band. Different from small-format frame cameras, for which filters are used to separate wavelength bands, scanners and line cameras use a prism and/or a grating as a *beam splitter*. A *grating* is a dispersion device used for splitting up SWIR and TIR radiation. Hyperspectral scanners also use gratings. A *prism* can split higher frequency radiation into red, green, blue, and NIR components. A simple

RGB and NIR scanner produces in one sweep of the mirror a single image line for each of the four channels.

Instead of using only one detector per band, space-borne scanners use several. The first civil space-borne remote sensor, Landsat MSS (launched in 1972), used six per band (thus, in total, 24; see Figure 2.18). ASTER uses 10 detectors for each of its five TIR channels. One sweep of the mirror of the ASTER thermal scanner produces, thus, 10 image lines for each of the five channels. If one channel should fail, only every 10th line of an image would be black. Section 5.2 treats the correcting of an image for periodic *line dropouts*.

Geometric aspects

At a particular instant, the detector of an across-track scanner observes an elliptical area on the ground, the ground resolution cell of the scanner. At nadir, the cell is circular, of diameter D. D depends on the IFOV, β, of the detector and the flying height.

$$D = \beta H \tag{4.2}$$

A scanner with β = 2.5 mrad operated at H = 4000 m would have, therefore, a ground resolution of 10 m at nadir. Towards the edge of a swath, the ground resolution cell becomes elongated and bigger (Figure 4.29).

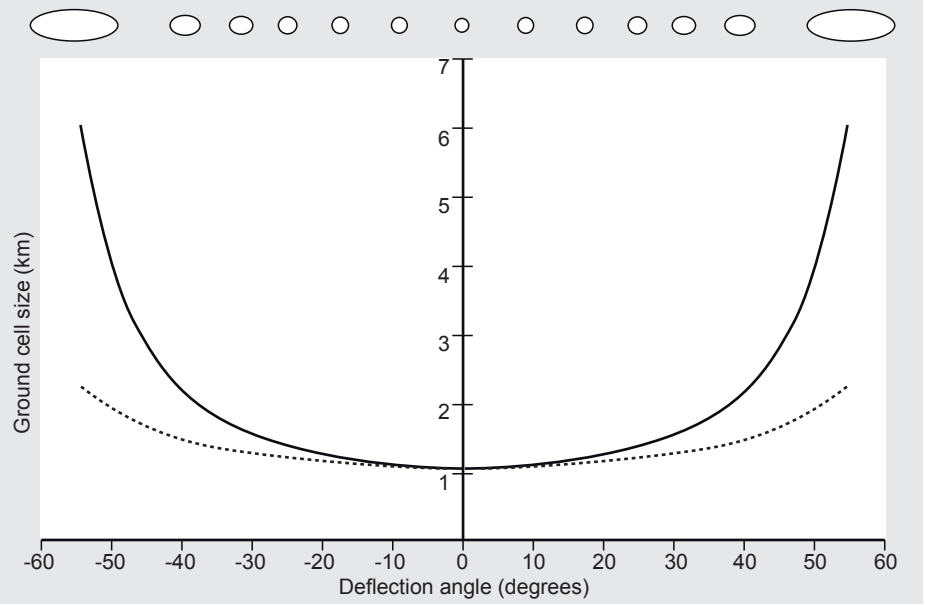


Figure 4.8
GRC of NOAA's AVHRR: at nadir the cell diameter is 1.1 km; at the edge the ellipse stretches to 6.1×2.3 km. The solid line shows the across-track resolution, the dashed line the along-track resolution. The ellipses at the top show the shape of the ground cells along a scanned line. NOAA processes the data ('resamples') to obtain a digital image with a pixel size on the ground of 1×1 km.

The width of the area that is covered by one sweep of the mirror, the *swath width*, depends on the FOV of the scanner. AVHRR has a very wide FOV of 110 °; easy geometry was not a concern in the AVHRR design. Landast-7 has an FOV of only 15 °, hence geometrically more homogeneous images result.

Reading out the detector is done at a fixed interval, the sampling interval. The sampling interval together with the speed of the moving mirror determines the GSD. The GSD can be smaller than D; we talk about *oversampling* if this is the case. The spatial resolution of the sensor is then not determined by the GSD but by the ground-resolution cell size (which is greater than or equal to D across the track).

4.1.4 Stereoscopy

Stereoscopy, the science of producing three-dimensional (3D) visual models using two-dimensional (2D) images, dates back to the 16th century. The astronomer Kepler was presumably the first person to define stereoscopic viewing. One of the main reasons for being able to perceive depth is that we have two eyes, which enables us to see a scene simultaneously from two viewpoints. The brain fuses the two 2D views into a three-dimensional impression. Judging which object is closer to us and which one is farther away with only one eye is only possible if we can use cues such as one object being partially obscured by the other one, or one appears smaller than the other although they are of the same size, etc. We can create the illusion of seeing three-dimensionally by taking two photographs or similar images and then displaying and viewing the pair simultaneously. Figure 4.30 illustrates the principle of stereoscopic viewing.

The advantage of stereoscopic viewing over monoscopic viewing (looking at a single image) is that image interpretation is easier, because we see the three-dimensional form of objects. Stereoscopy, moreover, has been the basis for 3D measurement by photogrammetry. Not just any two images can be viewed stereoscopically, they must fulfill several conditions. The same holds for making 3D measurements: we need at least two images and they must meet the preconditions. The basic requirements for a *stereo pair* are that the images of the same object or scene are taken from different positions, but not too far apart, and at a very similar scale. Different terms are used in stereoscopy, each with a slightly different meaning. A pair of images that meets the conditions of stereoscopic vision may be referred to as a stereo-image pair, a stereoscopic pair of images, stereo images, or simply as a stereo pair. A stereo pair arranged (on a computer monitor, on a table, or in a device) such that we can readily get a 3D visual impression may be called a stereograph, or stereogram). The 3D visual impression is called the stereo model, or stereoscopic model. We need special image-display techniques and stereoscopic viewing devices so that each eye sees only the image intended for it.

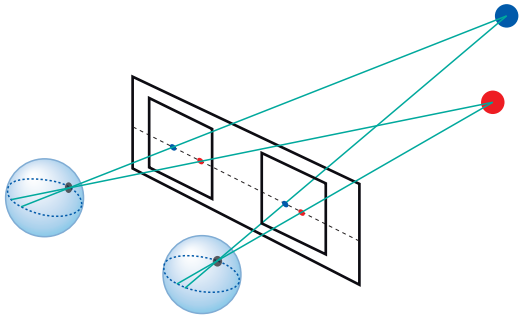


Figure 4.9
The principle of stereoscopy.

We have two options for obtaining a stereo pair with a space-borne sensor: a) use across-track pointing to image the same area from two different tracks, or b) apply along-track forward or backward viewing in addition to nadir viewing. The advantage of *in-track stereo* is that the two images are radiometrically very similar, because they are taken either at the same time or in quick succession; hence season, weather, scene illumination, and plant status are the same. In order to obtain a systematic coverage of an area with stereo images using an airborne frame camera, we need to take strips of vertical photos/images such that the images overlap by at least 60% (see Section 4.6).

4.2 Thermal remote sensing

Thermal remote sensing is based on the measuring of electromagnetic radiation in the infrared region of the spectrum. The wavelengths most commonly used are those in the intervals 3–5 µm and 8–14 µm, in which the atmosphere is fairly transparent and the signal is only slightly attenuated by atmospheric absorption. Since the source of the radiation is the heat of the imaged surface itself (see Figures 2.6 and 2.16), the handling and processing of TIR data is considerably different from remote sensing based on reflected sunlight:

- The surface temperature is the main factor that determines the amount of emitted radiation measured in the thermal wavelengths. The temperature of an object varies greatly depending on time of day, season, location, exposure to solar irradiation, etc. and is difficult to predict. In reflectance remote sensing, on the other hand, the incoming radiation from the Sun is considered constant and can be readily calculated, although atmospheric correction has to be taken into account.
- In reflectance remote sensing, the characteristic property we are interested in is the *reflectance* of the surface at different wavelengths. In thermal remote sensing, however, the one property we are interested in is, rather, how well radiation is *emitted* from the surface at different wavelengths.
- Since thermal remote sensing does not depend on reflected sunlight, it can also be done at night (for some applications this is even better than during the day).

4.2.1 Radiant and kinetic temperatures

The actual measurements by a TIR sensor will relate to the "spectral radiance" (measured in W m^{-2} sr^{-1} μm^{-1}) that reaches the sensor for a certain wavelength band. We know that the amount of radiation from an object depends on its temperature T and emissivity ϵ. That means that a cold object with high emissivity can radiate just as much radiation as a considerably hotter one with low emissivity. Often the emissivity of the object is unknown. If we assume that the emissivity of the object is equal to 1.0, then with the help of Planck's law we can calculate directly the ground temperature that is needed to create this amount of radiance in the specified wavelength band of the sensor for the object with a perfect emissivity. The temperature calculated in this way is the *radiant temperature* or T_{rad}. The terms *brightness* or "top-of-the-atmosphere" temperature are also frequently used.

radiant temperature

The radiant temperature calculated from the emitted radiation is in most cases lower than the true, *kinetic temperature* (T_{kin}) that we could measure on the ground with a contact thermometer. The reason for this is that most objects have an emissivity lower than 1.0 and radiate incompletely. To calculate the true T_{kin} from the T_{rad}, we need to know or estimate the emissivity. The relationship between T_{kin} and T_{rad} is:

kinetic temperature

$$T_{rad} = \epsilon^{1/4} T_{kin}. \tag{4.3}$$

With a single thermal band (e.g. Landsat-7 ETM+), ϵ has to be estimated from other sources. One way of doing this is to do a land cover classification with all available bands and then assign an ϵ value for each class from an emissivity table (e.g. 0.99 for water, 0.85 for granite).

In multispectral TIR, several bands of thermal wavelengths are available. With emissivity in each band, as well as the surface temperature (T_{kin}), unknown, we still have

an under-determined system of equations. For this reason, it is necessary to make certain assumptions about the shape of the emissivity spectrum we are trying to observe. Different algorithms exist to separate the influence of temperature from the emissivity.

4.2.2 Thermal applications

In general, applications of thermal remote sensing can be divided into two groups. In one group, the main interest is the study of surface composition by observing the surface emissivity in one or more wavelengths. In the other group, the focus is on surface temperature and its spatial and temporal distribution. The following discussion only concerns this second group.

Thermal hotspot detection Another application of thermal remote sensing is the detection and monitoring of small areas with thermal anomalies. The anomalies can be related to fires, such as forest fires or underground coal fires, or to volcanic activity, such as lava flows and geothermal fields. Figure 4.10 shows an ASTER scene that was acquired at night. The advantage of night images is that the Sun does not heat up the rocks surrounding the anomaly, as would be the case during the day. This results in higher contrast between the temperatures of the anomaly itself and surrounding rocks. This particular image was acquired over the Wuda coal-mining area in China in September 2002. Hotter temperatures are represented by brighter shades of grey. On the right side, the Yellow River is clearly visible, since water does not cool down as quickly as the land surface does, due to thermal inertia. Inside the mining area (white box in Figure 4.10), several hotspots, with higher temperatures compared to the surrounding rocks, are visible. The inset shows the same mining area slightly enlarged. The hottest pixels are orange and show the locations of coal fires. If images are taken several weeks, or even years, apart the development of these underground coal fires, as well as the effect of fire fighting efforts, can be monitored quite effectively with thermal remote sensing.

Glaciers monitoring With thermal remote sensing, studies of glaciers can go further than the plain observation of their extent. Understanding the dynamics of a glacier's state requires environmental variables. Ground surface temperature is obviously among the most important variables that affect glacier dynamics.

Urban heat islands The temperature of many urban areas is significantly higher than that of surrounding natural and rural areas. This phenomenon is referred to as an urban heat island. The temperature difference is usually larger at night than during the day and occurs mainly due to the change of matter covering the land as a result of urban development: land cover in built-up areas retains heat much better than land cover in natural and rural areas. This affects the environment in many ways: it modifies rainfall patterns, wind patterns, air quality, the seasonality of vegetation growth, and so on. Urban heat islands also affect the health of urban inhabitants: in particular, they can modify the duration and magnitude of heat waves in urban areas, leading to increases in mortality rates. There are several ways to mitigate the urban heat island effect, the most prominent ones being the use of highly reflective materials and increasing the amount of urban vegetation. To study the urban heat island effect we need to observe the temperature in urban and surrounding areas. Thermal remote sensing is a suitable tool as it provides temperature measurements that incorporate the spatial extent of cities and their surroundings.

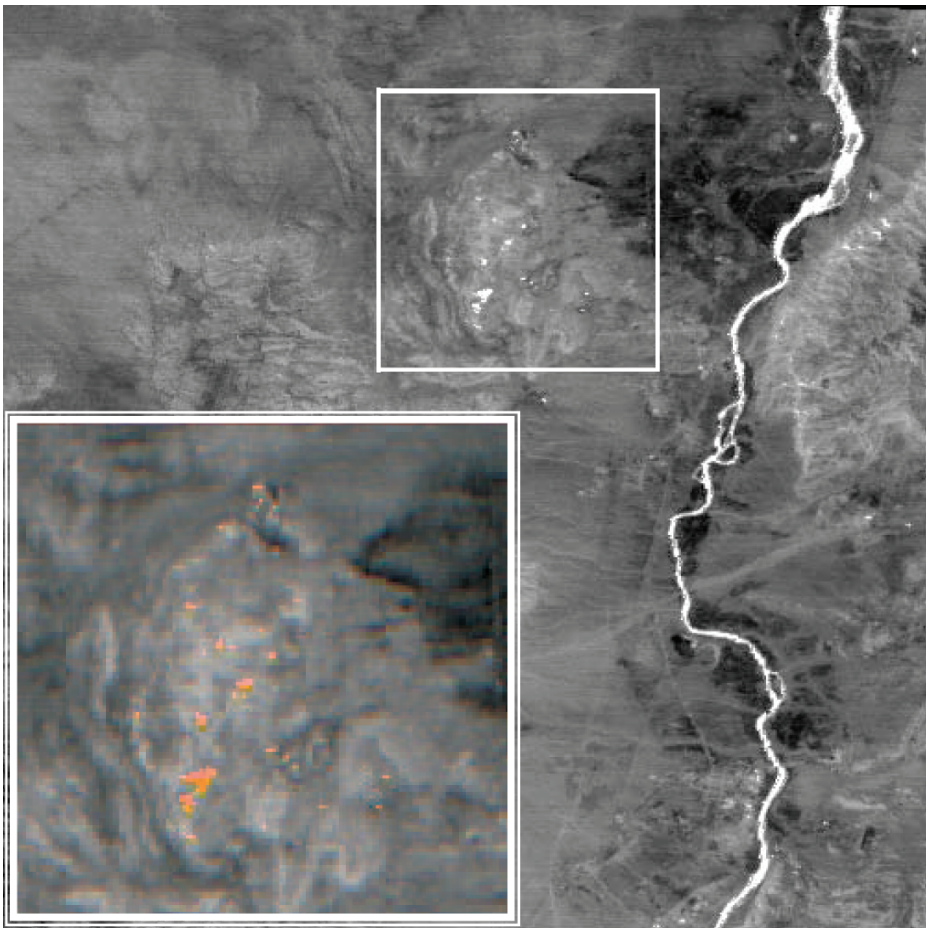


Figure 4.10
ASTER thermal band 10 over Wuda, China. Light coloured pixels inside the mining area (white box) are caused mainly by coal fires. Inset: pixels exceeding the background temperature of 18 °C are orange for better visibility of the fire locations. This scene is approximately 45 km wide.

4.3 Imaging Spectrometry

You have learnt in Section 2.5 that materials of interest may be distinguished by their spectral reflectance curves (e.g. Figure 2.14). In this section we will call spectral reflectance curves *reflectance spectra*. Most multispectral sensors that were discussed in Chapter 2 acquire data in a number of relatively broad wavelength bands. However, typical diagnostic absorption features, characterizing materials of interest in reflectance spectra, are in the order of 20–40 nm in width. Hence, broadband sensors under-sample this information and do not allow full exploitation of the spectral resolution potential available. Imaging spectrometers typically acquire images in a large number of spectral bands (more than 100). These bands are narrow (less than 10–20 nm in width) and contiguous (i.e. adjacent), which enables the extraction of reflectance spectra at pixel scale (Figure 4.11). Such narrow spectra enable the detection of diagnostic absorption features. Different names have been coined for this field of remote sensing, including imaging spectrometry, imaging spectroscopy and hyperspectral imaging.

Figure 4.12 illustrates the effect of spectral resolution for the mineral kaolinite. From top to bottom, the spectral resolution increases from 100–200 nm (Landsat), 20–30 nm (GERIS), 20 nm (HIRIS), 10 nm (AVIRIS), to 1–2 nm (USGS laboratory reference spec-

trum). With each improvement in spectral resolution, the diagnostic absorption features and, therefore, the unique shape of kaolinite's spectrum become more apparent.

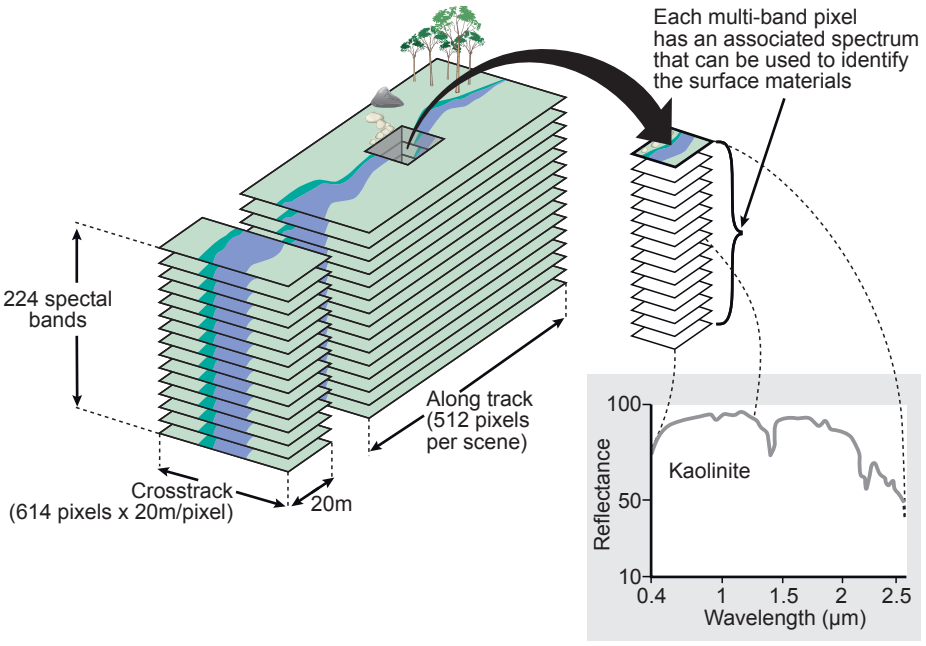


Figure 4.11
The concept of imaging spectrometry (adapted from [114]).

4.3.1 Reflection characteristics of rocks and minerals

reflectance spectra

Rocks and minerals reflect and absorb electromagnetic radiation as a function of the wavelength of the radiation. Reflectance spectra show these variations in reflection and absorption for various wavelengths (Figure 4.13). By studying the reflectance spectra of rocks, individual minerals and groups of minerals may be identified. In the Earth sciences, absorption in the wavelength region 0.4 μm–2.5 μm is commonly used to determine the mineralogical content of rocks. In this region, various groups of minerals have characteristic reflectance spectra; examples include phyllosilicates, carbonates, sulphates, and iron oxides and iron hydroxides. High-resolution reflectance spectra for mineralogy studies can easily be obtained in the field or the laboratory using field spectrometers.

Processes that cause absorption of electromagnetic radiation occur at the molecular and atomic levels. Two types of processes are important in the 0.4 μm–2.5 μm range: electronic processes; and vibrational processes ([21]). Depending on the molecular structure and composition, different absorption features can be identified. Reflectance spectra also correspond closely to the crystal structure of minerals and can, therefore, be used to obtain information about their crystallinity and chemical composition.

4.3.2 Pre-processing of imaging spectrometer data

radiometric calibration

atmospheric correction

Pre-processing of imaging spectrometer data involves radiometric calibration (see Section 5.2), which provides transfer functions to convert DN values to at-sensor radiance. The at-sensor radiance data have to be corrected by the user for atmospheric effects to obtain at-sensor or surface reflectance data. Section 5.2 contains an overview of the use of radiative transfer models for atmospheric correction. The correction provides absolute reflectance data, because the atmospheric influence is modelled and removed.

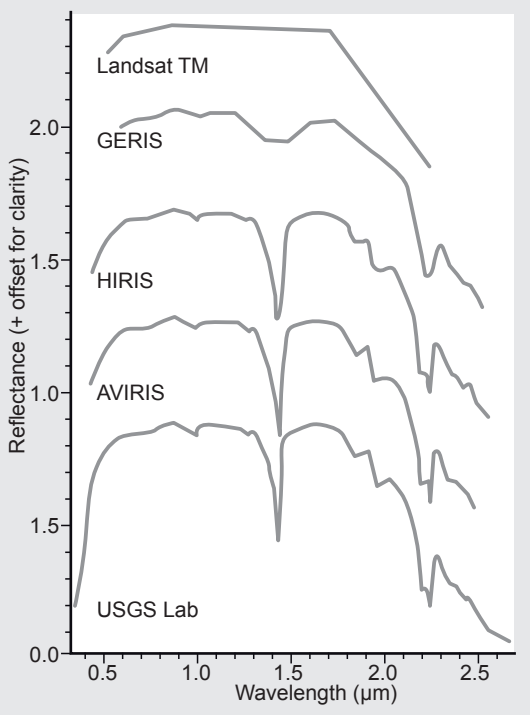


Figure 4.12
Example of a kaolinite spectrum at the original resolution (source: USGS laboratory) and at the spectral resolutions of various imaging devices. Note that the spectra are progressively offset upwards by 0.4 units for clarity (adapted from USGS).

Alternatively, users can make a scene-dependent relative atmospheric correction using empirically derived models for the radiance-reflectance conversion that are based on calibration targets found in the imaging spectrometer data set. Empirical models often used include techniques known as flat-field correction and empirical-line correction. Flat-field correction achieves radiance-reflectance conversion by dividing the whole data set on a pixel-by-pixel basis by the mean value of a target area within the scene that is spectrally and morphologically flat, spectrally homogeneous and has a high albedo. Conversion of raw imaging spectrometer data to reflectance data using the empirical-line method, on the other hand, requires selection and spectral characterization (in the field with a spectrometer) of two calibration targets (a dark and a bright target). This empirical correction uses a constant gain and offset for each band to force a best fit between sets of field and image spectra that characterize the same ground areas, thus removing atmospheric effects, residual instrument artefacts, and viewing geometry effects.

relative correction

4.3.3 Applications of imaging spectrometry data

A brief outline of current applications in various fields relevant to the thematic context of ITC are described in the remainder of this subsection.

Geology and resources exploration

Imaging spectrometry is used by the mining industry for surface mineralogy mapping, to aid in ore exploration. Other applications of this technology include lithological and structural mapping. The petroleum industry is also developing methods for using imaging spectrometry for reconnaissance surveys. The main targets are hydrocarbon seeps and microseeps.

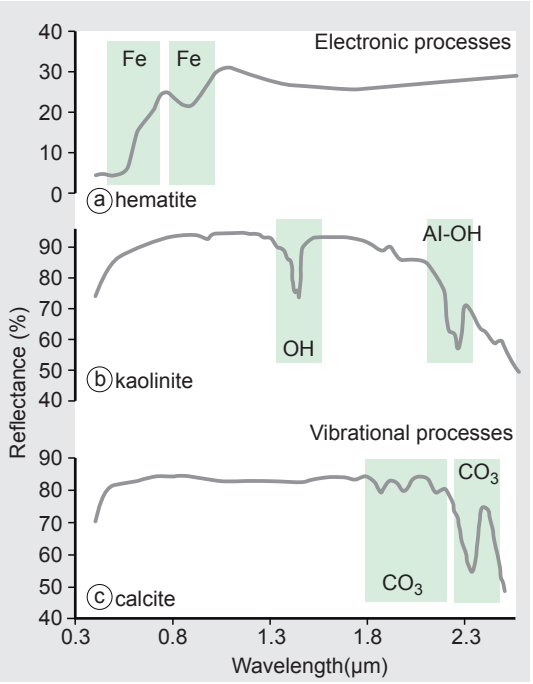


Figure 4.13
Effects of electronic and vibrational processes on absorption of electromagnetic radiation.

Other fields of application include environmental geology (and related geo-botany), in which currently much work is being done on acid mine drainage and mine-waste monitoring. Imaging of the atmospheric effects resulting from geological processes (e.g. sulfates emitted from volcanoes), to predict and quantify the presence of various gases for hazard assessment, is also an important field. In soil science, much emphasis has been placed on the use of spectrometry for the study of soil surface properties and soil composition analysis. Major elements such as iron and calcium, in addition to cation–anion exchange capacity, can be estimated from imaging spectrometry. In a more regional context, imaging spectrometry has been used to monitor agricultural areas (per-lot monitoring) and semi-nature areas. Recently, spectral identification from imaging spectrometers has been successfully applied to the mapping of the swelling clay minerals smectite, illite and kaolinite, in order to quantify the swelling potential of expansive soils. It should be noted that mining companies and, to a lesser extent, petroleum companies are already using imaging spectrometer data for reconnaissance-level exploration.

Vegetation sciences

Much research in vegetation studies has emphasized leaf biochemistry and leaf and canopy structure. Biophysical models for leaf constituents are currently available, as are soil–vegetation models. Estimates of plant material and structure, and biophysical variables, include carbon balance, yield/volume, nitrogen, cellulose, and chlorophyll. Leaf area index and vegetation indices have been extended to the hyperspectral domain and remain important physical variables for characterizing vegetation. One ultimate goal is the estimation of biomass and the monitoring of changes therein. Several research groups have been investigating the bi-directional reflectance function in relation to vegetation species analysis and floristics. Vegetation stress as a result of water deficiency, pollution (such as acid mine drainage) and geo-botanical anomalies in relation to ore deposits or petroleum and gas seepage links vegetation analysis to

exploration. Another upcoming field of application is precision agriculture, in which imaging spectrometry is being used to improve agricultural practices. An important factor in the health of vegetation is chlorophyll absorption and, in relation to that, the position of the red edge, determined using the red-edge index. Red edge is the name given to the steep increase in the reflectance spectrum of vegetation between visible red and near infrared wavelengths.

Hydrology

In hydrological sciences, the interaction of electromagnetic radiation with water, and the inherent and apparent optical properties of water are a central issue. Atmospheric correction and air–water interface corrections are very important in the imaging spectrometry of water bodies. Water quality of freshwater aquatic environments, estuarine environments and coastal zones usually has an important impact on national water bodies. Detection and identification of phytoplankton biomass, suspended sediments and other matter, coloured dissolved organic matter, and aquatic vegetation (i.e. macrophytes) are crucial variables in optical models of water quality. Much emphasis has been put on the mapping and monitoring of the state and growth or breaking down of coral reefs, as these are important for the CO_2 cycle. In general, many multisensor missions such as Terra and Envisat are directed towards integrated approaches for global climate change studies and global oceanography. Atmospheric models are important in global climate-change studies and aid in the correctin of optical data for scattering and absorption owing to trace gases in the atmosphere. In particular, the optical properties and absorption characteristics of ozone, oxygen, water vapour and other trace gases, and scattering by molecules and aerosols, are important variables in atmosphere studies. All these can be and are estimated from imaging spectrometry data.

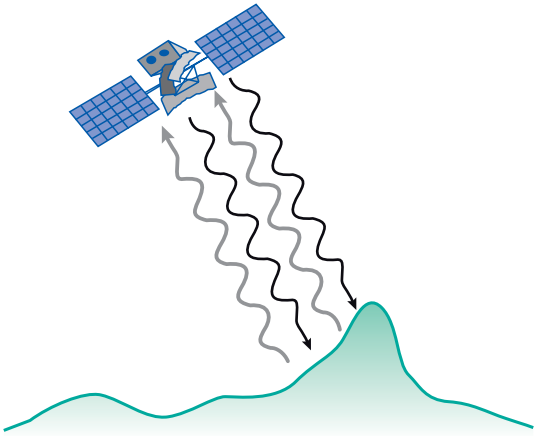


Figure 4.14
Principle of active microwave remote sensing.

4.4 Radar

4.4.1 What is radar?

Microwave remote sensing uses electromagnetic waves with wavelengths between 1 cm and 1 m (Figure 2.5). These relatively long wavelengths have the advantage that they can penetrate clouds and are not affected by atmospheric scattering. Although microwave remote sensing is primarily considered to be an active technique, passive sensors are also used. Microwave radiometers operate, similarly to thermal sensors, by detecting naturally emitted microwave radiation (either terrestrial or atmospheric). They are primarily used in meteorology, hydrology and oceanography.

microwave RS

In active systems, the antenna emits microwave signals to the Earth's surface, where they are backscattered. The part of the electromagnetic radiation that is scattered back in the direction of the antenna is detected by a sensor, as illustrated in Figure 4.14. There are several advantages to be gained from using active sensors, which have their own source of EM radiation:

- it is possible to acquire data at any time, also at night (similar to thermal remote sensing);
- since the waves are created by the sensor itself, the signal characteristics are fully controlled (wavelength, polarization, incidence angle, etc.) and can therefore be adjusted according to the desired application.

Active sensors can be divided into two types: imaging and non-imaging sensors. Radar sensors are typically active imaging microwave sensors. The term *radar* is an acronym for radio detection and ranging. *Radio* stands for the microwave component and *ranging* is another term for distance. Radar sensors were originally developed and used by the military. Nowadays, radar sensors are also widely used in civilian applications, such as environmental monitoring. Examples of non-imaging microwave instruments are *altimeters*, which collect distance information (e.g. sea-surface elevation), and *scatterometers*, which acquire information about object properties (e.g. wind speed).

non-imaging radar

The following subsection focuses on the principles of imaging radar and its applications. The interpretation of radar images is less intuitive than the interpretation of photographs and similar images. This is because of differences in the physical inter-

action of the waves with the Earth's surface. The interactions that take place and how radar images can be interpreted are also explained.

4.4.2 Principles of imaging radar

Imaging radar systems have a number of components: a transmitter, a receiver, an antenna, and a recorder. The transmitter is used to generate the microwave signal and transmit the energy to the antenna, from where it is emitted towards the Earth's surface. The receiver accepts the backscattered signal reaching the antenna and filters and amplifies it as required for recording. The recorder then stores the received signal.

Imaging radar acquires an image in which each pixel contains a digital number according to the strength of the backscattered radiation received from the ground. The radiation received from each emitted radar pulse can be expressed in terms of the physical variables and illumination geometry according to the *radar equation*:

backscattered radiation

$$P_r = \frac{G^2 \lambda^2 P_t \sigma}{(4\pi)^3 R^4}, \tag{4.4}$$

where

P_r	=	received radiance,
G	=	antenna gain,
λ	=	wavelength,
P_t	=	emitted radiance,
σ	=	*radar cross-section*, which is a function of the object characteristics and the size of the illuminated area, and
R	=	range from the sensor to the object.

This equation demonstrates that there are three main factors that influence the strength of the backscattered radiation received:

- radar system properties, i.e. wavelength, antenna and emitted power;
- radar imaging geometry, which defines the size of the illuminated area, which is in turn a function of, for example, beam width, incidence angle and range;
- the characteristics of interaction of the radar signal with objects, i.e. surface roughness and composition, and terrain relief (magnitude and orientation of slopes).

  These factors are explained in more detail below.

What exactly does a radar system measure? To interpret radar images correctly, it is important to understand what a radar sensor detects. Radar waves have the same physical properties as those explained in Chapter 2. Radar waves, too, have electric and magnetic fields that oscillate as a sine wave in perpendicular planes. In dealing with radar, the concepts of wavelength, period, frequency, amplitude, and phase are therefore relevant.

A radar transmitter creates microwave signals, i.e. *pulses* of microwaves at a fixed frequency (the *Pulse Repetition Frequency*), that are directed by the antenna into a beam. A pulse travels in this beam through the atmosphere, "illuminates" a portion of the Earth's surface, is backscattered and passes through the atmosphere back to the antenna, where the signal is received and its intensity measured. The signal needs to travel twice the distance between an object and the receiver/antenna. As we know

intensity

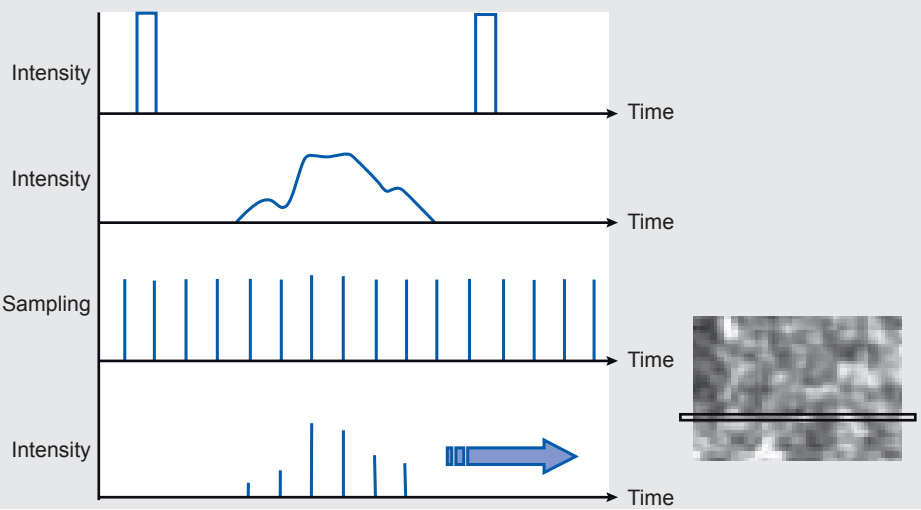


Figure 4.15
Illustration of how radar pixels result from pulses. For each sequence shown, one image line is generated.

the speed of light, we can calculate the distance (*range*) between sensor and object (see Formula 4.5).

To create an *image*, the return signal of each single pulse is sampled and samples stored in an image line (Figure 4.15). With the movement of the sensor while emitting pulses, a two-dimensional image is created (each pulse defines one line). The radar sensor therefore measures distances and backscattered signal intensities.

Commonly-used imaging radar bands Similarly to optical remote sensing, radar sensors operate within one or more different bands. For better identification, a standard has been established that defines various wavelength ranges using letters to distinguish them from each other (Figure 4.16); you can recognize the different wavelengths used in radar missions from the letters used. The European ERS mission and the Canadian Radarsat use, for example, C-band radar. Just like multispectral bands, different radar bands provide information about different object characteristics.

Band	P	L	S	C	X	K	Q	V	W
Frequency (GHz)	0.3	1.0	3.0		10.0	30.0			100.0
Wavelength (cm)	100	30	10		3	1			0.3

Figure 4.16
The microwave spectrum and band identification by letters.

Microwave polarizations The polarization of an electromagnetic wave is important in radar remote sensing. Depending on the orientation of the emitted and received radar wave, polarization will result in different images (see Figure 2.1, which shows a vertically polarized EM wave). It is possible to work with horizontally-, vertically- or cross-polarized radar waves. Using different polarizations and wavelengths, you can collect information that is useful for particular applications, e.g. to classify agricultural fields. In radar system descriptions you will come across the following abbreviations:

- HH: horizontal transmission and horizontal reception;
- VV: vertical transmission and vertical reception;

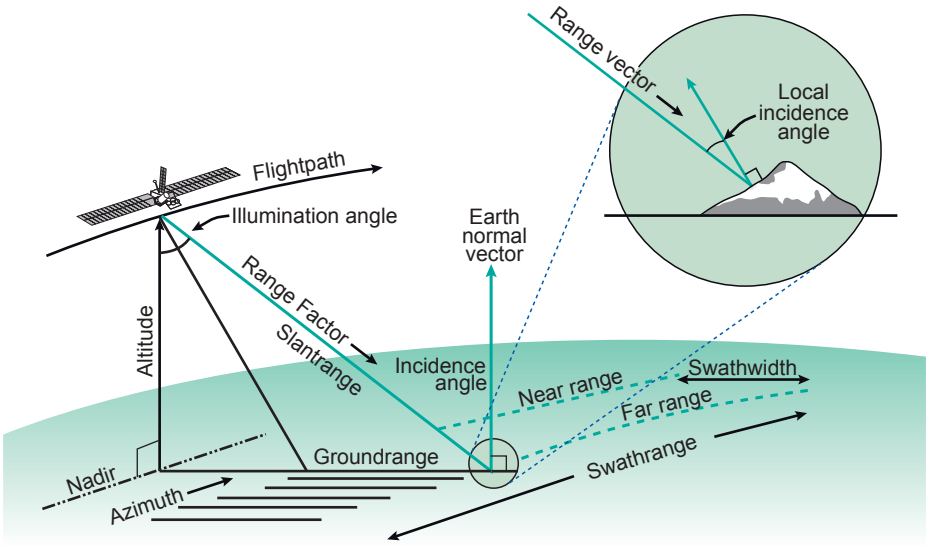


Figure 4.17
Radar remote-sensing geometry.

- HV: horizontal transmission and vertical reception;
- VH: vertical transmission and horizontal reception.

4.4.3 Geometric properties of radar

The platform carrying the radar sensor travels along its orbit or flight path (Figure 4.17). You can see the ground track of the orbit/flight path on the Earth's surface at nadir. The microwave beam illuminates an area, or *swath*, on the Earth's surface, with an offset from nadir, i.e. side-looking. The direction along-track is called *azimuth*and the direction perpendicular (across-track) is called *range*.

azimuth

Radar viewing geometry

Radar sensors are side-looking instruments. The portion of the image that is closest to the nadir track of the satellite carrying the radar is called *near range*. The part of the image that is farthest from nadir is called *far range* (Figure 4.17). The *incidence angle* of the system is defined as the angle between the radar beam and the local Earth normal vector. Moving from near range to far range, the incidence angle increases. It is important to distinguish between the incidence angle of the sensor and the *local incidence angle*, which differs depending on terrain slope and the curvature of the Earth (Figure 4.17). The local incidence angle is defined as the angle between the radar beam and the local surface normal vector. The radar sensor measures the distance between antenna and object. This line is called the *slant range*. The true horizontal distance along the ground corresponding to each point of measured slant range is called the *ground range* (Figure 4.17).

ranges

Spatial resolution

In radar remote sensing, the images are created from the backscattered portion of emitted signals. Without further sophisticated processing, the spatial resolutions of slant range and azimuth direction are defined by the pulse length and the antenna beam width, respectively. This setup is called *real aperture radar* (RAR). As different parameters determine the spatial resolution in range and azimuth, it is obvious that the spatial

RAR

resolution in each direction is different from the other. For radar image processing and interpretation it is useful to resample the data to the same GSD in both directions.

Slant range resolution For slant range, the spatial resolution is defined as the distance that two objects on the ground have to be apart to give two different echoes in the return signal. Two objects can be resolved in range direction if they are separated by at least half a pulse length. In that case, the return signals will not overlap. Slant range resolution is independent of the actual range (see Figure 4.18).

Azimuth resolution The spatial resolution in azimuth direction depends on the beam width and the actual range. The radar beam width is proportional to the wavelength and inversely proportional to the antenna length, i.e. *aperture*. This means the longer the antenna, the narrower the beam and the higher the spatial resolution in azimuth direction.

aperture

RAR systems have their limitations in getting useful spatial resolutions of images because there is a physical limit to the length of the antenna that can be carried on an aircraft or satellite. On the other hand, shortening the wavelength will reduce the capability of penetrating clouds. To improve the spatial resolution, a large antenna is synthesized by taking advantage of the forward motion of the platform. Using all the backscattered signals in which a contribution of the same object is present, a very long antenna can be synthesized. This length is equal to the part of the orbit or flight path in which the object is "visible". Most airborne and space-borne radar systems use this type of radar. Systems using this approach are referred to as *Synthetic Aperture Radar (SAR)*.

SAR

4.4.4 Distortions in radar images

Due to the side-looking geometry, radar images suffer from serious geometric and radiometric distortions. In a radar image, you encounter variations in scale (caused by slant range to ground range conversion), *foreshortening*, *layover* and *shadows* (due to terrain elevation; see Figure 4.19). Interference due to the coherence of the signal causes *speckle* effects.

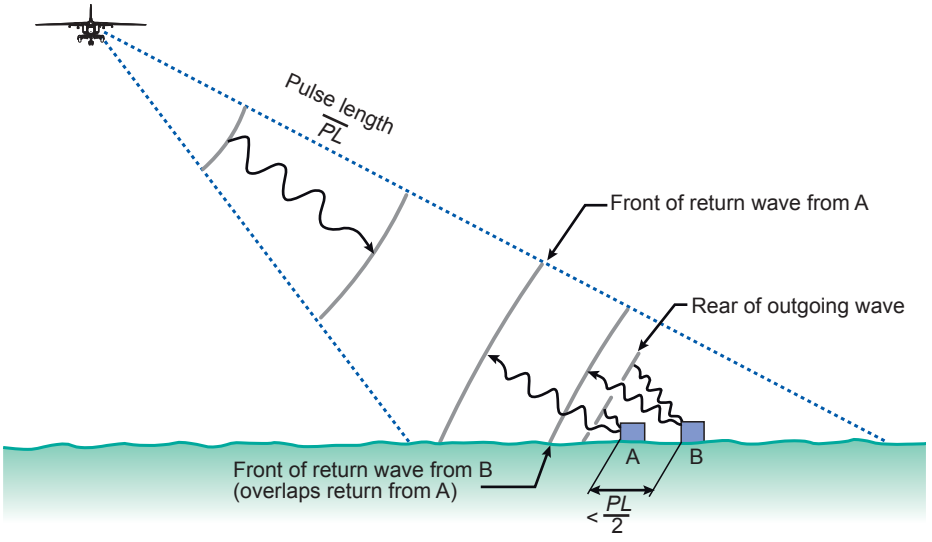


Figure 4.18
Illustration of the slant range resolution.

Scale distortions

Radar measures ranges to objects in slant range rather than true horizontal distances along the ground. Therefore the image has different scales moving from near to far range (Figure 4.17). This means that objects in near range are compressed as compared to objects in far range. For proper interpretation, the image has to be corrected and transformed into ground range geometry.

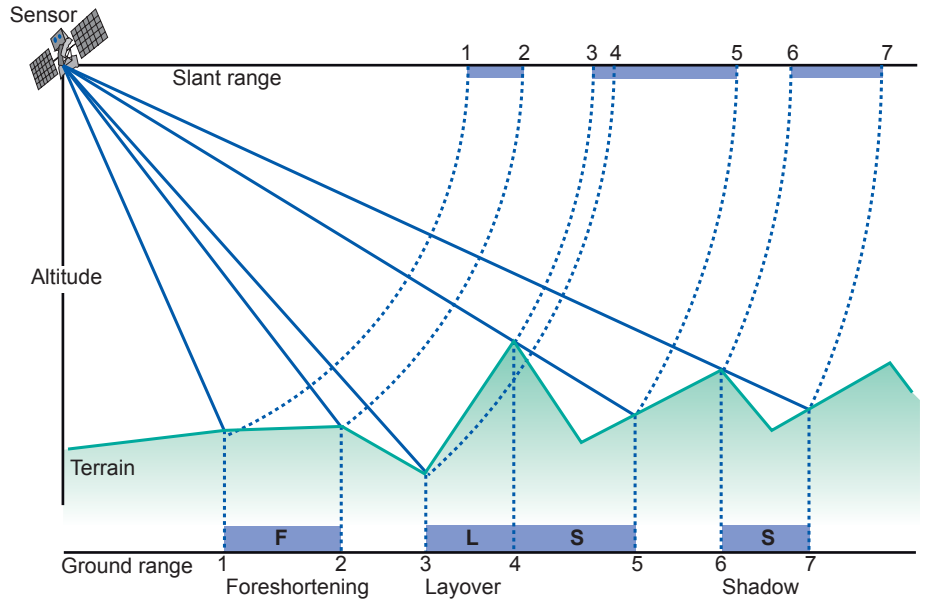


Figure 4.19
Geometric distortions in a radar image caused by varying terrain elevation.

Terrain-induced distortions

Similarly to optical sensors that can operate in an oblique manner (e.g. SPOT), radar images are subject to relief displacements. In the case of radar, these distortions can be severe. There are three effects that are typical for radar: *foreshortening*, *layover* and *shadow* (see Figure 4.19).

Radar measures distance in slant range. The slope area facing the radar is compressed in the image. The amount of shortening depends on the angle that the slope forms in relation to the incidence angle. The distortion is at its maximum if the radar beam is almost perpendicular to the slope. Foreshortened areas in the radar image are very bright.

foreshortening

If the radar beam reaches the top of the slope earlier than the bottom, the slope is imaged upside down, i.e. the slope "lays over". As you can understand from the definition of foreshortening, layover is an extreme case of foreshortening. Layover areas in the image are very bright.

layover

In the case of slopes that are facing away from the sensor, the radar beam cannot illuminate the area at all. Therefore, there is no radiation that can be backscattered to the sensor and so those regions remain dark in the image.

shadow

Radiometric distortions

Geometric distortions also influence the received radiation. Since backscattered radiation is collected in slant range, the received radiation coming from a slope facing the sensor is stored in a reduced area in the image, i.e. it is compressed into fewer pixels

than should be the case if obtained in ground range geometry. This results in high digital numbers because the radiation collected from different objects is combined. Slopes facing the radar appear bright. Unfortunately this effect cannot be corrected for. This is why especially layover and shadow areas in a radar image cannot be used for interpretation. However, they are useful in the sense that they contribute to a three-dimensional appearance of the image and therefore contribute to an understanding of surface structure and terrain relief.

speckle

interference

A typical property of radar images is *speckle*, which appears as grainy "salt and pepper" effects in the image (Figure 4.20). Speckle is caused by the interference of backscattered signals coming from an area that is encapsulated in one pixel. The wave interactions are called *interference*. Interference causes the return signals to be extinguished or amplified, resulting in dark and bright pixels in the image, even when the sensor observes a homogeneous area. Speckle degrades the quality of the image and makes the interpretation of radar images difficult.

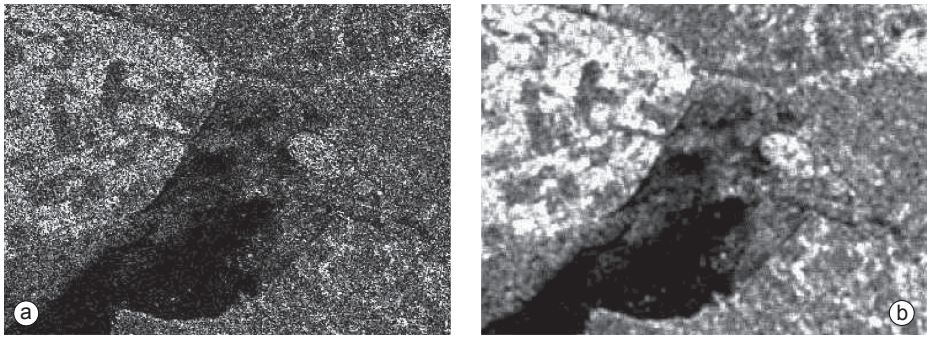


Figure 4.20
An original (a) and speckle filtered (b) radar image.

Speckle reduction

It is possible to reduce speckle by multi-look processing or spatial filtering. If you purchase an ERS SAR scene in "intensity (PRI)-format" you will receive a 3-look or 4-look image. Another way to reduce speckle is to apply spatial filters to the images. Speckle filters are designed to adapt to local image variations in order to smooth values, thus reducing speckle and enhancing lines and edges to maintain the sharpness of an image.

4.4.5 Interpretation of radar images

The brightness of features in a radar image depends on the strength of the backscattered signal. In turn, the amount of radiation that is backscattered depends on a number of factors. An understanding of these factors helps with the proper interpretation of radar images.

Microwave signal and object interactions

For those who are concerned with the visual interpretation of radar images, the degree to which they are able to interpret an image depends upon whether they can identify typical/representative tones related to surface characteristics. The amount of radiation that is received at the radar antenna depends on the illuminating signal (radar system variables such as wavelength, polarization and viewing geometry) and the characteristics of the illuminated object (e.g. roughness, shape, orientation, dielectric constant).

Surface roughness is the terrain property that most strongly influences the strength of radar backscatter. It is determined by textural features comparable to the size of the radar wavelength (typically between 5 and 40 cm), for example, leaves and twigs of vegetation and sand, gravel and cobble stones. A distinction should be made between surface roughness and terrain relief. Surface roughness occurs at the level of the radar wavelength (centimetres to decimetres). By terrain relief we mean the variation of elevation of the ground surface; relative to the resolution of radar images, only elevation change in the order of metres is relevant. *Snell's law* states that the angle of reflection is equal and opposite to the angle of incidence. A smooth surface reflects the radiation away from the antenna without returning a signal, thereby resulting in a black image. With an increase in surface roughness, the amount of radiation reflected away is reduced and there is an increase in the amount of signal returned to the antenna. This is known as the backscattered component. The greater the amount of radiation returned, the brighter the signal is shown on the image. A radar image is, therefore, a record of the backscatter component and is related to surface roughness.

surface roughness

terrain relief

Complex dielectric constant Microwave reflectivity is a function of the complex dielectric constant, which is a measure of the electrical properties of surface materials. The *dielectric constant* of a medium consists of a part referred to as permittivity and a part referred to as conductivity [112]. Both properties, permittivity and conductivity, are strongly dependent on the moisture or liquid-water content of a medium. Material with a high dielectric constant has a strongly reflective surface. Therefore the difference in the intensity of the radar return for two surfaces of equal roughness is an indication of the difference in their dielectric properties. In the case of soils, this could be due to differences in soil moisture content.

Surface Orientation Scattering is also related to the orientation of an object relative to the radar antenna. For example the roof of a building appears bright if it faces the antenna and dark if the incoming signal is reflected away from the antenna. Thus backscatter depends also on the local incidence angle.

Volume scattering is related to multiple scattering processes within a group of objects, such as the vegetation canopy of a wheat field or a forest. The cover may be all trees, as in a forested area, which may be of different species, with variations in leaf form and size; or grasses and bushes with variations in form, stalk size, leaf and angle, fruiting and a variable soil surface. Some of the radiation will be backscattered from the vegetation surface, but some, depending on the characteristics of radar system used and the object material, will penetrate the object and be backscattered from surfaces within the vegetation. Volume scattering is therefore dependent upon the heterogeneous nature of the object surface and the physical properties of the object, as well as the characteristics of the radar used, such as wavelength and its related effective penetration depth [8].

Point objects are objects of limited size that give a very strong radar return signal. Usually, the high level of backscatter is caused by *corner reflection*. An example of this is the dihedral corner reflector—a point object situation resulting from two flat surfaces intersecting at 90 ° and situated orthogonally to the incident radar beam. Common forms of dihedral configurations are man-made features such as transmission towers, railway tracks or the smooth side of buildings on a smooth ground surface. Another type of point object is a trihedral corner reflector, which is formed by the intersection of

corner reflection

three mutually perpendicular flat surfaces. Point objects that are corner reflectors are commonly used to identify known fixed points in an area in order to perform precise calibration measurements. Such objects can occur and are best seen in urban areas, where buildings can act as trihedral or dihedral corner reflectors. These objects give rise to intense bright spots on a radar image and are typical for urban areas. Point objects are examples of objects that are sometimes below the resolution of a radar system but, because they dominate the return signal from a cell, nevertheless give a clearly visible point; they may even dominate the surrounding cells.

4.4.6 Applications of radar

There are many useful applications of radar imaging. Radar data provide information complementary to visible and infrared remote-sensing data. In the case of forestry, radar images can be used to obtain information about forest canopy, biomass and different forest types. Radar images can also be used to distinguish between different types of land cover, e.g. urban areas, agricultural fields and water bodies. In urban areas, radar detects buildings (corner reflectors) and metal constructions, thus allowing the extent of urban areas to be delineated, which is a key observable for urban growth studies. In agricultural crop identification, the use of radar images acquired using different polarization (mainly airborne) is quite effective. It is crucial for agricultural applications to acquire data at a certain moment (season) to obtain the necessary parameters. This is possible because radar can operate independently of weather or light conditions. In geology and geomorphology, the fact that radar provides information about surface texture and roughness plays an important role in lineament detection and geological mapping. Since radar backscatter is sensitive to surface roughness, it helps to discriminate between ice and debris, thus making it potentially suitable for glaciers monitoring studies. Radar also allows the measurement of elevation and change in elevation by a technique called Interferometric SAR (INSAR). Radar has also been successfully applied in hydrological modelling and soil moisture estimation—based on the sensitivity of the microwave to the dielectric properties of the observed surface. The interaction of microwaves with ocean surfaces and ice provides useful data for oceanography and ice monitoring. Radar data is also used for oil-slick monitoring and environmental protection.

4.5 Laser scanning

4.5.1 Basic principles

Laser scanning—in functional terms—can be defined as a system that produces digital surface models. The system comprises an assembly of various sensors, recording devices and software. The core component is the laser mechanism. The laser measures distance, which is referred to as "laser ranging". When mounted on an aircraft, a laser rangefinder measures at very short time intervals the distance to the terrain. - Combining a laser rangefinder with sensors that can measure the position and attitude of the aircraft (GPS & IMU) makes it possible to create a model of the terrain surface in terms of a set of (X, Y, Z) coordinates, following the polar measuring principle; see Figure 4.21.

measuring by three sensors

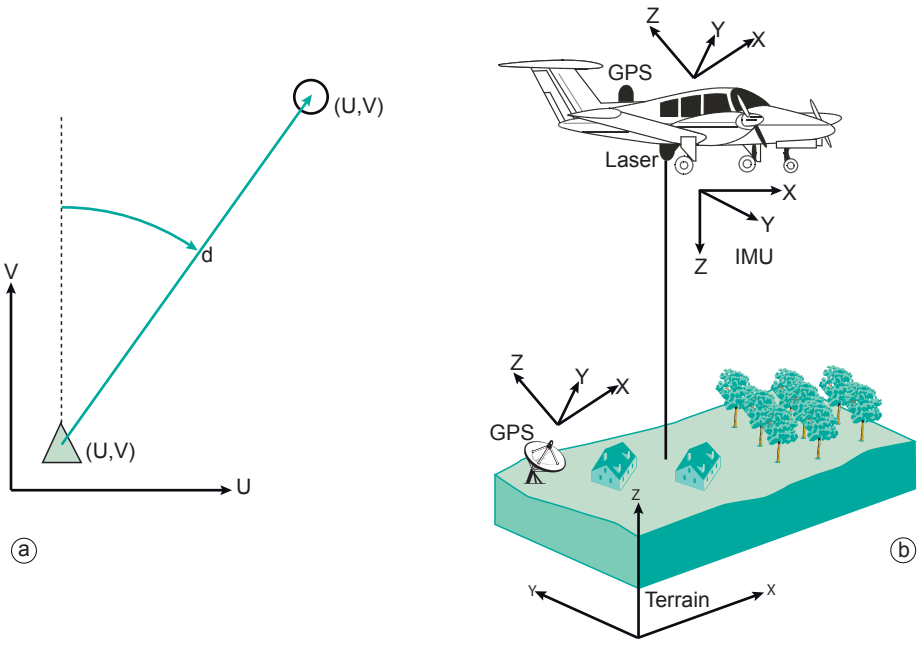


Figure 4.21
Polar measuring principle (a) and its application to ALS (b).

We can define the coordinate system in such a way that Z refers to elevation. The digital surface model (DSM) thus becomes a digital elevation model (DEM), i.e. we model the surface of interest by providing its elevation at many points, each with position coordinates (X, Y). Do the elevation values, which are produced by airborne laser scanning (ALS), refer to elevation of the *bare ground* above a predefined datum? Not necessarily, since the "raw DEM" gives us elevation of the surface the sensor "sees" (Figure 4.22). Post-processing is required to obtain a digital terrain model (DTM) from the DSM.

The key advantages of ALS are its high ranging precision, its ability to yield high resolution DSMs in near real-time, and its complete or nearly complete independence of weather, season or light conditions. Typical applications of ALS are, therefore, forest surveys, surveying of coastal areas and sand deserts, flood-plain mapping, power-line and pipeline mapping, monitoring open-pit mining and 3D city modelling.

applications of ALS

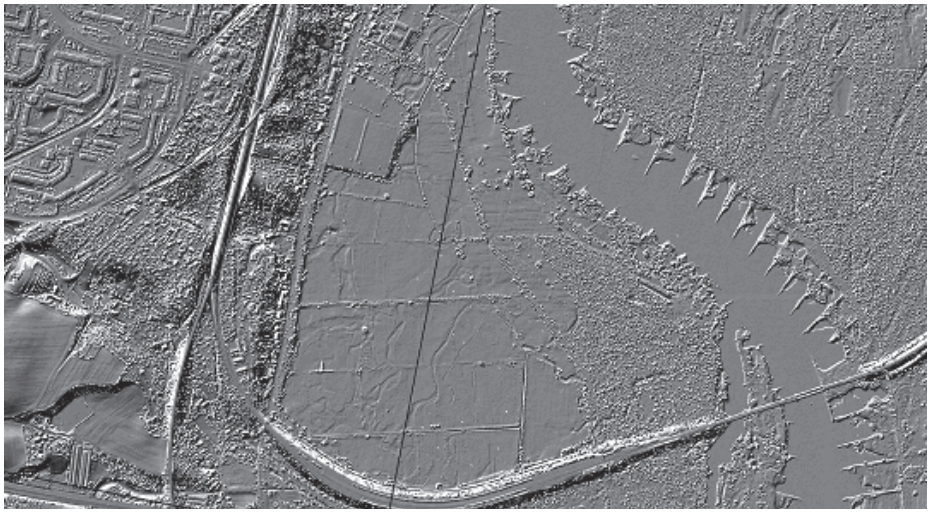


Figure 4.22
DSM of part of Frankfurt (Oder), Germany (1 m point spacing). Courtesy of TopoSys.

4.5.2 ALS components and processes

laser

LASER stands for Light Amplification by Stimulated Emission of Radiation. Although he did not invent it, Einstein can be considered the father of the laser. Roughly 85 years ago he postulated the phenomena of photons and stimulated emission; he won the Nobel prize for related research on the photoelectric effect. In 1960, Theodore Maiman, employed at Hughes Research Laboratories, developed a device to amplify light, thus building the first laser (instrument). A laser emits a beam of monochromatic light or radiation in the NIR range of the spectrum. The radiation is not really of a single wavelength, but it has a very narrow spectral band—smaller than 10 nm. Also specific for lasers is the very high intensity of radiation they emit. Today, lasers are used for many purposes, even human surgery. Lasers can damage cells (by boiling their water content), so they are a potential hazard for the eye. Users of laser rangefinders therefore have to attend safety classes; safety rules must be strictly observed when using lasers for surveying applications.

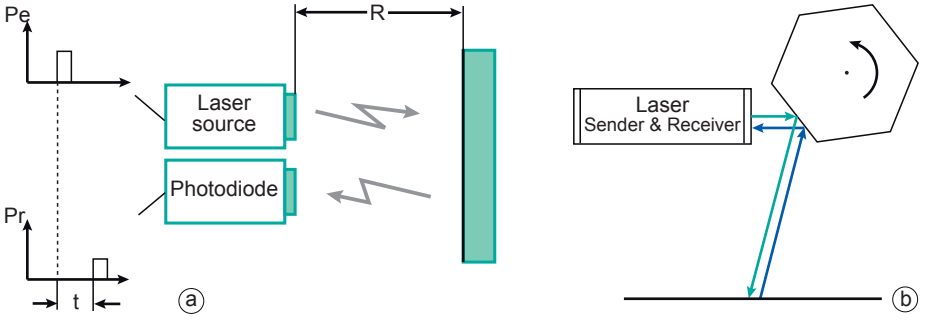


Figure 4.23
Concept of laser ranging and scanning [127].

laser rangefinders

Laser rangefinders and scanners come in various forms. Most airborne laser instruments are "pulse lasers". A pulse is a signal of very short duration that travels as a beam. Airborne laser rangefinders for topographic applications emit NIR radiation. An emitted pulse is reflected by the ground and its return signal is sensed by a photodiode (Figure 4.23). A time counter starts when a pulse is sent out and stops on its return. The elapse time is measured with a resolution of 0.1 ns. As we know the speed of light,c, the elapsed time can easily be converted to a distance:

$$R = \frac{1}{2}\, ct. \tag{4.5}$$

Modern laser scanners send out pulses at a very high frequency (up to 300,000 pulses per second). Across-track scanning is in most cases achieved by a moving mirror, which deflects the laser beam. The mirror can be of the oscillating, rotating, or nutating type. The Falcon system uses fiber optics to achieve scanning. Adding a scanning device to a ranging device has made surveying of a large area more efficient: a strip (a swath of points) can be captured on a single leg of , instead of just a line of points as was the case with earlier versions of laser systems (laser profilers).

laser scanner

Simple laser rangefinders register one return pulse for every emitted pulse. By contrast, modern laser rangefinders for airborne applications record multiple echoes from the same pulse. Multiple-return laser ranging is specifically relevant for aerial surveys of terrain covered by vegetation because it helps distinguish vegetation echoes from ground echoes. For a pulse that hits a leaf at the top of a tree, part of the pulse may be reflected, while another part of it may travel further, perhaps hitting a branch and, eventually, even the ground; see Figure 4.24. Many of the first return "echoes" will be from the tree canopy, while the last returns are more likely to come from the ground. Each return can be converted to an (X, Y, Z) of the illuminated target point. To figure out whether the point is on the ground or somewhere amongst the vegetation is far from trivial. Multiple return ranging does not give a direct answer but it helps find one. An example of a first return and last return DSM is shown in Figure 4.25. *Full waveform sensors* represent the further development of this approach. Instead of only detecting an echo if its intensity is above a certain threshold (Figure 4.24), full waveform scanners or altimeters (as on ICESat, see below) digitize the entire return signal of each emitted laser pulse. Full waveform laser rangefinders can provide information about surface roughness and more cues on vegetation cover.

multiple return ranging

full waveform sensors

As well as measuring the range, some laser-based instruments also measure the amplitude of the reflected signal to obtain an image (often referred to as "intensity image"). Imaging by laser scanner is different from imaging by radar instruments. While an image line of a microwave radar image stems from a single pulse, an image line of a laser intensity image stems from many pulses and is formed in the same way as for an across-track multispectral scanner. The benefit of "imaging lasers" is limited. The images obtained are monochromatic and are of lower quality than panchromatic images. A separate camera or multispectral scanner can produce much richer image content.

imaging laser

ALS provides 3D coordinates of terrain points. To calculate accurate coordinates of terrain points we must accurately observe all necessary elements. Measuring the distance from the aircraft to the terrain can be done very precisely by the laser rangefinder (accurate to within centimetres), and we can accurately determine the position and altitude of the aircraft using a POS (Section 4.1.1).

GPS and IMU

The most widely used platforms for ALS are airplanes and helicopters. Helicopters are better suited for very high-resolution surveys, because they can easily fly slowly. The minimum flying height is, among other things, dependent on the safe eye–laser distance for the instrument. The major limiting factor of the maximum flying height is energy loss of the laser beam. 1000 m and less are frequently used flying altitudes, although there are systems for which heights of 8000 m are feasible.

ALS platforms

Unlike aerial surveys for generating stereo coverage of photographs—for which each terrain point should be recorded at least twice—in ALS a terrain point is, in principle, only "collected" once, even if the strips flown overlap. This is an advantage when surveying urban areas and forests, but it has disadvantages for error detection.

After the flight, the recordings from the laser instrument and the POS are co-registered

co-registering the data

to the same time and then converted to (X, Y, Z) coordinates for each point that was hit by the laser beam. The resulting data set may still contain systematic errors and is often referred to as"raw data".

extracting information

Further data processing has then to solve the problem of extracting information from the un-interpreted set of (X, Y, Z) coordinates. Typical tasks are "extracting buildings", modelling trees (e.g. to compute timber volumes) and, in particular, filtering the DSM to obtain a DTM. Replacing the elevation value at non-ground points by an estimate of the elevation of the ground surface is also referred to as vegetation removal, or "devegging" for short, a term left over from the early days when ALS was primarily used for forested areas (Figure 4.26).

Proper system calibration, accurate flight planning and execution (including the GPS logistics), and adequate software are critical factors in ensuring one gets the right data at the right time.

4.5.3 System characteristics

ALS produces a DSM directly comparable with what is obtained by image matching of aerial photographs/images. *Image matching* is the core process of automatically generating a DSM from stereo images. Alternatively, we can also use microwave radar to generate DSMs and—eventually—DTMs. The question is then, why go for ALS? There are in fact several good reasons for using ALS for terrain modelling:

- A laser rangefinder measures distance by recoding the elapse time between emitting a pulse and receiving the reflected pulse from the terrain. Hence, the laser rangefinder is an active sensor and can be used both during daylight hours

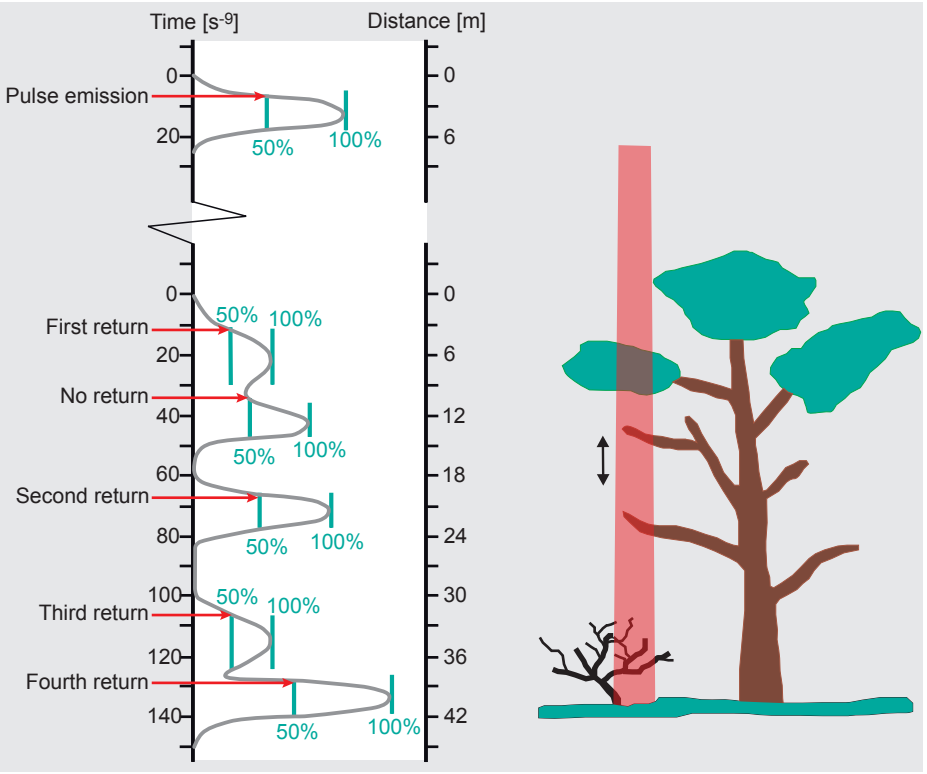
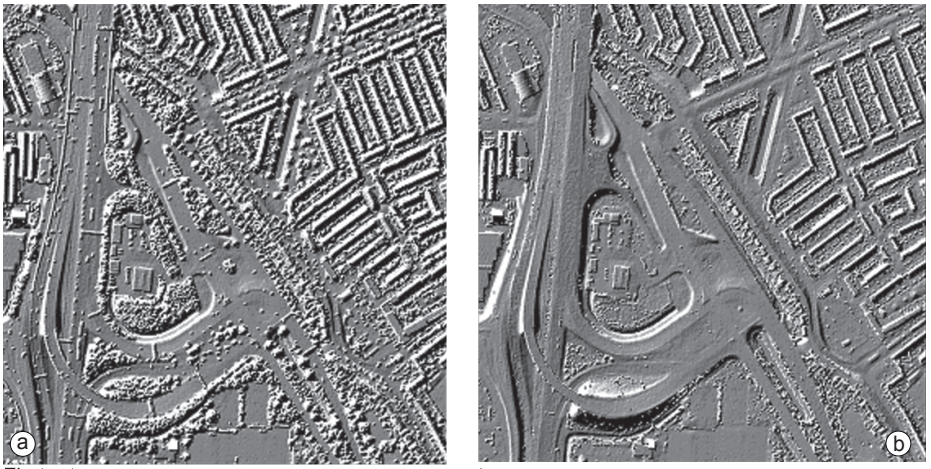


Figure 4.24
Multiple-return laser ranging. Adapted from Mosaic Mapping Systems, Inc.



First return Last return

Figure 4.25
First (a) and last (b) return DSMs of the same area. Courtesy of TopoSys.

and at night. The possibility of flying at night comes in handy when, for instance, surveying a busy airport.

- Unlike indirect distance measuring done using stereo images, laser ranging does not depend on surface/terrain texture.
- Laser ranging is less weather-dependent than passive optical sensors. A laser cannot penetrate clouds as microwave radar can, but it can be used at low altitudes, thus very often below the cloud ceiling.
- The laser beam is very narrow, with a beam divergence that can be less than 0.25 mrad; the area illuminated on the ground can, therefore, have a diameter smaller than 20 cm (depending on the laser type and flying height). The simplifying assumption of "measuring points" is thus closely approximated. ALS can "see" objects that are much smaller than the footprint of the laser beam, making it suitable for mapping power lines.
- A laser beam cannot penetrate leaves, but it can pass through the tree canopy, unless that is very dense.
- A laser rangefinder can measure distances very precisely and very frequently; therefore a DSM with a high density of points can be obtained with accurate elevation values. The attainable elevation (vertical coordinate) accuracy with ALS can be in the order of 3 cm for well-defined target surfaces.
- The multiple-return recording facility offers "feature extraction", especially for forest applications and urban mapping (building extraction), both attractive topics for researchers.
- The entire data collection process is digital, which allows it to be automated to a high degree, thus facilitating fast processing.
- Other than a calibration site, which can usually be set up near the airfield, ALS does not need any ground control.

There are two additional major advantages of laser ranging compared to microwave radar: high frequency X pulses can be generated at short intervals and highly directional beams can be emitted. The latter is possible because of the short wavelength of

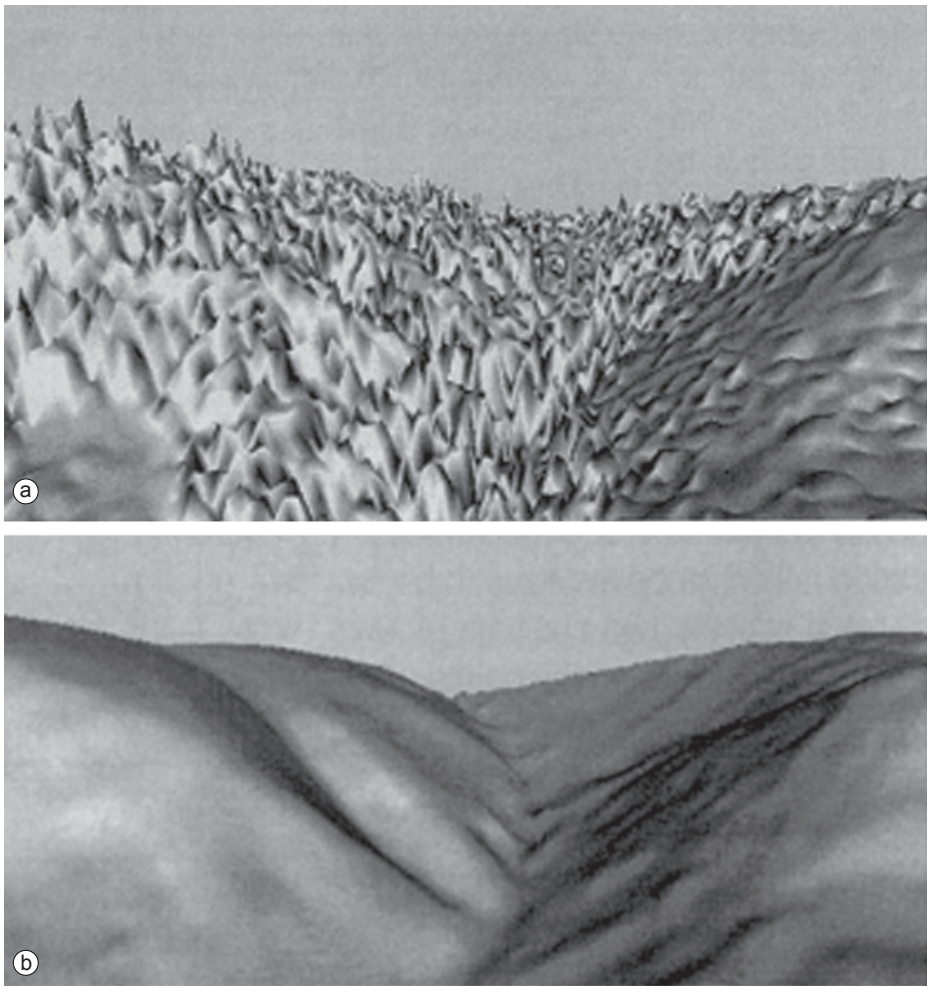


Figure 4.26
Devegging laser data: filtering a DSM (a) to create a DTM (b). From [61].

lasers (10,000 to 1,000,000 times shorter than microwaves). The consequence is much higher ranging accuracy.

Note that the term *radar* is often used as a short form for microwave radar. In the literature, however, you may also come across the term "laser radar", which is synonymous for laser ranging. A more frequently used synonym for laser ranging is LIDAR, although there are also LIDAR instruments that do not measure the distance to but, rather, the velocity of a target ('Doppler LIDARs').

Glacier monitoring Laser scanning provides information on surface elevation, thus making it a potentially useful tool for monitoring glaciers. However, modern laser scanners usually provide information at very fine spatial resolutions, which are not required in glacier studies. Furthermore, to differentiate glacier ice from debris, we require additional information from other sources.

Urban growth The sensitivity of laser scanning to the geometric properties of surfaces makes it a suitable tool for detecting objects of urban infrastructure. Laser scan-

ning is, therefore, a potentially useful tool for urbanization studies, especially when very detailed spatial information is required, such as for detecting informal settlements and city construction works. Detailed information about terrain is also relevant for the modelling and monitoring of city growth. Nevertheless, this technique is not currently being used in urban growth studies.

4.6 Aerial photography

Introduction

Aerial photographs have been used since the early 20th century to provide geospatial data for a wide range of applications. *Photography* is the process or art of producing images by directing light onto a light-sensitive surface. Taking and using photographs is the oldest, yet most commonly applied, remote sensing technique. *Photogrammetry* is the science and technique of making measurements on photos and converting these to quantities that are meaningful in the terrain. Some of ITC's early activities included photography and photogrammetry, the latter being, at that time, the most innovative and promising technique available for the topographic mapping of large areas. Aerial film cameras are typically mounted on aircraft, although a Russian satellite is known to have carried a photographic camera and NASA Space Shuttle missions have systematically photographed all aspects of their flights.

Aerial photographs and their digital variant, obtained by digital frame cameras, are today the prime data source for medium- to large-scale topographic mapping and for many cadastral surveys and civil engineering projects, as well as urban planning. Aerial photographs are also a useful source of information for foresters, ecologists, soil scientists, geologists and many others. Photographic film is a very mature medium and aerial survey cameras using film have reached vast operational maturity over the course of many years, so new, significant developments cannot be expected. Owners of aerial film cameras will continue to use them as long as Agfa and Kodak continue to produce film at affordable prices.

Two broad categories of aerial photographs can be distinguished: *vertical* and *oblique* photographs (Figure 4.27). For most mapping applications, vertical aerial photographs are required. A vertical aerial photograph is produced with a camera mounted into the floor of a survey aircraft. The resulting image is similar to a map and has a scale that is roughly constant throughout the image area. Vertical aerial photographs for mapping are usually taken such that they overlap in the flight direction by at least 60%. Two successive photos can form a stereo pair, thus enabling 3D measurement.

vertical photo

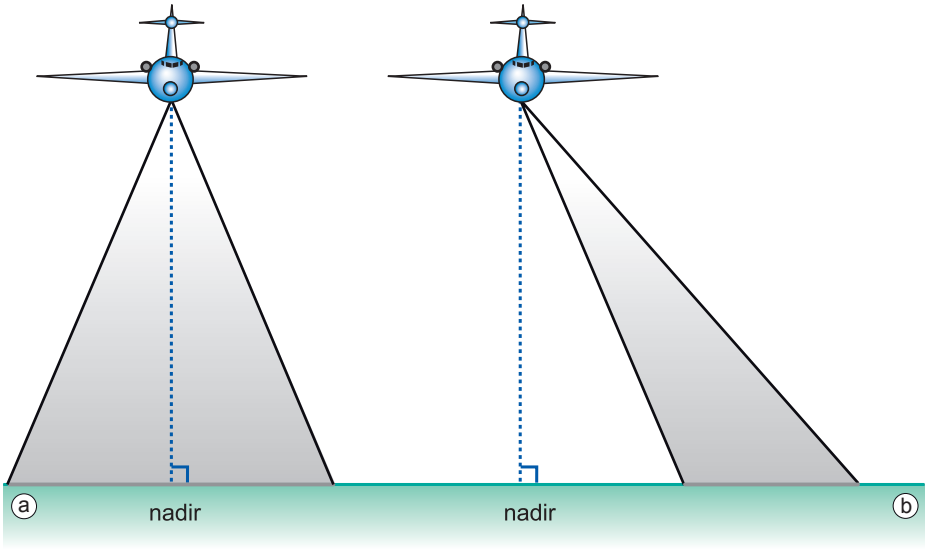


Figure 4.27
Vertical (a) and oblique (b) aerial photography.

Oblique photographs are obtained if the axis of the camera is not vertical. They can

Figure 4.28
A vertical (a) and oblique (b) aerial photograph of the ITC building, 1999.

also be made using a hand-held camera and shooting through an open window of an aircraft. The scale of an oblique photo varies from the foreground to the background, which complicates the measurement of positions from the image. For this reason, oblique photographs are rarely used for mapping. Nevertheless, oblique images can be useful for obtaining side views of objects such as buildings.

oblique photo

This section discusses the aerial photo camera, films and methods used for vertical aerial photography. Subsection 4.6.1 describes the aerial camera and its main components. In broad terms, photography is based on the exposure of a photographic film to light, the processing of the film, and the printing of photographs from the processed film. Subsection 4.6.2 discusses the basic geometric—i.e. spatial—characteristics of aerial photographs. Finally, the basics concepts of aerial photography missions are introduced in Subsection 4.6.3.

4.6.1 Aerial survey cameras

A camera used for vertical aerial photography for mapping purposes is called an *aerial survey camera*. Only two manufacturers of aerial survey cameras, namely Leica and Z/I Imaging, have continued to assemble aerial film cameras; their cameras are the RC-30 and the RMK-TOP, respectively. Aerial survey cameras contain a number of components that are also common to any typical hand-held camera, as well as a number of specialized components that are necessary for its specific role. The large size of aerial cameras results from the need to acquire images of large areas with a high spatial resolution. This is achieved by using very large-sized film. Modern aerial survey cameras produce negatives measuring 23 cm $\times$ 23 cm (9 inch $\times$ 9 inch); up to 600 photographs may be recorded on a single roll of film. To achieve the same degree of quality as an aerial film camera, a digital camera has to produce shots comprising about 200 million pixels.

4.6.2 Spatial characteristics

Two important properties of an aerial photograph are scale and spatial resolution. These properties are determined by sensor (lens cone and film) and platform (flying height) characteristics. Lens cones are available in different focal lengths.

Scale

The relationship between the photo scale factor, s, flying height, H, and focal length, f, is given by

$$s = \frac{H}{f}. \tag{4.6}$$

Obviously, the same scale can be achieved with different combinations of focal length and flying height. If a lens of smaller focal length is used, while the flying height remains constant, then (see also Figure 4.29):

- The *photo scale factor* will increase and the size of individual details in the image will become smaller. In the example shown in Figure 4.29, using a 150 mm and 300 mm lens at H =2000 m results in scale factors of 13,333 and 6,666, respectively.
- The ground coverage increases. A 23 cm $\times$ 23 cm negative covers an area of 3066 m $\times$ 3066 m if $f = 150$ mm. The width of the coverage reduces to 1533 m if $f = 300$ mm. Subsequent processing takes less time if we can cover a large area with fewer photos.
- The angular field of view increases and the image perspective changes. The FOV for a wide-angle lens is 74 °; for a normal angle lens (300 mm) it is 41 °. Using shorter focal lengths has the advantage of giving more precise elevation measurements in stereo images (see Section 5.3). Flying a camera with a wide-angle lens at low altitudes has the disadvantage of producing larger obscured areas: if there are tall buildings near the edges of a photographed scene, the areas behind the buildings become hidden because of the central perspective; we call this the *dead ground effect*.

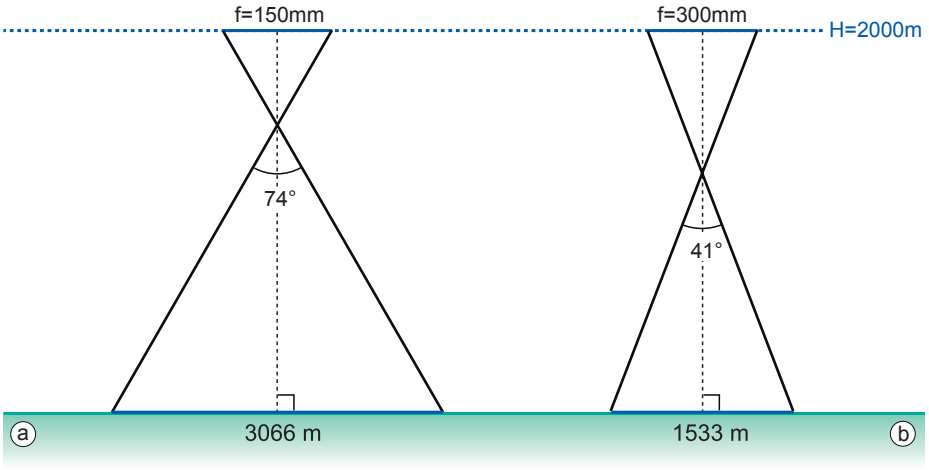


Figure 4.29
The effect of different focal lengths on ground coverage for the same flying height.

Spatial resolution

While scale is a generally understood and applied term, the use of *spatial resolution* in aerial photography is quite difficult. *Spatial resolution* refers to the ability to distinguish small adjacent objects in an image. The spatial resolution of B&W aerial photographs ranges from 40 to 800 line pairs per mm. The better the resolution of a recording system, the more easily the structure of objects on the ground can be viewed in the image. The spatial resolution of an aerial photograph depends on:

- the image scale factor—spatial resolution decreases as the scale factor increases;

- the quality of the optical system—expensive high-quality aerial lenses perform much better than the inexpensive lenses in amateur cameras;
- the grain structure of the photographic film—the larger the grains, the poorer the resolution;
- the contrast of the original objects—the higher the target contrast, the better the resolution,
- atmospheric scattering effects—this leads to loss of contrast and resolution;
- image motion—the relative motion between the camera and the ground causes blurring and loss of resolution.

From this list we can conclude that the actual value of resolution for an aerial photograph depends on quite a number of factors. The most variable factor is the atmospheric conditions, which can change from mission to mission and even during a mission.

4.6.3 Aerial photography missions

Mission planning When a mapping project requires aerial photographs, some of the first tasks to be done are to select the required photo scale factor, the type of lens to be used, the type of film to be used, and the required percentage of overlap (for stereo pairs). Forward overlap is usually around 60%, while sideways overlap is typically around 20%; Figure 4.30 shows a survey area covered by a number of flight lines. In addition, the date and time of acquisition should be considered with respect to growing season, light conditions and shadow effects.

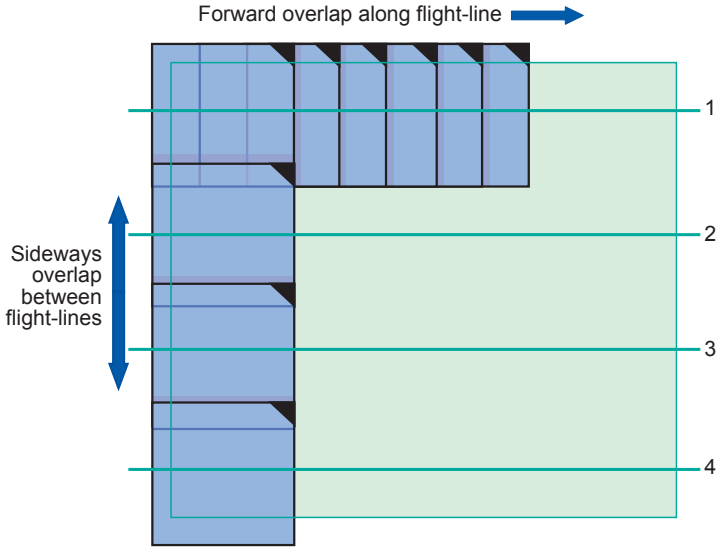


Figure 4.30
Arrangement of photos in a typical *aerial photo block*.

Once the required scale is defined, the following parameters can be determined:

- the required flying height,
- the ground coverage of a single photograph,
- the number of photos required along a flight line,

- the number of flight lines required.

After completion of the necessary calculations, either mission maps are prepared for use by the survey navigator, in the case of a conventional mission execution, or otherwise the data are fed into a mission guidance system.

Mission execution In current professional practice, we use a computer program to determine, after entering a number of relevant mission parameters and the area of interest, the (3D) coordinates of all positions from which photographs are to be taken. These are stored in a job database. On board, the camera operator/pilot can obtain all relevant information from that database, such as project area, camera and type of film to be used, the number of images required, and constraints regarding time of day or Sun angle, season, and atmospheric conditions.

With the camera positions loaded into a mission guidance system, the pilot is then guided—with the support of GPS—along the mission's flight lines such that deviation from the ideal line (horizontal and vertical) and time to the next exposure station is shown on a display (together with other relevant parameters). If the aircraft passes "close enough" to a predetermined exposure station, the camera fires automatically at the nearest position. This makes it possible to have the data of several projects on board, so that the pilot can choose a project (or part of a project) according to prevailing local weather conditions. If necessary, one can also abandon a project and resume it later.

In the absence of GPS guidance, the aircraft's navigator has to observe the terrain using the traditional viewing system of the aerial camera, check actual flight lines against the planned ones, which are shown graphically on topographic maps, give the required corrections (e.g. to the left or to the right) to the pilot, and tune the overlap regulator to the apparent forward speed of the airplane.

Satellite-based positioning systems and IMU provide a means for achieving accurate navigation. They offer precise positioning of the aircraft during a mission, ensuring that the photographs are taken at the correct points. Computer-controlled navigation and camera management is especially important in survey areas where topographic maps do not exist, are old, or are of small scale or poor quality. They are also helpful in areas where the terrain has few features (sand deserts, dense forests, etc.), because in these cases conventional visual navigation is particularly difficult. The major aerial camera manufacturers (as well as some independent suppliers) now offer complete software packages that enable the flight crew to plan, execute and evaluate an entire aerial survey mission.

4.7 Selection of sensors for a process study

4.7.1 Data selection criteria

For the selection of the appropriate data, it is necessary to fully understand the information requirements of a specific process study. In a nutshell, the questions to be answered concern coverage and resolution in space, time and spectrum. In addition, cost, availability or acquisition constraints, and quality will also be important. The surface characteristics of the object or objects under study determine which parts of the electromagnetic spectrum will be used for observation (spectral coverage), and whether a few broad bands are needed or many narrow bands (spectral resolution).

The level of detail determines the spatial resolution, whereas the size of the area, or the size of the area of the phenomenon, to be studied determines the spatial coverage, which corresponds to the area covered by one image. Of course, one can use several images to cover the area under study, which is often the case, but mosaicking images increases cost and processing time, and it often causes classification and interpretation problems at the seam between two images. Furthermore, the area of the Earth that can be observed by the sensor to be used is an important spatial aspect. For example, the geostationary MSG-SEVIRI covers only the Western Hemisphere, with very large distortions at the poles.

For the temporal aspect, we have to consider the speed of the process and the duration or the length of the period of observation. The speed of the process determines the frequency of observation within a given time (temporal resolution). When choosing the frequency and time of observation and the moments of observation, however, seasonality should be included in the considerations. For example, glaciers shrink in summer and expand in winter. If one wants to study long-term changes in glaciers, (trend versus cyclic changes), images should be recorded at comparable moments in the year, e.g. end-of-winter, at maximum size, end-of-summer, at minimum size, or images at several moments to get a more accurate estimate of seasonal fluctuations, to be able to separate them from long-term trends.

The temporal coverage needed depends on the duration of the process. For the past, temporal coverage is determined by the image archives of a sensor. Landsat archives date back to 1972, but aerial photographs may be available for many decades back. For the future, i.e. both current and planned satellite missions, continuity in the type of sensor are important. Landsat, NOAA, and Meteosat are examples of series of satellites that are each equipped with similar sensors, which guarantees the continuity of data. Security of continuity of data supply is a major issue for many institutes when deciding on which primary data sources to chose. The JERS-1, with its SAR sensor, has long been a typical example of a "one-off" research mission; JERS-1 operated between 1992 and 1998. It was finally followed up in 2006 with the PALSAR sensor on board ALOS.

The selection of data is further influenced by a number of acquisition constraints. Acquisition of optical data is hampered by clouds, so it is not always possible to acquire an image on a planned date, even if the satellite is in the appropriate orbit. Furthermore, not all sensors can record images continuously, because of power limitations, which means that the number of images recorded per orbit is limited. For stereo air photos, occurrence of optimal Sun elevation angles, resulting in enough shadow for the interpretation of height (but not so much that larger parts of the image are obscured), limits the number of days suitable for image acquisition.

Two quality aspects are especially important for process studies: radiometric quality, and calibration. Because of the shorter dwell time per pixel (ground resolution cell),

the radiometric quality of scanners (whiskbroom sensors) is usually less than that of comparable line cameras (pushbroom sensors). Over time, sensors change, so continuous calibration is needed to obtain unbiased observations of the process, and so that trends detected can be attributed to the phenomenon being studied rather than resulting from the aging of the sensor. Furthermore, similar sensors on different platforms in a constellation, or their successors on new platforms, need to be calibrated to make their measurements comparable.

Last but not least, cost plays a major role in image selection for process studies. Although the whole chain of images and processing should be included in cost calculations, in practice the focus tends to be on the cost of the images alone.

To illustrate all these aspects, let us have a look at some typical process studies and the type of data they frequently require. Studies of land processes on a regional or continental scale, for example drought or wildfires, typically use meteorological satellites such as the geostationary Meteosat Second Generation—SEVIRI; the polar orbiting NOAA-AVHRR; or the MODIS sensors of TERRA and AQUA satellites. Spatial resolutions range from 250 m to a few kilometres (depending on location), with the frequency of observations varying from twice a day (MODIS) to every 15 min.

Land use change Studies of land cover change, deforestation and urban expansion use sensors with spatial resolutions between 15 and 60 m, usually with a temporal coverage of more than a decade; observation once or twice a year is usually sufficient. Detailed change studies, focusing on changes within an urban environment or land cover changes in smaller but fragmented areas, use high resolution sensors. Since data from these sensors, with resolutions ranging from less than a metre up to a few metres, are only available for recent years, they are often combined with older aerial photographs.

Monitoring of glaciers Glaciers are dynamic objects: their spatial extent is continuously changing. Monitoring of glaciers can be performed daily, monthly, seasonally or yearly. For studies related to global climate change, one would most likely be interested in data of a longer temporal scale. For example, one could select images of the same season for several years in a row. When selecting images, one should keep weather conditions in mind: dense cloud cover or heavy snow covering the land surface will make delineation of the glacier impossible.

Monitoring of Urban growth The process of urban growth is related to change of land cover type, e.g. construction works, which often take quite some time. Given the typical time scales needed in urban growth studies, annual acquisition of images would be the most appropriate frequency. For studies on urban growth, it does not make sense to acquire images daily.

Chapter 5

Pre-processing

Wan Bakx
Ben Gorte
Wim Feringa
Karl Grabmaier
Lucas Janssen
Norman Kerle
Gabriel Parodi
Anupma Prakash
Ernst Schetselaar
Klaus Tempfli
Michael Weir

5.1 Visualization and radiometric operations

This section explains the processing of raw remote sensing data and the basic principles of visualization of data. The production of images on a computer monitor or paper print-out has always been a prime concern of RS. We use images for inspecting raw data and for performing various data rectification and restoration tasks. Once data are corrected, we convert them once more to images and use these for information extraction by visual interpretation or to support digital image analysis. Many RS applications make use of multispectral data; to visualize them we have to rely on the use of colour. Section 5.1.1 explains how we perceive colour, which can help us to understand how to produce optimal images from multispectral data and how to properly interpret them.

We try to build remote sensors SO that they faithfully image a scene, and we are increasingly successful in doing so. Consider as an example a vertical photograph (or nadir view) of high resolution from a space-borne sensor: it closely resembles a map of the scene and, if the scene was a city, urban planners would be able to readily recognize objects of interest. Taking a closer look, we know that RS images will be geometrically distorted as compared to a map. The degree and type of distortion depends on the type of sensor platform used. Geometrical correction of RS images will be treated in Section 5.3.

In Chapter 2 you have learned that remote sensors measure radiances. The results of

those measurements, however, are recorded as digital numbers, which have no direct physical meaning. The degree to which DNs directly correspond to radiances on the ground depends on many factors. Degradation with respect to what we would like to measure is caused, for example, by unfavourable scene illumination, atmospheric scattering and absorption, and detector-response characteristics. The need to perform radiometric correction in order to compensate for any or all types of degradation depends on the intended use of the data. Urban planners or topographic mappers do not need radiances of objects to be able to recognize them in images. Nevertheless, these images are likely to benefit from "haze correction" and contrast enhancement, to facilitate interpretation. Subsection 5.1.3 therefore briefly treats radiometric correction, only covering corrective measures that are of interest to a wider range of disciplines. (Image restoration and atmospheric correction are discussed further in Section 5.2.) A more detailed description of visualization for map production and spatial analysis is given in Chapter 10.

Elementary image processing techniques to improve the visual quality of an image—so that interpretation becomes easier—are introduced in Section 5.1.4. Image enhancement is not only useful for Earth observation: you may even find it handy for "touching up" your own digital photos.

5.1.1 Visualization

Perception of colour

The perception of colour takes place in the human eye and associated part of the brain. Colour perception concerns our ability to identify and distinguish colours, which in turn enables us to identify and distinguish entities in the real world. It is not completely known how human vision works, or what exactly happens in the eyes and brain before someone decides that an object is, for example, dark blue. Some theoretical models, supported by experimental results, are generally accepted, however. Colour perception theory is applied whenever colours are reproduced, for example in colour photography, TV broadcasting, printing and computer animation.

Tri-stimuli model

cones and rods

We experience light of different wavelengths as different colours. The retinas in our eyes have *cones* (light-sensitive receptors) that send signals to the brain when they are hit by photons that correspond to different wavelengths in the visible range of the electromagnetic spectrum. There are three different kinds of cones, responding predominantly to blue, green and red light (Figure 5.1). The signals sent to our brain by these cones give us sensations of colour. In addition to cones, we have *rods*, which sense brightness. The rods can operate with less light than the cones and do not contribute to colour vision. For this reason, objects appear less colourful in conditions of low light.

colour monitors

This knowledge of the three stimuli is important for displaying colour. Colour television screens and computer monitors are composed of a large number of small dots arranged in a regular pattern of groups of three: a red, a green and a blue dot. At a normal viewing distance from A TV screen, for example, we cannot distinguish the individual dots. We can individually trigger these dots and vary the amount of light emitted from each of them. All colours visible on such a screen are, therefore, created by mixing different amounts of red, green and blue. This mixing takes place in our brain. When we see a mixture of red (say, 700 nm) and green (530 nm) light, we get the same impression as when we see monochromatic yellow light (i.e. with a distinct wavelength of, say, 570 nm). In both cases, the cones are apparently stimulated in the same way. According to the tri-stimuli model, therefore, three different kinds of dots

are necessary and sufficient to recreate the sensation of all the colours of the rainbow.

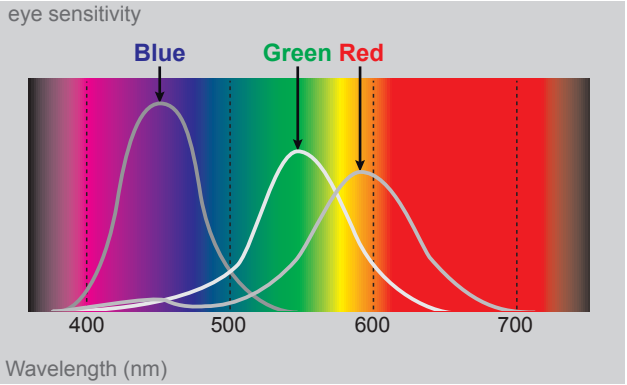


Figure 5.1
Visible range of the electromagnetic spectrum, including the sensitivity curves of cones in the human eye.

Colour spaces

The tri-stimuli model of colour perception is generally accepted. It states that there are three degrees of freedom in the description of a colour. Various three-dimensional spaces are used to describe and define colours. For our purposes, the following three are sufficient:

1. Red-Green-Blue (RGB) space, which is based on the additive mixing principle of colour;
2. Intensity-Hue-Saturation (IHS) space, which most closely resembles our intuitive perception of colour;
3. Yellow-Magenta-Cyan (YMC) space, which is based on the subtractive principle of colour.

RGB

The RGB definition of colour is directly related to the way in which computer and television screens function. Three channels directly related to the red, green and blue dots are the input to the screen. When we look at the result, our brain combines the stimuli from the red, green and blue dots and enables us to perceive all possible colours from the visible part of the spectrum. During the combination, the three colours are added. We see yellow when green dots are illuminated in addition to red ones. This principle is called the *additive colour scheme*. Figure 5.2 illustrates the additive colours caused by activating red, green and blue dots on a monitor. When only red and green light is emitted, the result is yellow. In the central area, there are equal amounts of light emitted from all three dots, so we experience *white*.

additive colour scheme

In the additive colour scheme, all visible colours can be expressed as combinations of red, green and blue, and can therefore be plotted in a three-dimensional space with R, G and B each being one of the axes. The space is bounded by minimum and maximum values for red, green and blue, thus defining what is known as the colour cube. Figure 5.3 shows the normalized colour cube; the maximum value for each colour is set to 1.

IHS

In day-to-day speech, we do not express colours using the RGB model. The IHS model more naturally reflects our perception of colour. *Intensity* in the colour space describes

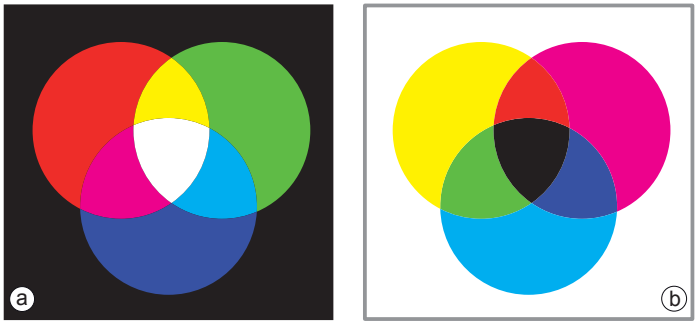


Figure 5.2
Comparison of the (a) additive and (b) subtractive colour schemes.

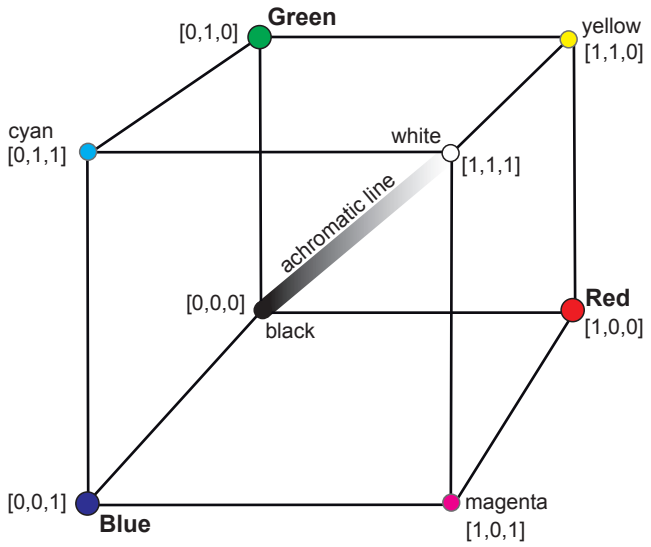


Figure 5.3
The RGB cube. Note the red, green and blue corner points.

intensity

hue

saturation

whether a colour is dark or light and we use for intensity the value range 0 to 1 (projection on the achromatic diagonal). *Hue* refers to the names that we give to colours: red, green, yellow, orange, purple, etc. We quantify hue by degrees in the range 0 to 360 around the achromatic line. *Saturation* describes a colour in terms of purity and we quantify it as the distance from the achromatic line. "Vivid" and "dull" are examples of common words in the English language that are used to describe colour of high and low saturation, respectively. A neutral grey has zero saturation. As is the case for the RGB system, again three values are sufficient to describe any colour.

Figure 5.4 illustrates the correspondence between the RGB and the IHS models. The IHS colour space cannot easily be transformed to the RGB space because they are completely different. The cube in Figure 5.3 must be converted to a double cone; the inset in Figure 5.4 illustrates this. Although the mathematical model for this description is tricky, the description itself is natural. For example, "light, pale red" is easier to imagine than "a lot of red with considerable amounts of green and blue". The result, however, is the same. Since the IHS model deals with colour perception, which is somewhat subjective, complete agreement of the definitions does not exist. Important for image fusion is the calculation of intensity values and, luckily, this is the simplest of all the calculations. Be aware that the values in the RGB model actually range from 0 to 255, while in the IHS model, intensity ranges from 0 to 1. The formula for intensity is:

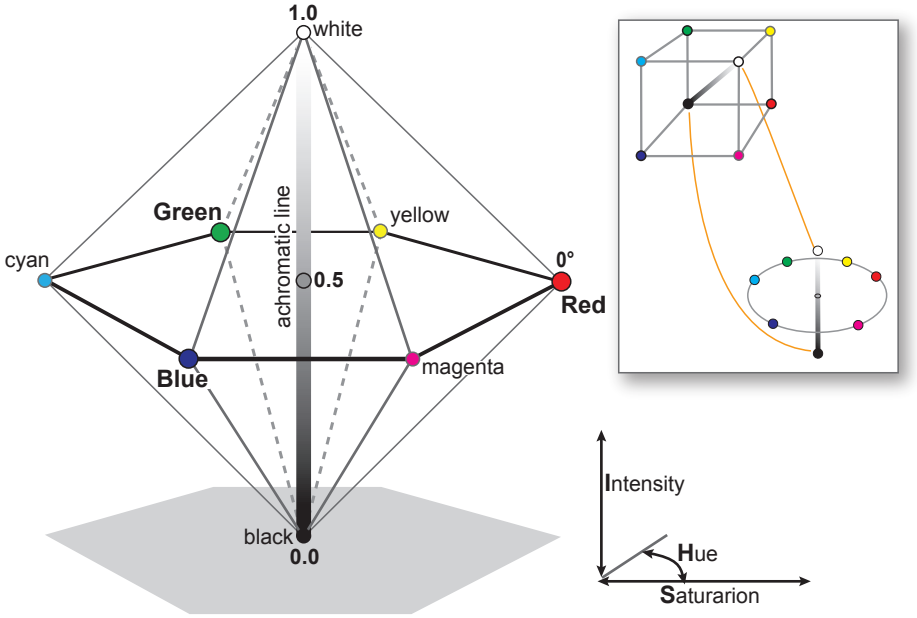


Figure 5.4
Relationship between RGB and IHS colour spaces.

$$I = \frac{R + G + B}{3 \cdot 255}, \tag{5.1}$$

an orthogonal projection on the achromatic line. For example: $(R, G, B) = (150, 200, 100)$. $I = ((150 + 200 + 100)/(3 \cdot 255)) = 0.59$.

YMC

Whereas RGB is used for computer and TV screens, the YMC colour model is used in colour definition for hardcopy media, such as printed pictures and photographic prints on paper. The principle of YMC colour definition is to consider each component as a coloured filter (Figure 5.2b). The filters are yellow, magenta and cyan. Each filter subtracts one primary colour from the white: the magenta filter subtracts green, so that only red and blue are left; the cyan filter subtracts red, and the yellow filter subtracts blue. Where the magenta filter overlaps the cyan filter, both green and red are subtracted and so we see blue. In the central area, all light is filtered so the result is black. Colour printing, which uses white paper and yellow, magenta and cyan ink, is based on the subtractive colour scheme. When sunlight falls on a colour-printed document, part of it is filtered out by the ink layers and the colour remaining is reflected from the underlying paper.

subtractive colour scheme

5.1.2 Image display

We normally display a digital image using a grey scale. A "digital image" can be raw data such as that obtained with a panchromatic camera, or data obtained by scanning a B&W photograph, or a single band of a multi-band image. For image display, standard computer monitors support 8 bits per pixel. Thus, if we have sensor recordings of 8 bits, each DN will correspond to exactly one grey value. A pixel having the value zero will be shown as black, a pixel having the value 255 as white. Any DN in between becomes, therefore, some shade of grey. One to one correspondence between DN and grey value used to be the standard, so we still often use "grey value" as a synonym for DN. A colour monitor has three input channels, so we have to feed each of them with

grey scale

Figure 5.5
Single-band and three-band image display using the red, green and blue input channels of the monitor.

the same DN to obtain a "grey scale image" (Figure 5.5).

pseudo-colour

An alternative way of displaying single-band data is to use a colour scale to obtain a *pseudo-colour* image. We can assign colours (ranging from blue via cyan, green and yellow to red) to different portions of the DN range 0–255 (Figure 5.5). The use of pseudo-colour is especially useful for displaying data that are not reflection measurements. With thermal infrared data, for example, the association of cold *versus* warm with blue *versus* red is more intuitive than with dark *versus* bright.

colour composites

When dealing with a multi-band image, any combination of three bands can, in principle, be used as input to the RGB channels of the monitor. The choice should be made based on the intended use of the image. Figure 5.5 indicates how we obtain a *false colour composite*.

true colour

Sometimes a *true colour composite* is made, where the RGB channels relate to the red, green and blue wavelength bands of a camera or multispectral scanner. An other popular choice is to link RGB to the near-infrared, red and green bands, respectively, to obtain a standard *false colour composite* (Figure 5.6). The most striking characteristic of false colour composites is that vegetation appears as a red-purple colour. In the visible part of the spectrum, plants reflect mostly green light, but their infrared reflection is even higher. Therefore, vegetation displays in a false colour composite as a combination of some blue and a lot of red, resulting in a reddish tint of purple.

false colour

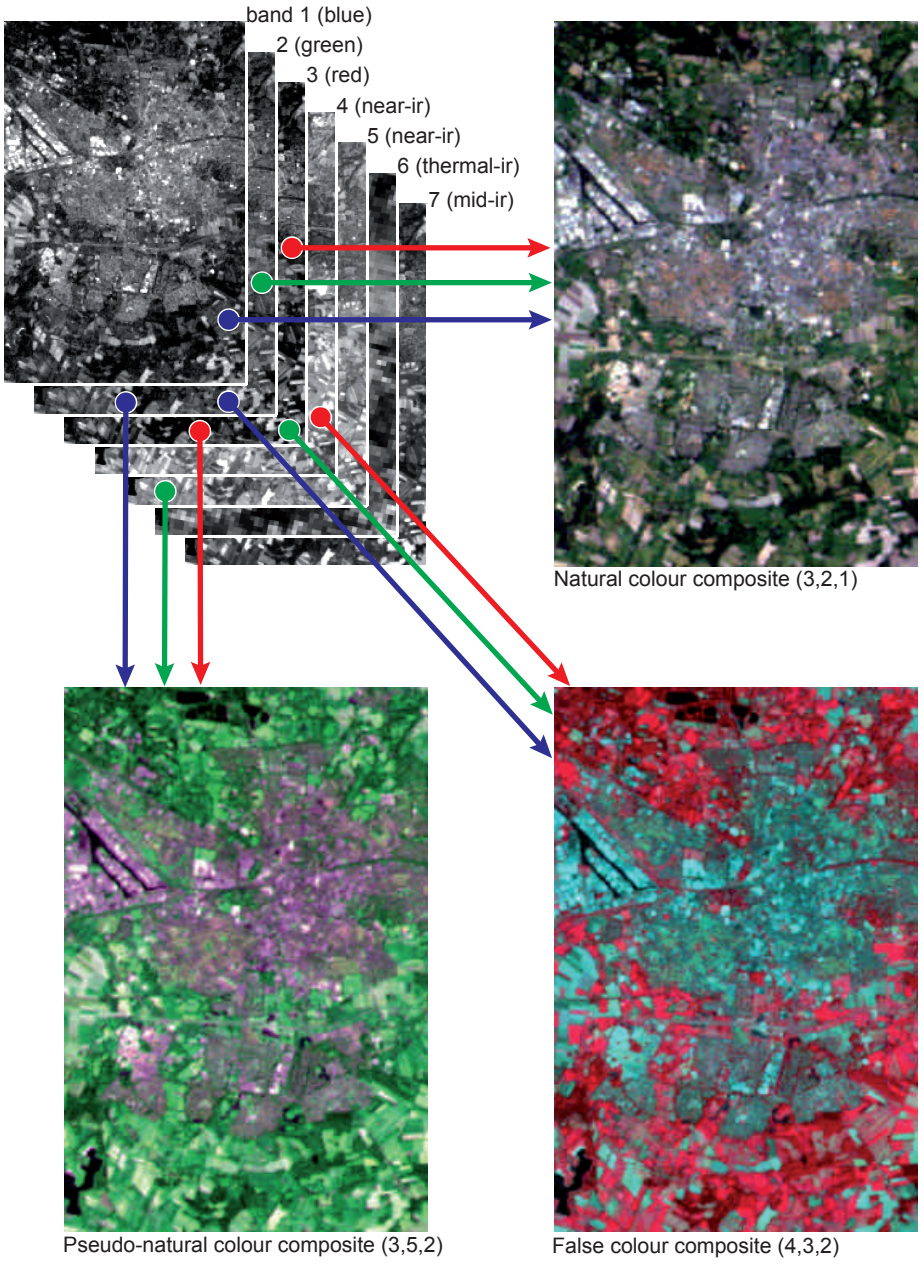


Figure 5.6
Landsat-5 TM false colour composite of Enschede and surroundings. Three different colour composites are shown: true colour, pseudo-natural colour and false colour composites.

Depending on the application, band combinations other than true or false colour may be used. Land use categories can often be distinguished quite well by assigning a combination of Landsat-5 TM bands 5–4–3 or 4–5–3 to RGB.

pseudo-natural colour

Combinations that display NIR as green show vegetation in a green colour and are, therefore, often called *pseudo-natural colour composites* (Figure 5.6). Note that there is no common consensus on the naming of certain composites ("true" may also be referred to as "natural"; "false colour" may also be used for other band combinations than green, red and NIR, etc.). Once you have become familiar with the additive mixing of the red, green and blue primaries, you can intuitively relate the colours—which you perceive on the computer monitor—to the digital numbers of the three input bands, thereby gaining a qualitative insight into the spectral properties of an imaged scene.

anaglyph stereograph

To obtain a 3D visual model from a stereo-image pair on a computer screen, we must combine the images into a stereograph (Figure 5.7) and then use a device that helps us to view the left image with the left eye and the right image with the right eye. There are various technical solutions for this problem, one of which is the *anaglyph* method. The left image is displayed in red, the right one in cyan and the two images are superimposed. For viewing, you need spectacles with a red glass for the left eye and a cyan glass for the right eye. High-end digital photogrammetric systems use polarization instead of colour coding. Polarized spectacles make the images visible to the appropriate eye. The advantage of using polarized images is that we can display a full-colour stereo model and superimpose the results of measurements in any colour. Yet another approach is to use a "split screen" display and a stereoscope in front of the monitor. A stereoscope is a device consisting of a pair of binoculars and two mirrors, which allows two images positioned next to each other to be viewed simultaneously, thus achieving stereoscopic vision. Stereoscopes can also be used to view paper prints of stereo photographs.

stereoscope

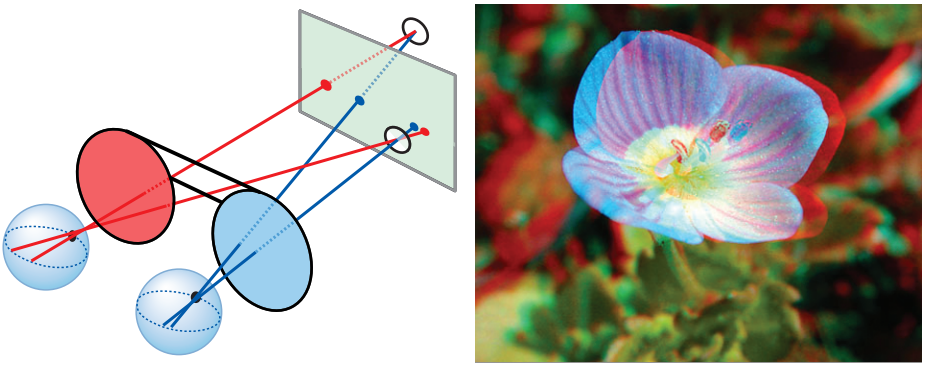


Figure 5.7
The Anaglyph principle and a stereograph.

5.1.3 Radiometric correction

Various techniques can be group under the heading radiometric correction, which aims to correct for various factors that cause degradation of raw RS data. Radiometrically correcting data should make them more suitable for information extraction. Techniques for modifying recorded DNs serve any of the three main purposes outlined below:

image enhancement

- Enhancing images so that they are better suited for visual interpretation. Image enhancement techniques are introduced in a separate section because they can be taken a step further, namely to "low-level image processing" for computer visualization.

- Correcting data for imperfections of the sensor. The detectors of a camera all have a slightly different response. We can determine the differences by radiometric calibration and, accordingly, apply radiometric correction later to the recordings of the camera. Scanners often use several detectors per channel instead of only one. Again, the detectors will each have (slightly) different radiometric responses, with the consequence that the resulting image may be striped. A *destriping* correction will normalize the detectors relatively, if calibration data is absent. A detector may also fail. We may then obtain an image in which, for example, every 10th line is black. A *line drop* correction will cosmetically fix the data. Another detector problem is random noise, which degrades radiometric information content and makes an RS image appear as if salt and pepper was sprinkled over the scene. Correcting all of these disturbances is fairly simple and explained in more detail in Subsection 5.1.8. There can be other degradations caused by the sensor-platform system that are not so easily corrected, such as compensating for image motion blur, which relies on a mathematically complex technique. We got used to referring to theses types of radiometric corrections as *image restoration*. Luckily, image restoration of new sensor data is usually done by the data providers, so you may only have to apply techniques such as destriping and dropped line correction when dealing with old data, e.g. from Landsat MSS. Image restoration should be applied before other corrections and enhancements.

  image restoration

- Correcting data for scene peculiarities and atmospheric disturbances. One scene peculiarity is how the scene is illuminated. Consider an area at an appreciable latitude, such as the Netherlands. The illumination of the area will be quite different in winter than in summer (overall brightness, shadows, etc.), because of differences in Sun elevation. Normalizing images taken in different seasons to make them comparable is briefly outlined below. An atmospheric degradation effect, which is already disturbing when extracting information from one RS image, is atmospheric scattering. Sky radiance at the detector causes haze in the image and reduces contrast. Haze correction is briefly described below. - Converting DNs to radiances on the ground (Section 5.2) becomes relevant if we want to compare RS data with ground measurements, or if we want to compare data acquired at different times by different sensors to detect change.

  scene normalization

  atmospheric correction

5.1.4 Elementary image enhancement

There are two approaches to elementary image processing to enhance an image: histogram operations and filtering. Histogram operations aim at global contrast enhancement, in order to increase the visual distinction between features, while filter operations aim at enhancing local contrast (edge enhancement) and suppressing unwanted image detail. Histogram operations look at DN values without considering where they occur in the image and assign new values from a look-up table based on image statistics. Filtering is a "local operation" in which the new value of a pixel is computed based on the values of the pixels in the local neighbourhood. Figure 5.8 shows the effect of contrast enhancement and edge enhancement for the same input image. Subsection 5.1.5 first explains the notion of histograms.

5.1.5 Histograms

The radiometric properties of a digital image are revealed by its *histogram*, which describes the distribution of the pixel values of the image. By changing the histogram, we change the visual quality of the image. Pixel values (DNs) for 8 bit data range from 0 to 255, so a histogram shows the number of pixels having each value in this

frequency distribution of DNs

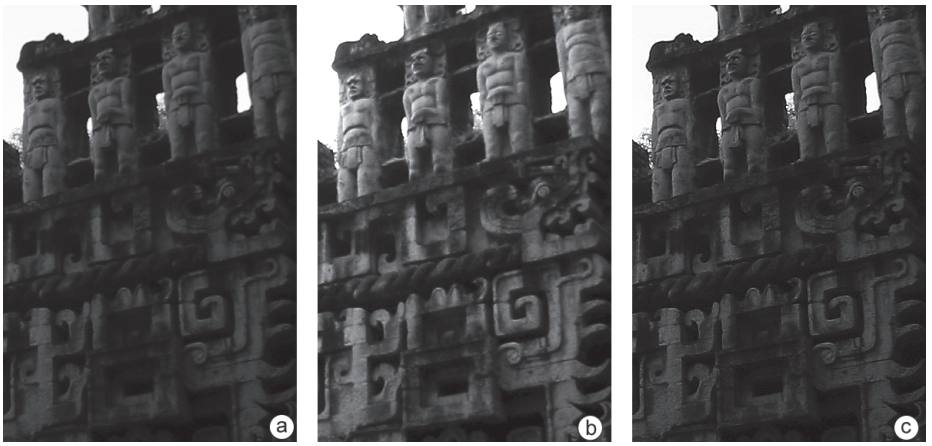


Figure 5.8
An (a) original, (b) contrast enhanced, and (c) edge enhanced image.

range, i.e. the frequency distribution of the DNs. Histogram data can be represented either in tabular form or graphically. The tabular representation (Table 5.1) usually shows five columns. From left to right these are:

- DN: Digital Numbers, in the range 0–255
- Npix: the number of pixels in the image with a particular DN (frequency)
- Perc: frequency as a percentage of the total number of image pixels
- CumNpix: cumulative number of pixels in the image with values less than or equal to a particular DN
- CumPerc: cumulative frequency as a percentage of the total number of image pixels

Figure 5.9 shows a plot of the columns 3 and 5 of Table 5.1 against column 1. More commonly, the histogram is displayed as a bar graph rather than as a line graph. The graphical representation of column 5 can readily be used to find the '1% value' and the '99% value'. The 1% value is the DN, below which only 1% of all the values are found. Similarly, there are only 1% of all the DNs of the image larger than the 99% value. The 1% and 99% values are often used in histogram operations as cut-off values for display, thus classifying very small and very large DNs as noise outliers rather than signals.

1% and 99% cut-off

A histogram can be "summarized" by descriptive statistics: mean, standard deviation, minimum and maximum, as well as the 1% and 99% values (see Table 5.2). The mean is the average of all the DNs of the image; note that it often does not coincide with the DN that appears most frequently (compare Table 5.1 and Table 5.2). The standard deviation indicates the spread of DNs around the mean.

A narrow histogram (thus, a small standard deviation) represents an image of low contrast, because all the DNs are very similar and are initially mapped to only a few grey values. Figure 5.11a shows the histogram of the image shown in Figure 5.8a. Notice the peak at the upper end (DN larger than 247), while most of DNs are smaller than 110. The peak for the white pixels stems from the sky. All other pixels are dark greyish; this narrow part of the histogram characterizes the poor contrast of the image (a Maya monument in Mexico). Remote sensors commonly use detectors with a

poor contrast

DN	Npix	Perc	CumNpix	CumPerc
0	0	0.00	0	0.00
13	0	0.00	0	0.00
14	1	0.00	1	0.00
15	3	0.00	4	0.01
16	2	0.00	6	0.01
51	55	0.08	627	0.86
52	59	0.08	686	0.94
53	94	0.13	780	1.07
54	138	0.19	918	1.26
102	1392	1.90	25118	34.36
103	1719	2.35	26837	36.71
104	1162	1.59	27999	38.30
105	1332	1.82	29331	40.12
106	1491	2.04	30822	42.16
107	1685	2.31	32507	44.47
108	1399	1.91	33906	46.38
109	1199	1.64	35105	48.02
110	1488	2.04	36593	50.06
111	1460	2.00	38053	52.06
163	720	0.98	71461	97.76
164	597	0.82	72058	98.57
165	416	0.57	72474	99.14
166	274	0.37	72748	99.52
173	3	0.00	73100	100.00
174	0	0.00	73100	100.00
255	0	0.00	73100	100.00

Table 5.1
Example of a histogram in a tabular format.

Mean	StdDev	Min	Max	1% value	99% value
113.79	27.84	14	173	53	165

Table 5.2
Statistics summarizing the histogram of the image represented in Table 5.1.

wide dynamic range, so that they can sense under a wide variety of different illumination and emission conditions. These differences, however, are not always present in one particular scene. In practice, therefore, we often obtain RS images that do not exploit the full range of DNs. A simple technique to achieve better visual quality is, then, to enhance the contrast by "grey scale transformation", which yields a histogram stretched over the entire grey range of pixels on the computer monitor (Figure 5.11c).

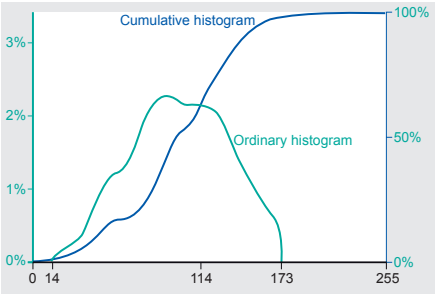


Figure 5.9
Standard histogram and cumulative histogram corresponding to Table 5.1.

5.1.6 Histogram operations

Although contrast enhancement may be done for different purposes, ultimately, in all cases, we will only do so to improve the visual interpretability of an image. In fact, there are two main purposes for applying contrast enhancement. The first is for "temporary enhancement", where we do not want to change the original data but only want to get a better image on the monitor so that we can carry out a specific interpretation task. Image display for geometric correction is an example of this. The second purpose is to generate new data that have a higher visual quality than the original data. This would be the case if an image is to be the output of our RS activities. Orthophoto production and image mapping are examples of such output (see Section 5.3).

Two techniques of contrast enhancement will now be described: *linear contrast stretch* and *histogram equalization* (occasionally also called histogram stretch). Both are "grey scale transformations", which convert input DNs (of raw data) to new brightness values (for a more appealing image) by a user-defined "transfer function".

linear contrast stretch

Linear contrast stretch is a simple grey scale transformation in which the lowest input DN of interest becomes zero and the highest DN of interest becomes 255 (Figure 5.10). The monitor will display zero as black and 255 as white. In practice, we often use the 1% and 99% values as the lowest and highest input DNs. The functional relationship between the input DNs and output pixel values is linear, as shown in Figure 5.11a. The functions shown in the first row (a, d, g) of Figure 5.11 (in the background of the histogram of each input image) are called the *transfer functions*. Many image processing software packages allow users to graphically manipulate the transfer function so that they can obtain an image that appears to their liking. The actual transformation is usually based on a look-up table.

transfer function

The transfer function to be used can be chosen in a number of ways. We can use linear grey-scale transformation to correct for haze and also for other purposes than contrast stretching, for instance to "invert an image" (convert a negative image to a positive one, or vice versa) or to produce a binary image (the simplest technique of image segmentation). Inverting an image is relevant, for example, after having scanned a negative of a photograph; see the last column of Figure 5.11.

Linear stretch is a straight-forward method of contrast enhancement that gives fair results when the input image has a narrow histogram that has a distribution close to

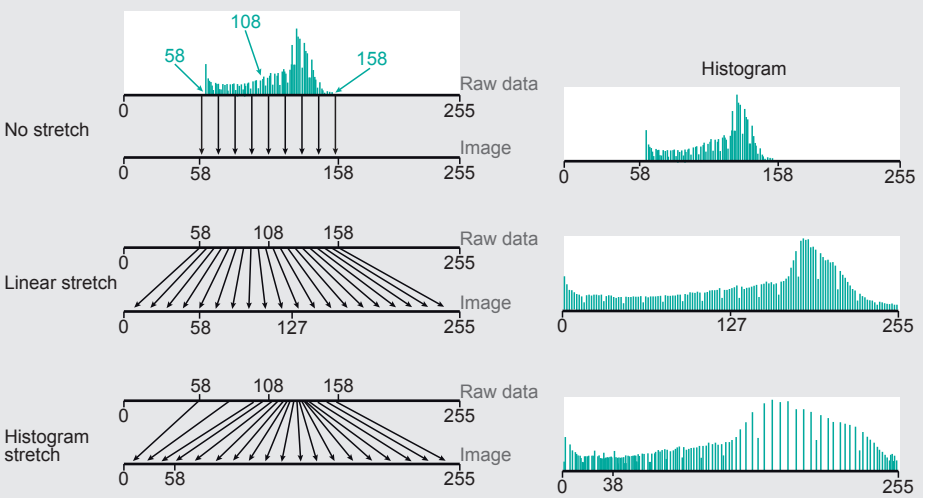


Figure 5.10
Linear contrast stretch versus histogram equalization.

being uniform. The histogram of our example image (Figure 5.8a) is asymmetric, with all DNs in a small range at the lower end, if the irrelevant sky pixels are not considered. Stretching this small range with high frequencies of occurrence (Figure 5.11d) at the expense of compressing the range with only few values (in our example, the range of high brightness) will make more detail (see Figure 5.11e) visible in the dark parts of the original.

histogram equalization

As the name suggests, *histogram equalization* aims at achieving a more uniform distribution in the histogram (see Figure 5.11f). Histogram equalization is a non-linear transformation; several image processing packages offer it as a default function. The idea can be generalized to achieve any desired target histogram.

It is important to note that contrast enhancement (by linear stretch or histogram equalization) merely amplifies small differences between DN values so that we can visually differentiate between features more easily. Contrast enhancement does not, however, increase the information content of the data and it does not consider a pixel neighbourhood. Histogram operations should be based on analysis of the shape and extent of the histogram of the raw data and understanding what is relevant for the intended interpretation. If this is not done, a decrease of information content could easily occur.

5.1.7 Filter operations

A further step in producing optimal images for interpretation is to apply filtering. Filtering is usually carried out for a single band. Filters—algorithms—can be used to enhance images by, for example, reducing noise ("smoothing an image") or sharpening a blurred image. Filter operations are also used to extract features from images, e.g. edges and lines, and to automatically recognize patterns and detect objects. There are two broad categories of filters: linear and non-linear filters.

Linear filters calculate the new value of a pixel as a linear combination of the given values of the pixel and those of neighbouring pixels. A simple example of the use of a linear smoothing filter is when the average of the pixel values in a 3×3 pixel neighbourhood is computed and that average is used as the new value of the central pixel in the neighbourhood (see Figure 5.12). We can conveniently define such a linear filter through a *kernel*. Figure 5.13a shows the kernel of the smoothing filter applied to the example of Figure 5.12. The kernel specifies the size of the neighbourhood that is considered (3×3, or 5×5, or 3×1, etc.) and the coefficients for the linear combination.

linear filter

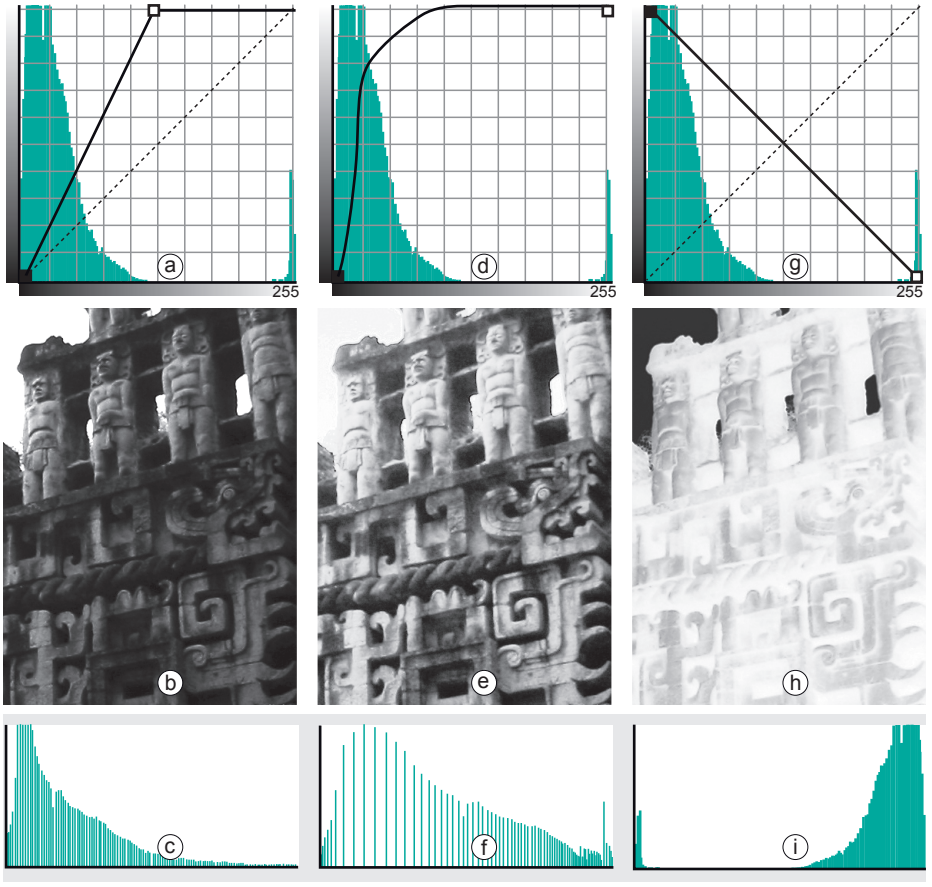


Figure 5.11
Effect of linear grey-scale transformations (b, c & h, i) and histogram equalization (e, f).

Image processing software allows us to select a kernel from a list or define our own kernel. The sum of the coefficients for a smoothing filter should be 1, otherwise an unwanted scaling of DN values will result. The filtering software will usually calculate the *gain*

$$gain = \frac{1}{\Sigma k_i} \tag{5.2}$$

and multiply the kernel values with it. The following two subsections give examples of kernels and their gain. Since there is only one way of using the kernel of a linear filter, the kernel completely defines the filter. The actual filter operation is to "move the kernel over the input image" line by line, thus calculating for each pixel a local combination of pixel values.

moving average

The significance of the gain factor is demonstrated in the following examples. Although in the examples only small neighbourhoods of 3×3 kernels are considered, in practice other kernel dimensions may be used.

Noise reduction

Consider the kernel shown in Figure 5.13a, in which all values equal 1. This means that the values of the nine pixels in the neighbourhood are summed. Subsequently, the result is divided by 9 to ensure that the overall pixel values in the output image

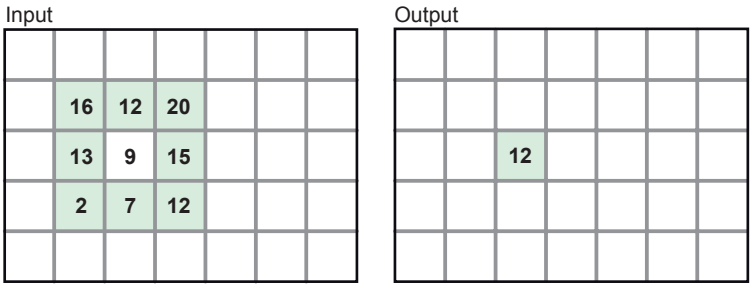


Figure 5.12
Input and output result of filtering: the neighbourhood in the original image and the filter kernel determine the value of the output. In this case, a smoothing filter was applied.

are in the same range as the input image. In this situation the gain is $1/9 = 0.11$. The effect of applying this *averaging filter* is that an image will become blurred or smoothed. This filter could be applied to radar images to reduce the effect of speckle.

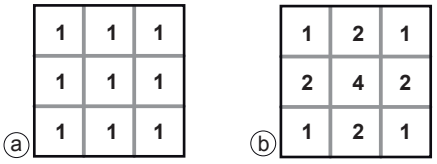


Figure 5.13
Filter kernels for smoothing: (a) equal weights, (b)distance-weighted smoothing.

In the kernel in Figure 5.13a, all pixels contribute equally to the result. It is also possible to define a distance-weighted average instead of an arithmetic mean: the larger the pixel's distance from the centre of the kernel, the smaller the weighting. As a result, less drastic blurring takes place. The resulting kernel, for which the gain is $1/16 = 0.0625$, is given in Figure 5.13b.

Edge detection

Filtering can also be used to detect the edges of objects in images. Such edges correspond to local differences in DN values. This is done using a *gradient* filter, which calculates the difference between neighbour pixels in some direction. Filters presented in Figures 5.14a and b, are called x- and $y-$gradient filters; they perform detection of vertical and horizontal edges, respectively. The filter shown in Figures 5.14c detects edges in all directions. Edge detection filtering produces small values in homogeneous areas of an image, while edges are represented by large positive or negative values. Edge detection filtering can be easily recognized by examining kernel elements: their sum must be zero. This applies to all filters shown in Figure 5.14.

gradient filter

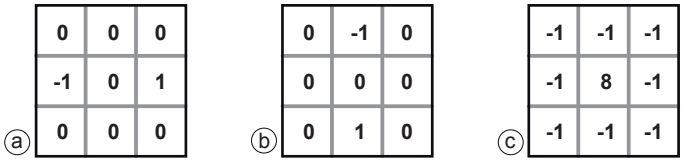


Figure 5.14
Filter kernels for edge detection: (a) x-gradient filter, (b) y-gradient filter, (c) all-directional filter.

Edge enhancement

Filtering can also be used to emphasize local differences in DN values by increasing contrast, for example for linear features such as roads, canals and geological faults. This is done using an *edge enhancing* filter, which calculates the difference between the central pixel and its neighbours. This is implemented using negative values for the non-central kernel elements. An example of an edge enhancement filter is given in Figure 5.15.

-1	-1	-1
-1	16	-1
-1	-1	-1

Figure 5.15
Filter kernel used for edge enhancement.

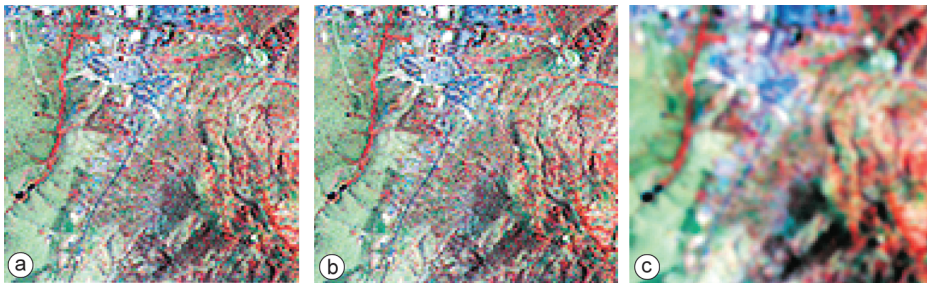


Figure 5.16
Original image (b), edge enhanced image (a) and smoothed image (c).

The gain is calculated as: $1/(16-8) = 1/8 = 0.125$. The sharpening effect can be made stronger by using smaller values for the centre pixel (with a minimum of 9). An example of the effect of using smoothing and edge enhancement is shown in Figure 5.16.

5.1.8 Correcting data for imperfections of the sensor

The objective of what is called here "cosmetics" is to correct visible errors and noise in the raw data. No atmospheric model of any kind is involved in these correction processes. Instead, corrections are achieved using especially designed filters and image stretching and enhancement procedures. These corrections are mostly executed (if required) at the station receiving the satellite data or at image pre-processing centres, i.e. before reaching the final user. All applications require this form of correction.

Typical problems requiring "cosmetic" corrections are:

- periodic line dropouts;
- line striping;
- random noise or spike.

These effects can be identified visually and automatically; Figure 5.17 illustrates this with Landsat-7 ETM image of Enschede.

Periodic line dropouts

Periodic line dropouts occur due to recording problems when one of the detectors of the sensor in question either gives wrong data or stops functioning. The Landsat-7 ETM, for example, has 16 detectors for each of its channels, except the thermal channel. A loss of one of the detectors would result in every sixteenth scan line being a string of zeros that would plot as a black line on the image (see Figure 5.18).

The first step in the restoration process is to calculate the average DN value per scan line for the entire scene. The average DN value for each scan line is then compared with this scene average. Any scan line deviating from the average by more than a designated threshold value is identified as defective. In regions of very diverse land cover, better results can be achieved by using the histogram for sub-scenes and processing these sub-scenes separately.

The next step is to replace the defective lines. For each pixel in a defective line, an average DN is calculated from the DNs for the corresponding pixel in the preceding and succeeding scan lines by using the principle of spatial autocorrelation. The average DN is then substituted for the defective pixel. The resulting image is a major improvement, although every sixteenth scan line (or every sixth scan line, in the case of Landsat MSS data) consists of artificial data (see Figure 5.19). This restoration program is equally effective for random line dropouts that do not follow a systematic pattern.

Line striping

Line striping is far more common than line dropouts. Line striping often occurs as a result of non-identical detector response. Although the detectors for all satellite sensors are carefully calibrated and matched before the launch of the satellite, with time the response of some detectors may drift to higher or lower levels. As a result, every scan line recorded by that detector is brighter or darker than the other lines (see Figure 5.20). It is important to understand that valid data are present in the defective lines, but these must be corrected to match the overall scene.

Though several procedures can be adopted to correct this effect, the most popular one is histogram matching. Separate histograms corresponding to each detector unit are constructed and matched. Taking one response as standard, the gain (rate of increase of DNs) and offset (relative shift of mean) for all other detector units are suitably adjusted, and new DNs are computed and assigned. This yields a destriped image in which all DN values conform to the reference level and scale.

Random noise or spike noise

Periodic line dropouts and striping are forms of non-random noise that may be recognized and restored by simple means. Random noise, on the other hand, requires a more sophisticated restoration method, such as digital filtering.

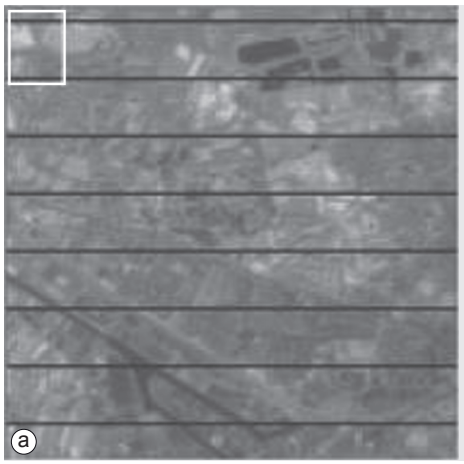
Random noise or spike noise may be caused by errors during transmission of data or a temporary disturbance. Here, individual pixels acquire DN values that are much higher or lower than the surrounding pixels (Figure 5.21). In the image, these pixels produce bright and dark spots that interfere with information extraction procedures.

A spike noise can be detected by mutually comparing neighbouring pixel values. If neighbouring pixel values differ by more than a specific threshold margin,it is desig-

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
114	97	93	86	96	100	92	88	88	92	93	92	105	110	106
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
91	91	85	87	80	97	115	105	94	87	78	70	73	65	85

b

Figure 5.17
Original Landsat ETM image of Enschede and and its surroundings (a), and corresponding DNs of the the indicated subset (b).



96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

b

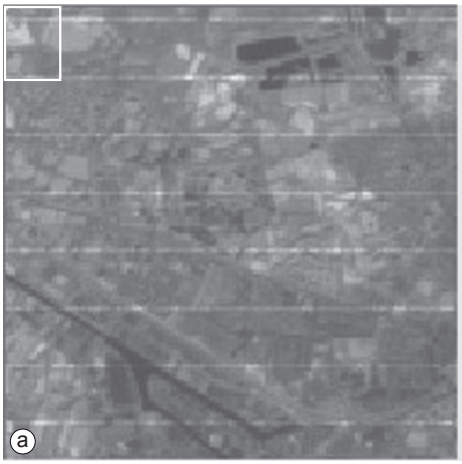
Figure 5.18
The image with periodic line dropouts (a) and the DNs (b). All erroneous DNs in these examples are shown in bold.

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
118	**126**	**100**	**97**	**110**	**111**	**87**	**84**	**81**	**87**	**91**	**84**	**100**	**116**	**124**
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
90	**93**	**93**	**90**	**102**	**95**	**110**	**101**	**87**	**81**	**77**	**74**	**74**	**81**	**91**

b

Figure 5.19
The image after correction for line dropouts (a) and the DNs (b).

nated as spike noise and the DN is replaced by an interpolated DN.



96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	83	91	75	75	110	131
150	**127**	**121**	**113**	**129**	**130**	**121**	**116**	**115**	**120**	**127**	**124**	**143**	**148**	**142**
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	164	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	80	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
125	**121**	**111**	**119**	**107**	**132**	**160**	**139**	**128**	**121**	**104**	**93**	**99**	**90**	**113**

Figure 5.20
The image with line striping (a) and the DNs (b). Note that the destriped image would look similar to the original image.

96	87	83	94	114	109	107	104	94	83	81	96	90	110	109
89	89	96	94	98	108	111	110	94	80	88	79	73	107	121
104	108	93	92	97	94	86	81	75	**255**	91	75	75	110	131
114	97	93	86	96	100	92	88	88	92	93	92	105	110	106
131	143	107	101	122	127	88	86	87	90	90	92	124	122	116
122	155	147	115	145	155	101	91	92	95	102	119	131	128	132
112	127	148	151	159	155	134	91	92	91	106	126	121	112	126
137	152	162	**0**	161	157	150	98	86	96	118	127	104	121	122
138	149	155	159	168	162	154	123	87	97	126	123	108	125	121
131	137	139	141	148	158	159	137	96	92	122	117	114	121	121
108	132	136	134	135	124	115	106	96	88	90	96	113	120	116
88	101	113	100	94	93	90	84	89	79	82	91	88	102	114
78	68	81	88	73	76	76	78	79	72	76	83	86	86	91
70	85	111	118	100	89	86	85	84	81	73	68	73	79	85
87	114	128	127	119	96	87	86	80	74	69	72	76	80	81
106	126	125	127	108	88	92	87	**255**	80	71	75	77	84	68
106	116	123	119	89	95	99	95	82	76	79	79	79	71	69
99	116	123	119	89	95	99	93	85	74	69	66	70	71	78
88	95	99	96	112	76	104	93	85	74	69	66	70	71	85
91	91	85	87	80	97	115	105	94	87	78	70	73	65	85

Figure 5.21
Image with spike errors (a) and the DNs (b).

5.2 Correction of atmospheric disturbance

Introduction

The radiance values of reflected polychromatic solar radiation and/or the emitted thermal and microwave radiances from a certain target area on the Earth's surface are for researchers the most valuable information obtainable from a remote sensor. In the absence of an atmosphere, the radiance for any wavelength on the ground would be the same as the radiance at the sensor. No atmosphere would make RS easier—but life impossible. So we have to figure out how we can convert remotely detected radiances to radiances at ground level.

In this section we will consider relative and absolute atmospheric correction. Relative atmospheric correction is based on ground reflectance properties, while absolute atmospheric correction is based on atmospheric process information. Before we explain how to correct, Subsection 5.2.1 will review the imaging process and the occurring disturbances.

5.2.1 From satellite to ground radiances: atmospheric correction

The presence of a heterogeneous, dense and layered terrestrial atmosphere composed of water vapour, aerosols and gases disturbs the signal reaching sensors in many ways. Therefore, methods of atmospheric corrections (AC) are needed to "clean" the images from these disturbances, in order to allow the retrieval of pure ground radiances from the target. The physics behind AC techniques in visible and thermal ranges is essentially the same, meaning that the same AC procedures that are applicable to one also apply to the other. However, there are a number of reasons for making a distinction between techniques applied to visible data and thermal data:

- Incident and reflected solar radiation and terrestrial thermal emissions belong to very different parts of the spectrum.
- Solar emission and reflection depends on the position of the Sun and the satellite at the time of image acquisition. Thermal emission is theoretically less dependent on this geometry.
- Solar rays travel twice through the atmosphere before they reach the sensor (Top of the Atmosphere (TOA)–ground–sensor), whereas ground thermal emissions only pass through the atmosphere once (ground–sensor; see Figure 2.6).
- Solar reflection at the Earth's surface depends on material reflectance (ρ). Thermal emission from the Earth depends on the emissivity of the surface materials (ϵ). Since solar reflection and Earth thermal emission occur at different wavelengths, the behaviour of one is not an indication of the other.
- The processes of atmospheric attenuation, i.e. scattering and absorption, are both wavelength dependent and affect the two sectors of the spectrum differently.
- As a result of the previous point, AC techniques are applied at a monochromatic level (individual wavelengths). This means that attenuation of radiation is calculated at every individual wavelength and then integrated across the spectrum of the sensor by mathematical integration.
- Atmospheric components affect different areas of the spectrum in different ways, meaning that some components can be neglected when dealing with data belonging to the thermal or the visible part of the spectrum.

A classification of different AC methods allows us to assess what kind of effort is needed to correct raw data for the particular application at hand. Some RS applications do not require AC procedures at all, except for some "cosmetics", while others call for rigourous and complex procedures. "Intermediate" solutions are sufficient for many applications.

In general, applications for which the actual radiance at ground level is not needed do not require atmospheric correction. Some "cosmetic" and/or image enhancement procedures may suffice: for instance, mapping applications where visual interpretation and image geometry are important, but not the chemical properties of surface material.

Applications that require the quantification of radiation at ground level must include rigourous atmospheric correction procedures. Quantification of evapotranspiration or CO_2 sequestration, or surface temperature and reflectivity mapping, are examples of such applications.

Applications concerned with the evolution of certain parameters or land properties over time, rather than their absolute quantification, are "intermediate" cases. For these, knowledge of the relative trend may suffice. Such procedures apply mainly when the mapping parameters do not really have a meaningful physical value, simply because they were designed primarily for multi-temporal relative comparison. Index evolution and correlation procedures, where radiances are associated with the evolution of a certain parameter (e.g. turbidity) are examples of this category. Be aware that some indexes such as NDVI typically require some absolute atmospheric correction.

The "effort" required is commensurate with the amount of information required to describe the components of the atmosphere at different altitudes (atmospheric profiling) at the moment and position at which the image is taken, and less so with sophistication of the AC procedure itself. State-of-the-art atmospheric models allow the "cleaning" of any cloudless image regardless of sensor type, as long as atmospheric profile data are available. Unfortunately, such detailed atmospheric information can only be obtained through atmospheric sounding procedures, which use a series of instruments to sample the atmosphere at fixed intervals while being transported vertically by a balloon, or sounding sensors on board a satellite. This kind of profiling is carried out daily (at fixed times) at some atmospheric centres, regardless of satellite overpass times. However, the atmosphere is dynamic. Atmospheric processes and composition change rapidly, mainly at low altitudes (water vapour and aerosols), meaning that soundings made somewhere close to the target and near the time of a satellite overpass might not be enough to ensure an adequate atmospheric description. As as rule of thumb regarding AC techniques, first consider the objectives of the project, then identify the appropriate AC procedure, and finally establish the effort, i.e. the required information to execute the chosen correction procedure.

5.2.2 Atmospheric corrections

Haze correction

Equation 2.2 shows that atmospheric scattering adds a "sky radiance". Haze correction aims at removing sky radiance effects from raw data, and doing so can be beneficial to many applications of space-borne RS. In Section 2.4 you have also learned that scattering depends on wavelength: Rayleigh scattering will hardly affect recordings in the red spectral band, while DNs in the blue band may become significantly larger. Reducing haze, therefore, must be done independently for each band of an RS image. How much to subtract from every DN from a particular band? We can find out if the scene is favourable and contains areas that should have zero reflectance (a spectral-band-specific black body). Deep clear water, for example, should yield pixel values of

subtraction per band

zero in the NIR band. If not, we attribute the minimum value found for "water pixels" to sky radiance and subtract this value from all DNs in this band. The alternative is less reliable, i.e. to look at the histogram (see Subsection 5.1.5) of the band and simply take the smallest DN found there as the haze correction constant.

Sun elevation correction

normalization by sine

Illumination differences will cause problems if we want to analyse sequences of images of a particular area that were taken on different dates (or images of the same date taken at different time), or if we would like to make mosaics of such images. We can apply a simple Sun elevation correction if the images stem from the same sensor. The trick is to normalize the images as if they were taken with the Sun at its zenith. We can achieve this normalization by dividing every pixel value of an image by the sine of the Sun elevation angle at the time of data acquisition. The Sun elevation angle is usually given in the meta-data file, which is supplied with an image. Obviously this is an approximate correction as it does not take into account the effect of elevation and height differences in the scene, nor atmospheric effects.

Relative AC methods based on ground reflectance

Relative AC methods avoid the evaluation of atmospheric components of any kind. They rely on the fact that, for one sensor channel, the relation between the radiances at TOA and at ground level follows a linear trend for the variety of Earth features present in the image. This linear relation is in fact an approximation of reality, but for practical purposes it is precise enough when there are other, more important sources of error. The AC methods are:

Two reflectance measurements: The output of this method is an absolute atmospherically corrected image, so it can be used on an individual basis for multi-temporal comparison or parameter evolution and also for flux quantification. "Absolute" means that the image output has physical units and that the calculated ground radiances are compatible with the actual atmospheric constituents. The application of this method requires the use of a portable radiometer able to measure in the same wavelength range as the image band to be corrected. If many bands are to be corrected, then the radiometer should have filters that allow measurement in all these individual bands separately.

Two reference surfaces: The output of this method is an image that matches a reflectance that is compatible with the atmosphere of a similar image taken on a previous date. No absolute values of radiances are obtained in any of the two images, only allowing comparative results. This method works on an individual band/channel basis and is valid for establishing a basis for a uniform comparison to study, for example, the evolution of non-flux related parameters such as indexes, or when certain convenient land properties can be derived directly or indirectly from the normalized radiance values in a band. The method relies on the existence of at least one dark and one bright invariant area. Normally, a sizable area should avoid mixed pixels (mixed land cover). As rule of thumb it should be a minimum of 2 or 3 times larger than the image spatial resolution. Reflective invariant areas are considered to retain their reflective properties over time. Deep reservoir lakes, sandy beaches or deserts, open quarries,large asphalted areas, and large salt deposits are examples of areas that are reflectively invariant. The supposition is that, for these pixels, the reflectance should always be the same since the reflective properties of the materials of which they are composed do not vary with time. If a difference in reflectance occurs for the reflective invariant area in the two date images, it can only be attributed to the different state of the atmosphere on those dates. The atmospheric composition is unknown in the two images, but its influence is measurable by analysing the change in radiance for the reflective invariant areas for the two dates.

Absolute AC methods based on atmospheric processes

These methods require a description of the components in the atmospheric profile. The output of these methods is an image that matches the reflectance of the ground pixels with a maximum estimated error of 10%, provided that atmospheric profiling is adequate enough. This image can be used for flux quantifications, parameter evolution assessments, etc., as mentioned above. The advantage of these methods is that ground reflectance can be evaluated for any atmospheric condition, altitude and relative geometry between the Sun and satellite. The disadvantage is that the atmospheric profiling required for these methods is rarely available. To address this inconvenience, various absolute AC methods have been developed that have different requirements in relation to the atmospheric profiling data—and differences in the accuracy of the results.

Radiative transfer models Radiative transfer models (RTMs) can be used for computing radiances for a wide variety of atmospheric and surface conditions. They require full descriptions of the atmospheric components at fixed altitudes throughout the atmosphere. RTMs are relatively easy to use when the complexity of the atmospheric input is simplified by using one standard atmosphere as input.

Because of the rapid dynamics of the atmosphere in terms of the temporal and spatial variation of its constituents, researchers have found the need to define some often-observed "common profiles" that correspond to average atmospheric conditions for different parts of the Earth. Compilation of these "fixed atmospheres" has been based on actual radio soundings carried out at different research sites, resulting in what are called "standard atmospheres", e.g. mid-latitude summer, mid-latitude winter, tropical, desert, arctic, US standard, and so on. Researchers use these well-defined standards to characterize typical on-site atmospherics. RTMs have these standards built into the system, allowing the influence of different constituents to be compared under strict simulations. For instance, the influence of water vapour in the thermal, or of aerosols and air molecules in the visible, part of the spectrum can be accurately predicted for different atmospheres, allowing sensitivity analyses for evaluating the importance of these constituents in attenuation processes at different wavelengths.

5.3 Geometric operations

Introduction

If you did not know so before, after reading Chapter 3 you will know that the Earth has a spherical shape and that several clever scientists have devised transformations to map the curved Earth's surface to a plane. Through a map projection (transformation) we can obtain an image of the Earth's surface that has convenient geometric properties. We can, for instance, measure angles on a map and use these for navigation in the real world, or for setting out a designed physical infrastructure. Or if, instead of a conformal projection such as UTM, we use an equivalent projection, we can determine the size of a parcel of land from the map—irrespective of where the parcel is on the map and at which elevation it is on the Earth. A remote sensor "images" the Earth's surface without knowledge of map projections, so we must not expect that remote sensing images have the same geometric properties as a map. Moreover, wherever a remote sensor detects, it merely records DNs. DNs do not come with a label that tell us where exactly on the Earth is the corresponding ground-resolution cell to be found (Figure 5.22).

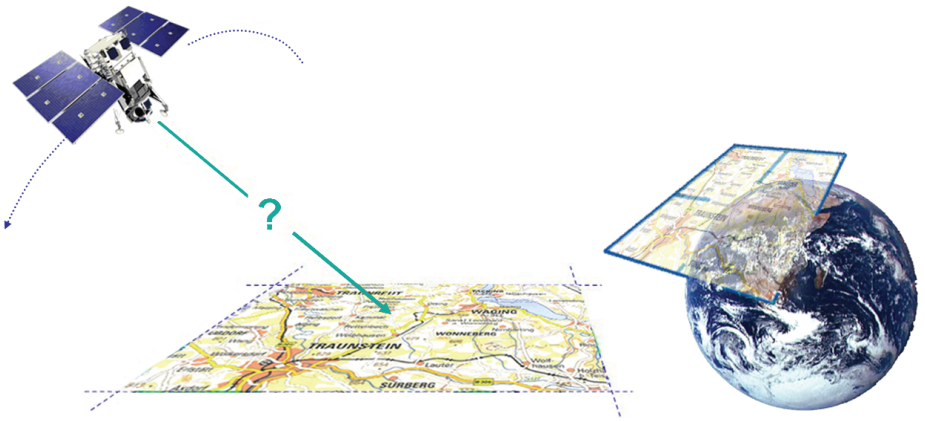


Figure 5.22
The problem of georeferencing an RS image.

Luckily, we know that the DNs are delivered in an orderly fashion, neatly arranged in rows and columns. The position of a point in the image is uniquely defined by the row and column numbers of the pixel that represents the point. Relating a pixel position in the image to the corresponding position on the ground is the purpose of *georeferencing the image*. Once we have figured out what the geometric relationship is between points on the ground and the corresponding point in the RS image, we can transform any pixel position to a position on the ground. "Position on the ground" can be defined either in a 3D terrain coordinate system or through a map projection in a 2D map coordinate system. By georeferencing an image we solve two problems at the same time: (1) we can get map coordinates of features that we identify in the image, and (2) we implicitly correct for geometric distortions of the RS image if we compute correct map coordinates for any pixel. It takes georeferencing to turn RS data into geospatial data. After georeferencing (or, speaking more generally, sensor orientation), we can:

- make measurements within images to obtain 2D and 3D object descriptions. Mapping the world in which we live has been man's concern for thousands of years, with navigation and military activities being the main triggers for topographic mapping. RS—photogrammetry, more specifically—has made that

mapping much more efficient. Mapping is still the prime application of RS, although environmental monitoring is catching up quickly because of the exponential damage we are causing to our environment. We are interested in mapping our environment at a variety of spatial and thematic resolutions and accuracies. For many applications, 2D representations (by points, lines and areas) of objects suffice. As long as certain conditions are met, we can obtain these representations from a single image and simple georeferencing, which directly relates the RS image to the digital map. We need stereo images or multiple images for applications requiring 3D coordinates, or for better 2D coordinates, such as for mapping scenes of large elevation differences or objects with large height differences. Sensor orientation must then use more rigourous approaches than for 2D georeferencing.

- combine an image with other images or vector (digital map) data (see also Chapter 11). Assume you would like to see how land property units relate to land cover. If you had a digital cadastral map, you could readily overlay the parcel boundaries on a RS image, e.g. a scanned photograph, which shows land cover nicely. Combining different RS images and/or map data can be done conveniently if all the data sets do not differ in their geometry, if they are all transformed to the same geometric reference system. From a computational perspective, producing a new image from an RS image such that it fits a specific map projection is often referred to as *geocoding*.

5.3.1 Elementary image distortions

Each sensor–platform combination is likely to have its own type of geometric image distortion. Here we only examine three very common types: (1) the effect of oblique viewing, (2) the effect of Earth rotation, and (3) "relief displacement". Tilting a camera (see Figure 4.27) leads to images of non-uniform scale (Figure 5.23). Objects in the foreground appear larger than those farther away from nadir. Earth rotation affects space-borne scanners and line camera images that have a large spatial coverage. The resulting geometric image distortion (Figure 5.23) can easily be corrected by 2D georeferencing. Relief displacement shows up specifically in large-scale camera images if there is significant terrain relief or if there are high, protruding objects.

oblique view
Earth rotation

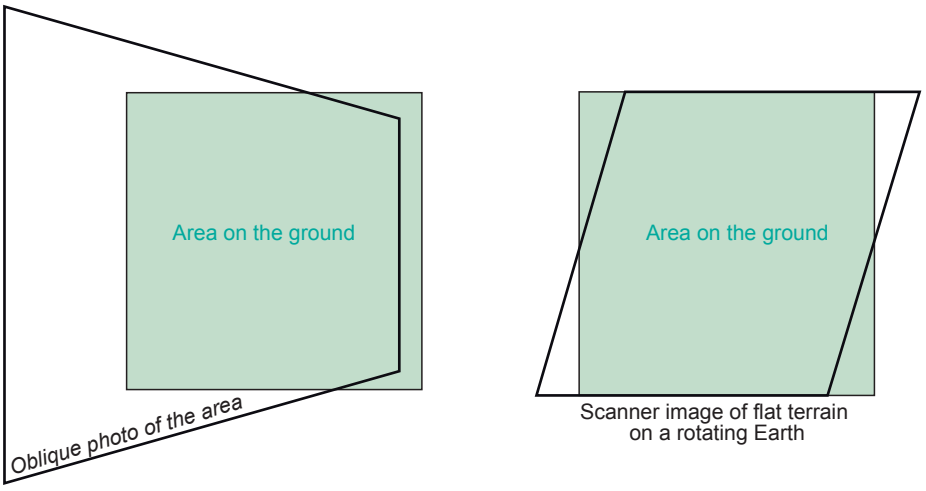


Figure 5.23
Examples of geometric image distortion.

Relief displacement

shifts due to height and elevation

A characteristic of most sensor systems is the occurrence of distortion of the geometric relationship between the image and a conventional map of the terrain due to elevation differences. This effect is most apparent in aerial photographs but also occurs in images from space-borne line cameras. The effect of relief displacement is illustrated in Figure 5.24 for a line camera. Consider the situation on the left (a), in which a true vertical image is taken of flat terrain. The distances $A - B$ and $a - b$ are proportional to the total width of the scene and its image size, respectively. In the situation on the left, by using the scale factor we can compute $A - B$ from a measurement of $a - b$ in the image. In the situation on the right (b), there is a significant difference in terrain elevation. As you can now observe, the distance between a and b in the image has become larger, although when measured in the terrain system, it is still the same as in the situation on the left. This phenomenon does not occur in the centre of a central projection image but becomes increasingly prominent towards the edges of a camera image. This effect is called *relief displacement*: terrain points whose elevation is above or below the reference elevation are displaced, respectively, away from or towards the nadir point, A, the point on the ground directly beneath the sensor. The magnitude of displacement, δr (in mm), is approximated by:

$$\delta r = \frac{rh}{H}. \tag{5.3}$$

In this equation, r is the radial distance (in mm) from nadir, h (in m) is the terrain elevation above the reference plane, and H (in m) is the flying height above the reference plane (where nadir intersects the terrain). The equation shows that the amount of relief displacement is zero at nadir (r =0) and largest at the edges of a line camera image and the corners of a frame camera image. Relief displacement is inversely proportional to the flying height.

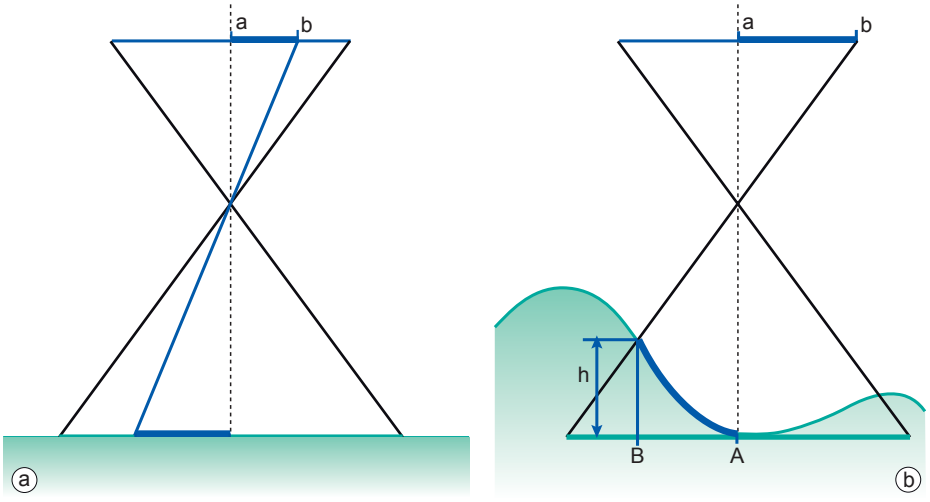


Figure 5.24
Illustration of the effect of terrain topography on the relationship between $A - B$ on the ground and $a - b$ in the image: (a) flat terrain, (b) significant elevation difference.

If the scene is just barren land, we cannot see any relief displacement. However, we can see relief displacement if there are protruding objects in the scene (occasionally referred to as height displacement). On large-scale photographs or very high-resolution space-borne images, buildings and trees appear to lean outwards, away from the nadir point (Figure 5.25).

The main effect of relief displacement is that inaccurate or wrong map coordinates will

Figure 5.25
Fragment of a large-scale aerial photograph of the centre of Enschede. Note the effect of height displacement on the higher buildings.

be obtained when, for example, digitizing images without further correction. Whether relief displacement should be considered in the geometric processing of RS data depends on its impact on the accuracy of the geometric information derived from the images. Relief displacement can be corrected for if information about terrain relief is available (in the form of a DTM); see Subsection 5.3.4 for more details. It is also useful to remember that it is relief displacement that allows us to perceive depth when looking at a stereograph and to extract 3D information from such images.

effect of relief displacement

Two-dimensional approaches

This subsection deals with the geometric processing of RS images in situations where relief displacement can be neglected, for example for a scanned aerial photograph of flat terrain. For practical purposes,"flat" may be considered as $h/H < 1/1000$, though this also depends on project accuracy requirements; h stands for relative terrain elevation, H for flying height. For space-borne images of medium spatial resolution, relief displacement is usually less than a few pixels in magnitude and thus less important, as long as near-vertical images are acquired. The objective of 2D georeferencing is to relate the image coordinate system to a specific map coordinate system (Figure 5.26).

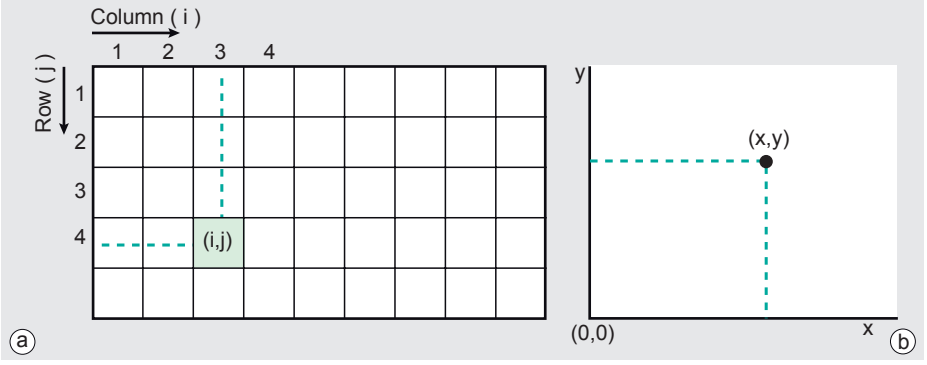


Figure 5.26
Coordinate system of (a) the image defined by rows and columns, and (b) the map with x- and y-axes.

5.3.2 Georeferencing

The simplest way to link image coordinates to map coordinates is to use a transformation formula. A *geometric transformation* is a function that relates the coordinates of two systems. A transformation relating (x, y) to (i, j) is commonly defined by linear equations, such as: $x = 3 + 5i$, and $y = -2 + 2.5j$.

transformation

Using the above transformation, for example, the image position $(i = 3, j = 4)$ corresponds to map coordinates $(x = 18, y = 8)$. Once the transformation parameters have been determined, the map coordinates for each pixel can be calculated. This implies that we can superimpose data that are given in the map coordinate system on the image vector, or that we can store features by map coordinates when applying on-screen digitizing. Note that the image in the case of georeferencing remains stored in the original (i, j) raster structure and that its geometry is not altered. As we will see in Subsection 5.3.3, transformations can also be used to change the actual shape of an image and thus make it geometrically equivalent to the map.

The process of georeferencing involves two steps: (1) selection of the appropriate type of transformation, and (2) determination of the transformation parameters. The type of transformation depends mainly on the sensor–platform system used. For aerial photographs (of flat terrain) what is known as "projective transformation" models well the effect of pitch and roll (see Figures 4.27, 5.23, and 5.28). Polynomial transformation, which enables 1st, 2nd to nth order transformations, is a more general type of transformation. In many situations a 1st order transformation is adequate. Such transformation relates map coordinates (x, y) with image coordinates (i, j) as follows:

type of transformation

$$x = a + bi + cj \tag{5.4}$$

$$y = d + ei + fj \tag{5.5}$$

Equations 5.4 and 5.5 require that six parameters (a to f) be determined. The transformation parameters can be determined by means of *ground control points* (GCPs). GCPs are points that can be clearly identified in the image and on the target map. The target map could be a topographic map or another image that has been transformed beforehand to the desired map projection system. The operator then needs to identify corresponding points on both images. The image and map scale determine which points are suitable. Typical examples of suitable points are road crossings, crossings of waterways and salient morphological structures. Another possibility is to identify points in the image and to measure the coordinates of these points in the field, for example by GPS, and then transform those to map coordinates. It is important to note that it can be quite difficult to identify good GCPs in an image, especially in lower-resolution space-borne images. Once a sufficient number of GCPs have been specified, software is used to determine the parameters a to f of the Equations 5.4 and 5.5 and quality indications.

ground control points

To solve the 1st order polynomial equations, only three GCPs are required; nevertheless, you should use more points than the strict minimum. Using merely the minimum number of points for solving the system of equations would obviously lead to a wrong transformation if you made an error in one of the measurements, whereas including more points for calculating the transformation parameters enables software to also compute the error of the transformation. Table 5.3 gives an example of the input and output of a georeferencing computation in which five GCPs have been used. Each GCP is listed with its image coordinates (i, j) and its map coordinates (x, y).

number of GCPs

Software performs a "least-squares adjustment" to determine the transformation pa-

GCP	i	j	x	y	x_c	y_c	d_x	d_y
1	254	68	958	155	958.552	154.935	0.552	-0.065
2	149	22	936	151	934.576	150.401	-1.424	-0.599
3	40	132	916	176	917.732	177.087	1.732	1.087
4	26	269	923	206	921.835	204.966	-1.165	-1.034
5	193	228	954	189	954.146	189.459	0.146	0.459

Table 5.3
A set of five ground control points, which are used to determine a 1st order transformation. x_c and y_c are calculated using the transformation, d_x and d_y are the residual errors.

rameters. The least squares adjustment ensures an overall best fit of the GCPs. We then use the computed parameter values to calculate coordinates (x_c, y_c) for any image point (pixel) of interest:

$$x_c = 902.76 + 0.206i + 0.051j,$$

and

$$y_c = 152.579 - 0.044i + 0.199j.$$

For example, for the pixel corresponding to GCP 1 ($i = 254$, $j = 68$) we can calculate the transformed image coordinates x_c and y_c as 958.552 and 154.935, respectively. These values deviate slightly from the input map coordinates (as measured on the map). Discrepancies between measured and transformed coordinates of GCPs are called residual errors (*residuals* for short). The residuals are listed in the table as d_x and d_y. Their magnitude is an indicator of the quality of the transformation. Residual errors can be used to analyse whether all GCPs have been correctly determined.

residuals

The overall accuracy of a transformation is either stated in the accuracy reportusually provided by software in terms of variances or as *Root Mean Square Error* (RMSE), which calculates a mean value from the residuals (at check points). The RMSE in the x-direction, m_x, is calculated using the following equation:

$$m_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \delta x_i{}^2}. \tag{5.6}$$

For the y-direction, a similar equation can be used to calculate m_y. The overall error, m_p, is calculated by:

$$m_p = \sqrt{m_x{}^2 + m_y{}^2}. \tag{5.7}$$

For the example data set given in Table 5.3, the residuals m_x, m_y and m_p are 1.159, 0.752 and 1.381, respectively. The RMSE is a convenient measure of overall accuracy, but it does not tell us which parts of the image are accurately transformed and which parts are not. Note also that the RMSE is only valid for the area that is bounded by the GCPs. In the selection of GCPs, therefore, points should be well distributed and include locations near the edges of the image.

5.3.3 Geocoding

The previous subsection explained that two-dimensional coordinate systems, e.g. an image system and a map system, can be related using geometric transformations. This georeferencing approach is useful in many situations. However, some situations a *geocoding* approach, in which the row–column structure of the image is also transformed, is required. Geocoding is required when different images need to be combined or when the images are used in a GIS environment that requires all data to be

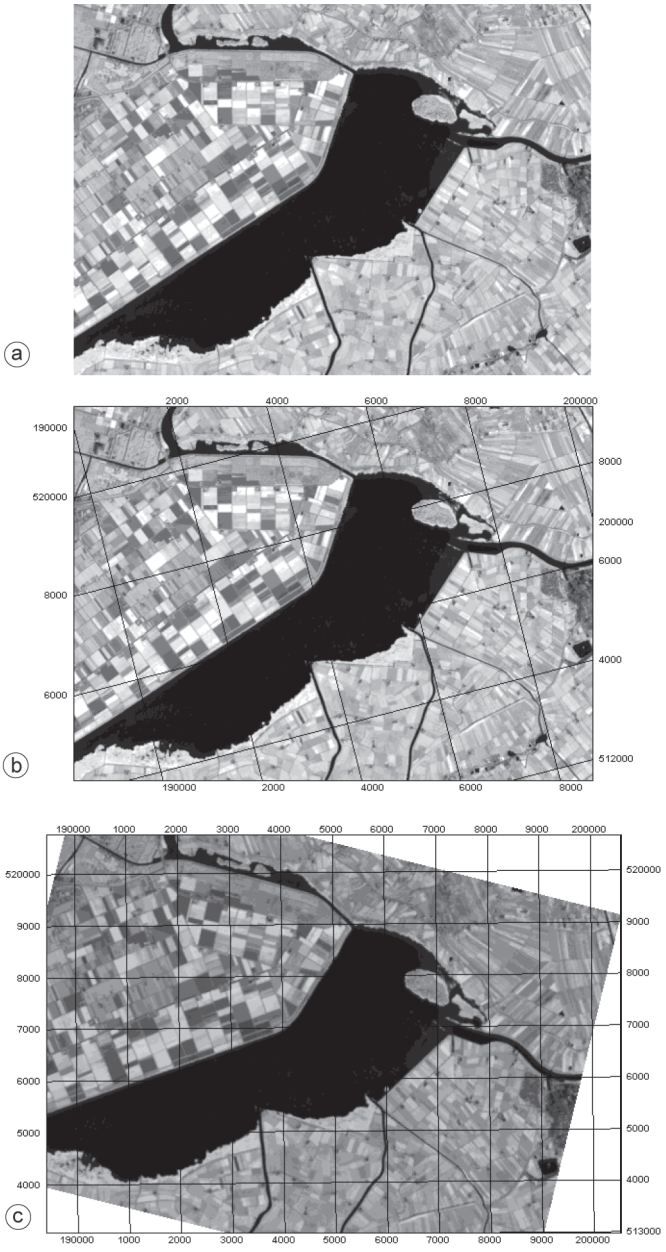


Figure 5.27
Original (top), georeferenced (middle) and geocoded (bottom) image of a part of Flevoland.

stored in the same map projection. The effect of georeferencing and geocoding is illustrated by Figure 5.27. Distinguishing georeferencing and geocoding is conceptually useful, but to the casual software user it is often not apparent whether only georeferencing has been applied in certain image manipulations or also geocoding.

Geocoding is georeferencing with subsequent *resampling* of the image raster. This means that a new image raster is defined along the $x-$ and y-axes of the selected map projection. The geocoding process comprises three main steps: (1) selection of a new grid spacing; (2) projection (using the transformation parameters) of each new raster element onto the original image; and (3) determination and storage of a DN for the

new pixel.

Figure 5.28 shows four transformation types that are frequently used in RS. The types shown increase from left to right in complexity and number of parameters required. In a conformal transformation, the image shape, including right angles, are retained. Therefore, only four parameters are needed to describe a shift along the x- and y-axes, a scale change, and the rotation. However, if you want to geocode an image to make it fit with another image or map, a higher-order transformation may be required, such as a projective or polynomial transformation.

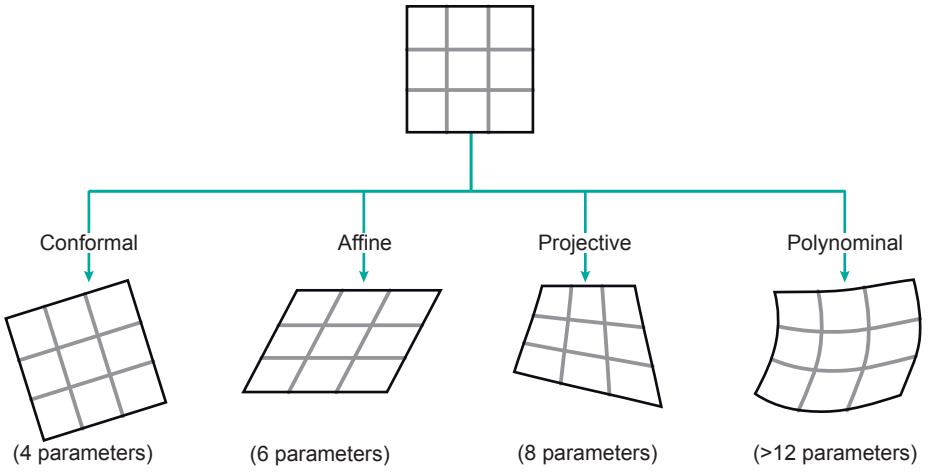


Figure 5.28
Illustration of different image transformation types and the number of required parameters.

Most of the projected raster elements of the new image will not match precisely with raster elements in the original image, as Figure 5.29 illustrates. Since raster data are stored in a regular row–column pattern, we need to calculate DNs for the pixel pattern of the corrected image intended. This calculation is performed by interpolation. This is called *resampling* of the original image.

resampling

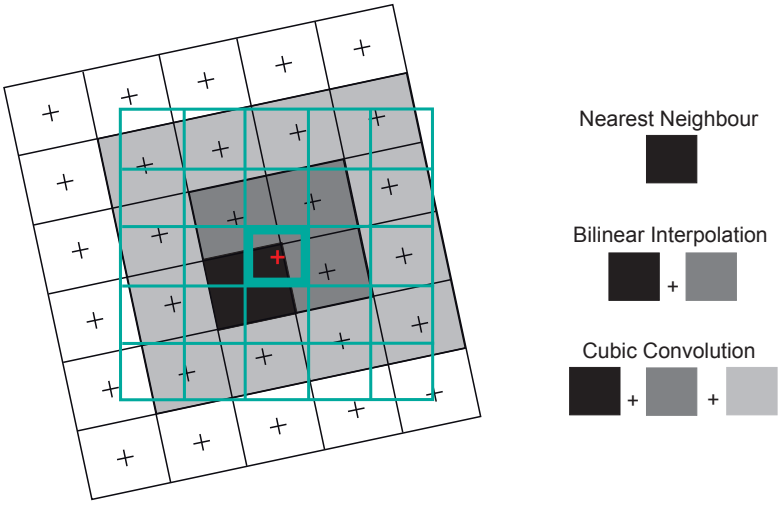


Figure 5.29
Principle of resampling using nearest neighbour, bilinear, and bicubic interpolation.

For resampling, we usually use very simple interpolation methods, the main ones being nearest neighbour, bilinear, and bicubic interpolation (Figure 5.29). Consider the green grid to be the output image to be created. To determine the value of the cen-

interpolation

tre pixel (bold), in *nearest neighbour interpolation* the value of the nearest original pixel is assigned, i.e. the value of the black pixel in this example. Note that the respective pixel centres, marked by small crosses, are always used for this process. In *bilinear interpolation*, a linear weighted average is calculated for the four nearest pixels in the original image (dark grey and black pixels). In *bicubic interpolation* a cubic weighted average of the values of 16 surrounding pixels (the black and all grey pixels) is calculated. Note that some software uses the terms "bilinear convolution" and "cubic convolution" instead of the terms introduced above.

choice of method

The choice of the resampling algorithm depends, among other things, on the ratio between input and output pixel size and the intended use of the resampled image. Nearest neighbour resampling can lead to the edges of features being offset in a step-like pattern. However, since the value of the original cell is assigned to a new cell without being changed, all spectral information is retained, which means that the resampled image is still useful in applications such as digital image classification (see Section 6.2). The spatial information, on the other hand, may be altered in this process, since some original pixels may be omitted from the output image, or appear twice. Bilinear and bicubic interpolation reduce this effect and lead to smoother images. However, because the values of a number of pixels are averaged, radiometric information is changed (Figure 5.30).

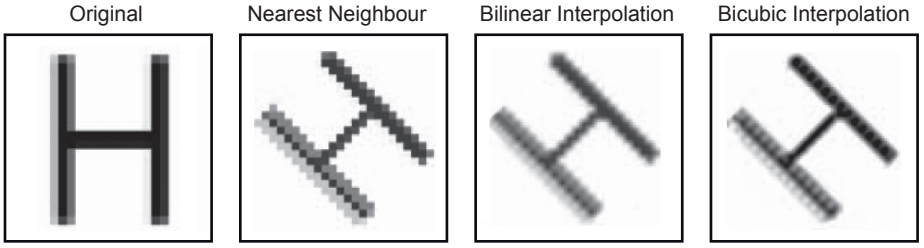


Figure 5.30
The effect of nearest neighbour and bilinear and bicubic resampling of the original data.

5.3.4 Three-dimensional approaches

In mapping terrain, we have to consider its vertical extent (elevation and height) in two types of situations:

- We want 2D geospatial data that describes the horizontal position of terrain features, but the terrain under consideration has large elevation differences. Elevation differences in the scene cause relief displacement in the image. Digitizing in the image without taking into account relief displacement causes errors in computed map coordinates. If the positional errors are larger than would be tolerated by the application (or map specifications), we should not use simple georeferencing and geocoding.
- We want 3D data.

When wishing to map terrain with an increasing degree of refinement, we have to first clarify what we mean by *terrain*. *Terrain* as described by a topographic map has two very different aspects: (1) there are agricultural fields and roads, forests and waterways, buildings and swamps, barren land and lakes; and (2) there is elevation changing with position—at least in most regions on Earth. We refer to land cover, topographic objects, etc. as *terrain features*; we show them on a (2D) map as areas, lines and point symbols. We refer to the shape of the ground surface as *terrain relief*, which we show on a topographic map by contour lines and/or relief shading. A *contour*

terrain relief

line on a topographic map is a line of constant elevation. Given contour lines, we can determine the elevation at any arbitrary point by interpolating between contour lines.

We could digitize the contour lines of a topographic map. The outcome would be a data set consisting of (X, Y, Z) coordinates of many points. Such a data set is the basis of a *digital terrain relief model* (DTM). We then need a computer program to utilize such a list of coordinate of points of the ground surface, to compute elevation at any horizontal position of interest and derive other terrain relief information such as slope and aspect. The idea of a DTM was developed in the late 1950s at MIT for computerizing highway design. Fifty years later, we use DTMs in all geosciences for a wide range of applications and we have remote sensors specifically built to supply us with data for DTMs (SRTM, SPOT-5 HRS, etc.). One of the applications of a DTM is to accomplish *digital monoplotting* and *orthoimage* production, as outlined later in this subsection.

A *DTM* is a digital representation of terrain relief, i.e. a model of the shape of the ground. We have a variety of sensors at our disposal that can provide us with 3D data: line cameras, frame cameras, laser scanners and microwave radar instruments. They can all produce (X, Y, Z) coordinates of terrain points, but not all the terrain points will be points on the ground surface. Consider a stereo pair of photographs, or a stereo pair from SPOT-5, of a tropical rainforest. Will you be able to see the ground? Coordinates obtained will pertain to points of the terrain relief. Since a model based on such data is not a DTM, we refer to it as digital surface model (DSM). The difference between a DTM and a DSM is illustrated by Figure 5.31; we need to filter DSM data to obtain DTM data.

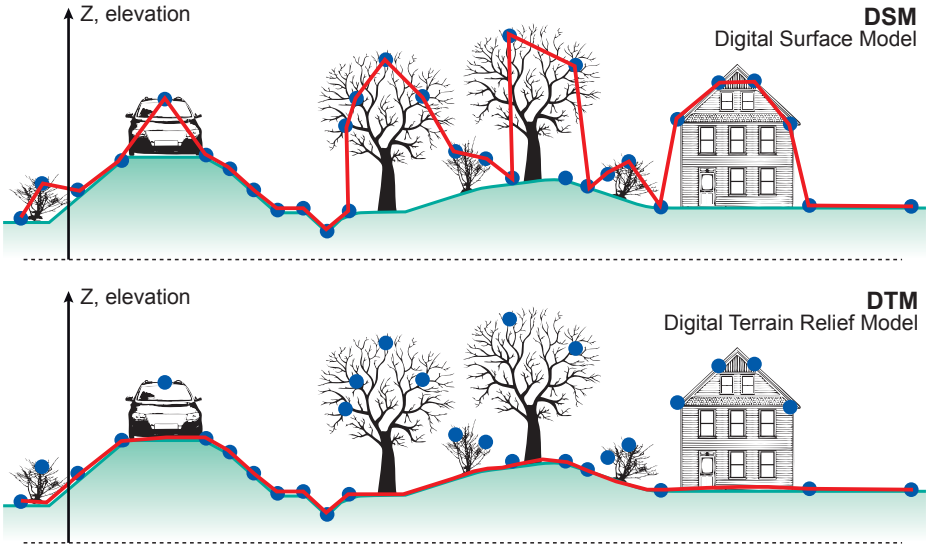


Figure 5.31
The difference between a DTM and a DSM.

In terrain modelling, it is handy to choose the coordinate system such that Z is the variable for elevation. If we model a surface digitally by nothing else than elevation values Z at horizontal positions (X, Y), why not call such a model a *digital elevation model* (DEM)? The term DEM was introduced in the 1970s with the purpose of distinguishing the simplest form of terrain relief modelling from more complex types of digital surface representation. Originally the term DEM was exclusively used for raster representations (thus elevation values given at the intersection nodes of a regular grid). Note that both a DTM and DSM can be a DEM and, moreover, "elevation" would not have to relate to terrain but could relate to some subsurface layer such as groundwater layers, soil layers or the ocean floor. Unfortunately, you will find in the

literature variations in the use of the terms introduced above. This is particularly so for DEM, which is often used carelessly. In this context, it is also worth mentioning the misuse of "topography" as synonym for terrain relief.

5.3.5 Orientation

The purpose of *camera orientation* is to obtain the parameter values for transforming terrain coordinates (X, Y, Z) to image coordinates, and vice versa. In solving the orientation problem we assume that any terrain point and its image lie on a straight line that passes through the projection centre (i.e. the lens). This assumption about the imaging geometry of a camera is called the *collinearity* relationship and it does not take into consideration that atmospheric refraction has the effect of slightly bending light rays. We solve the problem of orienting a single image with respect to the terrain in two steps: *interior orientation* and *exterior orientation*.

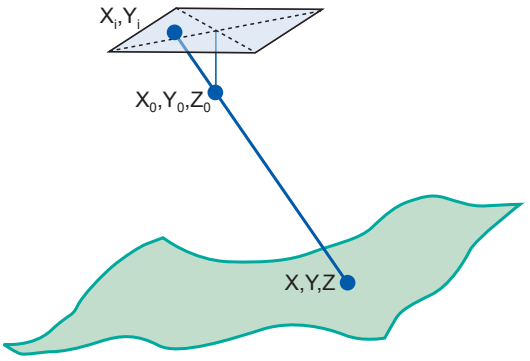


Figure 5.32
Illustration of the collinearity concept, where image point, lens centre and terrain point all lie on one straight line.

Orienting a single image as obtained by a line or frame camera

interior orientation

Interior orientation determines the position of the projection centre with respect to the image. The problem to solve is different for digital cameras and film cameras. In the case of a digital camera (line or frame camera, space-borne or aerial camera), the position of the projection centre with respect to the CCD array does not change from image to image, unless there are extreme temperature or pressure changes. For standard applications, we can assume that the position of the projection centre does not change when defined in the row–column system of the digital image. Two parameters are needed, the *principal distance* and the *principal point* (Figure 5.33); they are both determined by camera calibration. The *principal distance* is the mathematical abstraction of the focal length (which is a physical property of a lens). The *principal point* is the point of intersection of the perpendicular from the projection centre point with the image plane.

principal distance

principal point

The camera calibration report states the row (and column) number of the principal point and the principal distance. When using digital photogrammetric software for working with images from digital aerial cameras, the user only has to enter the principal distance, the position of the principal point and the pixel size to define the interior orientation. In the case of a film camera, the position of the principal point is only fixed in the camera and determined by camera calibration with respect to the fiducial marks. When scanning photographs, the principal point will have different positions in the row–column system of each image, because each image will be placed slightly differently in the scanner. Therefore, you have to measure for every image the imaged fiducial marks and calculate the transformation onto the fiducial marks as given by the calibration report. Digital photogrammetric software can then relate row–column

coordinates of any image point to image coordinates (x, y) at the time of exposure (Figure 5.33).

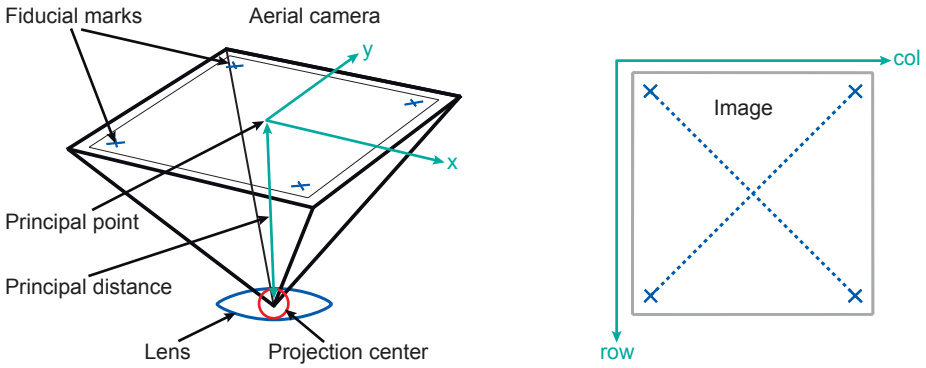


Figure 5.33
Inner geometry of a camera and the associated image.

The *exterior orientation* determines the position of the projection centre with respect to the terrain and also the attitude of the sensor. Frame cameras (film as well as digital ones) acquire the entire image at once, thus three coordinates for the projection centre (X, Y, Z) and three angles for the attitude of the sensor (angles relating to roll, pitch, and yaw) are sufficient to define the exterior orientation of an entire image. Line cameras yield images for which each image line has its own projection centre and the sensor attitude may change from one line to the next. Satellites have a very smooth motion and airborne line cameras are mounted on a stabilized platform so that any attitude change is gradual and small. Mathematically, we can model the variation of an attitude angle through a scene with a low-degree polynomial.

exterior orientation

sensor position and attitude

Images from modern, high-resolution line cameras on satellites come with *rational polynomial coefficients* (RPCs). The rational polynomials define by good approximation the relationship between the image coordinates of an entire frame (in terms of row–column pixel positions) and terrain coordinates. The nice thing is that RPCs are understood by RS software such as ERDAS and that they take care of interior and exterior orientation. For cases in which RPCs are not given or are considered not accurate enough, the exterior orientation needs to be solved in one of the following ways:

RPC

- *Indirect camera orientation*: identify GCPs in the image and measure the row and column coordinates; acquire (X, Y, Z) coordinates for these points, e.g. by GPS or a sufficiently accurate topographic map; use adequate software to calculate the exterior orientation parameters, after having completed the interior orientation.
- *Direct camera orientation*: during image acquisition, make use of GPS and IMU recordings by employing digital photogrammetric software.
- *Integrated camera orientation*, which is a combination of (a) and (b).

For high resolution satellite images such as Ikonos or QuickBird, adding one GCP can already considerably improve the exterior orientation as defined by the RPC. For Cartosat images, it is advisable to improve the exterior orientation by at least five GCPs. For orienting a frame camera image you need at least three GCPs (unless you also have GPS and IMU data). After orientation, you can use the terrain coordinates of any reference point and calculate its position in the image. The differences between measured and calculated image coordinates (the residuals) allow you to estimate the accuracy of orientation. As you may guess, advanced camera/sensor/image orientation is a topic for further study.

Orienting a stereo-image pair obtained by a line or frame camera

relative orientation

The standard procedure is to individually orient each image. In the case of images originating from a frame camera, we can, however, readily make use of the fact that both images partly cover the same area. Instead of doing two independent exterior orientations, we can better first do a *relative orientation* of the two images, followed by an *absolute orientation* of the pair to the terrain coordinate system. The *relative orientation* will cause the imaging rays of corresponding points to intersect more accurately than if you orient one image without the knowledge of the other. You do not have to measure points with known terrain coordinates to solve the relative orientation problem; you only need to measure image coordinates of corresponding points in the two images (after the individual interior orientations have been established). Measuring of corresponding points can even be done automatically (by *image matching*, see below).

absolute orientation

For absolute orientation, at least three GCPs are needed. The idea of splitting up the exterior orientation into a relative orientation and an absolute orientation is also used for orienting a whole block of overlapping images, not just two. The advantage is that we need a only a few GCPs (which are usually expensive to acquire) and we can still obtain accurate transformation parameters for each image. The method is known as *aerial triangulation*.

5.3.6 Monoplotting

Suppose you need to derive accurate planimetric coordinates of features expressed in a specific map projection from a single aerial photograph. For flat terrain, this can be achieved using a vertical photograph and a georeferencing approach. Recall from the earlier discussion on relief displacement (Figure 5.24) how elevation differences lead to distortions in the image, preventing the use of such data for direct measurements. Therefore, if there is significant terrain relief, the resulting relief displacement has to be corrected. The method of *monoplotting* was developed (with major research input from ITC) for just this purpose.

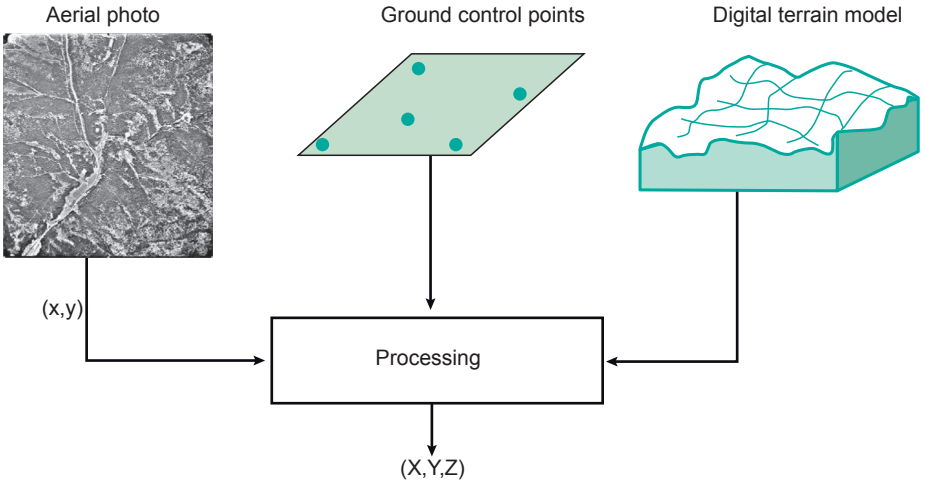


Figure 5.34
The process of digital monoplotting enables accurate determination of terrain coordinates from a single aerial photograph.

3D data from a single image

Monoplotting is based on the reconstruction of the position of the camera at the moment of image exposure with respect to the GCPs, i.e. the terrain. This is achieved by identifying several (at least four) GCPs for which both the photo and map coordinates are known. Information about the terrain relief is supplied by a DTM of adequate accuracy. The DTM should be given in the required map projection system and the elevations should be expressed in an adequate vertical reference system. When digitizing

features from the photograph, the computer uses the DTM to calculate the relief displacement for every point and corrects for it (Figure 5.34). A monoplotting approach is possible by using a hardcopy image on a digitizer tablet or by on-screen digitizing on a computer monitor. In the latter case, vector information can be superimposed on the image to update the changed features. Note that monoplotting is a (real-time) correction procedure and does not yield a new image, i.e. no resampling is carried out.

5.3.7 Orthoimage production

Monoplotting can be considered a georeferencing procedure that incorporates corrections for relief displacement without involving any resampling. For some applications, however, it is useful to actually correct the photograph or RS image, taking into account the effect of terrain relief. In such cases, the image should be transformed and resampled (making use of a DTM) into a product with the geometric properties of a specific map projection. Such an image is called an *orthophoto*.

geocoding an image of rough terrain

The production of orthophotos is quite similar to the process of monoplotting. Consider a scanned aerial photograph. First, the photo is oriented using ground control points. The terrain elevation differences are modelled by a DTM. The computer then calculates the position in the original photo for each output pixel. Using one of the resampling algorithms, the output value is determined and stored in the required raster. The result is geometrically equivalent to a map, i.e. direct distance or area measurements on the orthoimage can be carried out.

5.3.8 Stereo restitution

After relative orientation of a stereo pair, we can exploit the 3D impression gained from the stereo model to make measurements in 3D. The measurements made in a stereo model make use of a phenomenon known as *parallax* (Figure 5.35). Parallax refers to the fact that an object photographed from different camera locations (e.g. from a moving aircraft) has different relative positions in the two images. In other words, there is an apparent displacement of an object as it is observed from different locations. Figure 5.35 illustrates that points at two different elevations, regardless of whether it is the top and bottom of a hill or of a building, experience a relative shift.

3D data from a stereo pair

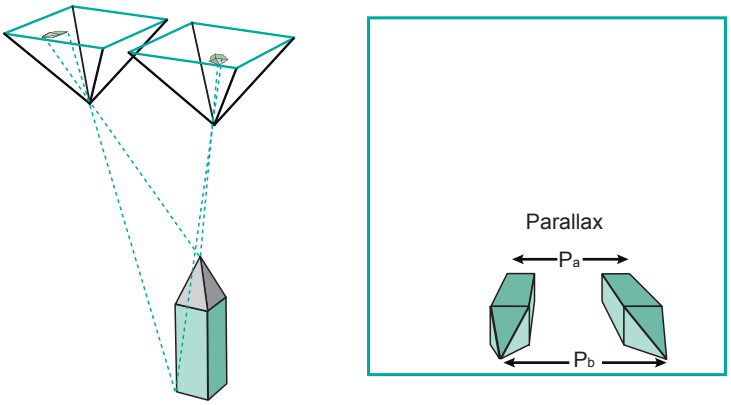


Figure 5.35
The same building is observed from two different positions. Because of the height of the building, the positions of the building top and base relative to the photo centres are different. This difference (parallax) can be used to calculate its height.

image matching

The measurement of the difference in position is basic input for elevation calculations. We could use stereo restitution to measure (X, Y, Z) coordinates of many points, in this way obtaining data for a DTM. This is both a boring and an error-prone process. Digital photogrammetric software can do this job automatically with a high degree of success—after some 30 years of research and development—using *image matching*. Manual measurements are then only needed as a supplement for difficult areas. The main purpose of stereo restitution is to collect 3D vector data of objects. Unlike mono-plotting, elevation is measured directly and not interpolated from a DTM, hence the coordinate accuracy can be higher. Another main advantage of stereo restitution is the better image interpretability obtained, because one can see and interpret in 3D.

analogue, analytical, digital systems

A stereo model enables parallax measurement using a special 3D cursor. If the stereo model is appropriately oriented, the parallax measurements yield (X, Y, Z) coordinates. *Analogue* systems use hardcopy images and perform the computation by mechanical, optical or electrical means. *Analytical* systems also use hardcopy images, but do the computation digitally, while in modern *digital* systems, both the images and the computation are digital. By using digital instruments, we cannot only make a few spot elevation measurements, but also generate automatically a complete DSM for the overlapping part of the two images. Recall, however, that reliable elevation values can only be extracted if the orientation steps were carried out accurately, using reliable ground control points.

Chapter 6

Image analysis

Wan Bakx
Lucas Janssen
Ernst Schetselaar
Klaus Tempfli
Valentyn Tolpekin
Eduard Westinga

6.1 Visual image interpretation

6.1.1 Introduction

How to extract information from images? In general, methods for extracting information from remote sensing images can be subdivided into two groups:

- Information extraction based on visual image interpretation. Typical examples of this approach are visual interpretation methods for land use or soil mapping. Acquisition of data from aerial photographs for topographic mapping is also based on visual interpretation.
- Information extraction based on semi-automatic processing by computer. Examples of this approach include automatic generation of DTMs, digital image classification and calculation of surface parameters.

The most intuitive way of extracting information from remote sensing images is by visual image interpretation, which is based on our ability to relate colours and patterns in an image to real world features. Chapter 5 explains the different methods used to visualize remote sensing data. We can interpret images displayed on a computer monitor or printed images, but how to convey our findings to somebody else? In everyday life we often do this verbally, but for thousands of years we have also been doing it by mapping. We used to overlay a transparency on a photograph and trace over the outline of areas that we recognized as having characteristics we were interested in. By doing so for all features of interest in a scene, we obtained a map. The digital variant of this approach is to digitize—either on-screen, or using a digitizer tablet if we only have a hardcopy image—points, lines and areas and label these geometric entities to convey thematic attributes. This way we obtain a map of, for example, all vineyards

mapping

in a certain area and the roads and tracks leading to them. Instead of interpreting and digitizing from a single image, we can also use a stereo-image pair. The interpretation process is the same, although we do need special devices for stereoscopic display and viewing, as well as equipment that allows us to measure properly in a stereogram.

Visual image interpretation is not as easy as it may seem at first glace; it requires training. Yet our eye–brain system is quite capable of doing the job. Visual interpretation is, in fact, an extremely complex process, as was discovered when we tried to let computers do image interpretation. Research on *image understanding* has helped us to conceptualize human vision and interpretation, and progress in this area continues to be made.

Subsection 6.1.2 explains the basics of how we recognize features and objects in images. Visual image interpretation is used to produce geospatial data in all of ITC's fields of interest: urban mapping, soil mapping, geomorphological mapping, forest mapping, natural vegetation mapping, cadastral mapping, land use mapping, and many others. Actual image interpretation is application specific, although it does follow a standard approach. Subsection 6.1.3 describes this general, practical approach. Aspects of assessing the quality of the outcome of an interpretation are treated in Subsection 6.1.4.

6.1.2 Interpretation fundamentals

Human vision

Human perception of colour is explained in Chapter 5. *Human vision* goes a step beyond the perception of colour: it deals with the ability of a person to draw conclusions from visual observations. When analysing an image, typically you find yourself somewhere between the following two processes: direct and *spontaneous recognition*; and *logical inference*, using clues to draw conclusions by a process of reasoning.

Spontaneous recognition refers to the ability of an interpreter to identify objects or features at first glance. Consider for a moment Figure 6.1. Agronomists would immediately recognize the pivot irrigation systems from their circular shape. They are able to do so because of earlier (professional) experience. Similarly, most people can directly relate what they see on an aerial photo to the terrain features of the place where they live (because of "scene knowledge"). The statement made by people that are shown an aerial photograph of their living environment for the first time, "I see because I know", is rooted in spontaneous recognition.

spontaneous recognition

As the term states, *Logical inference* means that the interpreter applies reasoning. In the reasoning, the interpreter uses acquired professional knowledge and experience. Logical inference is, for example, concluding that a rectangular shape is a swimming pool because of its location in a backyard garden near to a house. Sometimes logical inference alone is insufficient to interpret images; then field observations are required (see Subsection 6.1.3). Consider the aerial photograph in Figure 6.2. Are you able to interpret the material and function of the white mushroom-like objects? A field visit would be required for most of us to relate the different features to elements of a house or settlement.

logical inference

Interpretation elements

We need a set of terms to express the characteristics of an image that we can use when interpreting the image. These characteristics are called *interpretation elements* and are used, for example, to define *interpretation keys*, which provide guidelines on how to recognize certain objects.

The following seven interpretation elements can be distinguished: tone/hue, texture,

Figure 6.1
RS image of the Antequera area in Spain; the circular features are pivot irrigation systems. The area imaged is 5 km wide.

pattern, shape, size, height/elevation, and location/association.

- *Tone* is defined as the relative brightness in a B&W. *Hue* refers to the colour as defined in IHS colour space. Tonal variations are an important interpretation element. The tonal expression of objects in an image is directly related to the amount of light (or other forms of EM radiation) reflected (or emitted) from the surface. Different types of rock, soil or vegetation are most likely have different tones. Variations in moisture conditions are also reflected as tonal differences in an image: increasing moisture content gives darker grey tones. Variations in hue are primarily related to the spectral characteristics of the imaged terrain and also to the bands selected for visualization (see Chapter 5). The advantage of hue over tone is that the human eye has a much larger sensitivity for variations in colour (approximately 10,000 colours) than tone (approximately 200 grey levels).

- *Texture* relates to the frequency of tonal change. Texture may be described by terms such as coarse or fine, smooth or rough, even or uneven, mottled, speckled, granular, linear, woolly, etc. Texture can often be related to terrain surface roughness. Texture is strongly related to the spatial resolution of the sensor used. A pattern on a large-scale image may show as texture on a small-scale image of the same scene.

- *Pattern* refers to the spatial arrangement of objects and implies the characteristic repetition of certain forms or relationships. Pattern can be described by terms such as concentric, radial and checkerboard. Some land uses have specific and characteristic patterns when observed from the air or space. Different types of irrigation may spring to mind, or different types of housing on an urban fringe.

Figure 6.2
Mud huts of Labbezanga, near the Niger river (Photo by Georg Gerster, 1972).

Other typical examples include hydrological systems (a river and its tributaries) and patterns related to erosion.

- *Shape* or form characterizes many objects visible in an image. Both the two-dimensional projection of an object, as shown on a map, and the height of an object influence the shape of its image. The shape of objects often helps us to identify them (built-up areas, roads and railroads, agricultural fields, etc.).

- *Size* of objects can be considered in a relative or absolute sense. The width of a road can be estimated, for example, by comparing it to the size of the cars using it, which is generally known. Subsequently, the width determines the road type, e.g. primary road, secondary road, and so on.

- *Height* differences are important for distinguishing between different vegetation types, building types, etc. Elevation differences provide us with clues in geomorphological mapping. We need a stereogram and stereoscopic viewing to observe height and elevation. Stereoscopic viewing facilitates interpretation of both natural and man-made features.

- *Location/association* refers to situation of an object in the terrain or in relation to surroundings. A forest in the mountains is different from a forest close to the sea or one near a meandering river. A large building at the end of a number of converging railroads is likely to be a railway station—we would not expect a hospital at such a location.

With these seven interpretation elements, you may have noticed a relation with the spatial extent of the feature to which they relate. Tone or hue can be defined for a single pixel; texture is defined for a group of adjacent pixels, not for a single pixel. The other interpretation elements relate to individual objects or a combination of objects. The simultaneous and often intuitive use of all these elements is the strength of visual image interpretation. In standard digital image classification (Section 6.2) only hue is utilized, which explains the limitations of automated methods compared to visual image interpretation.

6.1.3 Mapping

Interpretation

The assumption in mapping with the help of remote sensing images is that areas that look homogeneous in the image will have similar features on the ground. The interpretation process consists of delineating areas that internally appear similar and at the same time different from other areas. Making an interpretation from only one aerial photograph or a small part of an image from a space-borne sensor seems quite simple. You have the overview of the entire area at all times and can easily compare one unit to another and decide if they are the same or different. Working with many photographs and also with several people will, in contrast, require a clear definition of the units to be delineated.

Definition of units is based on what can be observed in the image. Different interpretation units can be described according the interpretation elements. After establishing what the features are on the ground, '*interpretation elements* can be constructed, from which an interpretation of features can be made. These features are again described in terms of interpretation elements. If knowledge of the area is lacking (not yet available), you could also begin your interpretation based only on interpretation elements (Figure 6.3). After fieldwork, it will become clear what the units actually represent on the ground.

interpretation key

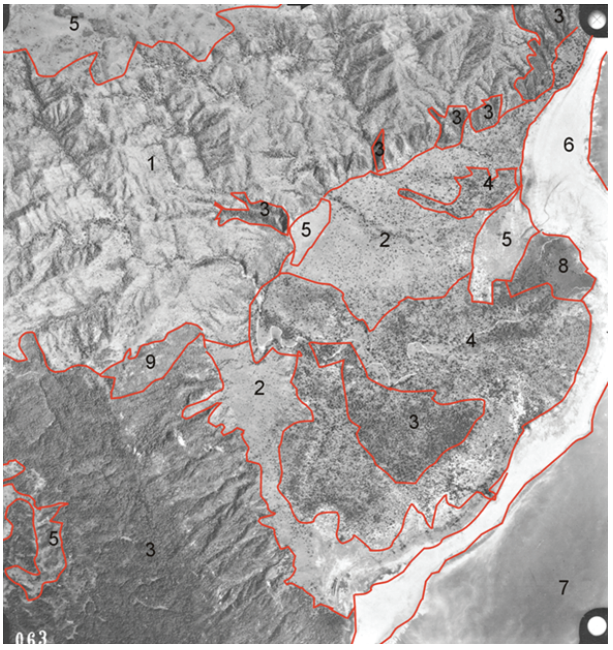


Figure 6.3
Example of an interpretation of Manyara, Tanzania.

Prior to the delineation of the units, a legend is constructed based on interpretation elements. The legend can be presented in the form of a table in which each element type is represented by a column. Table 6.1 presents a fictitious example of a legend. In this legend, the "unit number" represents an as yet unknown feature type; the corresponding row elements will be used to identify that feature type.

interpretation legend

When preparing a legend you need to consider that distinguishing units can be based on a difference in one element only or on differences of several elements. For example, consider Unit 1 in Table 6.1: its tone is black and all other units have a grey or white tone. In this case there is no need to define all the other elements for Unit 1. In the

Table 6.1
Fictitious example of an interpretation legend.

Unit	Tone	Texture	Shape	Size	Height	Location
1	black					
2	grey	smooth				
3	grey	rough			high	mountains
4	grey	rough			low	
5	grey	rough			high	sea + river
6	grey		field			
	white		line			
7	white		field			
	grey		line			
8	grey + black		field	square		
9	grey + black		field	rectangle 5 × 5		
10	grey + black		field	rectangle 20 × 20		

example of Figure 6.3, some units are different in texture. There are areas with smooth and rough texture. The rough texture areas are further differentiated according to height. Furthermore, rough, high areas are differentiated, depending on location in the mountains or along rivers or near the sea.

When delineating areas by hand, there is a limit to what can still be drawn. In practice, polygons smaller than 5 mm × 5 mm should not be drawn. This is called the smallest allowable unit. The scale of the image(s) used therefore limits the interpretation cell on the ground. When delineating areas by digitizing on-screen, one could zoom in—in principle to a monitor dot. However, you need to define the maximum scale at which the given remote sensing data are still reliable and then calculate the smallest allowable unit.

interpretation cell

In some cases an area may consist of two or three different types of too-small areas. Then, individual polygons for each small area cannot be drawn, even though at a larger scale the individual features could be mapped. The solution in such a case is to combine these areas to form a complex unit. The different features of such a complex can be described separately. In Table 6.1, Unit 6 and Unit 7 are two different complex units: in Unit 6 there are two features, namely grey fields and white lines, while in Unit 7 there are white fields and grey lines.

complex unit

Fieldwork

Maps and inventories should reflect what is actually on the ground. Field visits should, therefore, be made to observe what is there in reality. Field visits for ground observation are time-consuming and usually costly. Making observations everywhere in the entire area to be mapped is likely to take too much time. For reasons of efficiency, remote sensing data are used to extrapolate the results of a limited number of observations over the entire area being studied.

The selection of sample locations is a crucial step for cost-effective mapping. We can use the RS images to stratify the area. To do this, we make an preliminary interpretation of the area to be mapped based on its interpretation elements. The interpretation units are the strata to be sampled. For all strata, an equal number of samples are taken; this is known as *stratified sampling*. We can select the samples in such away that they are representative for the interpretation elements of that unit (strata). This is called *stratified representative sampling*. Stratified representative sampling is a very time-efficient and cost-effective method as compared to random or systematic sampling ([132]; [41]). If an interpretation unit occupies a very small area, many samples would be needed for random or systematic sampling, to make sure that small units are also sampled. When applying the stratified sampling approach, far fewer samples

stratified sampling

are needed.

Stratified representative sampling can only be applied if the data to be mapped are qualitative (i.e. nominal or ordinal). For mapping of quantitative data (i.e. interval or ratio data), *unbiased sampling* strategies (i.e. random or systematic sampling) should be applied to allow statistical analysis. Biomass measurements are an example of quantitative data. Then the entire area needs to be sampled and no prior interpretation is needed for the sampling strategy. Both stratified and unbiased sampling strategies will be used if quantitative data of certain strata are not required. For instance, we use *stratified random sampling* of grass biomass for livestock management if in the strata forest, water and urban areas no biomass measurements are needed. We do so to limit time-consuming, unbiased sampling procedures.

unbiased sampling

During fieldwork, the locations of boundaries of the interpretation are also verified. In addition, data is gathered about areas or features that cannot be derived from remote sensing images.

Analysing field data and map preparation

From the correlation between collected field data and the interpretation, the entire area can be mapped in terms of what is on the ground. If there is a good correlation, only recoding and renaming of the units will be required. If the correlation is poor, however, a complete re-interpretation might be needed, after carefully restudying the legend in terms of interpretation elements. For producing the final map, all aspects of map design and cartographic finishing should be observed; this is treated in Section 10.1).

6.1.4 Quality aspects

The quality of the result of image interpretation depends on three factors: the interpreter, the images used, and the guidelines provided.

- Professional experience, particularly experience with image interpretation, determines the skill of a photo-interpreter. A professional background is required: a geological interpretation, for example, can only be made by a geologist, since they are best able to related image features to geological phenomena. Local knowledge, derived by field visits, is needed to facilitate interpretation.

  trained interpreter

- The images used limit the phenomena that can be studied, both in a thematic and geometric sense. One cannot, for example, generate a reliable database on tertiary road systems using data from low resolution multispectral scanners. On the other hand, B&W aerial photos contain limited information about agricultural crops.

  adequate images

- The quality of the interpretation guidelines has a great deal of influence. Consider, for example, a project in which a group of persons is to carry out a mapping project. Ambiguous guidelines will prevent consistent mapping, for which a seamless database of consistent quality is required, despite individual input.

  clear guidelines

Especially in large projects and monitoring programmes, all three points just listed play an important role in ensuring the replicability of the work. *Replicability* refers to the degree of correspondence of results obtained by different persons for the same area or by the same person for the same area at different times. Replicability does not provide information on the accuracy (the relation with the real world) but it does give an indication of the quality of the class definition (crisp or ambiguous) and the instructions and methods used. Figure 6.4 and Figure 6.5 provide two examples of how this works. Figure 6.4 gives two interpretation results for the same area. Note

replicability

that both results differ in terms of the total number of objects (map units) and in terms of (line) generalization. Figure 6.5 compares 13 individual geomorphological interpretations. Similarly to Figure 6.4, large differences occur along the boundaries. In addition to this, you could also conclude that for some objects (map units) there was no agreement on the thematic attribute.

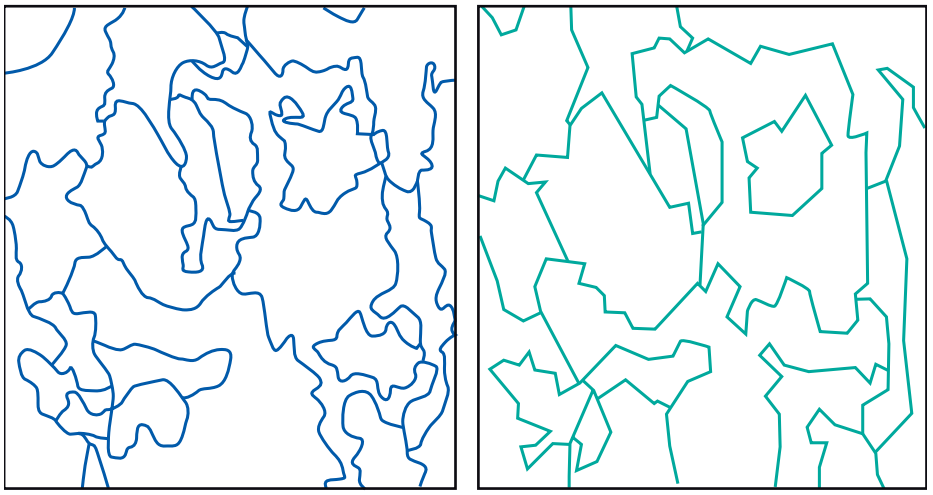


Figure 6.4
Two interpretation results derived by two photo-interpreters analysing the same image. Note the overall differences, but also differences in generalization of the lines. (From [73].)

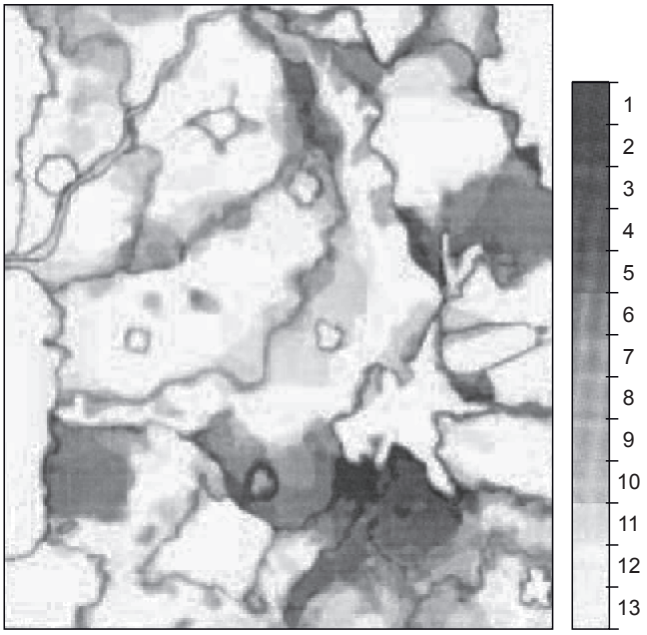


Figure 6.5
Comparison of 13 interpretations of the same image. The grey value represents the degree of correspondence: white indicates agreement of all 13 interpreters; black indicates that all 13 interpreters disagreed on the thematic class for that location. (From [73].)

6.2 Digital image classification

Introduction

The process of visual image interpretation has been explained in Section 6.1. In this process, human vision plays a crucial role in extracting information from images. Although computers may be used for visualization and digitization, the interpretation itself is done by the human operator.

This section introduces *digital image classification*. In this process, the human operator instructs the computer to perform an interpretation according to certain conditions, which are defined by the operator. Image classification is one of the techniques in the domain of digital image interpretation. Other techniques include automatic object recognition (for example, road detection) and scene reconstruction (for example, generation of 3D object models). Image classification is the most commonly applied technique in ITC's fields of interest.

digital image interpretation

Image classification is applied in many regional-scale projects. In Asia, the Asian Association of Remote Sensing (AARS) is generating various sets of land cover data based on supervised and unsupervised classification of multispectral satellite data. In the Africover project (an FAO initiative), techniques for digital image classification are being used to establish a pan-African land cover data set. The European Commission requires national governments to verify the claims of farmers related to crop subsidies. To meet these requirements, national governments employ companies to make an intial inventory, using image classification techniques, which is followed later by field checks.

regional scale projects

Image classification is based on the different spectral characteristics of different materials, as introduced in Section 2.5. Here, in section 6.2, the focus is on the classification of multispectral data. Subsection 6.2.1 explains the concepts of image space and feature space, where image classification (Subsection 6.2.2) takes place. Section 6.2.3 gives an overview of the classification process, the steps involved and the choices to be made. The results of image classifications need to be validated to assess their accuracy, the topic of Subsection 6.2.4). Subsection 6.2.5 discusses the problems of standard classification and introduces object-oriented classification.

6.2.1 Principles of image classification

Image space

A digital image is a 2D array of pixels. The value of a pixel, i.e. its DN, is in the case of an 8-bit record in the range 0 to 255. Each DN corresponds to the EM radiation reflected or emitted from a ground resolution cell—unless the image has been resampled. The spatial distribution of the DNs defines the image or *image space*. A multispectral sensor records the radiation from a particular GRC in different channels according to its spectral band separation. A sensor recording in three bands (Figure 2.24) yields three pixels with the same row and column tuple (i, j) since they stem from one and the same GRC.

Feature space

When we consider a two-band image, we can say that the two DNs for a GRC are components of a two-dimensional vector $[v_1, v_2]$, the *feature vector* (Figure 6.6). An example of a feature vector is $[13, 55]$, which indicates that the conjugate pixels of band 1 and band 2 have the DNs 13 and 55. This vector can be plotted in a two-dimensional graph.

feature vector

Similarly, we can visualize a three-dimensional feature vector $[v1, v2, v3]$ of a cell in a

three-band image found in a three-dimensional graph. A graph that shows the feature vectors is called a *feature space*, or *feature space plot* or *scatter plot*. Figure 6.6 illustrates how a feature vector (related to one GRC) is plotted in the feature space for two and three bands. Two-dimensional feature-space plots are the most common.

Note that plotting the values is difficult for a four- or more-dimensional case, even though the concept remains the same. A practical solution when dealing with four or more bands is to plot all possible combinations of two bands separately. For four bands, this already yields six combinations: bands 1 and 2, 1 and 3, 1 and 4, bands 2 and 3, 2 and 4, and bands 3 and 4.

scatterplot

Plotting all the feature vectors of a digital image pair yields a 2D scatterplot of many points (Figure 6.7). A 2D scatterplot provides information about pixel value pairs that occur within a two-band image. Note that some combinations will occur more frequently, which can be visualized by using intensity or colour (as introduced in Section 5.1).

Distances and clusters in the feature space

We use distance in the feature space to accomplish classification. Distance in the feature space is measured as *Euclidian distance* in the same units as the DNs (the unit of the axes). In a two-dimensional feature space, the distance between feature vectors $[v_{11}, v_{12}]$ and $[v_{21}, v_{22}]$ can be calculated according to Pythagoras' theorem:

$$d^2 = (v_{21} - v_{11})^2 + (v_{22} - v_{12})^2$$

For the situation shown in Figure 6.8, the distance between $[10, 10]$ and $[40, 30]$ is:

$$d = \sqrt{(40 - 10)^2 + (30 - 10)^2}$$

.

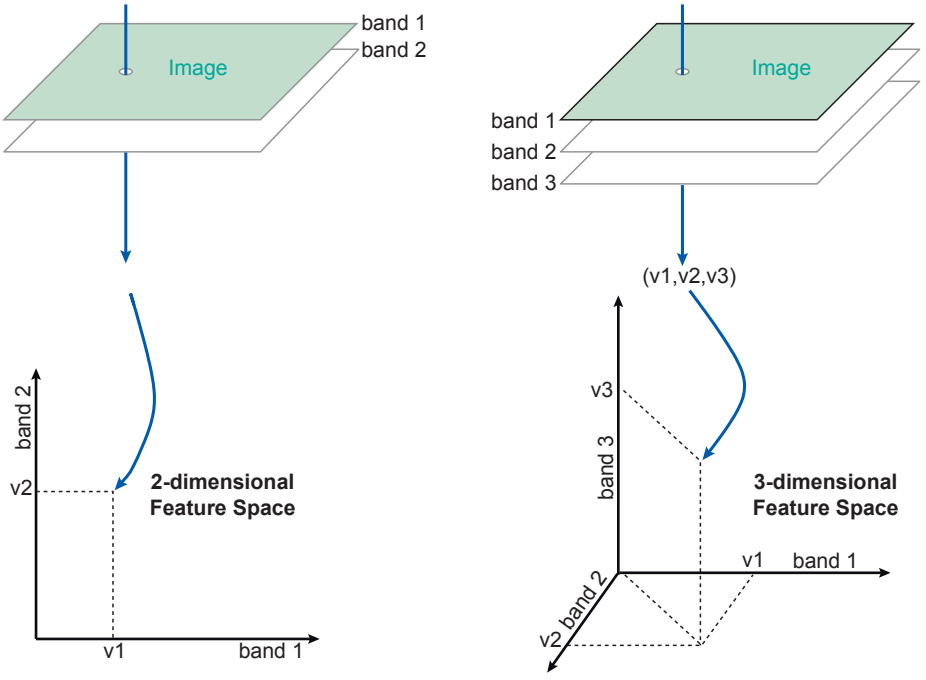


Figure 6.6
Plotting of the pixel values of a GRC in the feature space for a two-band and three-band image.

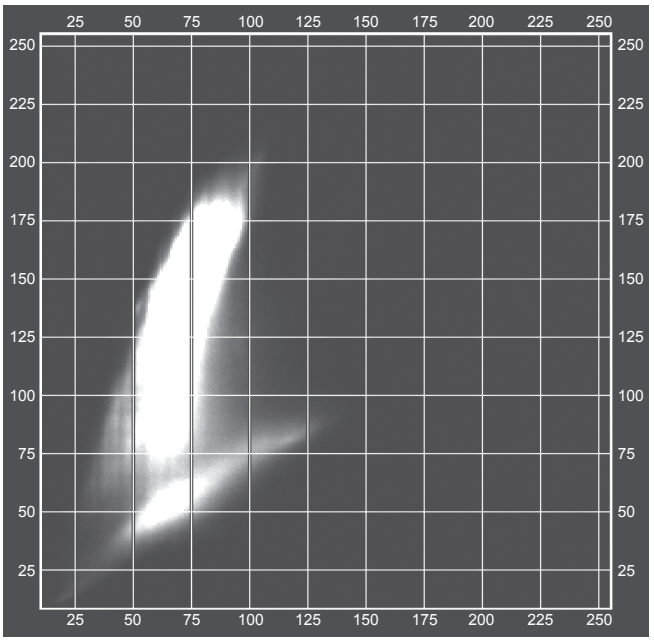


Figure 6.7
Scatterplot of two bands of a RS image. Note the units along the x- and y-axes. The intensity at a point in the feature space is related to the number of cells at that point.

For three or more dimensions, the distance is calculated in a similar manner.

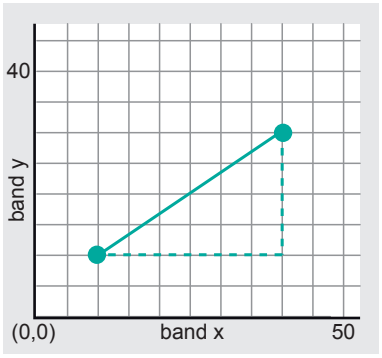


Figure 6.8
Euclidian distance between the two points is calculated using Pythagoras' theorem.

6.2.2 Image classification

The scatterplot shown in Figure 6.7 shows the distribution of conjugate pixel values of an actual two-band image. Figure 6.9 shows a feature space in which the feature vectors have been plotted of samples of five specific land cover classes (grass, water, trees, etc.). You can see that the feature vectors of GRCs that are water areas form a compact cluster. The feature vectors of the other land cover types (classes) are also clustered. Figure 6.9 illustrates the basic assumption for image classification: a specific part of the feature space corresponds to a specific class. Once the classes have been defined in the feature space, each feature vector of a multi-band image can be plotted and checked against these classes and assigned to the class where it fits best.

cluster

classes

Classes to be distinguished in an image classification need to have different spectral characteristics. This can, for example, be analysed by comparing spectral reflectance curves (Section 2.5). This brings us to an important limitation of image classification: if classes do not have distinct clusters in the feature space, image classification can only

spectral differentiation

give results to a certain level of reliability.

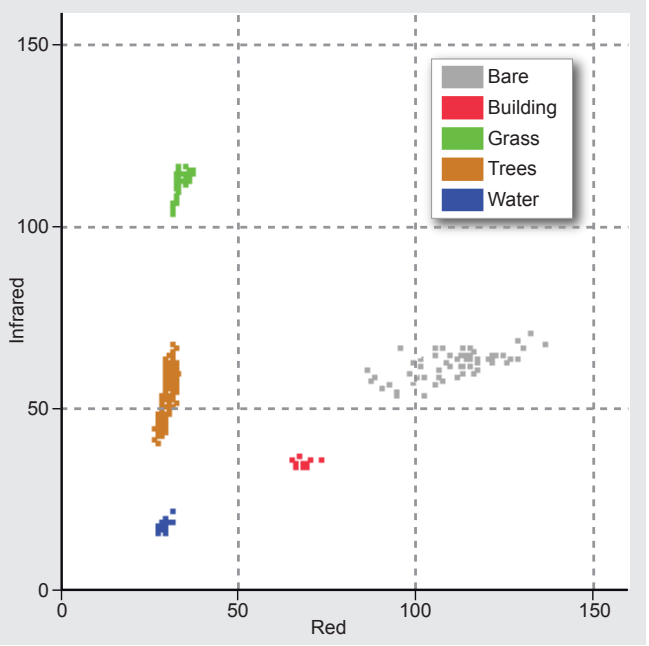


Figure 6.9
Feature space showing the respective clusters of five classes; note that each class occupies a limited area in the feature space.

The principle of image classification is that a pixel is assigned to a class based on its feature vector, by comparing it to predefined clusters in the feature space. Doing so for all pixels results in a classified image. The crux of image classification is in comparing it to predefined clusters, which requires definition of the clusters and methods for comparison. Definition of the clusters is an interactive process and is carried out during the *training process*. Comparison of the individual pixels with the clusters takes place using *classifier algorithms*. Both of these concepts are explained in the next subsection.

6.2.3 Image classification process

The process of image classification typically involves five steps (Figure 6.10):

data selection

1. Selection and preparation of the RS images. Depending on the land cover types or whatever needs to be classified, the most appropriate sensor, the most appropriate date(s) of acquisition and the most appropriate wavelength bands should be selected.

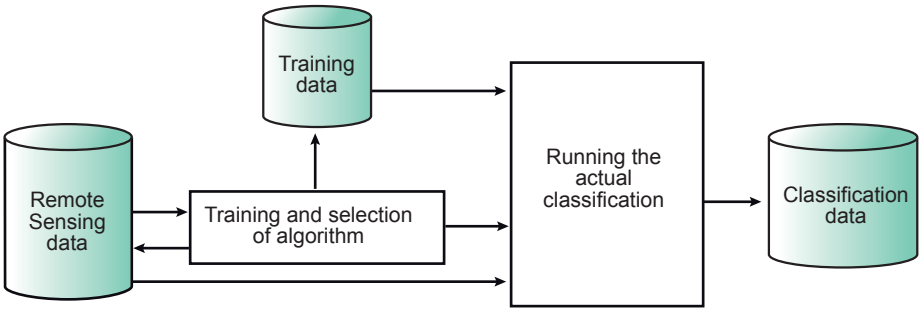


Figure 6.10
The classification process; the most important component is the training, in combination with selection of the algorithm.

2. Definition of the clusters in the feature space. Here two approaches are possible: *supervised classification* and *unsupervised classification*. In a supervised classification, the operator defines the clusters during the training process; in an unsupervised classification, a clustering algorithm automatically finds and defines the number of clusters in the feature space.

supervised or unsupervised

3. Selection of the classification algorithm. Once the spectral classes have been defined in the feature space, the operator needs to decide on how the pixels (based on their feature vectors) are to be assigned to the classes. The assignment can be based on different criteria.

algorithm selection

4. Running the actual classification. Once the training data have been established and the classifier algorithm selected, the actual classification can be carried out. This means that, based on its DNs, each "multi-band pixel" (cell) in the image is assigned to one of the predefined classes (Figure 6.11).

classifying

5. Validation of the result. Once the classified image has been produced its quality is assessed by comparing it to reference data (ground truth). This requires selection of a sampling technique, generation of an error matrix, and the calculation of error parameters (Subsection 6.2.4).

validating

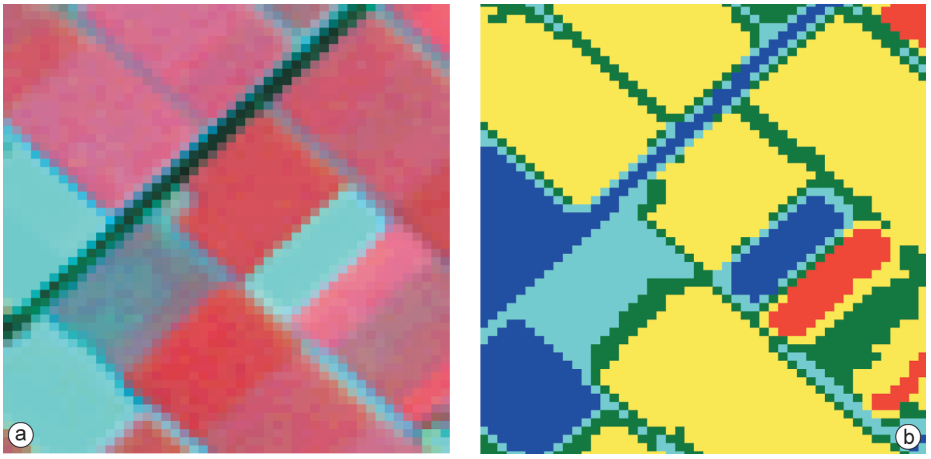


Figure 6.11
The result of classification of a multi-band image (a) is a raster in which each cell is assigned to some thematic class (b).

Each of the steps above are elaborated on in the next subsections. For simplicity and ease of visualization, most examples deal with a two-dimensional situation (two bands), although in principle image classification can be carried out on any n-dimensional data set. Visual image interpretation, however, limits itself to an image that is composed of a maximum of three bands.

Preparation for image classification

Image classification serves a specific goal: converting RS images to thematic data. In the context of a particular application, one is rather more interested in the thematic characteristics of an area (a GRC) than in its reflection values. Thematic characteristics such as land cover, land use, soil type or mineral type can be used for further analysis and input into models. In addition, image classification can also be considered as data reduction: the n multispectral bands result in a single-band image file.

thematic classes

With the particular application in mind, the information classes of interest need to be defined. The possibilities for the classification of land cover types depend on the date

data date

an image was acquired. This not only holds for crops, which have a certain growth cycle, but also for other applications such as snow cover or illumination by the Sun. In some situations, a multi-temporal data set is required. A non-trivial point is that the required images should be available at the required moment. Limited image acquisition and cloud cover may force one to make use of a less-optimal data set.

selection of bands

Before starting to work with the acquired data, a selection of the available spectral bands may be made. Reasons for not using all available bands (for example all seven bands of Landsat-5 TM) lie in the problem of band correlation and, sometimes, in limitations of hardware and software. Band correlation occurs when the spectral reflection is similar for two bands. The correlation between the green and red wavelength bands for vegetation is an example: a low reflectance in green correlates with a low reflectance in red. For classification purposes, correlated bands give redundant information and might disturb the classification process.

Supervised image classification

scene knowledge

One of the main steps in image classification is the "partitioning" of the feature space. In supervised classification this is done by an operator who defines the spectral characteristics of the classes by identifying sample areas (training areas). Supervised classification requires that the operator is familiar with the area of interest: the operator needs to know where to find the classes of interest in the scene. This information can be derived from general knowledge of the scene or from dedicated field observations.

A sample of a specific class, comprising a number of training cells, forms a cluster in the feature space (as portrayed in Figure 6.9). The clusters selected by the operator:

training set

- should form a representative data set for a given class. This means that the variability of a class within the image should be taken into account. Also, in an absolute sense, a minimum number of observations per cluster is required. Although it depends on the classifier algorithm to be used, a useful rule of thumb is $30 \times n$ (n = number of bands) observations.
- should not or only partially overlap with the other clusters, otherwise a reliable separation is not possible. For a specific data set, some classes may have significant spectral overlap, which, in principle, means that these classes cannot be discriminated by image classification. Solutions are to add other spectral bands, and/or add images acquired at other moments.

The resulting clusters can be characterized by simple statistics of the point distributions. These are for one cluster: the vector of mean values of the DNs (for band 1 and band 2) (see Figure 6.14), and the standard deviations of the DNs (for band 1 and band 2) (see Figure 6.15, where the standard deviations are plotted as crosses).

Unsupervised image classification

Supervised classification requires knowledge of the area of interest. If this knowledge is insufficiently available, or the classes of interest have not yet been defined, an unsupervised classification can be made. In an unsupervised classification, clustering algorithms are used to partition the feature space into a number of clusters.

number of classes

Several methods of unsupervised classification exist, their main purpose being to produce spectral groupings based on certain spectral similarities. In one of the most common approaches, the user has to define the maximum number of clusters in a data set. Based on this, the computer locates arbitrary mean vectors as the centre points of the clusters. Each pixel is then assigned to a cluster by the *minimum distance to cluster centroid* decision rule. Once all the cells have been labelled, recalculation of the cluster

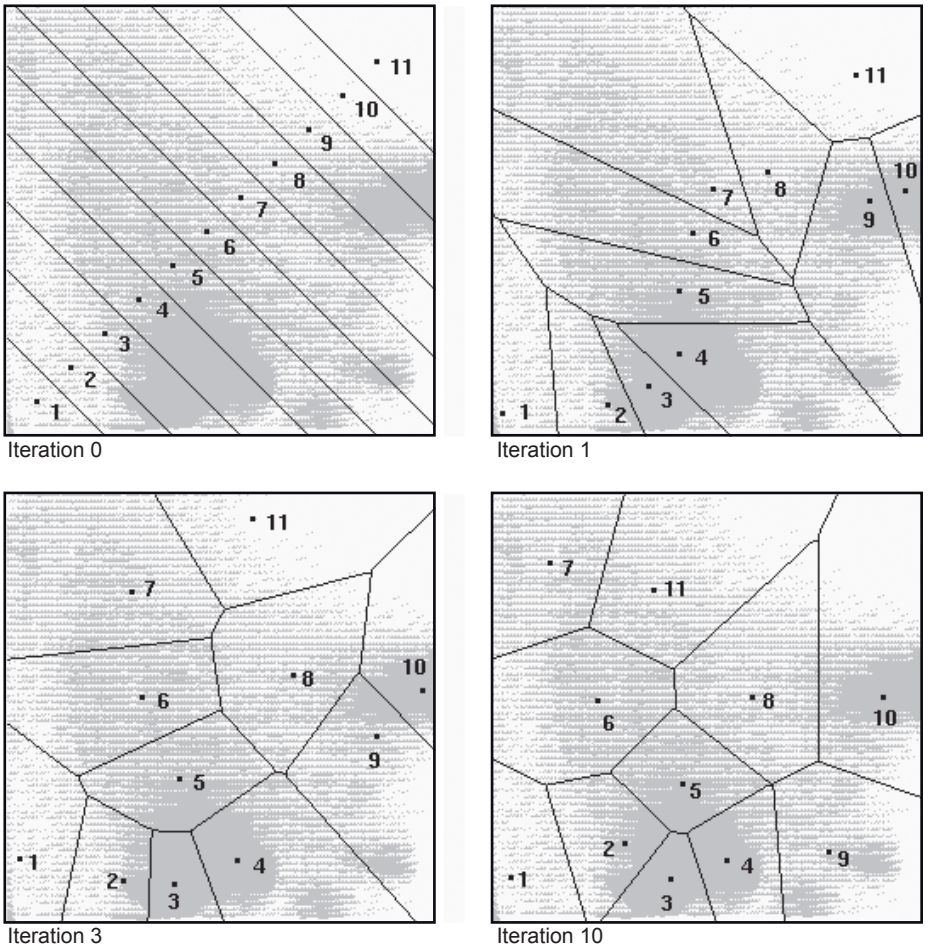


Figure 6.12
The subsequent results of an iterative clustering algorithm on a sample data set.

centre takes place and the process is repeated until the proper cluster centres are found and the cells are labelled accordingly.

iterative process

The iteration stops when the cluster centres no longer change. However, for any iteration, clusters with less than a specified number of cells are eliminated. Once the clustering is finished, analysis of the closeness or separability of the clusters takes place by means of inter-cluster distance or divergence measures.

Merging of clusters needs to be done to reduce the number of unnecessary subdivisions in the data set. This is be done using a pre-specified threshold value. The user has to define the maximum number of clusters/classes, the distance between two cluster centres, the radius of a cluster, and the minimum number of cells as a threshold for cluster elimination. Analysis of the cluster compactness around its centre point is done by means of the user-defined standard deviation for each spectral band. If a cluster is elongated, separation of the cluster will be done perpendicularly to the spectral axis of elongation.

Analysis of closeness of the clusters is carried out by measuring the distance between the two cluster centres. If the distance between two cluster centres is less than the pre-specified threshold, merging of the clusters takes place. The clusters that result after the last iteration are described by their statistics. Figure 6.12 shows the results of a

clustering algorithm on a data set. Note that the cluster centres coincide with the high density areas in the feature space.

Similarly to the supervised approach, the derived cluster statistics are then used to classify the complete image using a selected classification algorithm.

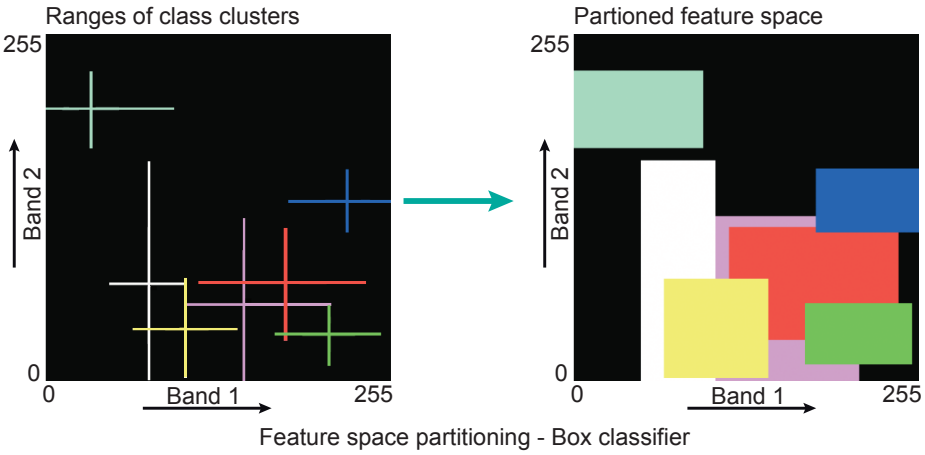


Figure 6.13
Principle of the box classification in a two-dimensional situation.

Classification algorithms

After the training sample sets have been defined, classification of the image can be carried out by applying a classification algorithm. Several classification algorithms exist. The choice of the algorithm depends on the purpose of the classification and the characteristics of the image and training data. The operator needs to decide if a *reject* or *unknown* class is allowed. In the following, three classifier algorithms are described. First the *box classifier* is explained—its simplicity helps in understanding the principle. In practice, the box classifier is hardly ever used, however; *minimum distance to mean* and the *maximum likelihood* classifiers are most frequently used.

Box classifier

The box classifier is the simplest classification method. For this purpose, upper and lower limits are defined for each band and each class. The limits may be based on the minimum and maximum values or on the mean and standard deviation per class. When the lower and the upper limits are used, they define a box-like area in the feature space (Figure 6.13). The number of boxes depends on the number of classes. During classification, every feature vector of an input (two-band) image will be checked to see if it falls in any of the boxes. If so, the cell will get the class label of the box it belongs to. Cells that do not fall inside any of the boxes will be assigned the "unknown class", sometimes also referred to as the "reject class".

The disadvantage of the box classifier is the overlap between classes. In such a case, a cell is arbitrarily assigned the label of the first box it encounters.

Minimum Distance to Mean classifier

The basis for the minimum distance to mean (MDM) classifier is the cluster centres. During classification, the Euclidean distances from a candidate feature vector to all the cluster centres are calculated. The candidate cell is assigned to the class that qualifies as the closest one. Figure 6.14 illustrates how a feature space is partitioned based on the cluster centres. One of the disadvantages of the MDM classifier is that points that are at a large distance from a cluster centre may still be assigned to this centre.

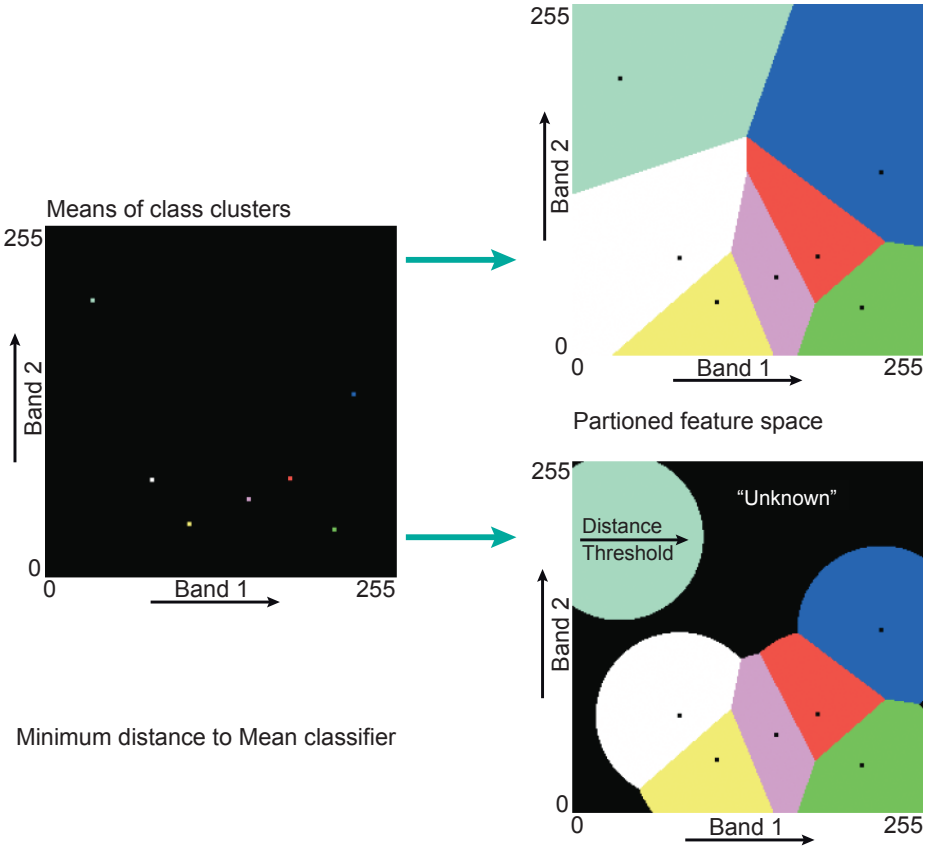


Figure 6.14
Principle of the *minimum distance to mean* classification in a two-dimensional situation. The decision boundaries are shown for a situation without threshold distance (upper right) and one with threshold distance (lower right).

This problem can be overcome by defining a threshold value that limits the search distance. Figure 6.14 illustrates the effect; the threshold distance to the centre is shown as a circle.

A further disadvantage of the MDM classifier is that it does not take the class variability into account: some clusters are small and dense, while others are large and dispersed. Maximum likelihood classification does, however, take class variability into account.

Maximum Likelihood classifier

The Maximum likelihood (ML) classifier considers not only the cluster centres but also the shape, size and orientation of the clusters. This is achieved by calculating a statistical distance based on the mean values and covariance matrix of the clusters. The statistical distance is a probability value: the probability that observation x belongs to specific cluster. A cell is assigned to the class (cluster) for which it has the highest probability. The assumption of most ML classifiers is that the statistics of the clusters follow a *normal* (Gaussian) distribution.

For each cluster, what are known as "equiprobability contours" can be drawn around the centres of the clusters. Maximum likelihood also allows the operator to define a threshold distance by defining a maximum probability value. A small ellipse centred on the mean defines the values with the highest probability of membership of a class. Progressively larger ellipses surrounding the centre represent contours of probability of membership to a class, with the probability decreasing the further away from the

centre. Figure 6.15 shows the decision boundaries for a situation with and without threshold distance.

6.2.4 Validation of the result

Image classification results in a raster file in which the individual raster elements are labelled by class. As image classification is based on samples of the classes, the actual quality of the classification result should be checked. This is usually done by a sampling approach in which a number of raster elements of the output are selected and both the classification result and the "true world class" are compared. Comparison is done by creating an *error matrix* from which different accuracy measures can be calculated. The true world class is preferably derived from field observations. Sometimes, sources for which higher accuracy can be assumed, such as aerial photos, are used as a reference.

sampling schemes

Various *sampling schemes* have been proposed for selecting pixels to test. Choices to be made relate to the design of the sampling strategy, the number of samples required, and the area of the samples. Recommended sampling strategies in the context of land cover data are simple random sampling or stratified random sampling. The number of samples may be related to two factors in accuracy assessment: (1) the number of samples that must be taken in order to reject a data set as being inaccurate; or (2) the number of samples required to determine the true accuracy, within some error bounds, of a data set. Sampling theory is used to determine the number of samples required. The number of samples must be traded-off against the area covered by a sample unit. A sample unit can be a point but it could also be an area of some size; it can be a single raster element but may also include surrounding raster elements. Among other

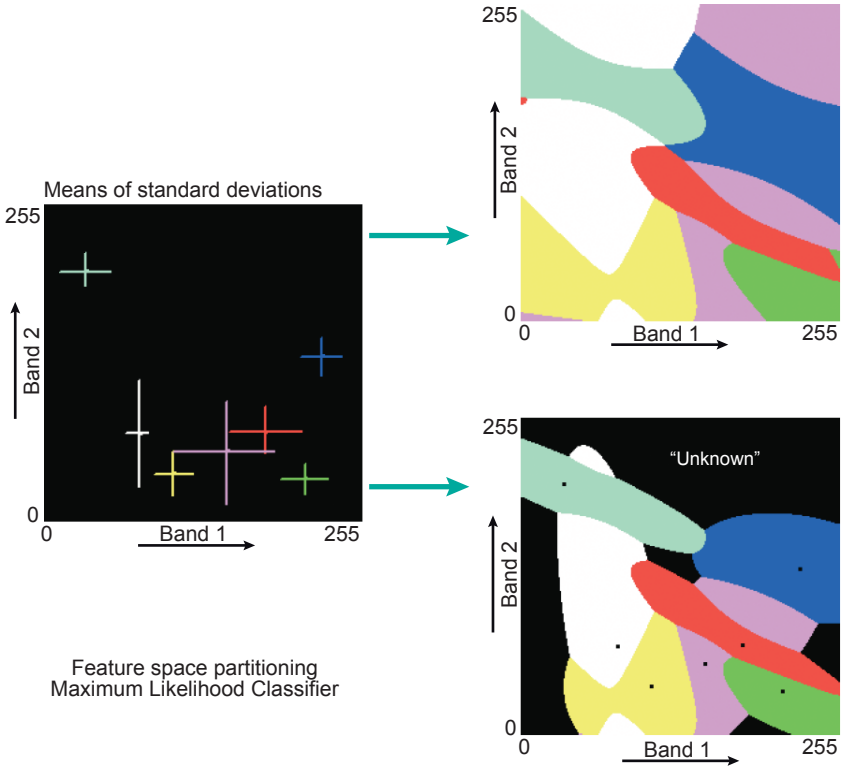


Figure 6.15
Principle of *maximum likelihood* classification. The decision boundaries are shown for a situation without threshold distance (upper right) and one with threshold distance (lower right).

considerations, the "optimal" sample-area size depends on the heterogeneity of the class.

Table 6.2 The error matrix with derived errors and accuracy expressed as percentages. A, B, C and D refer to the reference classes; a, b, c and d refer to the classes in the classification result. Overall accuracy is 53%.

	A	B	C	D	Total	Error of Commission (%)	User Accuracy (%)
a	35	14	11	1	61	43	57
b	4	11	3	0	18	39	61
c	12	9	38	4	63	40	60
d	2	5	12	2	21	90	10
Total	53	39	64	7	163		
Error of Omission	34	72	41	71			
Producer Accuracy	66	28	59	29			

Once sampling for validation has been carried out and the data collected, an error matrix, also sometimes called a *confusion matrix* or an *contingency matrix*, can be established (Table 6.2). In the table, four classes (A, B, C, D) are listed. A total of 163 samples were collected. The table shows that, for example, 53 cases of A were found in the real world ('reference'), while the classification result yielded 61 cases of a; in 35 cases they agree.

error matrix

The first and most commonly cited measure of mapping accuracy is the *overall accuracy*, or proportion correctly classified (PCC). Overall accuracy is the number of correctly classified pixels (i.e. the sum of the diagonal cells in the error matrix) divided by the total number of pixels checked. In Table 6.2 the overall accuracy is $(35 + 11 + 38 + 2)/163 = 53\%$. The overall accuracy yields one value for the result as a whole.

overall accuracy

Most other measures derived from the error matrix are calculated per class. *Error of omission* refers to those sample points that are omitted in the interpretation result. - Consider class A, for which 53 samples were taken. Eighteen out of the 53 samples were interpreted as b, c or d. This results in an error of omission of $18/53 = 34\%$. Error of omission starts from the reference data and therefore relates to the columns in the error matrix. The *error of commission* starts from the interpretation result and refers to the rows in the error matrix. The error of commission refers to incorrectly classified samples. Consider class d: only two of the 21 samples (10%) are correctly labelled. Errors of commission and omission are also referred to as Type I and Type II errors, respectively.

omission

commission

Omission error is the corollary of producer accuracy, while user accuracy is the corollary of commission error. The user accuracy is the probability that a certain reference class has indeed actually been labelled as that class. Similarly, producer accuracy is the probability that a sampled point on the map is indeed that particular class.

user and producer accuracies

Another widely used measure of map accuracy derived from the error matrix is the kappa or κ coefficient. Let x_{ij} denote the element of the error matrix in row i and column j, r denote number of classes and N total sum of all elements of the error matrix. Then kappa coefficient is computed as

kappa

$$\kappa = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+} x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+} x_{+i}} \tag{6.1}$$

where $x_{i+} = \sum_{j=1}^{r} x_{ij}$ and $x_{+i} = \sum_{j=1}^{r} x_{ji}$ are the sums of all elements in row i and column i, respectively.

Kappa coefficient takes into account the fact that even assigning labels at random will result in a certain degree of accuracy. Kappa statistics, based on kappa coefficient, can

be applied to test if two data sets, e.g. classification results, have different levels of accuracy. This type of testing is used to evaluate different RS data or methods for the generation of spatial data.

6.2.5 Pixel-based and object-oriented classification

Pixel-based image classification is a powerful technique to derive *thematic classes* from multi-band images. However, it has certain limitations that users should be aware of. The most important constraints of pixel-based image classification are that it results in (i) spectral classes, and that (ii) each pixel is assigned to one class only.

Spectral classes are those that are directly linked to the spectral bands used in the classification. In turn, these are linked to surface characteristics. In that respect, one can say that spectral classes correspond to land cover classes. In the classification process a *spectral class* may be represented by several *training classes*. Among other things, this is due to the variability within a spectral class. Consider, for example, a class such as "grass": there are different types of grass, each of which has different spectral characteristics. Furthermore, the same type of grass may have different spectral characteristics when considered over larger areas, owing to, for example, different soils and climatic conditions.

spectral classes

A related issue is that sometimes one is interested in land use classes rather than land cover classes. Sometimes, a land use class may comprise several land cover classes. Table 6.3 gives some examples of links between spectral land cover and land use classes. Note that between two columns there can be 1-to-1, 1-to-n, and n-to-1 relationships. The 1-to-n relationships are a serious problem and can only be solved by adding data and/or knowledge to the classification procedure. The data added can be other remote sensing images (other bands, other moments) or existing geospatial data, such as topographic maps, historical land inventories, road maps, and so on. Usually this is done in combination with adding expert knowledge to the process. An example would be using historical land cover data and defining the probability of certain land cover changes. Another example would be to use elevation, slope and aspect information. This will prove especially useful in mountainous regions, where elevation differences play an important role in variations of surface-cover types.

Table 6.3
Spectral classes distinguished during classification can be aggregated to land cover classes. 1-to-n and n-to-1 relationships can exist between land cover and land use classes.

Spectral class	Land cover class	Land use class
water	water	shrimp cultivation
grass1	grass	nature reserve
grass2	grass	nature reserve
grass3	grass	nature reserve
bare soil	bare soil	nature reserve
trees1	forest	nature reserve
trees2	forest	production forest
trees3	forest	city park

The other main problem and limitation of pixel-based image classification is that each pixel is only assigned to one class. This is not a problem when dealing with relatively small ground resolution cells. However, when dealing with relatively large GRCs, more land cover classes are likely to occur within a cell. As a result, the value of the pixel is an average of the reflectance of the land cover present within the GRC. In a standard classification, these contributions cannot be traced back and the pixel will be assigned to one of either classes or perhaps even to another class. This phenomenon is usually referred to as the *mixed pixel*, or *mixel* (Figure 6.16). This problem of mixed pixels is inherent to image classification: assigning a pixel to one thematic class. The

mixed pixels

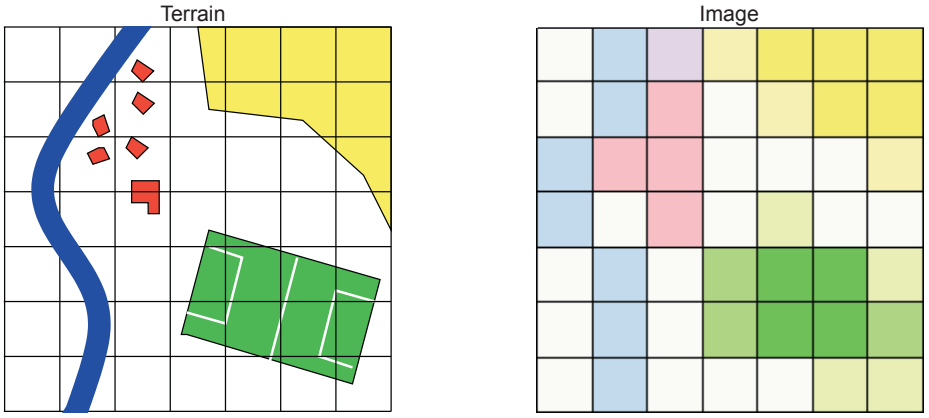


Figure 6.16
The origin of mixed pixels: different land cover types occur within one ground resolution cell. Note the relative abundance of mixed pixels.

solution to this is to use a different approach, for example, by assigning the pixel to more than one class. This brief introduction into the problem of mixed pixels also highlights the importance of using data with the appropriate spatial resolution.

We have seen that the choice of classification approach depends on the data available, but also on the knowledge we have about the area under investigation. Without knowledge of the land cover classes present, unsupervised classification can only give an overview of the variety of classes in an image. If knowledge is available, from field work or other sources, supervised classification may be superior. However, both methods only make use of spectral information, which becomes increasingly problematic for higher spatial resolutions. For example, a building that is made up of different materials leads to pixels with highly variable spectral characteristics, and thus a situation for which training of pixels is of little help. Similarly, a field may contain healthy vegetation pixels as well as some of bare soil.

object-oriented analysis

We are also increasingly interested in land use. However, to distinguish, for example, urban from rural woodland, or a swimming pool from a natural pond, an approach similar to visual interpretation (as described in Section 6.1) is needed. Object-oriented analysis (OOA), also called segmentation-based analysis, allows us to do that. Instead of trying to classify every pixel separately, and only based on spectral information, OOA breaks down an image into spectrally homogenous segments that correspond to fields, tree stands, buildings, etc. It is also possible to use auxiliary GIS layers, for example building footprints, to guide this segmentation. Similarly to the cognitive approach of visual image interpretation—where we consider each element in terms of its spectral appearance but also in terms of its shape and texture, and within its environment—in OOA we can then specify contextual relationships and more complex segment characteristics to classify the objects extracted in the segmentation process. For example, we can use object texture to distinguish two spectrally similar forest types, or distinguish a swimming pool from a pond, by considering its shape and perhaps the surrounding concrete instead of soil and vegetation. OOA is particularly suitable for images of high spatial resolution, but also for data obtained by ALS or microwave radar. It requires that we have substantial knowledge on what distinguishes a given land cover or land use type, as well as auxiliary data such as elevation, soil type or vector layers.