

VII

Surveying Engineering

Edward M. Mikhail
Purdue University

- 53 General Mathematical and Physical Concepts** *Edward M. Mikhail*
Coordinate Systems • Plane Geometry • Three-Dimensional Geometry • Vector Algebra • Matrix Algebra • Coordinate Transformations • Linearization of Nonlinear Functions • Map Projections • Observational Data Adjustment
- 54 Plane Surveying** *Steven D. Johnson and Wesley G. Crawford*
Introduction • Distance Measurement • Elevation Measurement • Angle Measurement • Plane Survey Computations • Horizontal Curves • Vertical Curves • Volume
- 55 Geodesy** *B. H. W. van Gelder*
Introduction • Coordinate Representations • Coordinate Frames Used in Geodesy and Some Additional Relationships • Mapping • Basic Concepts in Mechanics • Satellite Surveying • Gravity Field and Related Issues • Reference Systems and Datum Transformations
- 56 Photogrammetry and Remote Sensing** *J. S. Bethel*
Basic Concepts in Photogrammetry • Sensors and Platforms • Mathematics of Photogrammetry • Instruments and Equipment • Photogrammetric Products • Digital Photogrammetry • Photogrammetric Project Planning • Close-Range Metrology • Remote Sensing
- 57 Geographic Information Systems** *Jolyon D. Thurgood and J. S. Bethel*
Introduction • Geographic Information Components • Modeling Geographic Information • Building and Maintaining a GIS • Spatial Analysis • Information Extraction • Applications • Summary

SURVEYING IS ONE OF THE OLDEST ACTIVITIES of the civil engineer, and remains a primary component of civil engineering. It is also one field that continues to undergo phenomenal changes due to technological developments in digital imaging and satellite positioning. These modern surveying tools are not only revolutionizing regular surveying engineering tasks but are also impacting a myriad of applications in a variety of fields where near-real-time positioning is of great value.

Surveying and engineering are closely related professional activities. The area of surveying and mapping is in many countries a discipline by itself, and taken in total, it is almost as broad in scope as civil engineering. In the U.S., surveying engineering has been historically allied to civil engineering. Engineering surveying is defined as those activities involved in the planning and execution of surveys for the

location, design, construction, operation, and maintenance of civil and other engineered projects. Such activities include the preparation of survey and related mapping specifications; execution of photogrammetric and field surveys for the collection of required data, including topographic and hydrographic data; calculation, reduction, and plotting of survey data for use in engineering design; design and provision of horizontal and vertical control survey networks; provision of line and grade and other layout work for construction and mining activities; execution and certification of quality control measurements during construction; monitoring of ground and structural stability, including alignment observations, settlement levels, and related reports and certifications; measurement of material and other quantities for inventory, economic assessment, and cost accounting purposes; execution of as-built surveys and preparation of related maps and plans and profiles upon completion of construction; and analysis of errors and tolerances associated with the measurement, field layout, and mapping or other plots of survey measurements required in support of engineering projects. Engineering surveying may be regarded as a specialty within the broader professional practice of engineering and, with the exception of boundary, right-of-way, or other cadastral surveying, includes all surveying and mapping activities required to support the sound conception, planning, design, construction, maintenance, and operation of engineered projects. Engineering surveying does not include surveys for the retracement of existing land ownership boundaries or the creation of new boundaries.

Modern surveying engineering encompasses several specialty areas, each of which requires substantial knowledge and training in order to attain proper expertise. The most primary area perhaps is plane surveying because it is so widely applied in engineering and surveying practice. In plane surveying, we consider the fundamentals of measuring distance, angle, direction, and elevation. These measured quantities are then used to determine position, slope, area, and volume — the basic parameters of civil engineering design and construction. Plane surveying is applied in civil engineering projects of limited areal extent, where the effects of the earth's curvature are negligible relative to the positional accuracy required for the project.

Geodesy, or higher surveying, is an extensive discipline dealing with mathematical and physical aspects of modeling the size and shape of the earth, and its gravity field. Since the launch of earth-orbiting satellites, geodesy has become a truly three-dimensional science. Terrestrial and space geodetic measurement techniques, and particularly the relatively new technique of satellite surveying using the Global Positioning System (GPS), are applied in geodetic surveying. GPS surveying has not only revolutionized the art of navigation but has also brought about an efficient positioning technique for a variety of users, prominent among them the engineering community. GPS has had a profound impact on the fundamental problems of determining relative and absolute positions on the earth, including improvements in speed, timeliness, and accuracy. It is safe to say that any geometry-based data collection scheme profits to some degree from the full constellation of 24 GPS satellites. In addition to the obvious applications in geodesy, surveying, and photogrammetry, the use of GPS is applied in civil engineering areas such as transportation (truck and emergency vehicle monitoring, intelligent vehicle and highway systems, etc.) and structures (monitoring of deformation of structures such as water dams). Even in other areas such as forestry and agriculture (crop yield management) GPS provides the geometric backbone of modern (geographic) information systems.

Photogrammetry and remote sensing encompass all activities involved in deriving qualitative and quantitative information about objects and environments from their images. Such imagery may be acquired at close range, from aircraft, or from satellites. In addition to large-, medium-, and small-scale mapping, many other applications such as resource management and environmental assessment and monitoring rely on imageries of various types. Close-range applications include such tasks as accident reconstruction, mapping of complex piping systems, and shape determination for parabolic antennas. Large-scale mapping (including the capture of data on infrastructure) remains the primary civil engineering application of photogrammetry. Recent evolution toward working with digital imagery has brought about the increasing acceptance of the digital orthophoto to augment or supplant the planimetric map. Digital image processing tools offer the probability of great increases in mapping productivity through automation. For small- and medium-scale mapping, the increasing availability of satellite image

data offers an alternative to chemical photography. Commercially available satellite data with spatial resolutions of 1 to 3 meters, proposed for the near future, would have a profound impact on all mapping activities within civil engineering. Inclusion of GPS in photogrammetric and remote sensing acquisition platforms will lead to substantial improvements in accuracy, timeliness, and economy.

For centuries, maps have provided layered information in graphical form and have been used as legal documents and as tools to support decision making for applications such as urban planning. Recently, geographic information systems (GIS) have broadened the role played by all types of maps to encompass a total system of hardware, software, and procedures designed to capture, manage, manipulate, and produce information in a spatial context. GIS applications are broad indeed; they include land record management, base mapping, infrastructure maintenance, facilities management, and many others.

A driving force behind the move toward integrating mapping and other spatially oriented data has been the various utility industries and municipalities who need to plan and manage their infrastructure facilities and property assets. This automated mapping/facilities management, or AM/FM, concept is being used successfully today by many cities, counties, and utility industries, who may have embarked on the transition as much as 15 years ago. Successful practitioners of GIS can satisfy the needs of a broad spectrum of users with a single system, minimizing the duplication of resources required to support historically independent user groups.

All other components of surveying engineering contribute to the construction of a GIS. The range of survey methods, from classical to modern geodesic and space-based technologies, provide the required reference framework. Digital mapping provides an efficient technology to populate the GIS with spatial information. Remote sensing techniques applied to the earth and its environment provide the various thematic layers of information.

Scope of This Section of the Handbook

The scope of Section VII, Surveying Engineering, in this handbook is to present the reader with the basic information involved in the performance of different surveying engineering projects. As was mentioned earlier, this is a discipline of many areas, each of which will be covered in a separate chapter. The underlying mathematical concepts used by the different areas of surveying engineering are covered first in Chapter 53, followed by four chapters covering, in sequence, plane surveying, geodesy, photogrammetry and remote sensing, and geographic information systems. Of particular importance is the topic on measurements, their errors, and least squares adjustment of redundant data. Since surveying is fundamentally a measurement science, all phases are covered: preanalysis (design), data acquisition (observations), data preprocessing, data adjustment, and postadjustment analysis of the results (quality assessment). Each engineering surveying project must properly execute these phases.

53

General Mathematical and Physical Concepts

53.1 Coordinate Systems

Two-Dimensional Systems • Three-Dimensional Coordinate Systems

53.2 Plane Geometry

Distance between Two Points • Slope of a Line between Two Points • Azimuth • Bearing • Relation between Azimuth and Slope Angle • Internal and External Division of a Line Segment • Equation of a Line with Known X and Y Intercepts • General Equation of a Line • Lines Parallel to the Axes • Equation of a Line with Given Slope and Y Intercept • Equation of a Line with a Given Slope Passing through a Given Point • Equation of a Line Joining Two Points • Equation of a Line with Given Length and Slope of the Perpendicular from Origin • Perpendicular Distance from the Origin to a Line • Perpendicular Distance from a Point to a Line • Equation of a Line through a Point and Parallel to Another Line • Equation of a Line through a Point and Perpendicular to Another Line • Angle between Two Lines • Point of Intersection of Two Lines • Equation of a Circle • Intersection of a Line and a Circle • Areas

53.3 Three-Dimensional Geometry

Distance between Two Points • Equation of a Plane • Equation of a Straight Line • Equation of a Sphere

53.4 Vector Algebra

Definitions • Vector Operations • Planes and Lines

53.5 Matrix Algebra

Definition • Types of Matrices • Basic Matrix Operations • Matrix Inverse

53.6 Coordinate Transformations

Linear Transformations • Nonlinear Transformations

53.7 Linearization of Nonlinear Functions

One Function of Two Variables • Two Functions of One Variable • Two Functions of Two Variables Each • General Case of m Functions of n Variables • Differentiation of a Determinant • Differentiation of a Quotient

53.8 Map Projections

53.9 Observational Data Adjustment

Mathematical Model for Adjustment • Design/Prealysis • Data Acquisition • Data Preprocessing • Data Adjustment • Least Squares Adjustment • Techniques of Least Squares • Assessment of Adjustment Results • Confidence Region for Estimated Parameters • Applications in Surveying Engineering

Edward M. Mikhail

Purdue University

53.1 Coordinate Systems

Two-Dimensional Systems

Figure 53.1 depicts two commonly used coordinate systems, one polar (r, θ) and the other *Cartesian or rectangular* (x_1, x_2) . A point p can be located either by the angle θ , measured from the reference direction x_1 , and range r from the reference point 0, or by its two distances from two perpendicular axes, x_1, x_2 . The relationships between the two systems are given by

$$x_1 = r \cos \theta$$

$$x_2 = r \sin \theta \quad (53.1)$$

$$r = (x_1^2 + x_2^2)^{1/2}$$

$$\theta = \tan^{-1}(x_2/x_1) \quad (53.2)$$

Three-Dimensional Coordinate Systems

Figure 53.2 shows two systems of three-dimensional coordinates: *spherical* (α, β, r) and Cartesian or rectangular (x_1, x_2, x_3) . The Cartesian system depicted in Fig. 53.2 is *right-handed*, since a right-threaded screw rotated by an angle less than 90° from $+x_1$ to $+x_2$ would advance in the direction of $+x_3$. The relations between these two systems are as follows:

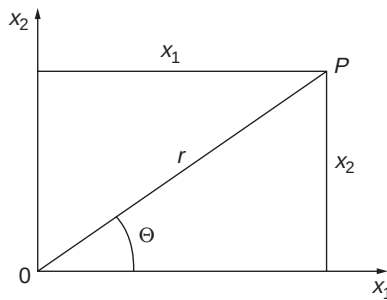


FIGURE 53.1 Two-dimensional coordinate systems.

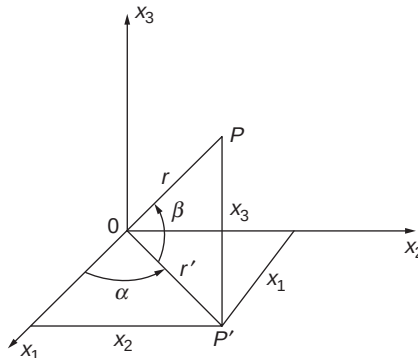


FIGURE 53.2 Three-dimensional right-handed Cartesian coordinate system.

$$\begin{aligned}
r' &= r \cos \beta \\
x_1 &= r' \cos \alpha \\
x_2 &= r' \sin \alpha \\
x_1 &= r \cos \alpha \cos \beta \\
x_2 &= r \sin \alpha \cos \beta \\
x_3 &= r \sin \beta \\
r &= (x_1^2 + x_2^2 + x_3^2)^{1/2} \\
\alpha &= \tan^{-1}(x_2/x_1) \\
\beta &= \sin^{-1}(x_3/r)
\end{aligned} \tag{53.3}$$

$$\begin{aligned}
\alpha &= \tan^{-1}(x_2/x_1) \\
\beta &= \sin^{-1}(x_3/r)
\end{aligned} \tag{53.4}$$

One example of spherical coordinates consists of latitude ϕ , longitude λ , and earth radius R , when the earth is considered as a sphere. A more accurate representation of the earth is an ellipsoid of revolution about its minor (shorter) axis. Coordinates referring to the earth ellipsoid, and other figures, are given in [Chapter 55](#), “Geodesy.”

Examples of rectangular systems include the geocentric and local space rectangular (LSR) coordinate systems. The geocentric system has its origin at the center of the earth, Z through the North Pole, X through the point of zero longitude, and the Y completing a right-handed system. The LSR varies with location, with its XY plane either tangent or secant to the ellipsoid and passing through a selected point λ_0 , ϕ_0 , and its Z along the local zenith.

53.2 Plane Geometry

Distance between Two Points

The distance between two points P_1 and P_2 , with coordinates X_1, Y_1 and X_2, Y_2 , is given by

$$L_{12} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{53.5}$$

Slope of a Line between Two Points

The slope of any line in a plane is the tangent of the angle it makes with the X axis. The slope angle is the counterclockwise angle from the $+X$ axis to the line in a *specified direction*. If the slope angle for the line P_1P_2 is θ_{12} , which is $<90^\circ$, and the slope angle for the line P_2P_1 is θ_{21} , which is $>180^\circ$, then

$$\theta_{21} = \theta_{12} + 180^\circ \tag{53.6}$$

The slope m_{12} of the line P_1P_2 is

$$m_{12} = \tan \theta_{12} = \frac{Y_2 - Y_1}{X_2 - X_1} \tag{53.7}$$

and that for the line P_2P_1 is

$$m_{21} = \tan \theta_{21} = \frac{Y_1 - Y_2}{X_1 - X_2} \quad (53.8)$$

Note that $m_{21} = \tan \theta_{21} = \tan(\theta_{12} + 180^\circ) = \tan \theta_{12}$.

Azimuth

Azimuth is a clockwise angle with a magnitude between 0 and 360°. It is measured either from North (or the +Y axis) or from South (or the −Y axis). Thus, for any line, α_N and α_S designate azimuths from North and from South, respectively. One azimuth angle is obtained from the other by simply adding 180° and dropping 360° whenever the sum exceeds 360°. Thus

$$\begin{aligned} \alpha_S &= \alpha_N + 180^\circ \\ \alpha_N &= \alpha_S + 180^\circ \end{aligned} \quad (53.9)$$

Bearing

Bearing is another form of expressing the direction of a line in surveying. It is always an *acute* angle, with a magnitude between 0 and 90°, and is a positive quantity. The bearing is the angle the line makes with either N (for North) or S (for South). The quadrant is indicated by specifying whether the angle is on the east or west side of the meridian (Y axis). Thus, the bearing angle is preceded by either *N* or *S* and succeeded by either *E* or *W*. If β designates a bearing, Table 53.1 shows how to convert bearing to azimuth. Table 53.2 shows how to convert azimuth to bearing.

TABLE 53.1 Conversion of Bearing to Azimuth

Bearing	α_N , deg	α_S , deg
<i>N</i> $\beta^\circ E$	β	$180 + \beta$
<i>S</i> $\beta^\circ E$	$180 - \beta$	$360 - \beta$
<i>S</i> $\beta^\circ W$	$\beta + 180$	β
<i>N</i> $\beta^\circ W$	$360 - \beta$	$180 - \beta$

Source: Tables 53.1 and 53.2 are from Anderson, J.M. and Mikhail, E.M. *Introduction to Surveying*, McGrawHill, Inc., New York, NY, 1985, p. 665. With permission.

TABLE 53.2 Conversion of Azimuth to Bearing

α_N , deg	α_S , deg	β		Bearing
0–90	180–270	α_N	or	$\alpha_S - 180^\circ$
90–180	270–360	$180^\circ - \alpha_N$	or	$360^\circ - \alpha_S$
180–270	0–90	$\alpha_N - 180^\circ$	or	α_S
270–360	90–180	$360^\circ - \alpha_N$	or	$180^\circ - \alpha_S$
				<i>N</i> $\beta^\circ E$
				<i>S</i> $\beta^\circ E$
				<i>S</i> $\beta^\circ W$
				<i>N</i> $\beta^\circ W$

Source: Tables 53.1 and 53.2 are from Anderson, J.M. and Mikhail, E.M. *Introduction to Surveying*, McGrawHill, Inc., New York, NY, 1985, p. 665. With permission.

Relation between Azimuth and Slope Angle

Considering the commonly used azimuth from North, α_N , the slope angle θ is obtained from

$$\theta = 90^\circ - \alpha_N \quad (53.10)$$

Internal and External Division of a Line Segment

Let points I and E divide a line segment in the proportion s_1/s_2 , where I is an internal and E is an external point. If (x_1, y_1) and (x_2, y_2) are the coordinates of the ends of the line segment, the coordinates of I and E are given by

$$\begin{aligned} X_I &= \frac{s_1 X_2 + s_2 X_1}{s_1 + s_2} \\ Y_I &= \frac{s_1 Y_2 + s_2 Y_1}{s_1 + s_2} \end{aligned} \quad (53.11)$$

$$\begin{aligned} X_E &= \frac{s_1 X_2 - s_2 X_1}{s_1 - s_2} \\ Y_E &= \frac{s_1 Y_2 - s_2 Y_1}{s_1 - s_2} \end{aligned} \quad (53.12)$$

Equation of a Line with Known X and Y Intercepts

$$\begin{aligned} \frac{X}{X_0} + \frac{Y}{Y_0} &= 1 \\ Y_0 X + X_0 Y &= X_0 Y_0 \end{aligned} \quad (53.13)$$

where X_0 and Y_0 are the X and Y intercepts, respectively.

General Equation of a Line

$$aX + bY + c = 0 \quad (53.14)$$

Since two points define a line, only two of the three coefficients a , b , c are independent. This can be shown by dividing by c , or

$$\frac{a}{c}X + \frac{b}{c}Y + 1 = 0 \quad \text{or} \quad a'X + b'Y + 1 = 0$$

The two intercepts are obtained from Eq. (53.14) as

$$X \text{ intercept} = -\frac{c}{a} \quad (53.15)$$

$$Y \text{ intercept} = -\frac{c}{b}$$

and the slope of the line is given by

$$m = \tan \theta = -\frac{a}{b} \quad (53.16)$$

Lines Parallel to the Axes

The equation of a line parallel to the X axis is simply

$$Y = k_1 \quad (53.17)$$

where k_1 is the distance of the line from the X axis. Similarly, the equation of a line parallel to the Y axis is

$$X = k_2 \quad (53.18)$$

where k_2 is the distance between the Y axis and the line.

Equation of a Line with Given Slope and Y Intercept

$$Y = mX + k \quad (53.19)$$

where m is the slope and k is the Y intercept.

Equation of a Line with a Given Slope Passing through a Given Point

$$Y - Y_p = m(X - X_p) \quad (53.20)$$

where m is the slope and X_p, Y_p are the coordinates of the point. In terms of azimuth α , the equation becomes

$$Y - Y_p = (X - X_p) \cot \alpha \quad (53.21)$$

Equation of a Line Joining Two Points

The equation of a line passing through two points P_1 and P_2 with coordinates X_1, Y_1 and X_2, Y_2 is given by

$$\frac{Y - Y_1}{X - X_1} = \frac{Y_2 - Y_1}{X_2 - X_1} \quad (53.22)$$

Equation of a Line with Given Length and Slope of the Perpendicular from Origin

$$X \cos \theta + Y \sin \theta = p \quad (53.23)$$

where p is the length of the perpendicular from the origin to the line, and θ is the angle it makes with the X axis.

Perpendicular Distance from the Origin to a Line

From the general form of the equation of a line, Eq. (53.14), the length of the perpendicular from the origin to the line is given by

$$p = \left| \frac{c}{\sqrt{a^2 + b^2}} \right| \quad (53.24)$$

Perpendicular Distance from a Point to a Line

$$s = \left| \frac{aX_1 + bY_1 + c}{\sqrt{a^2 + b^2}} \right| \quad (53.25)$$

where X_1, Y_1 are the coordinates of the point, and the line is given by the general equation $aX + bY + c = 0$.

Equation of a Line through a Point and Parallel to Another Line

$$X \cos \theta + Y \sin \theta = p + s \quad (53.26)$$

in which $X \cos \theta + Y \sin \theta = p$ is the equation of the given line and s is the perpendicular distance between the two lines. The value of s cannot be taken as its absolute value; its proper sign must be determined. This is done by realizing that the value of the left-hand side of Eq. (53.14) will always be positive for all points on one side of the line and negative for all points on the other side. (The value is of course zero for points falling on the line.) Thus,

$$\frac{a}{\sqrt{a^2 + b^2}}X + \frac{b}{\sqrt{a^2 + b^2}}Y = -\frac{c}{\sqrt{a^2 + b^2}} + \frac{aX_1 + bY_1 + c}{\sqrt{a^2 + b^2}}$$

which when clearing fractions becomes

$$aX + bY - (aX_1 + bY_1) = 0 \quad (53.27)$$

This is the equation sought, in which a, b belong to the given line and X_1, Y_1 are the coordinates of the given point.

Equation of a Line through a Point and Perpendicular to Another Line

The given line has the general equation $aX + bY + c = 0$, and the given point P has coordinates X_p, Y_p . The slope of the given line is

$$m = -\frac{a}{b}$$

The slope of the line perpendicular to the given line is b/a . Thus,

$$Y - Y_p = \frac{b}{a} (X - X_p)$$

or

$$bX - aY - (bX_p - aY_p) = 0 \quad (53.28)$$

Angle between Two Lines

The angle γ , between two lines is given by

$$\gamma = \theta_2 - \theta_1 \quad (53.29)$$

where θ_1 and θ_2 are the slope angles of the two lines. Then

$$\tan \gamma = \tan(\theta_2 - \theta_1)$$

or

$$\tan \gamma = \frac{\tan \theta_2 - \tan \theta_1}{1 + \tan \theta_1 \tan \theta_2}$$

With the line slopes $m_1 = \tan \theta_1$ and $m_2 = \tan \theta_2$,

$$\tan \gamma = \frac{m_2 - m_1}{1 + m_1 m_2} \quad (53.30)$$

If $\tan \gamma = 0$ or $m_1 = m_2$, the two lines are parallel. On the other hand, if $\tan \gamma = \infty$ or $m_1 m_2 = -1$, the two lines are perpendicular to each other.

Point of Intersection of Two Lines

If the point of intersection of the two lines is q , its coordinates X_q, Y_q satisfy their equations. Then, to get X_q, Y_q we simultaneously solve the two equations of the lines:

$$\begin{aligned} X_q &= \frac{c_1 b_2 - c_2 b_1}{a_1 b_2 - a_2 b_1} \\ Y_q &= \frac{c_1 a_2 - c_2 a_1}{a_1 b_2 - a_2 b_1} \end{aligned} \quad (53.31)$$

Equation of a Circle

Given a circle of radius r and center coordinates X_c, Y_c , its equation is

$$(X - X_c)^2 + (Y - Y_c)^2 = r^2 \quad (53.32)$$

If the circle's center is the origin of the coordinate system, $X_c = Y_c = 0$, its equation reduces to

$$X^2 + Y^2 = r^2 \quad (53.33)$$

Equation (53.32) may be expanded to the form

$$X^2 + Y^2 + 2dX + 2eY + f = 0 \quad (53.34)$$

which represents the general form of the equation of a circle. It contains three coefficients, d, e, f , which represent three geometric elements such as the radius and the two coordinates of its center.

Intersection of a Line and a Circle

In general, a straight line intersects a circle in two points. Given the equations of a circle and a line,

$$X^2 + Y^2 + 2dX + 2eY + f = 0$$

$$Y = mX + k$$

we can eliminate Y and get a general *quadratic* equation in X ,

$$AX^2 + BX + C = 0$$

Its two roots are in general given by

$$X_1 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \quad \text{and} \quad X_2 = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$$

for each of which a value for r is obtained from the equation of the line.

In addition to the case of two points of intersection, two other situations are possible, depending upon the quantity under the radical, $B^2 - 4AC$. If this quantity is zero, or $B^2 = 4AC$, then $X_1 = X_2$ and the line is tangent to the circle at one point. In the second situation, $B^2 < 4AC$, the quantity is negative, which means that one cannot take the square root. (This is usually referred to as the imaginary solution to the quadratic equation.) In this case, the line misses the circle and no intersection takes place.

Areas

Circle.

$$A = \pi r^2 = \frac{\pi d^2}{4} \quad (53.35)$$

where r and d are the radius and diameter of the circle, respectively.

Sector.

$$A = \frac{1}{2}ra = \frac{\pi r^2 \theta^\circ}{360^\circ} \quad (53.36)$$

where r is the radius, a is the arc length, and θ is the angle at the center in degrees.

Segment (Less Than a Semicircle).

$$A = \frac{1}{2}ra - \frac{1}{2}r^2 \sin \theta \quad (53.37)$$

or

$$A = \frac{\pi r^2 \theta^\circ}{360^\circ} - \frac{1}{2}r^2 \sin \theta \quad (53.38)$$

where r , a , and θ are as defined above. If the chord length c is given, then

$$\sin \frac{\theta}{2} = \frac{c}{2r} \quad (53.39)$$

Triangle. For a *general* triangle, the area T is given by

$$T = \frac{1}{2}bh \quad (h \text{ is perpendicular to } b) \quad (53.40)$$

or

$$T = \frac{1}{2}bc \sin A \quad (53.41)$$

or

$$T = \sqrt{s(s-a)(s-b)(s-c)} \quad (53.42)$$

with

$$s = \frac{1}{2}(a + b + c) \quad (53.43)$$

where a, b, c are the lengths of the sides, h is the height of the triangle (which is perpendicular to the base b), and A, B, C are the interior angles opposite to the side lengths a, b, c , respectively. Other useful relations for a plane triangle are

$$A + B + C = 180^\circ \quad (53.44a)$$

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c} \quad (53.44b)$$

$$a^2 = b^2 + c^2 - 2bc \cos A \quad (53.44c)$$

$$b^2 = c^2 + a^2 - 2ac \cos B \quad (53.44d)$$

$$c^2 = a^2 + b^2 - 2ab \cos C \quad (53.44e)$$

$$\tan \frac{A}{2} = \frac{1}{s-a} \sqrt{\frac{(s-a)(s-b)(s-c)}{s}} \quad (53.44f)$$

$$\cos \frac{A}{2} = \sqrt{\frac{s(s-a)}{bc}} \quad (53.44g)$$

$$\cos \frac{B}{2} = \sqrt{\frac{s(s-b)}{ca}} \quad (53.44h)$$

$$\cos \frac{C}{2} = \sqrt{\frac{s(s-c)}{ab}} \quad (53.44i)$$

$$\frac{a-b}{a+b} = \frac{\tan(A-B)/2}{\tan(A+B)/2} \quad (53.44j)$$

$$h = \frac{2}{b} \sqrt{s(s-a)(s-b)(s-c)} \quad (53.44k)$$

For an *equilateral* triangle, where sides $a = b = c$ and angles $A = B = C = 60^\circ$, then the area T becomes

$$T = \frac{a^2 \sqrt{3}}{4} \quad \left(\text{with } h = \frac{a\sqrt{3}}{2} \right) \quad (53.45)$$

Square.

$$A = a^2 \quad (53.46)$$

where a is the side length.

Rectangle.

$$A = ab \quad (53.47)$$

where a and b are its width and length.

Parallelogram. Let a and b be the sides, h the altitude upon side b , C the acute angle, and A the area. Then

$$A = bh = ab \sin C \quad (53.48)$$

Trapezoid. If b_1 and b_2 are the parallel sides, and h is the altitude between them, then the area A is

$$A = \frac{1}{2}h(b_1 + b_2) \quad (53.49)$$

For an *isosceles* trapezoid, if a is the length of one of the two nonparallel sides, and C is the acute angle between a and b_2 , then

$$A = \frac{1}{2}a(b_1 + b_2)\sin C \quad (53.50)$$

53.3 Three-Dimensional Geometry

Distance between Two Points

The distance between two points (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) is

$$d_{12} = [(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2]^{1/2} \quad (53.51)$$

Equation of a Plane

The general equation of a plane is given by

$$AX + BY + CZ + D = 0 \quad (53.52)$$

Only three of the four coefficients A, B, C, D are independent, since we can divide by D and get

$$\frac{A}{D}X + \frac{B}{D}Y + \frac{C}{D}Z + 1 = 0$$

or

$$EX + FY + GZ + 1 = 0 \quad (53.53)$$

Three noncollinear points determine a plane by writing three linear equations and solving them for E, F, G . When D in Eq. (53.52) is zero, or when no 1 is in Eq. (53.53), the plane passes through the origin. (For vector representation of a plane, see “Planes and Lines” in Section 53.4.)

Equation of a Straight Line

Since a straight line is the intersection of two planes, and since a plane is expressed by one linear equation, a straight line in three-dimensional space is represented by two linear equations. The two equations of a line passing through two points (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) are given by

$$\frac{X - X_1}{X_2 - X_1} = \frac{Y - Y_1}{Y_2 - Y_1} = \frac{Z - Z_1}{Z_2 - Z_1} \quad (53.54)$$

For vector representation of a straight line, see “Planes and Lines” in the following section.

Equation of a Sphere

If (X_c, Y_c, Z_c) represents the center of the sphere and R is its radius, its equation is given by

$$(X - X_c)^2 + (Y - Y_c)^2 + (Z - Z_c)^2 = R^2 \quad (53.55)$$

53.4 Vector Algebra

Definitions

A *vector* is an entity which has a magnitude and direction. In two- and three-dimensional spaces, it is a directed line segment from one point to another. The projections of the vector on the x_1, x_2 , and x_3 axes are a_1, a_2 , and a_3 and are called the vector components. It is represented by a column:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

The *length* of the vector is designated by $|\mathbf{a}|$ and is given by

$$|\mathbf{a}| = (a_1^2 + a_2^2 + a_3^2)^{1/2} \quad (53.56)$$

A vector's *direction* is given either by the angles α, β, γ it makes with the axes or by their cosines. The latter are called *direction cosines* and are given by

$$\cos \alpha = \frac{a_1}{|\mathbf{a}|} \quad \cos \beta = \frac{a_2}{|\mathbf{a}|} \quad \cos \gamma = \frac{a_3}{|\mathbf{a}|} \quad (53.57)$$

It is evident that

$$\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1 \quad (53.58)$$

Generalizing a vector to n dimensions, we write

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Vector Operations

Equality.

$$\mathbf{a} = \mathbf{b} \quad \text{when } a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$$

Addition/Subtraction.

$$\mathbf{c} = \mathbf{a} \pm \mathbf{b} \quad \text{or} \quad c_1 = a_1 \pm b_1, c_2 = a_2 \pm b_2, \dots, c_n = a_n \pm b_n$$

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$$

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$$

Multiplication by a Scalar. A *scalar* is a quantity which has magnitude but no direction, such as mass, temperature, time, etc., and will be designated by a lowercase Greek letter.

$$\lambda \mathbf{a} = \begin{bmatrix} \lambda a_1 \\ \lambda a_2 \\ \vdots \\ \lambda a_n \end{bmatrix}$$

$$\lambda(\mu \mathbf{a}) = (\lambda\mu)\mathbf{a} = \mu(\lambda \mathbf{a}) \quad (53.59)$$

$$(\lambda + \mu)\mathbf{a} = \lambda \mathbf{a} + \mu \mathbf{a}$$

$$\lambda(\mathbf{a} + \mathbf{b}) = \lambda \mathbf{a} + \lambda \mathbf{b}$$

$$|\lambda \mathbf{a}| = \lambda |\mathbf{a}|$$

Any vector $\mathbf{a}$ is reduced to a unit vector $\mathbf{a}^\circ$ when dividing its components by its length, which is a scalar, or $\mathbf{a}^\circ = \mathbf{a}/|\mathbf{a}|$. The components of $\mathbf{a}^\circ$ are the direction cosines of $\mathbf{a}$. Unit vectors along the coordinate axes are called base or basis vectors and are given by

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (53.60)$$

Any vector in 3-space is uniquely expressed as

$$\mathbf{a} = a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k} \quad (53.61)$$

The right-handed system introduced in Section 53.1 can be generalized for three vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. If they are not coplanar, and they have the same initial point, then they are said to form a right-handed system if a right-threaded screw rotated through an angle *less than* 180° from $\mathbf{a}$ to $\mathbf{b}$ would advance in the direction $\mathbf{c}$.

Vector Products

Dot (or Scalar) Product.

$$\mathbf{a} \cdot \mathbf{b} = \sum_{p=1}^n a_p b_p = a_1 b_1 + a_2 b_2 + \dots + a_n b_n \quad (53.62)$$

This is also called the *inner product*. It is a scalar and has the following properties:

$$\begin{aligned}
\mathbf{a} \cdot \mathbf{b} &= \mathbf{b} \cdot \mathbf{a} \\
\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} \\
\lambda(\mathbf{a} \cdot \mathbf{b}) &= (\lambda \mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (\lambda \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b})\lambda \\
\mathbf{i} \cdot \mathbf{i} &= \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = 1 \\
\mathbf{i} \cdot \mathbf{j} &= \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0
\end{aligned} \tag{53.63}$$

The dot product of a vector with itself is equal to the square of its length, or

$$\mathbf{a} \cdot \mathbf{a} = a_1^2 + a_2^2 + \cdots a_n^2 = |\mathbf{a}|^2 \tag{53.64}$$

If θ is the angle between two vectors $\mathbf{a}$ and $\mathbf{b}$ (in two- or three-dimensional space), it can be shown that

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \theta \tag{53.65}$$

It follows that if $\mathbf{a}$ is perpendicular to $\mathbf{b}$, then $\mathbf{a} \cdot \mathbf{b} = 0$.

Cross (or Vector) Product. $\mathbf{a} \times \mathbf{b}$ (read “ $\mathbf{a}$ cross $\mathbf{b}$ ”) is another vector $\mathbf{c}$, which is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ and in a direction such that $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ (in this order) form a right-handed system. The length of $\mathbf{c}$ is given by

$$|\mathbf{c}| = |\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin \theta \tag{53.66}$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. This quantity is the area of the parallelogram determined by $\mathbf{a}$ and $\mathbf{b}$. If $\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$, and $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$, then $\mathbf{c}$ is given by the determinant

$$\mathbf{c} = \mathbf{a} \times \mathbf{b} = \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \tag{53.67}$$

It has the following properties

$$\begin{aligned}
\mathbf{a} \times \mathbf{b} &= -(\mathbf{b} \times \mathbf{a}) \\
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \quad (\text{observing the order}) \\
\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) &= 0 \\
|\mathbf{a} \times \mathbf{b}|^2 &= |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\
\mathbf{i} \times \mathbf{i} &= \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = 0 \\
\mathbf{i} \times \mathbf{j} &= \mathbf{k}; \mathbf{j} \times \mathbf{k} = \mathbf{i}; \mathbf{k} \times \mathbf{i} = \mathbf{j}
\end{aligned} \tag{53.68}$$

For two nonzero vectors, if $\mathbf{a} \times \mathbf{b} = \mathbf{0}$, then $\mathbf{a}$ and $\mathbf{b}$ are parallel.

Scalar Triple Product.

$$\mathbf{a} \times \mathbf{b} \cdot \mathbf{c} = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} \tag{53.69}$$

is a scalar which is equal to the volume of the parallelepiped determined by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. If it is zero, then the three vectors are coplanar. It has the following properties:

$$\begin{aligned}
\mathbf{a} \times \mathbf{b} \cdot \mathbf{c} &= \mathbf{b} \times \mathbf{c} \cdot \mathbf{a} = \mathbf{c} \times \mathbf{a} \cdot \mathbf{b} \\
\mathbf{a} \times \mathbf{b} \cdot \mathbf{c} &= \mathbf{a} \cdot \mathbf{b} \times \mathbf{c}
\end{aligned}$$

Planes and Lines

If $\mathbf{p}_0$ is a given point in a plane, $\mathbf{n}$ is a nonzero vector normal to the plane, and $\mathbf{p}$ is any point in the plane, then the equation of the plane takes the form

$$(\mathbf{p} - \mathbf{p}_0) \cdot \mathbf{n} = 0 \quad \text{or} \quad \mathbf{p} \cdot \mathbf{n} - \mathbf{p}_0 \cdot \mathbf{n} = 0 \quad (53.70)$$

Let $\mathbf{n} = A\mathbf{i} + B\mathbf{j} + C\mathbf{k}$, $\mathbf{p}_0 = X_0\mathbf{i} + Y_0\mathbf{j} + Z_0\mathbf{k}$, and $\mathbf{p} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$. Then Eq. (53.70) becomes

$$A(X - X_0) + B(Y - Y_0) + C(Z - Z_0) = 0 \quad (53.71)$$

or

$$AX + BY + CZ + D = 0$$

where $D = -(AX_0 + BY_0 + CZ_0)$. Two planes are parallel when they have a common normal vector $\mathbf{n}$, and are perpendicular when their normals are, or $\mathbf{n}_1 \cdot \mathbf{n}_2 = 0$.

If $\mathbf{p}_0$ represents a given point on a line, $\mathbf{p}$ any other point on the line, and $\mathbf{v}$ is a given nonzero vector parallel to the line, then

$$\mathbf{p} = \mathbf{p}_0 + \lambda \mathbf{v} \quad (53.72)$$

is an equation of the line. In component form, it yields three scalar equations describing the parametric form (λ is the *running parameter*):

$$\begin{aligned} X &= X_0 + \lambda v_x \\ Y &= Y_0 + \lambda v_y \\ Z &= Z_0 + \lambda v_z \end{aligned} \quad (53.73)$$

If λ is eliminated, one gets the usual two-equation form of a straight line in space; see Eq. (53.54).

53.5 Matrix Algebra

Definition

A *matrix* is a group of numbers or scalar functions collected in two-dimensional (rectangular) array. A matrix is designated by a boldface capital Roman letter. Thus, an $m \times n$ matrix can be symbolically written as

$${}_{m,n}\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Types of Matrices

Square Matrix. This is a matrix in which the number of rows equals the number of columns. In this case, ${}_{m,m}\mathbf{A}$ is a square matrix of order m . The *principal* (or *main*) diagonal of a square matrix is composed of all elements a_{ij} for which $i = j$.

Row Matrix.

$${}_{1,n}\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

Column Matrix or Vector.

$$\mathbf{b}_{m,1} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Diagonal Matrix.

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & d_{mm} \end{bmatrix}$$

That is, $d_{ij} = 0$ for all $i \neq j$.

Scalar Matrix.

$$\mathbf{A} = \begin{bmatrix} a & 0 & \cdots & 0 \\ 0 & a & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a \end{bmatrix}$$

$$\begin{aligned} a_{ij} &= 0 && \text{for all } i \neq j \\ a_{ij} &= a && \text{for all } i = j \end{aligned}$$

Identity or Unit Matrix.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\begin{aligned} a_{ij} &= 0 && \text{for all } i \neq j \\ a_{ij} &= 1 && \text{for all } i = j \end{aligned}$$

Null Matrix. A null or zero matrix is a matrix whose elements are all zero. It is denoted by a boldface zero, $\mathbf{0}$.

Upper Triangular Matrix.

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ 0 & u_{22} & & u_{2m} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & u_{mm} \end{bmatrix}$$

with $u_{ij} = 0$ for $i > j$.

Lower Triangular Matrix.

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix}$$

where $l_{ij} = 0$ for $i > j$.

Basic Matrix Operations

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are equal if they are of the same dimensions and each element $a_{ij} = b_{ij}$ for all i and j . The sum of two matrices $\mathbf{A}$ and $\mathbf{B}$ is possible only if they are of equal dimensions, and the elements of the resulting matrix $\mathbf{C}$ are $c_{ij} = a_{ij} + b_{ij}$ for all i, j . The following relations apply to addition (and subtraction) of matrices:

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \mathbf{B} + \mathbf{A} \\ \mathbf{A} + (\mathbf{B} + \mathbf{C}) &= (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + \mathbf{B} + \mathbf{C} \\ \mathbf{A} + (-\mathbf{A}) &= \mathbf{0} \end{aligned} \tag{53.74}$$

with $\mathbf{0}$ being the zero or null matrix, and $-\mathbf{A}$ is the matrix composed of $-a_{ij}$ as elements.

Multiplication of a matrix by a scalar α results in another $\mathbf{B} = \alpha\mathbf{A}$ whose elements are $b_{ij} = \alpha a_{ij}$ for all i and j .

The following relations hold for scalar multiplication (λ, μ are scalars):

$$\begin{aligned} \lambda(\mathbf{A} + \mathbf{B}) &= \lambda\mathbf{A} + \lambda\mathbf{B} \\ (\lambda + \mu)\mathbf{A} &= \lambda\mathbf{A} + \mu\mathbf{A} \\ \lambda(\mathbf{AB}) + (\lambda\mathbf{A})\mathbf{B} &= \mathbf{A}(\lambda\mathbf{B}) \\ \lambda(\mu\mathbf{A}) &= (\lambda\mu)\mathbf{A} \end{aligned} \tag{53.75}$$

The product of two matrices is another matrix. The two matrices must be *conformable for multiplication*, i.e., the number of columns of the first matrix must equal the number of rows of the second matrix. Thus, if $\mathbf{A}$ is an $m \times q$ matrix and $\mathbf{B}$ is a $q \times n$ matrix, the product $\mathbf{AB}$, *in that order*, is another matrix $\mathbf{C}$ with m rows (as in $\mathbf{A}$) and n columns (as in $\mathbf{B}$). Each element c_{ij} in $\mathbf{C}$ is obtained by multiplying each one of the q elements in the i th row in $\mathbf{A}$ by the corresponding element in the j th column in $\mathbf{B}$ and adding. Algebraically, this is written as

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^q a_{ik}b_{kj} \tag{53.76}$$

To illustrate matrix multiplication:

$$\begin{aligned} \mathbf{C}_{2,1} &= \mathbf{A}_{2,3} \mathbf{B}_{3,1} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} (1 \times 1) + (0 \times 5) + (2 \times 3) \\ (2 \times 1) + (1 \times 5) + (0 \times 3) \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \end{bmatrix} \end{aligned}$$

Matrix multiplication is not commutative; that is, in general $\mathbf{FG} \neq \mathbf{GF}$, even if the dimensions of the matrices allow multiplication in both directions (e.g., $m \times n$ and $n \times m$, or square matrices). The following relationships regarding matrix multiplication hold:

$\mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A}$ in which $\mathbf{I}$ is the unit or identity matrix

$\mathbf{AB} \neq \mathbf{BA}$

$$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} = \mathbf{ABC} \quad (\text{associative law}) \quad (53.77)$$

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \quad (\text{distributive laws})$$

$$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} \quad (\text{distributive laws})$$

$\mathbf{AB} = \mathbf{0}$ is possible without either $\mathbf{A}$ or $\mathbf{B}$ equaling $\mathbf{0}$. Also, $\mathbf{AB} = \mathbf{AC}$ does not imply $\mathbf{B} = \mathbf{C}$.

The transpose of the $m \times n$ matrix $\mathbf{A}$ is an $n \times m$ matrix formed from $\mathbf{A}$ by interchanging rows and columns such that the i th row of $\mathbf{A}$ becomes the i th column of the transposed matrix. We denote the transpose of $\mathbf{A}$ by $\mathbf{A}^T$. If $\mathbf{B} = \mathbf{A}^T$, it follows that $b_{ij} = a_{ji}$ for all i and j . The following relationships apply to the transpose of a matrix:

$$\begin{aligned} (\mathbf{A} + \mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\ (\mathbf{AB})^T &= \mathbf{B}^T \mathbf{A}^T \quad (\text{note reverse order}) \\ (\alpha \mathbf{A})^T &= \alpha \mathbf{A}^T \\ (\mathbf{A}^T)^T &= \mathbf{A} \end{aligned} \quad (53.78)$$

A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A}^T = \mathbf{A}$. Diagonal, scalar, and identity matrices are symmetric, since each is equal to its transpose. For any matrix $\mathbf{A}$ (not necessarily square), both $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric. If $\mathbf{B}$ is a symmetric matrix of suitable dimensions, then for any matrix $\mathbf{A}$, both $\mathbf{ABA}^T$ and $\mathbf{A}^T\mathbf{BA}$ are also symmetric.

If $\mathbf{a}$ is a column matrix (or vector), then $\mathbf{a}^T\mathbf{a}$ is a positive scalar which is equal to the sum of the squares of its elements; for example, the square of the vector's length.

A square matrix $\mathbf{A}$ is *skew-symmetric* if $\mathbf{A}^T = -\mathbf{A}$ and $a_{ij} = -a_{ji}$ for all i, j . For any square matrix $\mathbf{A}$, the matrix $(\mathbf{A} + \mathbf{A}^T)$ is symmetric and $(\mathbf{A} - \mathbf{A}^T)$ is skew-symmetric.

The *trace of a square matrix* is the scalar which is equal to the sum of its main diagonal elements. It is denoted by $\text{tr}(\mathbf{A})$; thus $\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn}$. The following are properties of the trace:

$$\begin{aligned} \text{tr}(\mathbf{A}) &= \text{tr}(\mathbf{A}^T) \\ \text{tr}(\lambda \mathbf{A}) &= \lambda \text{tr}(\mathbf{A}^T) \\ \text{tr}(\mathbf{A} + \mathbf{B}) &= \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}) \\ \text{tr}(\mathbf{AB}) &= \text{tr}(\mathbf{BA}) \\ \text{tr}(\mathbf{FAF}^{-1}) &= \text{tr}(\mathbf{A}) \quad (\mathbf{F} \text{ nonsingular matrix}) \end{aligned} \quad (53.79)$$

Matrix Inverse

Division of matrices is not defined. Instead, the *inverse of a square matrix* $\mathbf{A}$, if it exists, is the unique matrix $\mathbf{A}^{-1}$ with the following property:

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \quad (53.80)$$

where $\mathbf{I}$ is the identity matrix.

The properties of the inverse are

$$\begin{aligned} (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \quad (\text{note reverse order}) \\ (\mathbf{A}^{-1})^{-1} &= \mathbf{A} \\ (\mathbf{A}^T)^{-1} &= (\mathbf{A}^{-1})^T \\ (\lambda \mathbf{A})^{-1} &= \frac{1}{\lambda} \mathbf{A}^{-1} \end{aligned} \quad (53.81)$$

A square matrix which has an inverse is called *nonsingular*, whereas a matrix which does not have an inverse is called *singular*.

It was stated previously that $\mathbf{AB}$ can equal $\mathbf{0}$ without either $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$. If, however, either $\mathbf{A}$ or $\mathbf{B}$ is nonsingular, then the other matrix must be a null matrix. Hence, the product of two nonsingular matrices cannot be a null or zero matrix.

Associated with each *square matrix* $\mathbf{A}$ is a unique scalar called the *determinant* of $\mathbf{A}$. It is denoted either by $\det \mathbf{A}$ or by $|\mathbf{A}|$. Thus, for

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

the determinant is expressed as

$$|\mathbf{A}| = \begin{vmatrix} 3 & 1 \\ 1 & 2 \end{vmatrix}$$

The determinant of order n (for an $n \times n$ square matrix) can be defined in terms of determinants of order $n - 1$ and less. The determinant of a 1×1 matrix is defined as the value of that one element, i.e., for $\mathbf{A} = [a_{11}]$, $|\mathbf{A}| = \det \mathbf{A} = a_{11}$.

If $\mathbf{A}$ is an $n \times n$ matrix, and one row and one column of $\mathbf{A}$ are deleted, the resulting matrix is an $(n - 1) \times (n - 1)$ *submatrix* of $\mathbf{A}$. The determinant of such a submatrix is called a *minor* of $\mathbf{A}$, and it is designated by m_{ij} , where i and j correspond to the deleted row and column, respectively. More specifically, m_{ij} is known as the *minor of the element* a_{ij} in $\mathbf{A}$. Thus, each element of $\mathbf{A}$ has a minor.

The *cofactor* c_{ij} of an element a_{ij} is defined as

$$c_{ij} = (-1)^{i+j} m_{ij} \quad (53.82)$$

The determinant of an $n \times n$ matrix $\mathbf{A}$ can now be defined as

$$|\mathbf{A}| = a_{11}c_{11} + a_{12}c_{12} + \cdots + a_{1n}c_{1n} \quad (53.83)$$

which states that the determinant of $\mathbf{A}$ is the sum of the products of the elements of the first row of $\mathbf{A}$ and their corresponding cofactors. (It is equally possible to define $|\mathbf{A}|$ in terms of any other row or column, but for simplicity we used the first row.) On the basis of this definition, the 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

has cofactors $c_{11} = |a_{22}| = a_{22}$ and $c_{12} = -|a_{21}| = -a_{21}$, and the determinant of $\mathbf{A}$ is

$$|\mathbf{A}| = a_{11}c_{11} + a_{12}c_{12} = a_{11}a_{22} - a_{12}a_{21}$$

The *cofactor matrix* $\mathbf{C}$ of a matrix $\mathbf{A}$ is the square matrix of the same order as $\mathbf{A}$ in which each element a_{ij} is replaced by its cofactor c_{ij} .

The *adjoint matrix* of $\mathbf{A}$, denoted by $\text{adj } \mathbf{A}$, is the transpose of its cofactor matrix, i.e.,

$$\text{adj } \mathbf{A} = \mathbf{C}^T \quad (53.84)$$

It can be shown that

$$\mathbf{A}(\text{adj } \mathbf{A}) = (\text{adj } \mathbf{A})\mathbf{A} = |\mathbf{A}|\mathbf{I} \quad (53.85)$$

Comparison of Eqs. (53.80) and (53.85) leads directly to a procedure for evaluating the inverse from the adjoint matrix, namely,

$$\mathbf{A}^{-1} = \frac{\text{adj } \mathbf{A}}{|\mathbf{A}|} \quad (53.86)$$

It is easy to show that for a 2×2 matrix, the adjoint matrix is simply

$$\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

A square matrix is called orthogonal if its inverse is equal to its transpose, or $\mathbf{A}^{-1} = \mathbf{A}^T$. Thus, a matrix $\mathbf{M}$ is orthogonal when

$$\mathbf{M}^T \mathbf{M} = \mathbf{M} \mathbf{M}^T = \mathbf{I} \quad (53.87)$$

The columns of an orthogonal matrix are mutually orthogonal vectors of unit length. Also,

$$|\mathbf{M}| = \pm 1 \quad (53.88)$$

when $|\mathbf{M}| = +1$, $\mathbf{M}$ is called “proper orthogonal”; otherwise it is termed “improper orthogonal.” The product of two orthogonal matrices is also an orthogonal matrix.

Matrix Inverse by Partitioning

Let $\mathbf{A}$ be an $n \times n$ square nonsingular matrix whose inverse is to be evaluated. We *partition* $\mathbf{A}$ in the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{matrix} s & m \\ m & m \end{matrix}$$

where $\mathbf{A}_{11}$ is $s \times s$, $\mathbf{A}_{12}$ is $s \times m$, $\mathbf{A}_{21}$ is $m \times s$, $\mathbf{A}_{22}$ is $m \times m$, and $m + s = n$. The inverse $\mathbf{A}^{-1}$ exists, and we shall denote it, in the correspondingly partitioned form, by

$$\mathbf{A}^{-1} = \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

From the basic definition of an inverse we have $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}\mathbf{B} = \mathbf{I}$, or in the partitioned form,

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix}$$

which, when multiplied out, leads to four matrix equations in the four $\mathbf{B}_{ij}$ submatrices as unknowns, the solution of which, when $\mathbf{A}_{11}^{-1}$ exists, is given by

$$\begin{aligned} \mathbf{B}_{11} &= \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{B}_{21} \\ \mathbf{B}_{12} &= -\mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{B}_{22} \\ \mathbf{B}_{21} &= -\mathbf{B}_{22} \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \\ \mathbf{B}_{22} &= (\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12})^{-1} \end{aligned} \quad (53.89)$$

Alternatively, when $\mathbf{A}_{22}^{-1}$ exists, the solution is

$$\begin{aligned}\mathbf{B}_{11} &= [\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} \\ \mathbf{B}_{12} &= -\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{B}_{21} &= -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11} \\ \mathbf{B}_{22} &= \mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{12}\end{aligned}\tag{53.90}$$

If $\mathbf{A}$ is originally a symmetric matrix, then $\mathbf{A}_{21} = \mathbf{A}_{12}^T$ correspondingly $\mathbf{B}_{21} = \mathbf{B}_{11}^T$.

The *rank* of a matrix is the order of the largest nonzero determinant that can be formed from the elements of the matrix by appropriate deletion of rows or columns (or both). Thus, a matrix is said to be of *rank* m if and only if it has *at least one nonsingular submatrix of order* m , but has no nonsingular submatrix of order more than m . A nonsingular matrix of order n has a rank n . A matrix with zero rank has elements that must all be zero.

The inverse $\mathbf{A}^{-1}$ is defined only for square matrices, and it exists when the rank of $\mathbf{A}$ is equal to its order. A more general inverse may be defined for rectangular matrices with arbitrary rank. It is called the generalized inverse, denoted by $\mathbf{A}^-$, and satisfies the relation

$$\mathbf{A}\mathbf{A}^- = \mathbf{A}\tag{53.91}$$

This condition is not sufficient to define a unique $\mathbf{A}^-$. Additional conditions may be imposed on $\mathbf{A}^-$, such as

$$\begin{aligned}\mathbf{A}^- \mathbf{A} \mathbf{A}^- &= \mathbf{A}^- \\ (\mathbf{A} \mathbf{A}^-)^T &= \mathbf{A} \mathbf{A}^- \\ (\mathbf{A}^- \mathbf{A})^T &= \mathbf{A}^- \mathbf{A}\end{aligned}\tag{53.92}$$

If we impose all four conditions in Eqs. (53.91) and (53.92), the inverse is called the *pseudo inverse* or the Moore–Penrose inverse, and is denoted by $\mathbf{A}^+$.

The Eigenvalue Problem

For a square matrix $\mathbf{A}$ of order n , we seek a nonzero vector $\mathbf{x}$ and a scalar λ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}\tag{53.93}$$

which is called the “eigenvalue problem.” A solution λ_0 and $\mathbf{x}_0$ to this problem is called an *eigenvalue* (proper value, characteristic value) and the corresponding *eigenvector* (proper vector, characteristic vector) of the matrix $\mathbf{A}$. An eigenvector, if one exists, can be determined only to a scalar multiplication, for if $\lambda_0, \mathbf{x}_0$ satisfy Eq. (53.93), then $\lambda_0, \alpha\mathbf{x}_0$, where α is an arbitrary scalar, will also.

Equation (53.93) can be rewritten as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}\tag{53.94}$$

which represents a set of homogenous linear equations. For a nontrivial solution to this set the following condition must be satisfied:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0\tag{53.95}$$

Equation (53.95) represents a real polynomial equation of degree n :

$$b_n(-\lambda)^n + b_{n-1}(-\lambda)^{n-1} + \dots + b_0 = 0\tag{53.96}$$

where

$$\begin{aligned}
 b_n &= 1 \\
 b_{n-1} &= a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^n a_{ii} = \text{tr}(\mathbf{A}) = \text{trace of } \mathbf{A} \\
 &\vdots \\
 b_{n-r} &= \text{sum of all principal minors of order } r \text{ of } \mathbf{A} \\
 &\vdots \\
 b_0 &= |\mathbf{A}| = \text{determinant of } \mathbf{A}
 \end{aligned} \tag{53.97}$$

Equation (53.96) is called the *characteristic equation* of $\mathbf{A}$, or the *eigenvalue equation*. The matrix $(\mathbf{A} - \lambda \mathbf{I})$ is called the *characteristic matrix*. There are n roots for Eq. (53.96), counting multiplicity. These are the n eigenvalues of $\mathbf{A}$, $\lambda_1, \lambda_2, \dots, \lambda_n$. For an eigenvalue λ_i , we solve the set of (homogeneous) linear equations $(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x} = 0$ to determine the components of the corresponding eigenvector $\mathbf{x}_i$. In general, λ_i and $\mathbf{x}_i$ are either real or complex numbers and vectors, respectively.

If the matrix $\mathbf{A}$ is *symmetric*, then:

1. The eigenvalues are real.
2. The eigenvectors are all mutually orthogonal; that is,

$$\mathbf{x}_i^T \mathbf{x}_j = \mathbf{x}_j^T \mathbf{x}_i = 0$$

Bilinear and Quadratic Forms

If $\mathbf{A}$ is a square matrix of order n and $\mathbf{x}$ and $\mathbf{y}$ are two arbitrary n vectors, then the scalar

$$u = \mathbf{x}^T \mathbf{A} \mathbf{y} \tag{53.98}$$

is called a *bilinear form*. If, however, the matrix $\mathbf{A}$ is also *symmetric*, then

$$v = \mathbf{x}^T \mathbf{A} \mathbf{x} \tag{53.99}$$

is called a *quadratic form* with the kernel $\mathbf{A}$.

The matrix $\mathbf{A}$ is called *positive definite* if $v > 0$ for all $\mathbf{x} \neq \mathbf{0}$, and we write $\mathbf{A} > \mathbf{0}$. If $v \geq 0$ for all $\mathbf{x}$ and there exists a nonzero vector $\mathbf{x}$ for which equality holds, we say $\mathbf{A}$ is *positive semidefinite* (or *nonnegative definite*) and write $\mathbf{A} \geq \mathbf{0}$. There are corresponding definitions for *negative definite* (or *nonpositive definite*). If there exist vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 > 0$ and $\mathbf{x}_2^T \mathbf{A} \mathbf{x}_2 < 0$, we say $\mathbf{A}$ is *indefinite*.

For a positive definite matrix $\mathbf{A}$ it is necessary and sufficient that

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \dots, \quad |\mathbf{A}| > 0$$

A quadratic form represents, in general, a conic section of some kind. Considering the two-dimensional case for simplicity, we write

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = b \quad \text{with } \mathbf{A} \text{ symmetric} \tag{53.100}$$

or

$$a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = b$$

which is the equation of an ellipse.

53.6 Coordinate Transformations

Linear Transformations

A general *linear transformation* of a vector $\mathbf{x}$ to another vector $\mathbf{y}$ takes the form

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{t} \quad (53.101)$$

Each element of the $\mathbf{y}$ vector is a linear combination of the elements of $\mathbf{x}$ plus a translation or shift represented by an element of the $\mathbf{t}$ vector. The matrix $\mathbf{M}$ is called the *transformation matrix*, which is in general rectangular, and $\mathbf{t}$ is called the translation vector. For our use we restrict $\mathbf{M}$ to being square nonsingular; thus, the inverse relation exists, or

$$\mathbf{x} = \mathbf{M}^{-1}(\mathbf{y} - \mathbf{t}) \quad (53.102)$$

in which case it is called *affine transformation*. Although both Eqs. (53.101) and (53.102) apply to higher-dimension vectors, we will limit our discussions, without loss of generality, to the more practical two- and three-dimensional spaces, where the elements of the transformations can be depicted geometrically.

Two-Dimensional Linear Transformations

There are six *elementary* transformations, each representing a single effect, which are geometrically represented in Fig. 53.3. Initially, four vectors (1,3) (1,5), (3,3) (3,5) representing the corners of a square (solid lines in Fig. 53.3) are referred to the x_1, x_2 coordinate system. Each of the six elementary transformations operates on the square, and the resulting y_1, y_2 coordinates are plotted to show the effect on the location, orientation, size, and shape of the square after the transformation (dashed lines in Fig. 53.3). In displaying the effects of the transformations, we either display the new figure (dashed lines) in the same coordinate system, or we change the coordinate system. It is easier for the student to visualize these transformations if the new figure is drawn without changing the coordinate system, which we did in Fig. 53.3. However, as we discuss each elementary transformation, we will comment on the second interpretation when appropriate.

1. Translation

$$\mathbf{y} = \mathbf{x} + \mathbf{t} \quad \text{where } \mathbf{M} = \mathbf{I} \quad (53.103)$$

The square is shifted 3 units in x_1 direction and 1 unit in x_2 direction, as shown in Fig. 53.3(a). Alternatively, the solid square remains and the coordinate axes shifted (in the opposite direction and shown in dashed lines).

2. Uniform Scale

$$\mathbf{y} = \mathbf{M}\mathbf{x} \quad \mathbf{M} = \mathbf{U} = \begin{bmatrix} u & 0 \\ 0 & u \end{bmatrix} = u\mathbf{I} \quad (53.104)$$

The (dotted) square is enlarged by the uniform scale u ($= 1.5$ in Fig. 53.3(b)), which results from all four point coordinate pairs multiplied by u . Alternatively, the solid square is referred to the same coordinate system, except that the units along the axes are now $1/u$ of the original units.

3. Rotation

$$\mathbf{y} = \mathbf{M}\mathbf{x} \quad \mathbf{M} = \mathbf{R} = \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \quad (53.105)$$

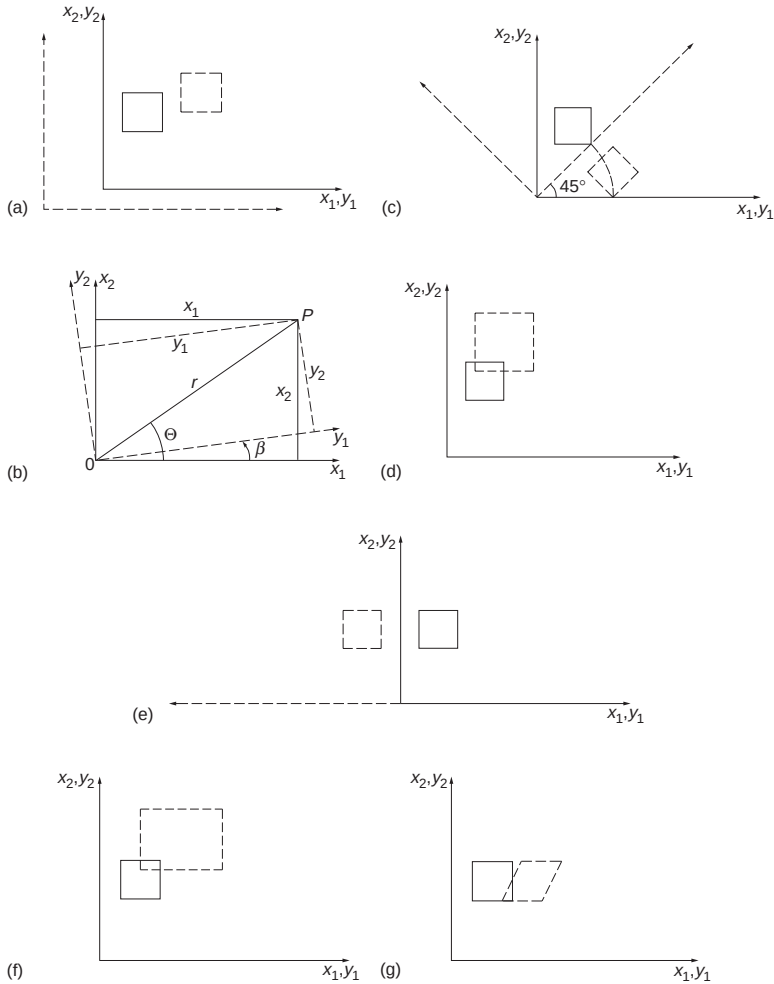


FIGURE 53.3 (a) Translation. (b) Uniform scale. (c) Rotation. (d) Rotation of a two-dimensional coordinate system. (e) Reflection. (f) Stretch (nonuniform scale). (g) Skew (nonperpendicularity of axes).

The square retains its shape, but is rotated through β about the origin of the coordinate system. In Fig. 53.3(c), the coordinate system is also rotated (45°). The elements of $\mathbf{R}$ are derived from Fig. 53.3(d) as follows:

$$\begin{aligned} y_1 &= r \cos(\theta - \beta) = r \cos \theta \cos \beta + r \sin \theta \sin \beta \\ y_2 &= r \sin(\theta - \beta) = r \sin \theta \cos \beta - r \cos \theta \sin \beta \end{aligned}$$

or

$$\begin{aligned} y_1 &= x_1 \cos \beta + x_2 \sin \beta \\ y_2 &= -x_1 \sin \beta + x_2 \cos \beta \end{aligned}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (53.106)$$

The matrix $\mathbf{R}$ is proper orthogonal, $\mathbf{R}^{-1} = \mathbf{R}^T$ and $|\mathbf{R}| = +1$. Rotation matrices do not change the length of the vector, i.e., $|\mathbf{x}| = |\mathbf{y}|$. Considering the square of the vector length,

$$\mathbf{y}^T \mathbf{y} = (\mathbf{M}\mathbf{x})^T \mathbf{M}\mathbf{x} = \mathbf{x}^T \mathbf{M}^T \mathbf{M}\mathbf{x} = \mathbf{x}^T \mathbf{x}$$

or

$$\mathbf{x}^T (\mathbf{M}^T \mathbf{M} - \mathbf{I}) \mathbf{x} = \mathbf{0}$$

which for a nontrivial solution means that $\mathbf{M}^T \mathbf{M} = \mathbf{I}$, thus showing that $\mathbf{M}$ is an orthogonal matrix.

4. Reflection

$$\mathbf{y} = \mathbf{M}\mathbf{x} \quad \mathbf{M} = \mathbf{F} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

Figure 53.3(e) shows reflection of the x_1 axis (i.e., about the x_2 axis). $\mathbf{F}$ is improper orthogonal, $\mathbf{F}^{-1} = \mathbf{F}^T = \mathbf{F}$ and $|\mathbf{F}| = -1$.

5. Stretch (Two Scale Factors)

$$\mathbf{y} = \mathbf{M}\mathbf{x} \quad \mathbf{M} = \mathbf{S} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \quad (53.107)$$

The square is transformed into a rectangle as shown in Fig. 53.3(f), in which

$$\mathbf{S} = \begin{bmatrix} 2 & 0 \\ 0 & 1.5 \end{bmatrix}$$

6. Skew (Shear)

$$\mathbf{y} = \mathbf{M}\mathbf{x} \quad \mathbf{M} = \mathbf{K} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \quad (53.108)$$

The square is transformed into a parallelogram as shown in Fig. 53.3(g), where

$$\mathbf{K} = \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix}$$

From these elementary transformations several affine transformations may be constructed using various sequences. The following are two of the commonly used transformations in photogrammetry.

Four-Parameter Transformation.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} u & 0 \\ 0 & u \end{bmatrix} \begin{bmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \quad (53.109a)$$

or

$$\begin{aligned} y_1 &= ux_1 \cos \beta + ux_2 \sin \beta + t_1 \\ y_2 &= -ux_1 \sin \beta + ux_2 \cos \beta + t_2 \end{aligned} \quad (53.109b)$$

or

$$\begin{aligned} y_1 &= ax_1 + bx_2 + c \\ y_2 &= -bx_1 + ax_2 + d \end{aligned} \quad (53.109c)$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix} \quad (53.109d)$$

The inverse transformation is given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{u} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} y_1 - c \\ y_2 - d \end{bmatrix} \quad (53.109e)$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{a^2 + b^2} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} y_1 - c \\ y_2 - d \end{bmatrix} \quad (53.109f)$$

This transformation has four parameters: a uniform scale, a rotation, and two translations. It is a conformal transformation.

Six-Parameter Transformation.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \quad (53.110a)$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a & b \\ d & e \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} c \\ f \end{bmatrix} \quad (53.110b)$$

The six parameters of this transformation consist of two scales, one skew factor (lack of perpendicularity of the axes), one rotation, and two shifts. The inverse transformation is given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{ae - bd} \begin{bmatrix} e & -b \\ -d & a \end{bmatrix} \begin{bmatrix} y_1 - c \\ y_2 - f \end{bmatrix} \quad (53.110c)$$

Three-Dimensional Linear Transformations

As in the two-dimensional case, affine transformation in three dimensions can be factored out in several elementary transformations: translation, uniform scale, nonuniform scale, rotations, reflections, etc. Consideration, however, is limited to the seven-parameter transformation, which is composed of a uniform scale change, three translations, and three rotations.

We first consider rotations in three-dimensional space.

Rotations of a Three-Dimensional Coordinate System. There are three elementary rotations, one about each of the three axes. They are frequently performed in sequence one after the other. A set of three of these is as follows, where $\mathbf{x}$ is the original system, $\mathbf{x}'$ is once rotated, and $\mathbf{x}''$ is twice rotated:

1. β_1 about x_1 axis, positive rotation advances $+x_2$ to $+x_3$
2. β_2 about x'_2 axis, positive rotation advances $+x'_3$ to $+x'_1$
3. β_3 about x''_3 axis, positive rotation advances $+x''_1$ to $+x''_2$

Each of the three elementary rotations is represented in matrix form by

$$\begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \beta_1 & \sin \beta_1 \\ 0 & -\sin \beta_1 & \cos \beta_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{M}_{\beta_1} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (53.111a)$$

where x_1, x_2, x_3 are the coordinates before rotation and x'_1, x'_2, x'_3 are the coordinates after rotation. Similarly, rotations of $+\beta_2$ about the x'_2 axis and $+\beta_3$ about the x''_3 axis are given by

$$\begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \end{bmatrix} = \begin{bmatrix} \cos \beta_2 & 0 & -\sin \beta_2 \\ 0 & 1 & 0 \\ \sin \beta_2 & 0 & \cos \beta_1 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} = \mathbf{M}_{\beta_2} \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} \quad (53.111b)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \end{bmatrix} = \begin{bmatrix} \cos \beta_3 & \sin \beta_3 & 0 \\ -\sin \beta_3 & \cos \beta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \end{bmatrix} = \mathbf{M}_{\beta_3} \begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \end{bmatrix} \quad (53.111c)$$

The three rotations in Eq. (53.111) are often referred to as *elementary* rotations, since they may be used to construct any required set of sequential rotations. By successive substitution, the total rotation matrix is obtained:

$$\mathbf{y} = \mathbf{x}'' = \mathbf{M}_{\beta_3} \mathbf{M}_{\beta_2} \mathbf{M}_{\beta_1} \mathbf{x} = \mathbf{M} \mathbf{x} \quad (53.112)$$

in which $\mathbf{M}$ is now a function of the three rotation angles $\beta_1, \beta_2, \beta_3$. The most commonly used set of sequential rotations (in photogrammetry) is given the symbols ω, ϕ, κ where $\omega \equiv \beta_1; \phi \equiv \beta_2; \kappa \equiv \beta_3$. In this case, the matrix $\mathbf{M}$ which rotates the object coordinate system (X, Y, Z) parallel to the photo coordinate system (x, y, z) is given by

$$\mathbf{M} = \begin{bmatrix} \cos \phi \cos \kappa & \cos \omega \sin \kappa + \sin \omega \sin \phi \cos \kappa & \sin \omega \sin \kappa - \cos \omega \sin \phi \cos \kappa \\ -\cos \phi \sin \kappa & \cos \omega \cos \kappa - \sin \omega \sin \phi \sin \kappa & \sin \omega \cos \kappa + \cos \omega \sin \phi \sin \kappa \\ \sin \phi & -\sin \omega \cos \phi & \cos \omega \cos \phi \end{bmatrix} \quad (53.113)$$

in which ω is about the X axis, ϕ is about the once-rotated Y axis, and κ is about the twice-rotated Z axis. The matrix $\mathbf{M}$ is orthogonal, since $\mathbf{M}_\omega, \mathbf{M}_\phi$, and $\mathbf{M}_\kappa$ are each orthogonal.

Seven-Parameter Transformation. This transformation contains seven parameters: a uniform scale change u , three rotations $\beta_1, \beta_2, \beta_3$, and three translations t_1, t_2, t_3 . It takes the general form

$$\mathbf{y} = u\mathbf{M}\mathbf{x} + \mathbf{t} \quad (53.114)$$

The orthogonal matrix $\mathbf{M}$ is a function of only three independent parameters, in this case the angles $\beta_1, \beta_2, \beta_3$. This transformation is useful for different applications, such as absolute orientation, model connection, etc.

The orthogonal matrix $\mathbf{M}$ may be constructed by other methods besides sequential rotations. Two such methods follow.

Constructing M by One Rotation about a Line. This is also often referred to as the *solid body rotation*. Given a three-dimensional object in two different orientations, there exists a line in space about which the object may be rotated by a finite angle to change it from one orientation to the other. If the said line has λ, μ, ν as direction cosines and the angle of rotation is designated by α , the rotation matrix is given by

$$\mathbf{M} = \begin{bmatrix} \lambda^2(1 - \cos \alpha) + \cos \alpha & \lambda\mu(1 - \cos \alpha) - \nu \sin \alpha & \lambda\nu(1 - \cos \alpha) + \mu \sin \alpha \\ \lambda\mu(1 - \cos \alpha) + \nu \sin \alpha & \mu^2(1 - \cos \alpha) + \cos \alpha & \mu\nu(1 - \cos \alpha) - \lambda \sin \alpha \\ \lambda\nu(1 - \cos \alpha) - \mu \sin \alpha & \mu\nu(1 - \cos \alpha) + \lambda \sin \alpha & \nu^2(1 - \cos \alpha) + \cos \alpha \end{bmatrix} \quad (53.115)$$

A Purely Algebraic Derivation of M . The following skew-symmetric matrix contains only three parameters a, b, c :

$$\mathbf{S} = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix} \quad (53.116a)$$

An orthogonal matrix $\mathbf{M}$ can be obtained from $\mathbf{M}$ using

$$\mathbf{M} = (\mathbf{I} + \mathbf{S})(\mathbf{I} - \mathbf{S})^{-1} = (\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S}) \quad (53.116b)$$

in which $\mathbf{M}$ is the identity matrix. Then

$$\begin{aligned} \mathbf{M} &= (\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S}) \\ &= \frac{1}{1 + a^2 + b^2 + c^2} \begin{bmatrix} 1 + a^2 - b^2 - c^2 & 2ab - 2c & 2ac + 2b \\ 2ab + 2c & 1 - a^2 + b^2 - c^2 & 2bc - 2a \\ 2ac - 2b & 2bc + 2a & 1 - a^2 - b^2 + c^2 \end{bmatrix} \end{aligned} \quad (53.117)$$

Nonlinear Transformations

In addition to the linear transformations discussed so far, we use nonlinear transformations both in two and three dimensions. In two dimensions we have the following two transformations:

Eight-Parameter Transformation. The equations

$$\begin{aligned} y_1 &= \frac{a_1x_1 + b_1x_2 + c_1}{a_0x_1 + b_0x_2 + 1} \\ y_2 &= \frac{a_2x_1 + b_2x_2 + c_2}{a_0x_1 + b_0x_2 + 1} \end{aligned} \quad (53.118a)$$

represent the projective transformation from the $\mathbf{x}$ to the $\mathbf{y}$ coordinate systems, with the eight transformation parameters being $a_0, b_0, a_1, \dots, c_2$. Its inverse is given by

$$\begin{aligned} x_1 &= \frac{(c_1 - y_1)(b_0y_2 - b_2) - (c_2 - y_2)(b_0y_1 - b_1)}{(a_0y_1 - a_1)(b_0y_2 - b_2) - (a_2y_2 - a_2)(b_0y_1 - b_1)} \\ x_2 &= \frac{(a_0y_1 - a_1)(c_2 - y_2) - (a_0y_2 - a_2)(c_1 - y_1)}{(a_0y_1 - a_1)(b_0y_2 - b_2) - (a_0y_2 - a_2)(b_0y_1 - b_1)} \end{aligned} \quad (53.118b)$$

These equations describe the central projectivity between two planes.

Two-Dimensional General Polynomials.

$$\begin{aligned} y_1 &= a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 + a_4x_1^2 + a_5x_2^2 + \dots \\ y_2 &= b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 + \dots \end{aligned} \quad (53.119a)$$

These polynomials can obviously be extended to higher powers in x_1, x_2 . A special case of these is the conformal form given in the following section.

Two-Dimensional Conformal Polynomials. The conformal property preserves the angles between intersecting lines after the transformation. If we impose the two conditions

$$\frac{\partial y_1}{\partial x_1} = \frac{\partial y_2}{\partial x_2} \quad \text{and} \quad \frac{\partial y_1}{\partial x_2} = -\frac{\partial y_2}{\partial x_1} \quad (53.119b)$$

on the general polynomials in Eq. (53.119a), we get

$$\begin{aligned} y_1 &= A_0 + A_1x_1 + A_2x_2 + A_3(x_1^2 - x_2^2) + A_4(2x_1x_2) + \dots \\ y_2 &= B_0 - A_2x_1 + A_1x_2 - A_4(x_1^2 - x_2^2) + A_3(2x_1x_2) + \dots \end{aligned} \quad (53.119c)$$

Note that the first three terms after the equal signs are the same as those in the four-parameter transformation given in Eq. (53.109c). Equation (53.119c) can also be derived using complex numbers by writing

$$(y_1 + y_2i) = (a_0 + b_0i) + (a_1 + b_1i)(x_1 + x_2i) + (a_3 + b_3i)(x_1 + x_2i)^2 + \dots$$

in which $i = \sqrt{-1}$. Expanding and equating y_1 to the real part and y_2 to the imaginary part (multiplier of i) on the right-hand side leads to Eq. (53.119c).

Three-dimensional General Polynomials.

$$\begin{aligned} y_1 &= a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1^2 + a_5x_2^2 + a_6x_1x_2 + a_7x_2x_3 + a_8x_1x_3 + \dots \\ y_2 &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1^2 + b_5x_2^2 + b_6x_1x_2 + b_7x_2x_3 + b_8x_1x_3 + \dots \\ y_3 &= c_0 + c_1x_1 + c_2x_2 + c_3x_3 + c_4x_1^2 + c_5x_2^2 + c_6x_1x_2 + c_7x_2x_3 + c_8x_1x_3 + \dots \end{aligned} \quad (53.120a)$$

We can extend these polynomials to higher order. Unlike the two-dimensional case, conformal transformation does not exist in three dimensions beyond the first-order (or linear) case given by the seven-parameter transformation, Eq. (53.114). A close approximation, which exists for only second-degree terms, is derived by imposing conditions similar to those in Eq. (53.119b) on every pair of coordinates in Eq. (53.120a). This makes the projections of the 3-space onto each of the three planes conformal. Thus, imposing the following on the general polynomials in Eq. (53.120a)

$$\begin{aligned} \frac{\partial y_1}{\partial x_1} &= \frac{\partial y_2}{\partial x_2} = \frac{\partial y_3}{\partial x_3} \\ \frac{\partial y_1}{\partial x_2} &= -\frac{\partial y_2}{\partial x_1}, \quad \frac{\partial y_2}{\partial x_3} = -\frac{\partial y_3}{\partial x_2}, \quad \frac{\partial y_1}{\partial x_3} = -\frac{\partial y_3}{\partial x_1} \end{aligned} \quad (53.120b)$$

leads to

$$\begin{aligned} y_1 &= A_0 + Ax_1 + Bx_2 - Cx_3 + E(x_1^2 - x_2^2 - x_3^2) + 0 + 2Gx_3x_1 + 2Fx_1x_2 + \dots \\ y_2 &= B_0 - Bx_1 + Ax_2 + Dx_3 + F(-x_1^2 + x_2^2 - x_3^2) + 2Gx_2x_3 + 0 + 2Ex_1x_2 + \dots \\ y_3 &= C_0 + Cx_1 - Dx_2 + Ax_3 + G(-x_1^2 - x_2^2 + x_3^2) + 2Fx_2x_3 + 2Ex_3x_1 + 0 + \dots \end{aligned} \quad (53.120c)$$

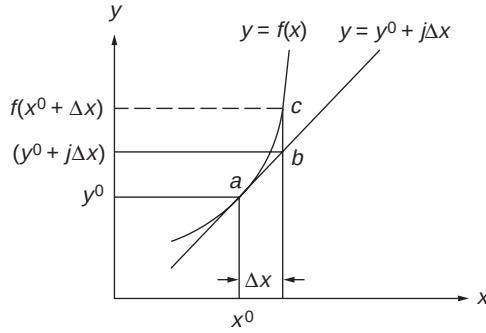


FIGURE 53.4 Linearization.

53.7 Linearization of Nonlinear Functions

Frequently, the equations expressing the geometric and physical conditions of a problem are nonlinear, which makes their direct solution difficult and uneconomical. We linearize these equations using series expansion, usually Taylor's series, which in general is given by the following for $y = f(x)$:

$$y = f(x^0) + \left. \frac{df}{dx} \right|_{x^0} \Delta x + \frac{1}{2!} \left. \frac{d^2 y}{dx^2} \right|_{x^0} (\Delta x)^2 + \dots + \frac{1}{n!} \left. \frac{d^n y}{dx^n} \right|_{x^0} (\Delta x)^n + \dots \quad (53.121)$$

This gives the value of y at $(x^0 + \Delta x)$, given the value of the function $f(x^0)$ at x^0 . Equation (53.121) includes still higher-order terms, and therefore we usually drop the second- and higher-order terms and use the approximation

$$y \approx f(x^0) + \left. \frac{dy}{dx} \right|_{x^0} \Delta x \approx y^0 + j\Delta x \quad (53.122)$$

with obvious correspondence in terms.

The technique of linearization is demonstrated in Fig. 53.4. The curve represents the original nonlinear function $f(x)$, whereas the straight line represents the linearized form, Eq. (53.122).

That line is tangent to the curve at the given point a , (x^0, y^0) . When Δx is given (or evaluated), the value of the function would be approximated by point b , whose ordinate is $(y^0 + j\Delta x)$, and the exact value from the nonlinear function is point c , with ordinate $f(x^0 + \Delta x)$. The error arising from using the linear form is the line segment bc .

One Function of Two Variables

$$\begin{aligned} y &= f(x_1, x_2) \\ &= f(x_1^0, x_2^0) + \left. \frac{\partial y}{\partial x_1} \right|_{x_1^0, x_2^0} \Delta x_1 + \left. \frac{\partial y}{\partial x_2} \right|_{x_1^0, x_2^0} \Delta x_2 \\ &\quad + \frac{1}{2!} \left. \frac{\partial^2 y}{\partial x_1^2} \right|_{x_1^0, x_2^0} (\Delta x_1)^2 + \frac{1}{2!} \left. \frac{\partial^2 y}{\partial x_2^2} \right|_{x_1^0, x_2^0} (\Delta x_2)^2 \\ &\quad + \left. \frac{\partial y}{\partial x_1} \right|_{x_1^0, x_2^0} \left. \frac{\partial y}{\partial x_2} \right|_{x_1^0, x_2^0} (\Delta x_1)(\Delta x_2) + \dots \end{aligned} \quad (53.123)$$

For the linearized form, Eq. (53.123) is truncated to

$$y = y^0 + j_1 \Delta x_1 + j_2 \Delta x_2 \quad (53.124)$$

where

$$y^0 = f(x_1^0, x_2^0) \quad j_1 = \left. \frac{\partial y}{\partial x_1} \right|_{x_1^0, x_2^0} \quad j_2 = \left. \frac{\partial y}{\partial x_2} \right|_{x_1^0, x_2^0}$$

Equation (53.124) can be rewritten in matrix form as

$$y = y^0 + \begin{bmatrix} j_1 & j_2 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix}$$

or

$$y = y^0 + \mathbf{J}_{yx} \Delta \mathbf{x} \quad (53.125)$$

where

$$\mathbf{J}_{yx} = \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} \end{bmatrix}$$

is the Jacobian of y with respect to $\mathbf{x}$.

Two Functions of One Variable

$$\begin{aligned} y_1 &= f_1(x) \approx y_1^0 + j_1 \Delta x \\ y_2 &= f_2(x) \approx y_2^0 + j_2 \Delta x \end{aligned} \quad (53.126)$$

or

$$\mathbf{y} = \mathbf{y}^0 + \mathbf{J}_{yx} \Delta x$$

with

$$\begin{aligned} y_1^0 &= f_1(x^0) \\ y_2^0 &= f_2(x^0) \\ \mathbf{J}_{yx} &= \begin{bmatrix} j_1 & j_2 \end{bmatrix}^T = \left[\left. \frac{dy_1}{dx} \right|_{x^0} \quad \left. \frac{dy_2}{dx} \right|_{x^0} \right]^T \end{aligned}$$

Two Functions of Two Variables Each

$$\begin{aligned} y_1 &= f_1(x_1, x_2) \approx y_1^0 + j_{11} \Delta x_1 + j_{12} \Delta x_2 \\ y_2 &= f_2(x_1, x_2) \approx y_2^0 + j_{21} \Delta x_1 + j_{22} \Delta x_2 \end{aligned} \quad (53.127a)$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \approx \begin{bmatrix} y_1^0 \\ y_2^0 \end{bmatrix} + \begin{bmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix} \quad (53.127b)$$

or

$$\mathbf{y} = \mathbf{y}^0 + \mathbf{J}_{yx} \Delta \mathbf{x} \quad (53.127c)$$

where

$$\mathbf{y}^0 = \begin{bmatrix} y_1^0 \\ y_2^0 \end{bmatrix} = \begin{bmatrix} f_1(x_1^0, x_2^0) \\ f_2(x_1^0, x_2^0) \end{bmatrix}$$

and

$$\mathbf{J}_{xy} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix}$$

evaluated at x_1^0, x_2^0 .

General Case of m Functions of n Variables

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \dots, x_n) \\ y_2 &= f_2(x_1, x_2, \dots, x_n) \\ &\vdots \\ y_m &= f_m(x_1, x_2, \dots, x_n) \end{aligned} \quad (53.128a)$$

With the auxiliaries,

$$\begin{aligned} \mathbf{y}^0 &= \begin{bmatrix} y_1^0 \\ y_2^0 \\ \vdots \\ y_m^0 \end{bmatrix} = \begin{bmatrix} f_1(x_1^0, x_2^0, \dots, x_n^0) \\ f_2(x_1^0, x_2^0, \dots, x_n^0) \\ \vdots \\ f_m(x_1^0, x_2^0, \dots, x_n^0) \end{bmatrix} \\ \mathbf{J}_{yx} &= \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \text{ evaluated at } \mathbf{x}^0 \\ \Delta \mathbf{x} &= \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_n \end{bmatrix} \end{aligned}$$

the linearized form of Eq. (53.128a) becomes

$$\mathbf{y} \approx \mathbf{y}^0 + \mathbf{J}_{yx} \Delta \mathbf{x} \quad (53.128b)$$

which represents the general form, with $\mathbf{y}, \mathbf{y}^0$ being $m \times 1$ vectors, $\mathbf{J}$ an $m \times n$ Jacobian matrix, and $\Delta \mathbf{x}$ an $n \times 1$ vector. Equations (53.122), (53.125), and (53.127c) are special cases of Eq. (53.128b).

Differentiation of a Determinant

The partial derivative of a $p \times p$ determinant with respect to a scalar is composed of the sum of p determinants, each having the elements of only one row or one column replaced by their derivatives. Thus, given the determinant $d = |\mathbf{D}_1 \mathbf{D}_2 \cdots \mathbf{D}_p|$ in which $\mathbf{D}_i, i = 1, 2 \cdots p$ represents its p columns, then

$$\frac{\partial d}{\partial x} = \left| \frac{\partial \mathbf{D}_1}{\partial x} \mathbf{D}_2 \cdots \mathbf{D}_p \right| + \left| \mathbf{D}_1 \frac{\partial \mathbf{D}_2}{\partial x} \cdots \mathbf{D}_p \right| + \cdots + \left| \mathbf{D}_1 \mathbf{D}_2 \cdots \frac{\partial \mathbf{D}_p}{\partial x} \right| \quad (53.129)$$

An expression similar to Eq. (53.129) can be written in which rows instead of the columns of d are partially differentiated.

Differentiation of a Quotient

The partial derivative of $g = U/W$ with respect to a variable x is given by

$$\frac{\partial g}{\partial x} = \frac{1}{W} \left[\frac{\partial U}{\partial x} - \frac{U}{W} \frac{\partial W}{\partial x} \right] \quad (53.130)$$

Both U and W can be general functions, including determinants, of several variables.

53.8 Map Projections

Map projection is concerned with the theory and techniques of proper representation of the curved earth surface on the plane of a map. When the map is of such large scale as to represent a very limited area, the earth curvature is insignificant, and field survey measurements can be directly represented on the map. On the other hand, as the surface area of the earth gets larger, this curvature becomes significant and must be dealt with. The earth is an ellipsoid and is also sometimes approximated by a sphere; neither of these surfaces can accurately be developed into a plane. Therefore, all map projection methods must by necessity contain some distortion. Various methods are selected to fit best the shape of the area to be mapped and to minimize the effects of particular distortions.

Locations on the earth are represented by meridians of longitude, λ , and parallels of latitude, ϕ . On the map these are represented by scaled linear distances X, Y , using the dimensions of the earth ellipsoid and selected criteria which the specific map projection must satisfy. These are obtained from transformation equations taking the general functional form of

$$\begin{aligned} X &= f_x(\lambda, \phi) \\ Y &= f_y(\lambda, \phi) \end{aligned} \quad (53.131)$$

Although all modern map projections are performed by computer programs, several are based on geometric projection of the earth onto one of three surfaces: a plane, a cylinder, or a cone. It is clear that the cylinder and cone are chosen because they can be developed into a plane — that of the map. When a plane is used, it is tangent to the earth's surface at a point and the projection center is either the center of the earth, as in the *gnomonic* projection shown in Fig. 53.5(a), or the point diametrically opposite to the tangent point, as in the *stereographic* projection shown in Fig. 53.5(b). If the projection lines are perpendicular to the plane, we have an *orthographic* projection, Fig. 53.5(c).

A cone is usually selected with its axis coincident with the earth's polar axis. It may be tangent to the earth at one small circle, called *standard parallel*, or intersect it in two standard parallels. When developed, the scale will be true (i.e., without any distortion or error) at the standard parallels; see Fig. 53.6. Polyconic projections use a series of frustums of cones, each from a separate cone.

Like a cone, a cylinder may be selected to be tangent to the earth or secant to it. It may be *regular*, with its axis being the polar axis as in Fig. 53.7(a); *transverse*, with one tangent meridian as in Fig. 53.7(b); *secant*, with two meridians of intersection; or *oblique* cylindrical, shown in Fig. 53.7(c).

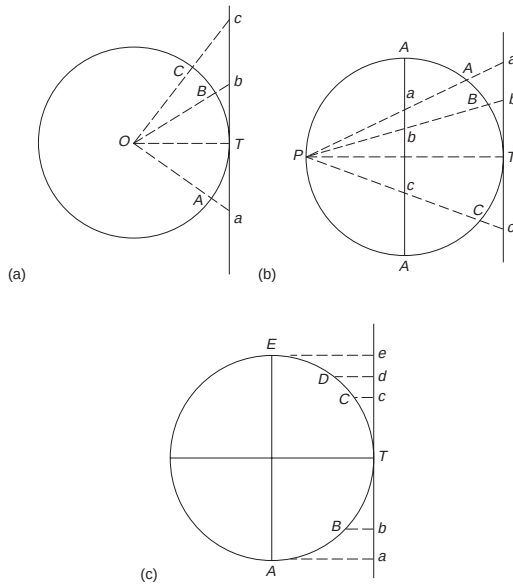


FIGURE 53.5 (a) Gnomonic projection. (b) Stereographic projection. (c) Orthographic projection.

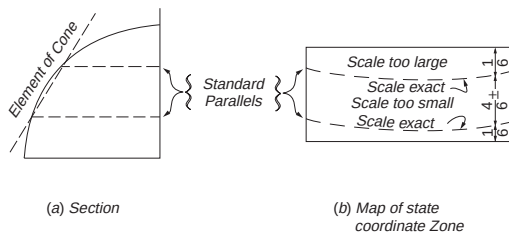


FIGURE 53.6 Lambert conformal conic projection. (Source: Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying: Theory and Practice*, 6th ed., p. 570. McGraw-Hill, New York. With permission.)

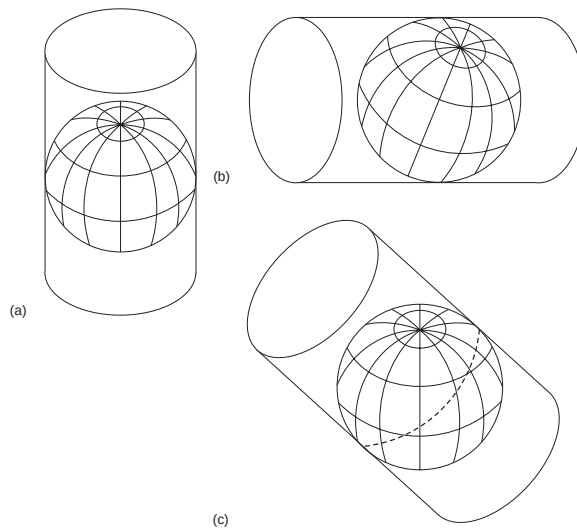


FIGURE 53.7 (a) Vertical cylinder. (b) Transverse horizontal cylinder. (c) Oblique cylinder.

Equation (53.131) will produce a *perfect* map — without any distortions — if it satisfies *all* of the following conditions:

1. All distances and areas have correct relative magnitudes.
2. All azimuths and angles are correctly represented on the map.
3. All great circles are shown on the map as straight lines.
4. Geodetic longitudes and latitudes are correctly shown on the map.

No one map projection can satisfy *all* these conditions. However, each class can satisfy some selected conditions. The following are four classes:

1. **Conformal or orthomorphic projection** results in a map showing the correct angle between any pair of short intersecting lines, thus making small areas appear in correct *shape*. As the scale varies from point to point, the shapes of larger areas are incorrect.
2. An **equal-area projection** results in a map showing all areas in proper relative *size*, although these areas may be much out of shape and the map may have other defects.
3. In an **equidistant projection** distances are correctly represented from one central point to other points on the map.
4. In an **azimuthal projection** the map shows the correct *direction* or azimuth of any point relative to one central point.

For conformal mapping, a new latitude, ψ , called the isometric latitude, is used in place of ϕ , where

$$\psi = \ln \left[\left(\frac{1 - e \sin \phi}{1 + e \sin \phi} \right)^{e/2} \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right) \right] \quad (53.132)$$

in which $e^2 = (a^2 - b^2)/a^2$, with a , b being the semimajor and semiminor axes of the earth ellipsoid, respectively. Then, Eq. (53.131) is replaced by

$$\begin{aligned} X &= f_1(\lambda, \psi) \\ Y &= f_2(\lambda, \psi) \end{aligned} \quad (53.133)$$

In order for the mapping in Eq. (53.133) to be conformal, the following Cauchy–Riemann conditions must be satisfied:

$$\frac{\partial X}{\partial \lambda} = \frac{\partial Y}{\partial \psi} \quad \text{and} \quad \frac{\partial X}{\partial \psi} = -\frac{\partial Y}{\partial \lambda} \quad (53.134)$$

Two commonly used conformal projections are the Lambert conformal conic projection and the transverse Mercator projection. A figure of the former is shown in Fig. 53.6, where the projection cone intersects the ellipsoid in two standard parallels. It is very widely used in the U.S., particularly as a State Plane Coordinate System for those states, or zones thereof, with greater east–west extent than north–south. The transverse Mercator projection is shown in Fig. 53.8, where the cylinder is either tangent or secant to the ellipsoid. When it is tangent, the scale at the central meridian is 1:1. But when it is not, the scale at the central meridian is less than 1:1, as shown in Fig. 53.8. The central meridian is the origin of the map X coordinate, while the origin of the map Y coordinate is the equator. This projection is used as a State Plane Coordinate System for states with greater north–south extent.

An extensively used map projection system is the *Universal* Transverse Mercator, or UTM, schematically shown in Fig. 53.9. It is in 6° wide zones, with the scale at each central meridian of a zone being 0.9996. A false easting for each central meridian is 500,000 m. A transverse Mercator projection with 3° wide zones is possible, where the scale at the central meridian is improved to 0.9999.

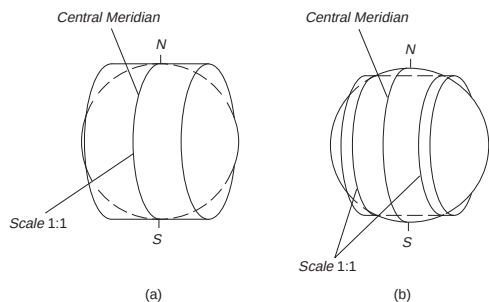


FIGURE 53.8 Transverse Mercator projection. (a) Cylinder with one standard line. (b) Cylinder with two standard lines.

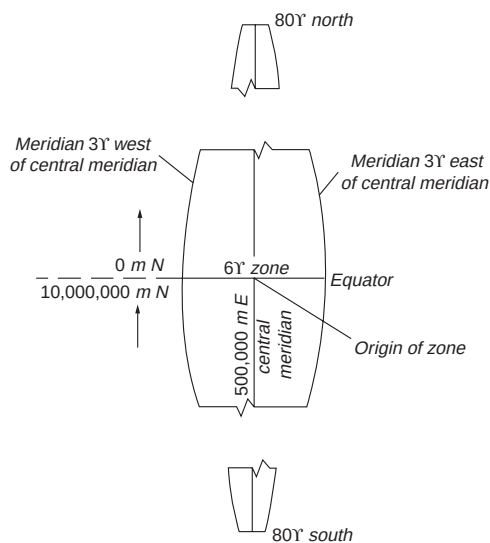
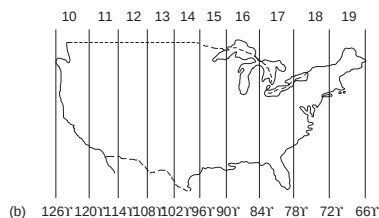
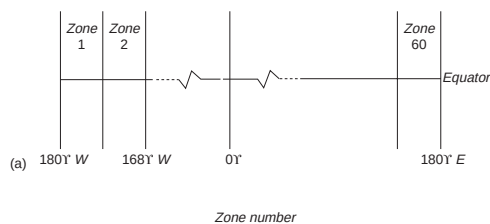


FIGURE 53.9 (a) Universal Transverse Mercator zones. (b) UTM zones in the United States. (c) X and Y coordinates of the origin of a UTM grid zone. (Source: U.S. Army Field Manual, *Map Reading*, FM 21-26.)

53.9 Observational Data Adjustment

Mathematical Model for Adjustment

In surveying engineering, measurements are rarely used directly as the required information. They are frequently used in subsequent operations to derive other quantities, often computationally, such as directions, lengths, relative positions, areas, shapes, and volumes. The relationships applied in the computational effort are the mathematical representations of the geometric and physical conditions of the problem, which, together with the quality of the measurements, are called the *mathematical model*.

This mathematical model is composed of two parts: a functional model and a stochastic model. The *functional model* is the part which describes the geometric or physical characteristics of the survey problem and the resulting mathematical relationships. The *stochastic model* is the part of the mathematical model that describes the statistical properties of all the elements involved in the functional model. It designates which parts are the observables, which are constants, and which are unknown parameters to be estimated in the **adjustment**. It also provides the information necessary to properly describe the quality of the observations to be used in the adjustment.

As a simple example, consider the size and shape of a plane triangle. While the shape depends only on angles, its size requires at least one side. Therefore, this model has three angular elements, the interior angles, and three linear (or distance) elements, the triangle sides. Two angles and one side will be the minimum number of measurements necessary to uniquely fix the triangle. If more measurements than these three are obtained, *redundancy* will exist, thus leading to inconsistency, which is resolved through an adjustment technique. Once the number of measurements is decided upon, the required set of independent condition equations can be written to express the functional model. The stochastic model will denote those elements (of the total of six) which are observed, and the quality of the observations.

The *a priori* quality of an observed angle or distance is usually expressed by a *standard deviation*, σ , or its square, the *variance*, σ^2 . Correlation between observations is represented by the *covariance*. Thus, for two observables, or *random variables*, say $\bar{x}$ and $\bar{y}$, the variances, σ_x^2 and σ_y^2 , and the covariance, σ_{xy} , are collected in a single square symmetric matrix called the variance-covariance matrix, or simply the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \quad (53.135)$$

where the variances are along the main diagonal and the covariance off the diagonal. The concept of the covariance matrix can be extended to the multidimensional case by considering n random variables $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$ and writing

$$\Sigma_{xx} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} \quad (53.136)$$

which is an $n \times n$ square symmetric matrix.

Often in practice, the variances and covariances are not known in absolute terms but only to a scale factor. The scale factor is given the symbol σ_0^2 and is termed the *reference variance*, although other names, such as “variance factor” and “variance associated with weight unity,” have also been used. The square

root σ_0 , of σ_0^2 , is called the “reference standard deviation” and was classically known as the “standard error of unit weight.” The relative variances and covariances are called *cofactors* and are given by

$$q_{ii} = \frac{\sigma_i^2}{\sigma_0^2} \quad \text{and} \quad q_{ij} = \frac{\sigma_{ij}}{\sigma_0^2} \quad (53.137)$$

Collecting the cofactors in a square symmetric matrix produces the *cofactor matrix* $\mathbf{Q}$, with the obvious relationship with the covariance matrix.

$$\mathbf{Q} = \frac{1}{\sigma_0^2} \Sigma \quad (53.138)$$

When $\mathbf{Q}$ is nonsingular, its inverse is called the *weight matrix* and designated by $\mathbf{W}$; thus,

$$\mathbf{W} = \mathbf{Q}^{-1} = \sigma_0^2 \Sigma^{-1} \quad (53.139)$$

If σ_0^2 is equal to 1, or, in other words, if the covariance matrix is known, the weight matrix becomes its inverse.

Design/Prealysis

Engineering *design* is more frequently known as *preanalysis* in surveying engineering. It refers to the task of determining the observations to be made and their required accuracy so that the required accuracy of the final product is met. This is usually done by an iterative procedure, often using interactive graphics, and applying the established mathematical model of the problem. The physical limitations of the project, such as visibility and accessibility problems, are first imposed on the design. Next, what is considered to be an adequate set of measurements, with suitable accuracy estimates, is input in a design program to estimate the required unknown parameters and their expected accuracy. The overall accuracy of the estimated parameters is reflected by the posterior covariance matrix Σ . (This is why preanalysis is equivalent to *covariance analysis* used in the mathematical literature.) From Σ , individual confidence measures, such as **error ellipses** and **ellipsoids**, are computed and compared to the design requirements. If they are too large, additional observations or measurements of increased quality are attempted and the process repeated. On the other hand, if they are too small (i.e., too good), reduced observations or observations of decreased quality (using less expensive equipment/techniques) are attempted instead. The procedure is iterated until an optimum design results.

Data Acquisition

The results of preanalysis provide the information required to set up the specifications for data acquisition, particularly the quantities to be measured and their required accuracy. This leads to deciding on equipment to be used and observational techniques to be applied. It is important that the selected instruments be properly calibrated and in good working order, and that the procedures specified be rigorously followed. Good field practices must be followed to minimize blunders. In fact, all field activities are carefully planned and monitored so that the following operation, preprocessing of data, can be effectively carried out to yield the most suitable data entering the adjustment.

Data Preprocessing

Preprocessing of survey data involves the elimination of *blunders* and the correction for all known systematic errors. The resulting preprocessed measurements should have essentially nothing but random errors before they are used in the adjustment. Any uncorrected systematic errors known to still exist in the measurements must then be modeled mathematically and accounted for during the adjustment.

Because of the significant cost of the field operations, it is becoming more of a requirement to perform preprocessing in the field so as to minimize the cost of any remeasurement. Progressively more sophisticated preprocessing programs are available for use with portable computers in conjunction with electronic data collectors.

There are several techniques for checking the observations which apply to the different types of surveys, such as triangulation, trilateration, traverse, leveling, etc. These techniques depend on the shape of the network and various minimum combinations of observations.

Data Adjustment

When redundant measurements exist, an *adjustment* of the observed data becomes necessary in order to resolve the inconsistency between the observations and the model. As an illustration, consider the size and shape of a plane triangle in which the three interior angles A, B, C and two sides a, b (opposite to A and B , respectively) are measured. Suppose that we are interested in the length of the side c . We obviously have more information than needed, since any one side, a or b , and any two angles suffice to solve the triangle and hence determine c . However, due to random measurement errors, every combination selected would be expected to lead to a slightly different value for c . A choice from all possible combinations would be essentially arbitrary. More important is the fact that one should take advantage of using *all* the available information in one operation, in which each measurement contributes relative to its role in the mathematical model and commensurate with its quality (variance) relative to the quality of all other measurements. One of the most commonly used techniques of adjustment of redundant survey data is the method of **least squares**.

Least Squares Adjustment

Although least squares is an estimation procedure, it has been traditionally referred to in surveying as an *adjustment* technique. This stems from the fact that after the adjustment the original observations are replaced by a set of *adjusted* observations that *are* consistent with the model. Thus, after least squares adjustment the five observations in the triangle example are replaced by a consistent set, $\hat{A}, \hat{B}, \hat{C}, \hat{a}, \hat{b}$, in that *any* minimum combination of these would yield the *same* triangle solution. Each adjusted observation is the sum of the original measurement and a *residual*, v , which is calculated in the adjustment.

The method of least squares is based on the observational residuals. If all the residuals in a given set of n observations, $\mathbf{l}$ are denoted by the vector $\mathbf{v}$, and the weight matrix associated with these observations is $\mathbf{W}$, the least squares criterion is given by

$$\phi = \mathbf{v}^T \mathbf{W} \mathbf{v} \rightarrow \text{minimum} \quad (53.140)$$

Note that ϕ is a scalar, for which a minimum is obtained by equating to zero its partial derivatives with respect to v . In Eq. (53.140) the weight matrix of the observations, $\mathbf{W}$, may be full, implying that the observations are correlated. If the observations are uncorrelated, $\mathbf{W}$ will be a diagonal matrix, and the criterion simplifies to

$$\phi = \sum_{i=1}^n w_i v_i^2 = w_1 v_1^2 + w_2 v_2^2 + \cdots w_n v_n^2 \rightarrow \text{minimum} \quad (53.141)$$

which says that the sum of the weighted squares of the residuals is a minimum. Another and simpler case involves observations which are uncorrelated and of equal weight (precision), for which $\mathbf{W} = \mathbf{I}$, and ϕ becomes

$$\phi = \sum_{i=1}^n v_i^2 = v_1^2 + v_2^2 + \cdots v_n^2 \rightarrow \text{minimum} \quad (53.142)$$

The case covered by Eq. (53.142) is the oldest and may have accounted for the name “least squares,” since it seeks the “least” sum of the squares of the residuals.

If we denote by n_0 the minimum number of independent variables needed to determine the selected model uniquely, then the *redundancy*, r , is given by

$$r = n - n_0 \quad (53.143)$$

As illustrations, consider the following examples.

1. The shape of a plane triangle is uniquely determined by a minimum of two interior angles, or $n_0 = 2$. If three interior angles are measured, then, with $n = 3$, redundancy is $r = 1$.
2. The size and shape of a plane triangle require a minimum of three observations, at least one of which is the length of one side, or $n_0 = 3$. If three interior angles and two side lengths are available, then with $n = 5$ the redundancy is $r = 2$.

After the redundancy r is determined, the adjustment proceeds by writing equations that relate the model variables in order to reflect the existing redundancy. Such equations will be referred to either as *condition equations* or simply as *conditions*. The number of independent conditions, c , will be equal to r if only observational variables and constants are involved. In many situations, however, additional unknown variables, called *parameters*, are carried in the adjustment. In such a case, if the number of unknown parameters is u , then a total of

$$c = r + u \quad (53.144)$$

independent condition equations in terms of both the n observations and u parameters must be written. In order for the parameter to be functionally independent, number, u , should not exceed the minimum number of variables, n_0 , necessary to specify the model. Hence, the following relation must be satisfied:

$$0 \leq u \leq n_0 \quad (53.145)$$

Similarly, for the formulated condition equations to be independent, their number, c , should not be larger than the total number of observations n . Hence,

$$r \leq c \leq n \quad (53.146)$$

Techniques of Least Squares

Although there is only one least squares criterion, there are several techniques by which least squares may be applied. Regardless of which technique is applied, the results of an adjustment of a given set of measurements associated with a specified model *must* be the same. The choice of a technique, therefore, is mostly a matter of convenience and computational economy. The first technique is called *adjustment of observations only*. The condition equations take the form

$$\begin{matrix} \mathbf{A} & \mathbf{v} \\ r, n & n, 1 \end{matrix} = \begin{matrix} \mathbf{f} \\ r, 1 \end{matrix} \quad (53.147)$$

For linear adjustment problems the vector $\mathbf{f}$ is given by

$$\mathbf{f} = \mathbf{d} = \mathbf{A}\mathbf{1} \quad (53.148)$$

in which $\mathbf{d}$ is a vector of numerical constants and $\mathbf{1}$ is the vector of given numerical values of the measurements.

Let the cofactor matrix of the observations be denoted by $\mathbf{Q} (= \mathbf{W}^{-1})$. Then

$$\mathbf{k} = (\mathbf{A}\mathbf{Q}\mathbf{A}^T)^{-1}\mathbf{f} = \mathbf{Q}_e^{-1}\mathbf{f} = \mathbf{W}_e\mathbf{f} \quad (53.149)$$

$$\mathbf{v} = \mathbf{Q}\mathbf{A}^T\mathbf{k} \quad (53.150)$$

$$\hat{\mathbf{I}} = \mathbf{I} + \mathbf{v} \quad (53.151)$$

Error propagation:

$$\mathbf{Q}_{vv} = \mathbf{Q}\mathbf{A}^T\mathbf{W}_e\mathbf{A}\mathbf{Q} \quad (53.152)$$

$$\mathbf{Q}_{\hat{I}\hat{I}} = \mathbf{Q} - \mathbf{Q}_{vv} \quad (53.153)$$

$$\hat{\sigma}_0^2 = \frac{\mathbf{v}^T\mathbf{W}\mathbf{v}}{r} \quad (53.154)$$

The second technique is called *adjustment of indirect observations*. The condition equations are of the general (nonlinear) form:

$$\mathbf{I} + \mathbf{v} + \mathbf{F}(\mathbf{x}) = 0 \quad (53.155)$$

When linearized at approximations $\mathbf{x}^0$ for the $u = n_0$ parameters $\mathbf{x}$, they become

$$\mathbf{v} + \mathbf{B}\Delta = \mathbf{f} \quad (53.156)$$

where Δ is a vector of unknown parameter corrections, $\mathbf{B} = \left. \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \right|_{\mathbf{x}^0}$ is an n by u coefficient matrix, and $\mathbf{f} = -\mathbf{F}(\mathbf{x}^0) - \mathbf{I}$. With $\mathbf{W} = \mathbf{Q}^{-1} (= \Sigma^{-1}$ if $\sigma_0 = 1$) as the weight matrix of the observations, then

$$\mathbf{N} = \mathbf{B}^T\mathbf{W}\mathbf{B} \quad (53.157)$$

$$\mathbf{t} = \mathbf{B}^T\mathbf{W}\mathbf{f} \quad (53.158)$$

$$\Delta = \mathbf{N}^{-1}\mathbf{t} \quad (53.159)$$

$$\hat{\mathbf{x}} = \mathbf{x}^0 + \Sigma\Delta \quad (\text{iterate}) \quad (53.160)$$

Error propagation:

$$\mathbf{Q}_{\hat{\hat{x}}\hat{\hat{x}}} = \mathbf{Q}_{\Delta\Delta} = \mathbf{N}^{-1} \quad (53.161)$$

$$\mathbf{Q}_{vv} = \mathbf{B}\mathbf{N}^{-1}\mathbf{B}^T \quad (53.162)$$

$$\mathbf{Q}_{\hat{I}\hat{I}} = \mathbf{Q} - \mathbf{Q}_{vv} \quad (53.163)$$

$$\hat{\sigma}_0^2 = \frac{\mathbf{v}^T\mathbf{W}\mathbf{v}}{r} \quad (53.164)$$

In the preceding technique, the largest number allowed for the parameters, $u = n_0$, must be carried in the adjustment so that each condition equation contains one and only one observation. In many applications,

it is more economical to carry in the adjustment only the parameters of interest, which are fewer in number: $u < n_o$. This technique is called *combined adjustment of observations and parameters*. The general (nonlinear) condition equations are of the form

$$\mathbf{F}(\mathbf{l}, \mathbf{x}) = \mathbf{0} \quad (53.165)$$

which in linearized form becomes

$$\mathbf{A}\mathbf{v} + \mathbf{B}\Delta = \mathbf{f} \quad (53.166)$$

where

$$\mathbf{A} = \left. \frac{\partial \mathbf{F}}{\partial \mathbf{l}} \right|_{l^0, x^0} \quad \mathbf{B} = \left. \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \right|_{l^0, x^0}$$

and

$$\mathbf{f} = -[\mathbf{F}(\mathbf{l}^0, \mathbf{x}^0) + \mathbf{A}(\mathbf{l} - \mathbf{l}^0)] \quad (53.167)$$

Then

$$\mathbf{Q}_e = \mathbf{A}\mathbf{Q}\mathbf{A}^T \quad (53.168)$$

$$\mathbf{W}_e = \mathbf{Q}_e^{-1} = (\mathbf{A}\mathbf{Q}\mathbf{A}^T)^{-1} \quad (53.169)$$

$$\mathbf{N} = \mathbf{B}^T \mathbf{W}_e \mathbf{B} \quad (53.170)$$

$$\mathbf{t} = \mathbf{B}^T \mathbf{W}_e \mathbf{f} \quad (53.171)$$

$$\Delta = \mathbf{N}^{-1} \mathbf{t} \quad (53.172)$$

$$\hat{\mathbf{x}} = \mathbf{x}^0 + \Sigma \Delta \quad (\text{iterate}) \quad (53.173)$$

$$\mathbf{v} = \mathbf{Q}\mathbf{A}^T \mathbf{W}_e (\mathbf{f} - \mathbf{B}\Delta) \quad (53.174)$$

$$\hat{\sigma}_0^2 = \frac{\mathbf{v}^T \mathbf{W} \mathbf{v}}{r} \quad (53.175)$$

Error propagation:

$$\mathbf{Q}_{\hat{x}\hat{x}} = \mathbf{Q}_{\Delta\Delta} = \mathbf{N}^{-1} \quad (53.176)$$

$$\mathbf{Q}_w = \mathbf{Q}\mathbf{A}^T (\mathbf{W}_e - \mathbf{W}_e \mathbf{B} \mathbf{Q}_{\Delta\Delta} \mathbf{B}^T \mathbf{W}_e) \mathbf{A} \mathbf{Q} \quad (53.177)$$

$$\mathbf{Q}_{\hat{w}} = \mathbf{Q} - \mathbf{Q}_{vv} \quad (53.178)$$

The linear set $\mathbf{N}\Delta = \mathbf{t}$ is called the *normal equations*.

In the three techniques above, the parameters $\mathbf{x}$ are functionally independent. In many surveying engineering applications, functions may exist between the parameters in the adjustment. Examples include points on geometric forms (lines, planes, surfaces), known distances, angles, etc., to be fixed in the adjustment. These functions are called *constraint equations* or simply *constraints*, to distinguish them from conditions. They are characterized by not containing any observations. The technique when constraints exist is called *adjustment with functional constraints*; of course, the condition equations to be

combined with the constraints can take any of the three situations discussed. It is more general to consider the combined adjustment technique, thus:

$$\mathbf{F}(\mathbf{1}, \mathbf{x}) = \mathbf{0} \quad (53.179)$$

$$\mathbf{G}(\mathbf{x}) = \mathbf{0} \quad (53.180)$$

are the general (nonlinear) conditions and constraints. The linearized form is

$$\mathbf{A}\mathbf{v} + \mathbf{B}\Delta = \mathbf{f} \quad (53.181)$$

$$\mathbf{C}\Delta = \mathbf{g} \quad (53.182)$$

where $\mathbf{C} = \left. \frac{\partial \mathbf{G}}{\partial \mathbf{x}} \right|_{\mathbf{x}^0}$ and $\mathbf{g} = -\mathbf{G}(\mathbf{x}^0)$

$$\mathbf{N} = \mathbf{B}^T \mathbf{W}_e \mathbf{B} \quad \mathbf{t} = \mathbf{B}^T \mathbf{W}_e \mathbf{f} \quad (53.183)$$

$$\mathbf{M} = \mathbf{C} \mathbf{N}^{-1} \mathbf{C}^T \quad (53.184)$$

$$\mathbf{k}_c = \mathbf{M}^{-1}(\mathbf{g} - \mathbf{C} \mathbf{N}^{-1} \mathbf{t}) \quad (53.185)$$

$$\Delta = \mathbf{N}^{-1}(\mathbf{t} + \mathbf{C}^T \mathbf{k}_c) \quad (53.186)$$

$$\mathbf{x} = \mathbf{x}^0 + \Sigma \Delta \quad (\text{iterate}) \quad (53.187)$$

$$\mathbf{Q}_{\Delta\Delta} = \mathbf{N}^{-1}(\mathbf{I} - \mathbf{C}^T \mathbf{M}^{-1} \mathbf{C} \mathbf{N}^{-1}) \quad (53.188)$$

The quantities v , $\hat{\sigma}_0^2$, $\mathbf{Q}_{vv}$, and $\mathbf{Q}_{\eta}$ are as given by Eqs. (53.174), (53.175), (53.177), and (53.178), respectively.

Another technique is to perform the adjustment sequentially. For a fixed number of parameters, both the inverse of the normal equations coefficient matrix $\mathbf{N}$ and the constant terms vector $\mathbf{t}$ are sequentially updated due to either addition or deletion of a set of condition equations. The relations for *sequential adjustment* are as follows:

$$\mathbf{A}_i \mathbf{v}_i + \mathbf{B}_i \Delta = \mathbf{f}_i \quad \text{with } \mathbf{Q}_i \quad (53.189)$$

are the conditions to be added or subtracted. Let $\mathbf{N}_{I-1}$, $\mathbf{t}_{I-1}$ and $\mathbf{N}_I$, $\mathbf{t}_I$ designate the matrices *before* and *after*, respectively, incorporating the effects of Eq. (53.189). Then

$$\mathbf{N}_I^{-1} = \mathbf{N}_{I-1}^{-1} [\mathbf{I} \mp \mathbf{B}_i^T (\mathbf{Q}_{e_i} \pm \mathbf{B}_i \mathbf{N}_{I-1}^{-1} \mathbf{B}_i^T)^{-1} \mathbf{B}_i \mathbf{N}_{I-1}^{-1}] \quad (53.190)$$

$$\mathbf{t}_I = \mathbf{t}_{I-1} \pm \mathbf{B}_i^T \mathbf{W}_{e_i} \mathbf{f}_i \quad (53.191)$$

in which $\mathbf{Q}_{e_i} = \mathbf{A}_i \mathbf{Q}_i \mathbf{A}_i^T$ and $\mathbf{W}_{e_i} = \mathbf{Q}_{e_i}^{-1}$. In Eqs. (53.190) and (53.191) the *upper signs* refer to condition *addition* and the *lower signs* to condition *deletion*.

Finally, a very flexible technique is used in which all the variables in the *mathematical model* are *considered as observables*. This requires *a priori* estimates for all model variables, but more importantly, estimates of their *a priori* weights. The *a priori* weights provide the mechanism for effectively distinguishing between different groups of variables in the model. Very large weights leave a variable essentially as a constant in the adjustment, while very small or even zero weight leaves it as an unknown parameter that can freely adjust. This is referred to as the *unified least squares* technique and applies to *all* of the techniques presented above. As an example, we consider the combined adjustment of observations and

parameters with the linear (or linearized) conditions: $\mathbf{A}\mathbf{v} + \mathbf{B}\Delta = \mathbf{f}$. Let $\mathbf{x}$ be the prior estimates of the parameters and $\mathbf{W}_{xx}$ its corresponding prior weight matrix. The solution given by Eqs. (53.168) to (53.178) essentially apply, except that

$$\Delta = (\mathbf{N} + \mathbf{W}_{xx})^{-1}(\mathbf{t} - \mathbf{W}_{xx}\mathbf{f}_x) \quad (53.192)$$

$$\mathbf{f}_x = \mathbf{x}^0 - \mathbf{x} \quad (53.193)$$

$$\mathbf{Q}_{\Delta\Delta} = (\mathbf{N} + \mathbf{W}_{xx})^{-1} \quad (53.194)$$

Assessment of Adjustment Results

After least squares adjustment it is quite important in surveying to analyze the results and provide a statement regarding the quality of the estimates. This operation is often referred to as *postadjustment analysis*, which applies various statistical techniques.

Test on Reference Variance

The first test is on the estimated reference variance, $\hat{\sigma}_0^2$. Let the *a priori* reference variance be σ_0^2 ; let r be the degrees of freedom (redundancy) in the adjustment, and assume that the residuals v_i are normally distributed. The statistic $r\hat{\sigma}_0^2/\sigma_0^2$ has a χ^2 distribution with r degrees of freedom. The two-tailed $100(1 - \alpha)$ confidence region for σ_0^2 is given by

$$(r\hat{\sigma}_0^2/\chi_{r, \alpha/2}^2) < \sigma_0^2 < (r\hat{\sigma}_0^2/\chi_{r, 1-\alpha/2}^2) \quad (53.195)$$

If σ_0^2 is incorrect or the mathematical model used is improper or incomplete (does not adequately account for systematic errors), then $\hat{\sigma}_0^2$ will fall outside this interval.

Test for Blunders or Outliers

If v_i is the i th residual and σ_{v_i} is its standard deviation, then

$$\bar{v}_i = v_i/\sigma_{v_i} \quad (53.196)$$

is called the *standardized residual*. Frequently, the effort involved in computing Σ_{vv} is quite extensive, and therefore an approximate estimate of σ_{v_i} may be obtained from

$$\hat{\sigma}_{v_i} = [(n - u)/n]^{1/2} \hat{\sigma}_0 \sigma_{l_i} / \sigma_0 \quad (53.197)$$

in which n is the number of observations, u is the number of parameters (thus $n - u = r$, the redundancy), and σ_{l_i} is the *a priori* standard deviation of observation l_i . When σ_0^2 is known, $\bar{v}_i$ has a probability density function, or pdf, $N(0, \sigma_{v_i}^2)$ and

$$\bar{v}_i = \left| \frac{v_i}{\sigma_{v_i}} \right| < N_{1-\alpha/2} \quad (53.198)$$

If σ_0^2 is not known, then

$$\bar{v}_i = \left| v_i / \hat{\sigma}_{v_i} \right| < \tau_{r, 1-\alpha/2} \quad (53.199)$$

in which $\hat{\sigma}_{v_i}$ is computed from Eq. (53.197), and τ_r has a Tau pdf with r degrees of freedom. If r is large, as in surveying, photogrammetric, or geodetic nets with extensive observations, τ_r may be replaced by Student t_r pdf or even normal pdf.

Confidence Region for Estimated Parameters

The covariance matrix for the parameters as evaluated from the least squares is given by (see, for example, Eq. (53.161))

$$\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = \sigma_0^2 N^{-1} \quad (53.200)$$

It can be shown that a region of constant probability is bounded by a u -dimensional hyperellipsoid centered at $\hat{\mathbf{x}}$, if the parameters are assumed to have a multivariate normal pdf. The function

$$k^2 = (\mathbf{x} - \hat{\mathbf{x}})^T \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) \quad (53.201)$$

describes the hyperellipsoid. The quadratic k^2 has a χ_u^2 distribution, with the probability for a point estimate being

$$P(\chi_u^2 < k^2) = 1 - \alpha$$

For the two-dimensional case (error ellipses), typical values are

p	0.394	0.500	0.900	0.950	0.990
k	1.000	1.177	2.146	2.447	3.035

and for the three-dimensional case, they are

P	0.199	0.500	0.900	0.950	0.990
k	1.000	1.538	2.500	2.700	3.368

when $k = 1$, we usually call it the *standard* region, *standard error ellipse* (for 2-D), or *standard error ellipsoid* (for 3-D). The standard regions for several dimensions are

Dimension	1	2	3	4	5	6
P	0.683	0.394	0.199	0.090	0.037	0.014

Given Σ , for example, for a point in a plane, the semimajor axis, a , and semiminor axis, b , of the standard error ellipse are computed from the eigenvalues and eigenvectors (see the discussion of “Basic Matrix Operations,” in Section 53.5). If one is interested in the 90% confidence region (i.e., significance level of $\alpha = 0.10$), the a, b are multiplied by 2.146. For the standard regions, there is a 0.683 probability that an adjusted point falls in a one-dimensional interval, a 0.394 probability that it falls inside the standard error ellipse, and only a 0.199 probability that it falls within the standard error ellipsoid.

Applications in Surveying Engineering

Level Net

Let l_{ij} represent the observed difference in elevation between two points whose (unknown) elevations are x_i and x_j . If l_{ij} is from point i to point j , then the condition equation is given by

$$x_i + l_{ij} + v_{ij} - x_j = 0$$

or

$$v_{ij} + x_i - x_j = -l_{ij} \quad (53.202)$$

This condition equation is in the form of adjustment of indirect observations. At least one benchmark (a point of known elevation) is needed for any given level net.

Traverse

There are two conditions, one for a measured angle α_i , and one for a measured distance d_{ij} . If α_i is at station i from the line $i - 1$ to i clockwise to the line from i to $i + 1$, then

$$\alpha_i = A_{i, i+1} - A_{i, i-1}$$

or

$$\nu_i + \tan^{-1}\left(\frac{x_{i-1} - x_i}{y_{i-1} - y_i}\right) - \tan^{-1}\left(\frac{x_{i+1} - x_i}{y_{i+1} - y_i}\right) = -\alpha_i \quad (53.203)$$

where A represents the azimuth and $(x, y)_{i-1, i, i+1}$ are the coordinates of the three points involved. The distance condition is given by

$$\nu_{ij} - [(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2} = -d_{ij} \quad (53.204)$$

Again, this is in the form of adjustment of indirect observations. If any of the points involved is a fixed point (i.e., with known coordinates), its coordinates are not carried as unknown parameters in the adjustment.

Trilateration

This is the operation in which *only* distances are measured in the network. Therefore, the only condition used is that given by Eq. (53.204).

Triangulation

This is the operation in which chains of triangles are connected together. The fundamental measurement is the angle, and the adjustment unit is the triangle. The single condition for one triangle is

$$\sum l_i - \pi = 0 \quad (53.205)$$

where l_i represents all the measured angles inside a single triangle. For a quadrilateral, there is another condition called the *side condition*, the composition of which can be found in the literature.

Defining Terms

Adjustment of observations — The mathematical technique used to resolve the inconsistency between the measurements collected and the underlying mathematical model when *redundancy* exists, or when the measurements exceed the minimum necessary to uniquely define the model.

Azimuthal projection — A map projection that yields a map where correct *direction* or *azimuth* of any point relative to one central point is shown.

Conformal or orthomorphic projection — A map projection technique that preserves *angles* between short intersecting lines, thus making small areas appear in their correct *shape* on the map. Scale varies from point to point, and thus larger areas are incorrect.

Equal-area projection — A map projection technique that results in a map showing all areas in proper relative *size*; however, they may be distorted in *shape*.

Equidistant projection — A map projection technique in which distances are correctly represented from one central point to other points on the map.

Error ellipses and ellipsoids — Confidence regions about estimated survey points in two dimensions (ellipses) or three dimensions (ellipsoids) for specified probabilities.

Error propagation — The general technique of determining the covariance matrix of a set of quantities, which are estimated from functions of another set of quantities of known values and covariance matrix.

Least squares adjustment — The most common adjustment technique used in surveying engineering; it is based on minimizing the sum of the weighted squares of the observational residuals.

Map projection — The theory and techniques of proper representation of the curved earth surface on the plane of a map.

References

- Anderson, J.M. and Mikhail, E.M. 1985. *Introduction to Surveying*. McGraw-Hill, New York.
- Davis, R.E. et al. 1981. *Surveying: Theory and Practice*, 6th ed. McGraw-Hill, New York.
- Krakiwsky, E.J., ed. 1983. *Papers for the CIS Adjustment and Analysis Seminars*. The Canadian Institute of Surveying, Ottawa.
- Mikhail, E.M. 1976. *Observations and Least Squares*. University Press of America, Lanham, MD.
- Mikhail, E.M. and Gracie, G. 1981. *Analysis and Adjustment of Survey Measurements*. Van Nostrand Reinhold, New York.
- Vanicek, P. and Krakiwsky, E.J. 1982. *Geodesy: The Concepts*. North-Holland, Amsterdam.

Further Information

Articles on advances in the general field, and particularly in observational data adjustment, can be found in the following periodicals:

Bulletin Géodésique, published by Springer International

Manuscripta Geodetica, published by Springer International

Photogrammetric Engineering and Remote Sensing, published by the American Society for Photogrammetry and Remote Sensing, Bethesda, MD

Surveying and Land Information Systems, published by the American Congress on Surveying and Mapping, Bethesda, MD

The Photogrammetric Record, published by The Photogrammetric Society, London, England

Photogrammetria, Journal of the International Society for Photogrammetry and Remote Sensing, published by Elsevier, Amsterdam, The Netherlands

Geomatica, Journal of the Canadian Institute of Geomatics, Ottawa, Canada

54

Plane Surveying

54.1 Introduction

54.2 Distance Measurement

Tacheometry • Taping • Electronic Distance Measurement

54.3 Elevation Measurement

Benchmark (BM) • Turning Point (TP) • Types of Instruments • Fundamental Relationships • Ordinary Differential Leveling • Leveling Closure • Precise Leveling • Trigonometric Leveling

54.4 Angle Measurement

Horizontal and Vertical Angles • Direction Angles • Types of Instruments • Instrument Components • Fundamental Relationships • Instrument Operation

54.5 Plane Survey Computations

Traverse • Partitioning Land

54.6 Horizontal Curves

54.7 Vertical Curves

54.8 Volume

General • Field Measurements for Volume Computations • Area • Volume Computations — Road Construction • Volume Computations — Building Excavation • Summary

Steven D. Johnson

Purdue University

Wesley G. Crawford

Purdue University

54.1 Introduction

The roots of surveying are contained in plane-surveying techniques that have developed since the very first line was measured. Though the fundamental processes haven't changed, the technology used to make the measurements has improved tremendously. The basic methods of distance measurement, angle measurement, determining elevation, etc., are becoming easier, faster, and more accurate.

Even though electronic measurement is becoming commonplace, distances are still being taped. While many transits and optical theodolites are still being used for layout purposes, electronic instruments are rapidly becoming the instruments of choice by the engineer for measurement of angles. Although establishing elevations on the building site is still being performed using levels, the laser is everywhere. It is the responsibility of the engineer in the field to choose the measuring method and the technology that most efficiently meet the accuracy needed.

The requirements of construction layout are more extensive than ever, with the complex designs of today's projects. The engineer must be exact in measurement procedures and must check and double-check to ensure that mistakes are eliminated and errors are reduced to acceptable limits to meet the tolerances required.

This chapter is directed to the engineer who will be working in the field on a construction project. Often that individual is called the *field engineer*, and that term will be used throughout this chapter. The field engineer should be aware of fundamental measuring techniques as well as advances in the technology of measurement.

Although measurement concepts and calculation procedures are introduced and discussed in this chapter, only the basics are presented. It is the responsibility of the field engineer to review other sources for more detailed information to become more competent in these procedures. See the further information and references at the end of this chapter.

54.2 Distance Measurement

Distance may be measured by indirect and direct measurement procedures. Indirect methods include odometers, optical rangefinders, and tacheometry. Direct methods include pacing, taping, and electronic distance measurement.

When approximate distances are appropriate, pacing can be used over short distances and odometers or optical rangefinders can be used over longer or inaccessible distances. These methods can yield accuracies in the range of 1 part in 50 to 1 part in 100 over modest distances. The accuracy of pacing and optical distance measuring methods decreases rapidly as the distance increases. However, these approximate methods can be useful when checking for gross blunders, narrowing search areas in the field, and making preliminary estimates for quantities or future surveying work. Applications where these lower-order accuracy methods can be used to obtain satisfactory distance observations occur regularly in surveying, construction and engineering, forestry, agriculture, and geology.

Tacheometry

Tacheometry uses the relationship between the angle subtended by a short base distance perpendicular to the bisector of the line and the length of the bisecting line. Stadia and subtense bar are two tacheometric methods capable of accuracies in the range of 1 part in 500 to 1 part in 1000.

In the stadia method the angle is fixed by the spacing of the stadia cross hairs (i) on the telescope reticule and the focal length (f) of the telescope. Then the distance on the rod (d) is observed, and the distance is computed by

$$\frac{D}{d} = \frac{f}{i}$$

$$D = d \left(\frac{f}{i} \right)$$

When the telescope is horizontal, as in a level, a horizontal distance is obtained. When the telescope is inclined, as in a transit or theodolite, a slope distance is obtained. Stadia may be applied with level or transit, plane table and alidade, or self-reducing tacheometers.

The subtense bar is a tacheometric method in which the base distance, d , is fixed by the length of the bar and the angle, α , is measured precisely using a theodolite. The horizontal distance to the midpoint of the bar is computed by

$$\tan \frac{\alpha}{2} = \frac{d/2}{D}$$

$$D = \frac{d}{2 \tan \left(\frac{\alpha}{2} \right)}$$

Since a horizontal angle is measured, this method yields a horizontal distance and no elevation information is obtained.

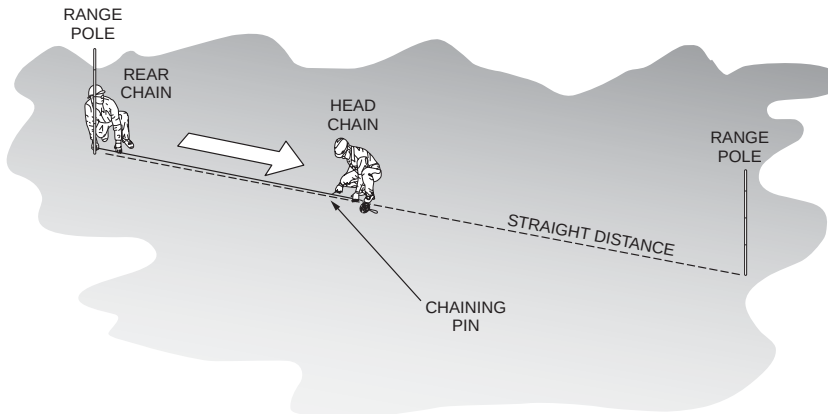


FIGURE 54.1 Taping on level terrain.

Taping

Taping is a direct method of measuring distance in which a tape of known length and graduated at intervals throughout its length is stretched along the line to be measured. [Figure 54.1](#) illustrates taping on level terrain. Low-order accuracy measurements can be obtained using tapes made of cloth or fiberglass material. These tapes are available in a variety of lengths, and they may be graduated in feet, meters, or both. Accurate taped distances are obtained using calibrated metal ribbon tapes made of tempered steel. Tapes made of invar steel for temperature stability are also available. [Table 54.1](#) summarizes the sources of errors and the procedures required to measure distances accurately.

TABLE 51.1 Taping Errors and Procedures

Error	Source	Type	Makes Tape	Importance	Procedure to Eliminate
Tape length	Instrumental	Systematic	Long or short	Direct impact to recorded measurement, always check.	Calibrate tape and apply adjustment.
Temperature	Natural	Systematic or random	Long or short	For a chain standardized at 68°, 0.01 per 15° per 100 ft chain.	Observe temperature of chain, calculate, and apply adjustment.
Tension	Personal	Systematic or random	Long or short	Often not important.	Apply the proper tension. When in doubt, PULL HARD!
Tape not level	Personal or natural	Systematic	Short	Negligible on slopes less than 1%; must be calculated for greater than 1%.	Break chain by using a hand level and plumb bob to determine horizontal; or correct by formula for slope or elevation difference.
Alignment	Personal	Systematic	Short	Minor if less than 1 ft off of line in 100 ft; major if 2 or more ft off of line.	Stay on line; or determine amount off of line and calculate adjustment by formula.
Sag	Natural or personal	Systematic	Short	Large impact on recorded measurement.	Apply proper tension; or calculate adjustment by formula.
Plumbing	Personal	Random	Long or short	Direct impact to recorded measurement.	Avoid plumbing if possible; or plumb at one end of chain only.
Interpolation	Personal	Random	Long or short	Direct impact to recorded measurement.	Check and recheck any measurement that requires estimating a reading.
Improper marking	Personal	Random	Long or short	Direct impact to recorded measurement.	Mark all points so that they are distinct.

TABLE 54.2 Tape Corrections for Systematic Errors

Systematic Error	Correction Formula	0.01-ft Correction Caused by
Tape length	$C_l = \frac{L_{\text{tape}} - 100.00}{100.00}$	0.01 per 50-ft tape length
Tension	$C_p = \frac{(P - P_0)L}{AE}$	±9 lb
Temperature	$C_t = \alpha(T - T_0)L$	±15°F
Sag	$C_s = \frac{-w^2 L_s^3}{24P^2}$	−5 lb in 100-ft tape length
Alignment	$C_a = -\frac{h}{2L}$	±1.4 ft

Accuracy in taping requires a **calibrated** tape that has been compared to a length standard under known conditions. A tape calibration report will include the true length of the tape between marked intervals on the tape. The tape is calibrated when it is fully supported throughout its length at a specified pull or tension applied to the tape and at a specified temperature of the tape. The calibration report should also include the cross-sectional area of the tape, the weight of the tape per unit length, the modulus of elasticity for the tape material, and the coefficient of thermal expansion for the tape material. When a distance is observed in the field and the conditions of tape support, alignment, temperature, and tension are recorded, the **systematic errors** affecting the measurement can be corrected and a true distance obtained. The tape correction formulas are summarized in the following example.

Example 54.1

A typical 100-foot surveyor's steel tape is calibrated while supported throughout its length. The calibration report lists the following:

$$\text{True length, 0-ft to 100-ft mark} \quad L_0 = 99.98 \text{ ft}$$

Other length intervals may also be reported (and the tape characteristic constants):

Tension	$P_0 = 25 \text{ lb}$
Temperature	$T_0 = 68^\circ\text{F}$
Cross-sectional area	$A = 0.003 \text{ in.}^2$
Weight	$w = 0.01 \text{ lb/ft}$
Coefficient of thermal expansion	$\alpha = 0.00000645/^\circ\text{F}$
Modulus of elasticity	$E = 29(10^6) \text{ lb/in.}^2$

Systematic errors in taping, the correction formula, and the change required to cause a 0.01-foot tape correction in a full 100-foot length of the example tape are summarized in [Table 54.2](#). When a distance is measured directly by holding the tape horizontal, the effect of alignment error is present in both the alignment of the tape along the direction of the line to be measured and the vertical alignment necessary to keep the tape truly horizontal.

Normal taping procedures require that the tape be held horizontal for each interval measured. When taping on sloping terrain or raising the tape to clear obstacles, a plumb bob is required to transfer the tape mark to the ground. Steadying the hand-held tape and plumb bob will be the largest source of **random error** in normal taping. Accuracies of 1 part in 2000 to 1 part in 5000 can be expected.

Precise taping procedures eliminate the use of the plumb bob by measuring a slope distance between fixed marks that either are on the ground or are supported on tripods or taping stands. Taping the slope distance, S , will require that the horizontal distance, H , be computed by

$$H = S \cos \alpha = S \sin z$$

where α = **vertical** angle of the tape and z = **zenith** angle of the tape, or by

$$H = \sqrt{S^2 - h^2}$$

where h = difference in elevation between the ends of the tape. If all tape corrections are carefully applied, accuracies of 1 part in 10,000 to 1 part in 20,000 can be obtained.

Electronic Distance Measurement

Distance can be measured electronically if the velocity and travel time of electromagnetic energy propagated along a survey line are determined. Terrestrial electronic distance measurement instruments (EDMIs) measure the travel time by comparing the phase of the outgoing measurement signal to the phase of the signal returning from the remote end of the line. The phase difference is thus a function of the double path travel of the measurement signal. The distance is given by

$$D = \frac{VT}{2}$$

where V = the velocity of electromagnetic energy in the atmosphere and T = the double path travel time determined using the phase difference.

If $\Delta\phi$ is the phase difference observed for the fine or shortest wavelength measurement signal, then the total travel time is found from the equation

$$T = k \left(\frac{1}{f} \right) + \frac{\Delta\phi}{2\pi} \left(\frac{1}{f} \right)$$

where f = the frequency of the measurement signal and k = the integer number of full cycles in the double path distance. The integer k is ambiguous for the fine measurement since the phase difference only determines the fractional part of the last cycle in the double path distance, as illustrated in Fig. 54.2. The value of k can be determined by measuring the phase difference of one or more coarse- or long-wavelength signals to resolve the distance to the nearest full cycle of the fine wavelength.

EDMIs may be classified by type of energy used to carry the measurement signal or by the maximum measurement range of the system. Visible (white) light, infrared light, laser (red) light, and microwaves have been used as carrier energy.

The velocity of electromagnetic energy in the atmosphere is given by the expression

$$V = \frac{c}{n}$$

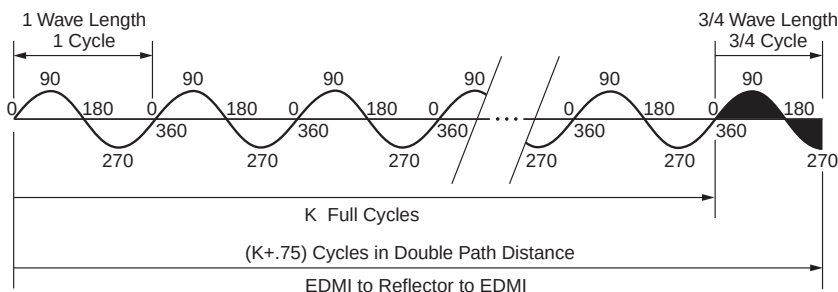


FIGURE 54.2 Phase shift principle of distance measurement.

where c is the velocity in a vacuum, 299,792,158 meters/second, and n is the atmospheric index of refraction for the conditions at the time of observation. The atmospheric index of refraction is a function of the wavelength of the electromagnetic energy propagated and the existing conditions of atmospheric temperature, pressure, and water vapor pressure. For EDMIs using visible, laser, or infrared light carrier wavelengths, the effect of water vapor pressure is negligible, and it is often ignored. For EDMIs using microwave carrier wavelengths, the effect of water vapor pressure is more significant.

Refraction causes a scale error in the observation. The correction is usually expressed in terms of parts per million, ppm, of the distance measured.

$$D_{\text{corrected}} = D + S \left(\frac{D}{10^6} \right)$$

The ppm correction for electro-optical instruments can be expressed in the form

$$S = A + \frac{Bp}{t + 273.2}$$

where p is the atmospheric pressure and t is the atmospheric temperature. The constants A and B are functions of the wavelength of the carrier and the precise frequency used for the highest-resolution measuring signal. The values of A and B can be obtained from the instrument manufacturer. Typically, the ppm value, S , is determined graphically using pressure and temperature read in the field, and the value is entered into the EDM. On many modern digital instruments, the pressure and temperature readings can be entered directly, and the instrument computes and applies the refraction correction.

The measurement accuracy of an EDM is expressed as

$$\sigma = \pm (c + s_{\text{ppm}})$$

In this expression, c is a constant that represents the contribution of uncertainty in the offset between the instrument's measurement reference point and the geometric reference point centered over the survey station. The ppm term is a distance-dependent contribution representing the uncertainty caused by measurement frequency drift and atmospheric refraction modeling. An EDM should be checked periodically to verify that it is operating within its specified error tolerance, σ . A quick check can be done by measuring a line of known length or at least a line that has been measured previously to see if the observed distance changes. When a more rigorous instrument calibration is warranted, use a calibrated base line and follow the procedures recommended in *NOAA Technical Memorandum NOS NGS-10* [Fronczek, 1980]. EDMs should be sent to the manufacturer for final calibration and adjustment.

54.3 Elevation Measurement

Elevation is measured with respect to a datum surface that is everywhere perpendicular to the direction of gravity. The datum surface most often chosen is called the *geoid*. The geoid is an equipotential surface that closely coincides with **mean sea level**. Elevations measured with respect to the geoid are called orthometric heights. The relationships between the mean sea level geoid, a level surface, and a horizontal line at a point are illustrated in [Fig. 54.3](#).

Benchmark (BM)

A benchmark is best described as a permanent, solid point of known elevation. Benchmarks can be concrete **monuments** with a brass disk in the middle, iron stakes driven into the ground, or railroad spikes driven into a tree, etc.

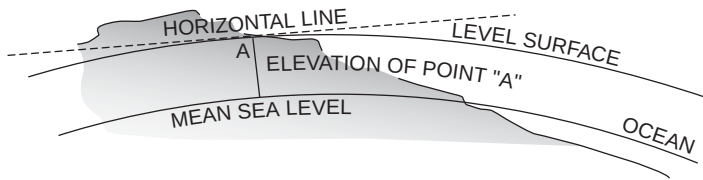


FIGURE 54.3 Elevation datum and level surface.

Turning Point (TP)

A turning point is a point used in the differential leveling process to temporarily transfer the elevation from one setup to the next.

Types of Instruments

Elevation may be measured by several methods. Some of these methods measure elevation directly, whereas some measure the difference in elevation from a reference benchmark to the point to be determined. Many types of instruments have been used for leveling purposes. The instruments have ranged from water troughs and hoses to instruments such as the barometer, dumpy level, automatic level, and laser. The field engineer should realize how each of these can be used to determine elevations for vertical control and construction.

Altimeters

Surveying altimeters are precise aneroid barometers that are graduated in feet or meters. As the altimeter is raised in elevation, the barometer senses the atmospheric pressure drop. The elevation is read directly on the face of the instrument. Although the surveying altimeter may be considered to measure elevation directly, best results are obtained if a difference in elevation is observed by subtracting readings between a base altimeter kept at a point of known elevation and a roving altimeter read at unknown points in the area to be surveyed. The difference in altimeter readings is a better estimate of the difference in elevation because local weather changes, temperature, and humidity that affect altimeter readings are canceled in the subtraction process. By limiting the distance between base and roving altimeters, accuracies of 3 to 5 feet are possible. Other survey configurations utilizing low and high base stations, and/or leap-frogging roving altimeters, can yield good results over large areas.

Level Bubble Instruments

Level bubble instruments include the builder's transit level, the transit, and the dumpy level. Each of these instruments contains a level vial with a bubble that must be centered to be used for leveling. Each instrument consists of three main components: a four-screw leveling head, a level vial attached to the telescope, and a telescope for magnification of the objective.

Instruments that use a level bubble to orient the axes to the direction of gravity depend on the bubble's sensitivity for accuracy. Level bubble sensitivity is defined as the central angle subtended by an arc of one division on the bubble tube. The smaller the angle subtended, the more sensitive the bubble is to dislevelment. A bubble division is typically 2 millimeters long, and bubble sensitivity typically ranges from 60 seconds to 1 second.

Builder's Level. The builder's level is one of the most inexpensive and versatile instruments that is used by field engineers. In addition to being able to perform leveling operations, it can be used to turn angles, and the scope can be tilted for sighting. Many residential builders use this instrument because it serves their purpose of laying out a building. See [Fig. 54.4](#).

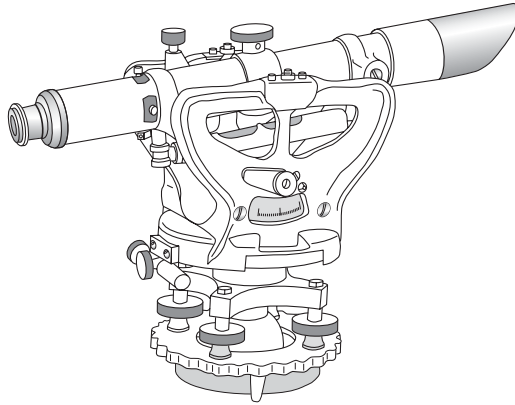


FIGURE 54.4 Builder level.

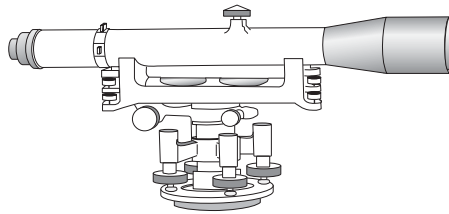


FIGURE 54.5 Engineer's dumpy level.

Transit. Although the primary functions of the transit are for angle measurement and layout, it can also be used for leveling because there is a bubble attached to the telescope. In fact, many construction companies who don't have both an angle measuring instrument and a dumpy or automatic level will do all of their leveling work with a transit. Some people prefer to use the transit for leveling because they are comfortable with its operation. However, the field engineer should be aware that the transit may not be as sensitive and stable as a quality level.

Dumpy Level. The engineer's dumpy level shown in Fig. 54.5 has been the workhorse of leveling instruments for more than 150 years. It has been used extensively for many of the great railroad, canal, bridge, tunnel, building, and harbor projects for the last century and a half. Even with advancements in other leveling instruments such as the automatic level and the laser, the dumpy is still the instrument of choice for a number of persons in construction because of its stability. On any type of project where there is going to be a great deal of vibration — such as pile driving, heavy-equipment usage, or high-rise construction — the dumpy may be the best choice for a leveling instrument.

Automatic Compensator Instruments

Compensator instruments illustrated in Fig. 54.6 were developed about 50 years ago. Although each manufacturer may have developed a unique compensator, all compensators serve the same purpose — maintaining a fixed relationship between the line of sight and the direction of gravity. If the instrument is in adjustment, the line of sight will be maintained as a horizontal line. The operation of a compensator is illustrated in Fig. 54.7. Compensator instruments are extremely fast to set up and level. An experienced person can easily have an automatic level ready for a backsight in less than ten seconds, compared to a minute or more with a bubble-based leveling system. Compensators are available in several styles. Some are constructed

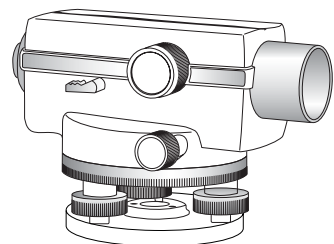


FIGURE 54.6 Automatic level.

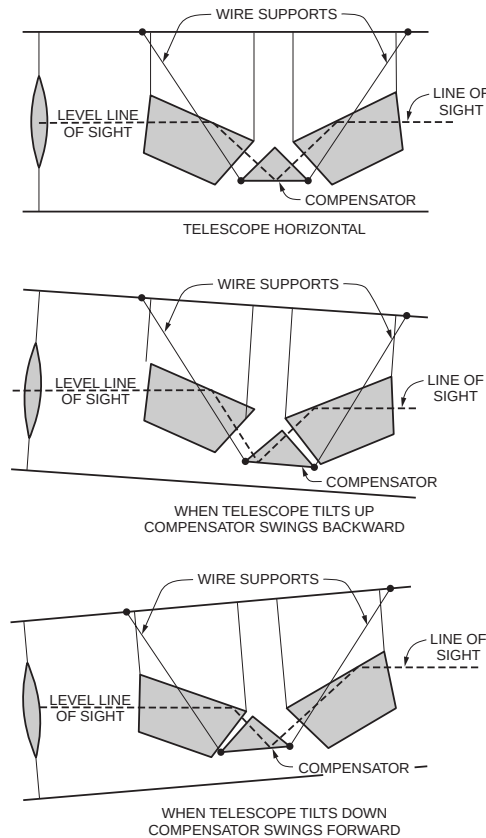


FIGURE 54.7 Leveling the line of sight by compensator.

by suspending a prism on wires. Some have a prism that is contained within a magnetic dampening system. Note that the instrument must be manually leveled to within the working range of the compensator by centering a bull's-eye bubble.

Laser Levels

A laser level uses a laser beam directed at a spinning optical reflector. The reflector is oriented so that the rotating laser beam sweeps out a horizontal reference plane. The level rod is equipped with a sensor to detect the rotating beam. By sliding the detector on the rod, a vertical reading can be obtained. Laser levels are especially useful on construction sites. As the laser beam continuously sweeps out a constant elevation reference plane, anyone with a detector can get a rod reading to set or check grade elevations. Laser levels can be equipped with automatic leveling devices to maintain level orientation. The spinning optics can also be oriented to produce a vertical reference plane.

Digital Levels

Digital levels are electronic levels that can be used to more quickly obtain a rod reading and make the reading process more reliable. The length scale on the level rod is replaced by a bar code. The digital level senses the bar code pattern and compares it to a copy of the code held in its internal memory. By matching the bar code pattern, a rod-reading length can be obtained.

Level Rods

In addition to the leveling instrument, a level rod is required to be able to transfer elevations from one point to another. The level rod is a graduated length scale affixed to a rod and held vertically on a turning

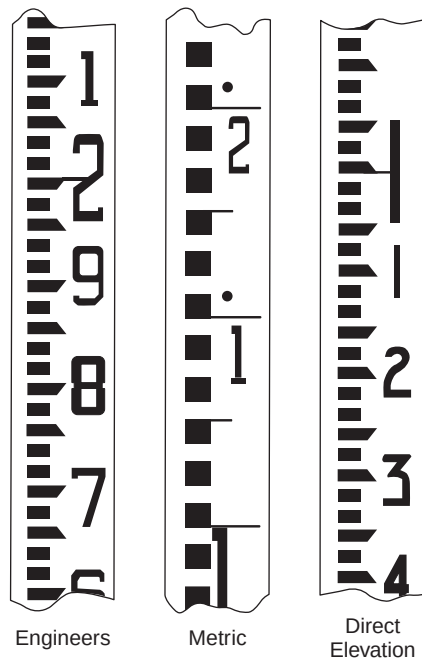


FIGURE 54.8 Typical level rod graduations.

point or benchmark. The scale is read by the person at the instrument. The reading taken is the vertical distance from the point to the line of sight.

Level rods are available in many sizes, shapes, and colors. They are made of wood, fiberglass, metal, or a combination of these materials. There are one-piece rods, two-piece rods, three-piece rods, six-piece rods, etc. Some have a square cross section and others are round or oval. Some are less than 10 feet long, while others are up to 30 feet long. Practically whatever type of rod a field engineer needs is available. Level rods have been named after the cities where they have been manufactured. The Philadelphia rod, for example, is a two-piece rod that can be extended to approximately 13 feet. It is a popular rod used in leveling surveys.

A popular rod that doesn't seem to have a proper name is the telescoping rod. Because a telescoping rod is 25 or 30 feet long, it is the rod of choice for field engineers working on projects where there is a great deal of change in elevation. Great rod lengths increase the elevation difference that can be transferred at one time. It has been argued that these types of rods wear rapidly and therefore aren't as accurate as the more traditional rods. If telescoping rods are well cared for, they are excellent for construction use.

Level rods are graduated in feet, inches, and fractions; feet, tenths, and hundredths; or meters and centimeters. The method of representing units of measurement onto the face of the rod also varies. Typical level rod graduations are shown in Fig. 54.8. The field engineer should, after studying the graduations on the face of the rod, be able to use any rod available. The rod illustrated on the left in Fig. 54.8 shows the markings on a typical "engineer's rod" that is widely used on the construction site. Note that the rod is graduated to the nearest foot with large numbers that are usually painted red. The feet are then graduated to the nearest tenth from 1 to 9. Each tenth is then graduated to the nearest one hundredth, which is the width of the smallest mark on the rod.

Rod Targets

Rod targets are useful devices for several purposes. The rod target can be used as a target by the person looking through the instrument to help locate the rod when visibility conditions are poor. If the rod target has a vernier, it can also be used to obtain a reading on the rod to the nearest thousandth of a foot. In this case the instrument person communicates to the rod person to move the target until it has

centered on the horizontal cross hair. The rod person can then read the rod and the vernier to obtain thousandths of a foot. This accuracy is sometimes required on very precise leveling work.

Rod Levels

Rod levels are used to keep the level rod plumb while the reading is being taken. Rod levels are simple devices made of metal or plastic and have a bull's-eye bubble attached. They are held along the edge of the level rod while the level rod is moved until the bubble is centered. If the rod level is in proper adjustment, the rod is plumb. When a rod level is not used, the rod person should slowly rock the rod through the vertical position. The instrument person watches the cross hair appear to move on the rod and records the lowest reading — the point when the rod was vertical.

Turning Point Pin

If a solid natural point is not available during the leveling process to be used as a turning point, then the field engineer will have to create one. This is accomplished by carrying something such as a railroad spike, piece of rebar, wooden stake, or plumb bob that can be inserted solidly into the ground. The rod is placed on top of the solid point while the foresight and backsight readings are taken. These solid points are removed after each backsight to be used the next time a solid turning point is needed.

Fundamental Relationships

All level instruments are designed around the same fundamental relationships and lines. The principal relationships among these lines are described as follows:

- The **axis of the level bubble** (or compensator) should be perpendicular to the vertical axis.
- The **line of sight** should be parallel to the axis of the level bubble.

When these instrument adjustment relationships are true and the instrument is properly set up, the line of sight will sweep out a horizontal plane that is perpendicular to gravity at the instrument location. However, several effects must be considered if the instrument is to be used for differential leveling.

Earth Curvature

The curved shape of the earth means that the level surface through the telescope will depart from the horizontal plane through the telescope as the line of sight proceeds to the horizon. This effect makes actual level rod readings too large by

$$C = 0.0239D^2$$

where D is the sight distance in thousands of feet.

Atmospheric Refraction

The atmosphere refracts the horizontal line of sight downward, making the level rod reading smaller. The typical effect of refraction is equal to about 14% of the effect of earth curvature. Thus, the combined effect of curvature and refraction is approximately

$$(C - r) = 0.0206D^2$$

Instrument Adjustment

If the geometric relationships defined above are not correct in the leveling instrument, the line of sight will slope upward or downward with respect to the horizontal plane through the telescope. The method of testing the line of sight of the level to ensure that it is horizontal is called the *two-peg test*. It requires setting up the level exactly between two points about 200 feet apart and subtracting readings taken on the points to determine the true difference in elevation between them. The instrument is then moved and placed adjacent to one of the points, and rod readings are again taken on the two points. If the line

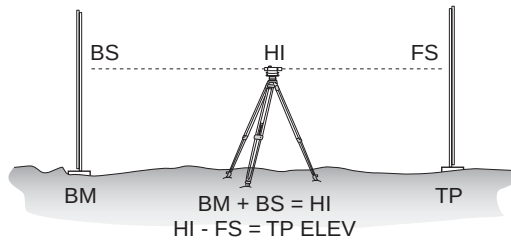


FIGURE 54.9 Differential leveling.

of sight of the instrument is truly horizontal, the difference in elevation obtained from the two setups will be the same. If the line of sight is inclined, the difference in elevation obtained from the two setups will not be equal. Either the instrument must be adjusted, or the slope of the line of sight must be calculated. The slope is expressed as a collimation factor, C , in terms of rod-reading correction per unit sight distance. It may be applied to each sight by

$$\text{Corrected rod reading} = \text{Rod reading} + (C_{\text{Factor}} \cdot D_{\text{Sight}})$$

In ordinary differential leveling discussed next, these effects are canceled in the field procedure by always setting up so that the backsight distance and foresight distance are equal. The errors are canceled in the subtraction process. If long unequal sight distances are used, the rod readings should be corrected for curvature and refraction and for collimation error.

Ordinary Differential Leveling

Determining or establishing elevations is, at times, the most essential activity of the field engineer. Elevations are needed to set slope stakes, grade stakes, footings, anchor bolts, slabs, decks, sidewalks, curbs, etc. Just about everything located on the project requires elevation. Differential leveling is the process used to determine or establish those elevations.

Differential leveling is a very simple process based on the measurement of vertical distances from a horizontal line. Elevations are transferred from one point to another through the process of using a leveling instrument to read a rod held vertically on, first, a point of known elevation and, then, on the point of unknown elevation. Simple addition and subtraction are used to calculate the unknown elevations.

A single-level setup is illustrated in Fig. 54.9. A backsight reading is taken on a rod held on a point of known elevation. That elevation is transferred vertically to the line of sight by reading the rod and then adding the known elevation and the backsight reading. The elevation of the line of sight is the height of instrument, HI . By definition, the line of sight is horizontal; therefore, the line of sight elevation can then be transferred down to the unknown elevation point by turning the telescope to the foresight and reading the rod. The elevation of the foresight station is found by subtracting the rod reading from the height of instrument. Note that the difference in elevation from the backsight station to the foresight station is determined by subtracting the foresight rod reading from the backsight rod reading.

A level route consists of several level setups, each one carrying the elevation forward to the next foresight using the differential-leveling method. Figure 54.10 shows a short level route and illustrates the typical format used in the field for differential level notes.

Leveling Closure

Level route **closure** is obtained by taking the last foresight on a benchmark of known elevation. If a second benchmark is not available near the end of the level route, the route should be looped back to the starting benchmark to obtain closure. At the closing benchmark,

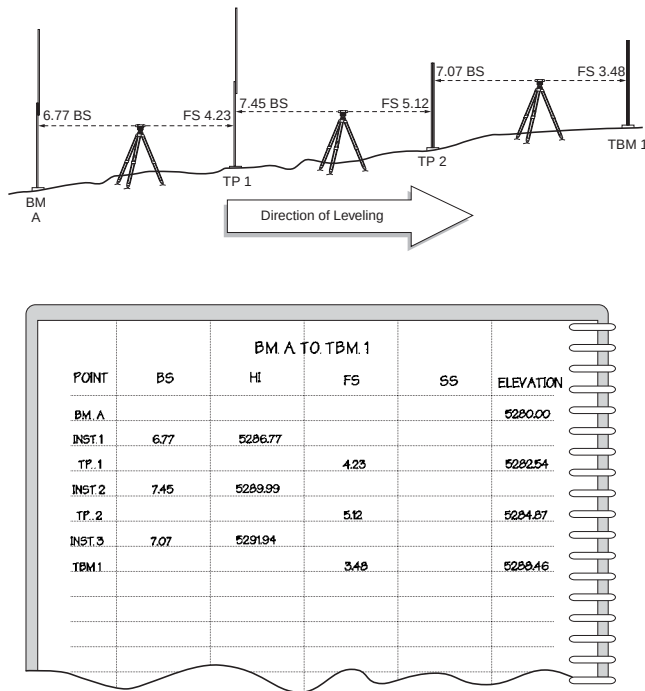


FIGURE 54.10 Level route and field note form.

$$\text{Closure} = \text{Computed elevation} - \text{Known elevation}$$

Since differential leveling is usually performed with approximately equal setup distances between turning points, the level route is adjusted by distributing the closure equally to each setup:

$$\text{Adjustment} = \left(-\frac{\text{Closure}}{n} \right) \text{per setup}$$

where n is the number of setups in the route. When a network of interconnected routes is surveyed, a **least-squares** adjustment is warranted.

Precise Leveling

Precise leveling methods are required for engineering work that requires extreme accuracy. The process requires that special equipment and methods be used. Instruments used in precise leveling are specifically designed to obtain a high degree of accuracy in leveling. Improved optics in the telescope, improved level sensitivity, and carefully calibrated rod scales are all incorporated into the differential-leveling process. Methods have been developed to ensure that mistakes are eliminated and errors are minimized.

Typically when performing precise leveling, a method of leveling called *three-wire leveling* is used. This involves reading the center cross hair as well as the upper and lower “stadia” cross hairs. The basic process of leveling is the same, except that the three cross hair readings are averaged to improve the **precision** of each backsight and foresight value.

Another method that can be used to improve the precision of the level rod reading involves using an optical micrometer on the telescope. The optical micrometer is a rotating parallel-plate prism attached in front of the objective lens of the level. The prism enables the observer to displace the line of sight parallel with itself and set the horizontal cross hair exactly on the nearest rod graduation. The observer

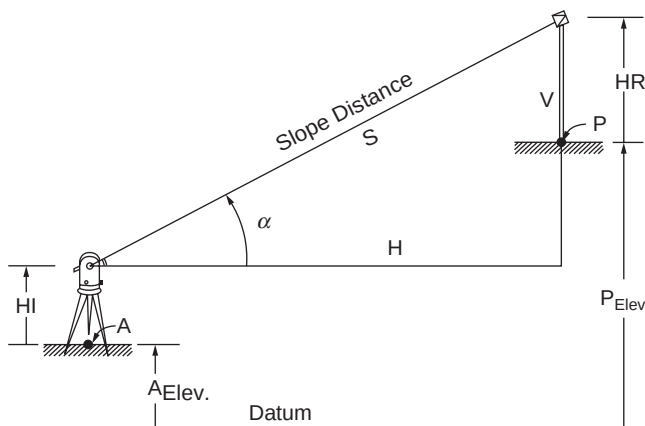


FIGURE 54.11 Trigonometric leveling in plane surveying.

adds the middle cross hair rod reading and the displacement reading on the micrometer to obtain a rod reading precise to the nearest 0.1 millimeter.

When long precise level routes are surveyed, it is necessary to account for the fact that the level surfaces converge as the survey proceeds north. The correction to be applied for convergence of level surfaces at different elevations can be calculated by

$$\text{Correction} = -0.0053 \sin 2\phi H \Delta\phi_{\text{rad}}$$

where ϕ is the latitude at the beginning point, H is the elevation at the beginning point, and $\Delta\phi$ is the change in latitude from beginning point to end point in radians.

Trigonometric Leveling

Trigonometric leveling is a method usually applied when a total station is used to measure the slope distance and the vertical angle to a point. This method is illustrated in Fig. 54.11. Assuming the total station is set up on a station of known elevation, the elevation of the unknown station is

$$V = S \sin \alpha$$

$$P_{\text{Elev}} = A_{\text{Elev}} + HI + V - HR$$

The precision of trigonometric elevations is determined by the uncertainty in the vertical angle measurement and the uncertainty in the atmospheric refraction effects. For long lines the effects of earth curvature and atmospheric refraction must be included.

54.4 Angle Measurement

Horizontal and Vertical Angles

The angular orientation of a line is expressed in terms of horizontal angles and vertical angles. Survey angles illustrated in Fig. 54.12 are defined by specifying the plane that contains the angle, the reference line or plane where the angle starts, the direction of turning, and the terminal line or plane where the angle ends.

Field angles that are measured in the horizontal plane include clockwise angles, counterclockwise angles, and deflection angles. The clockwise angle or “angle to the right,” shown in Fig. 54.13, is measured from a backsight line, clockwise in the horizontal plane, to a foresight line. All transits or theodolites

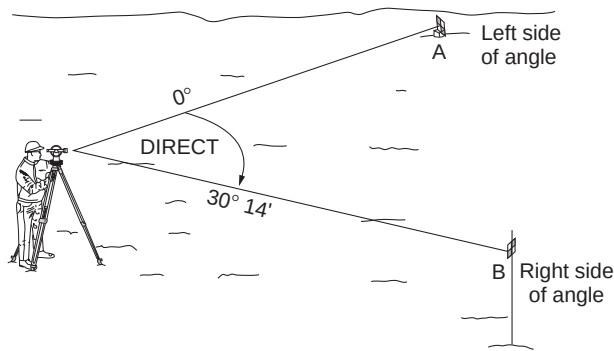


FIGURE 54.12 Horizontal survey angle.

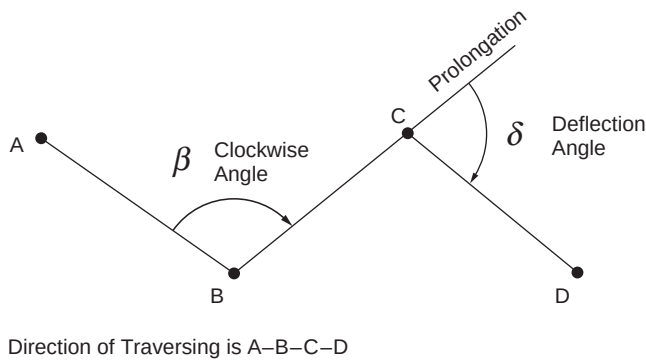


FIGURE 54.13 Horizontal field angles.

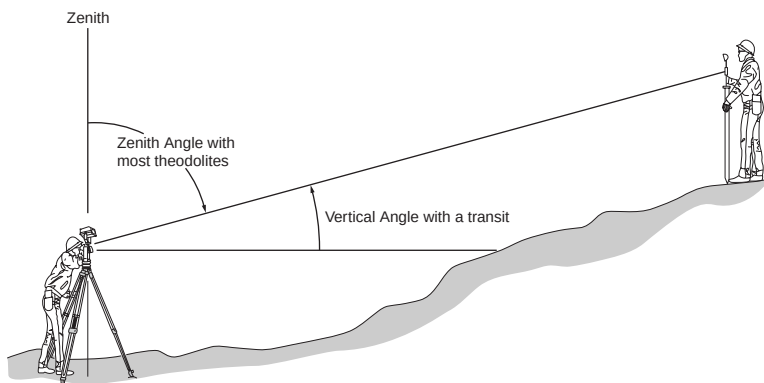


FIGURE 54.14 Vertical field angles.

have horizontal circles marked to read clockwise angles. Transits used to stake out projects typically include counterclockwise markings to facilitate laying out angles to either side of a reference line. A deflection angle, also shown in Fig. 54.13, is measured from the prolongation of the backsight line either left or right to the foresight line. It is commonly used in route centerline surveys.

Vertical angles are measured in the vertical plane as shown in Fig. 54.14. An instrument with a vertical circle graduated to measure vertical angles reads 0° when the instrument is level and the telescope is in the horizontal plane. The vertical angle increases to $+90^\circ$ above the horizontal and to -90° below the

horizontal plane. Many instruments today have vertical circles graduated to measure the zenith angle. Zero degrees is at the zenith point directly overhead on the vertical axis. The zenith angle will increase from 0° at the zenith to 90° at the horizon. The zenith angle is greater than 90° if the telescope is pointing below the horizon. The circle graduations continue to 180° at the nadir, to 270° at the horizon with the telescope reversed, and to 360° closing the circle at the zenith.

Direction Angles

Direction angles are horizontal angles from a north–south reference meridian to a survey line. Direction angles that are computed from field angles include bearings and azimuths. A bearing, as shown in Fig. 54.15, is the acute angle from the reference meridian to the survey line. It always includes the quadrant designation. An azimuth, as shown in Fig. 54.16, is the angle from the north end of the reference meridian clockwise to the survey line. It can have a magnitude from 0° to 360° , and a quadrant designation is not necessary.

Several different meridians can be used as a reference for direction. The direction of north at a point can be defined as any of the following:

- Astronomic north
- Geodetic north
- Magnetic north
- Grid north
- Assumed north

The first three types of meridians converge to their respective north poles on the earth. Therefore, they do not form the basis for a Cartesian coordinate system to be used for plane surveying. There is only a small angular difference between astronomic and geodetic north, but astronomic north is typically taken to be synonymous with the term “true north.”

Magnetic north is the direction of the earth’s magnetic field at a point. A declination angle measured from the geodetic meridian to the magnetic meridian defines the relationship between the two systems. However, since the direction of magnetic north changes with time, proper use of magnetic north must account for variation in the magnetic declination angle.

Grid north is used in plane surveying. The grid should be defined on an appropriate map projection, such as the state plane coordinate systems available in each state. Then a defined relationship exists between the geodetic and grid meridians.

Assumed north may be used for preliminary surveys, but it is not recommended for permanent work. If the stations on the ground are lost or destroyed, the basis of direction is lost and the survey cannot be retraced.

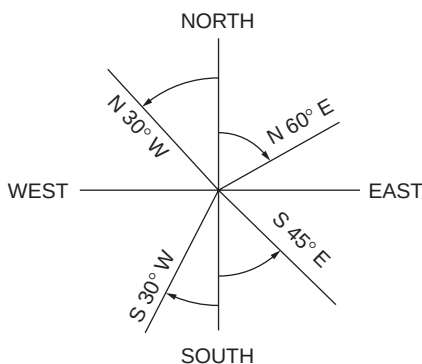


FIGURE 54.15 Bearing direction angles.

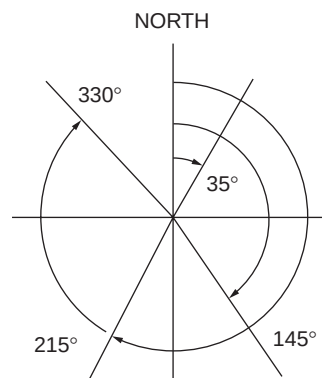


FIGURE 54.16 Azimuth direction angles.

Types of Instruments

There are many types of angle-measuring instruments used in surveying and construction. They include transits, theodolites, digital theodolites, and total stations. Even though the names are different, they look different, and the operation is slightly different, they all are used for the same purpose — angle measurement and layout.

Although there are a number of differences in the construction and features of transits, optical theodolites, and digital theodolites, the major difference is in the method of reading angles. Instruments are often classified according to the smallest interval, or so-called least count, that can be read directly in the instrument. The least count may range from 1 minute for a construction transit to 0.1 seconds for a first-order theodolite. Least counts from 10 seconds to 1 second are common in surveying and engineering instruments. The field engineer should become familiar with the instrument being used on the project.

Transit

The transit was developed to its present form during the 1800s. It has been used on construction projects from railroads across the wild west to the skyscrapers of the modern city. A good, solid, and reliable instrument, the transit is still used by many contractors today. However, optical theodolite technology, and now digital electronic technology, have passed it by. The major companies who manufactured transits have dropped them from their product line. They are slower to use than modern instruments, and they are not as easily adaptable to having an EDM attached to them. Its major contribution to the field engineer today is that it is a good tool for learning the fundamentals of angle measurement. Its parts are exposed, making it easier to see what is going on in the manipulation of the clamps during the angle measurement process. By understanding the transit, one can easily move on to any other type of angle-measuring equipment.

Repeating Optical Theodolite

The repeating theodolite contains the same upper and lower clamp system as the transit; however, the reading of the angle is different because of the optical-reading capability. *Optical theodolite* is a term that was originally applied in Europe to instruments similar to the transit. However, as instrument technology progressed, *theodolite* became synonymous with a style of instrument that was enclosed, used a magnified optical system to read the angles, had a detachable tribrach with an optical plummet, used a three-screw leveling system, and was more precise than the transit. These features have made it much easier to use than the transit. The better optical theodolites have been “delicate” workhorses since they were introduced. That is, if they are properly cared for, they seem as though they will last forever because of their excellent construction and quality materials. However, they must be handled gently and carefully. A typical optical theodolite may have as many as 20 prisms or lenses as part of the optical angle-reading system. With a sharp bump, these can get out of alignment, which may render the instrument unusable. As with any surveying instrument, the theodolite cannot be exposed to inclement weather because of the optical system.

Scale-Reading Optical Theodolite. The typical scale-reading theodolite has a glass circle with a simple scale that is read directly. The scale is read where it is intersected by the degree readings from the circle. See [Fig. 54.17](#) for an example scale reading. Simply read the degree that shows up in the window and observe where the degree index mark intersects the scale. Both the horizontal circle and the vertical circle are generally observed at the same time.

Micrometer-Reading Optical Theodolite. The micrometer-reading instrument also has a glass circle, but it does not have a scale. An adjustable micrometer is used to precisely read the circle and subdivide the degree intervals into minutes and seconds. See [Fig. 54.18](#) for an example micrometer reading. The operator points the instrument and then uses the micrometer to align the degree index marks. The readings from the degree window and the micrometer window are added together to obtain the angle.

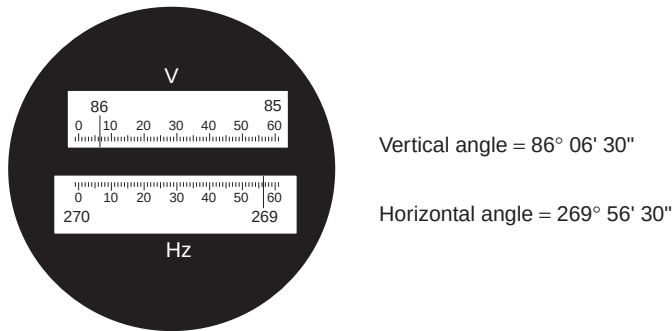


FIGURE 54.17 Scale angle reading display.

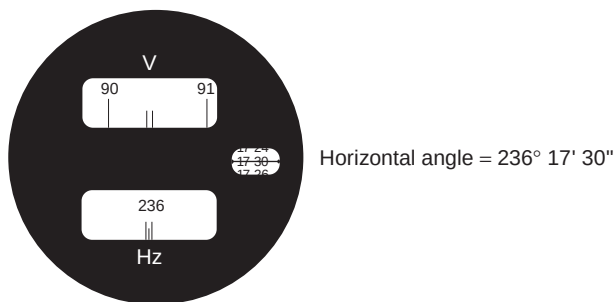


FIGURE 54.18 Micrometer angle reading display.

Directional Optical Theodolite

The directional theodolite is different from other instruments because it does not have a lower motion clamp and tangent screw. Directions (circle readings) are observed and recorded, and then the directions are subtracted to obtain the angle between the backsight and foresight lines of sight. This type of system has generally been used only on the most precise instruments. Zero is usually not set on the instrument. A micrometer is used to optically read the circle. Optical theodolites are excellent instruments, but, just like the transit, they are being surpassed by the technology of electronics.

Digital Theodolite

Digital electronics has recently entered the area of surveying instruments. Angles are no longer read optically, they are displayed on a screen in degree, minute, and second format. The micrometer has been replaced with electronic sensors that determine the angle quickly and precisely. The digital theodolite has the appearance of an optical theodolite in size and overall shape. The telescope, the clamping system, the tribrach, and the optical plummet are the same. Only the angle-measuring and -reading system is different. Digital theodolites are easy to read and fewer reading and recording blunders occur in the field. Because it is electronically based, the digital theodolite is like other electronic equipment — it either works or it doesn't work. If a circuit goes out or the battery is not charged, the instrument is unusable.

Most digital theodolites are designed to be interfaced with top-mounted EDMs. This essentially has the impact of turning them into what are commonly called *semitotal stations*, which measure distances and angles and can be connected to a data collector for recording measurements. Sighting by both the instrument telescope and the EDM scope are accomplished separately with this configuration.

Total Stations

The electronic total station is the ultimate in surveying measurement instruments. It is a combination digital theodolite and EDM that allows the user to measure distances and angles electronically, calculate

coordinates of points, and attach an electronic field book to collect and record the data. Since the total station is a combination digital theodolite and EDM, its cost is quite high compared to a single instrument. However, its capabilities are simply phenomenal in comparison to the way surveyors and engineers had to measure just a few years ago. Total stations with data collectors are especially effective when a large number of points are to be located in the field, as in topographic mapping. The data collector can be used to transfer the points to a computer for final map preparation. Conversely, complex projects can be calculated on a computer in the office and the data uploaded to the data collector. The data collector is then taken to the field and connected to the total station, where hundreds of points can be rapidly established by radial layout methods.

Instrument Components

Although there are differences between the transit, optical theodolite, and electronic theodolite, the construction and operation of all these instruments is basically the same. The three major components of any instrument are illustrated, using a transit, in Fig. 54.19. The upper plate assembly or the alidade, the lower plate assembly or the horizontal circle, and the leveling head are shown.

Alidade Assembly

The alidade assembly consists of the telescope, the vertical circle, the vertical clamp and vertical tangent, the standards or structure that holds everything together, the verniers, plate bubbles, telescope bubble, and the upper tangent screw. A spindle at the bottom of the assembly fits down into a hollow spindle on the horizontal circle assembly.

Horizontal Circle Assembly

The horizontal circle assembly comprises the horizontal circle, the upper clamp that clamps the alidade and horizontal circle together, and the hollow spindle that accepts the spindle from the alidade and fits into the leveling head.

Leveling Head

The leveling head is the foundation that attaches the instrument assemblies to the tripod. It consists of leveling screws, the lower clamp that clamps the horizontal circle and the leveling head together, and a threaded bracket for attaching to the tripod.

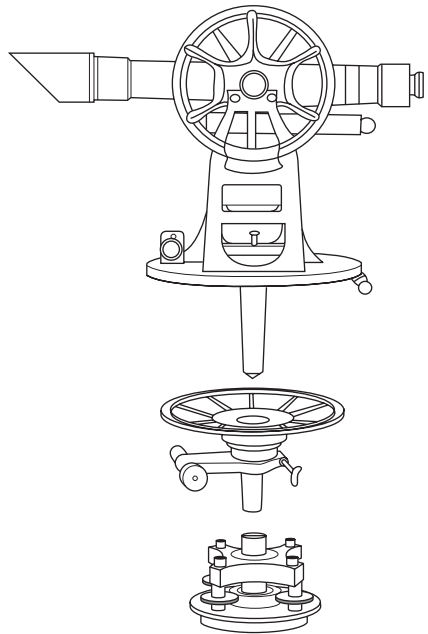


FIGURE 54.19 Transit or theodolite components.

Fundamental Relationships

All transit or theodolite instruments are designed around the same fundamental relationships and lines. These lines are shown in Fig. 54.20, again illustrated with a standard transit. The principal relationships between these lines are described as follows:

- The axis of the plate level(s) should be perpendicular to the vertical axis.
- The line of sight should be perpendicular to the horizontal axis.
- The horizontal axis should be perpendicular to the vertical axis.
- The vertical circle should read 0° when the instrument is leveled and the telescope is horizontal.

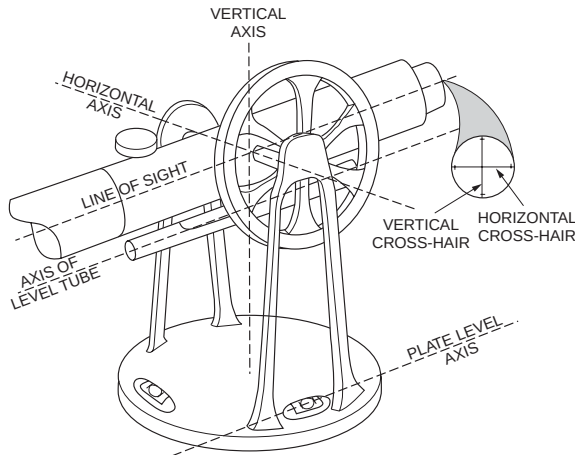


FIGURE 54.20 Principal adjustment relationships.

These four relationships are necessary to measure horizontal and vertical angles correctly. If the instrument will be used as a level, then the following must also be true: The axis of the telescope level should be parallel to the line of sight. Incorrect instrument adjustment in any of the relationships above will be readily apparent during field use of the instrument. Less apparent will be errors in the following relationships:

- The line of sight should be coincident with the telescope optical axis.
- The circles should be mounted concentrically on the axis of rotation.
- The circles should be accurately graduated throughout their circumference.
- The optical plummet should correctly position the instrument vertically over the field station.
- The vertical crosshair should lie in a vertical plane perpendicular to the horizontal axis.

Instrument Operation

The fundamental principle in using any transit or theodolite is the principle of reversion. That is, all operations should be performed in pairs, once with the telescope in the direct position and again with the telescope inverted on the horizontal axis or the reverse position. The correct value is the average of the two observations. Instrument operators should always use the double-centering ability of the instrument to eliminate the instrumental errors listed above. All of the principal instrument adjustment errors — except the vertical axis not being truly vertical — will be compensated for by averaging direct and reversed pairs of observations.

Prolonging a Straight Line

Prolonging a straight line from a backsight station through the instrument station to set a foresight station is a basic instrument operation that illustrates the principle of reversion. Often referred to as double centering, the steps to be performed are outlined as follows:

- Set up and level on the instrument station. With the telescope in the direct position, sight to the backsight station and clamp all horizontal motions.
- Rotate the telescope on the horizontal axis to the reverse position. Sight and set a point P_1 in the foresight direction as illustrated in [Fig. 54.21](#).
- Revolve the instrument on the vertical axis and sight the backsight station again. The telescope will be in the reverse position.

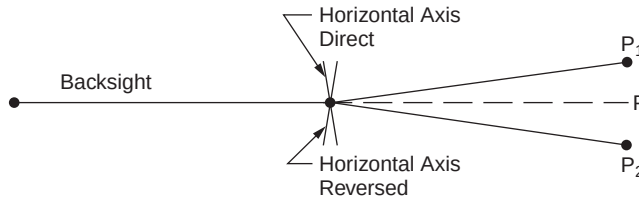


FIGURE 54.21 Double-centering to prolong a straight line.

- Rotate the telescope on the horizontal axis to the direct position. Sight and set a point P_1 in the foresight direction.
- Set the final point, P , on the true extension of the backsight line at the midpoint between P_1 and P_2 .

If the stations are on nearly level terrain, the instrument error apparent in the distance between P_1 and P_2 is the line of sight not being perpendicular to the horizontal axis. When an instrument is severely out of adjustment, it is inconvenient to use it in the field. The instrument should be cleaned and adjusted periodically by a qualified instrument service technician.

Horizontal Angles by Repetition

The clamping system is probably the most important feature on an instrument because it determines the angle measurement procedures that can be used with the instrument. Transits and some optical theodolites have two horizontal motion clamps. They are commonly called the *upper motion* and *lower motion*. These types of instruments are called *repeating instruments* since the angle can be repeated and accumulated on the instrument circle. A procedure for measuring angles by repetition is outlined below.

1. Set the horizontal angle to read zero with the upper clamp and upper tangent screws. This is for convenience and any initial angle can be used.
2. Point on the backsight station with the telescope in the direct position using the lower motion. Check the circle reading and record the value in the notes.
3. Loosen the upper clamp (the lower clamp remains fixed), turn the instrument to the foresight station, and point using the upper motion clamp and upper tangent screw.
4. Note the circle reading. An approximate value of the angle can be obtained from this first turning.
5. Loosen the lower clamp, keeping the angle reading on the circle, and repeat steps 2 and 3. The number of repetitions must be an even number with half of the turnings with the telescope in the direct position and half of the turnings with the telescope in the reversed position.
6. With each turning, the circle reading will be incremented by the value of the angle. Thus, the average angle can be determined quickly by dividing the total angle read from the circle by the number of times the angle was measured.

Horizontal Angles by Direction

Many theodolites have only the upper horizontal motion clamp and the fine adjustment screw for pointing. A backsight cannot be made while holding an angle on the circle because there is no lower-motion fine adjustment for pointing. These types of instruments are called *direction instruments* since the circle reading is a clockwise direction angle from an arbitrary orientation of the 0° mark on the circle. A procedure for measuring angles by reading directions is outlined below.

1. Set the horizontal circle to read approximately 0° when pointed toward the backsight point with the telescope in the direct position.
2. Point on the backsight station with the telescope in the direct position using the upper-motion clamp and tangent screw. Read the horizontal circle and record the value in the notes.
3. Loosen the upper clamp and turn the instrument to point on the foresight station using the upper-motion clamp and tangent screw. Read the horizontal circle and record the value in the notes.

4. The clockwise angle is the difference between the foresight reading and the backsight reading.
5. Invert the telescope to the reversed position and repeat steps 2, 3, and 4. The circle readings should be exactly 180° from the telescope direct readings if there are no instrument adjustment, pointing, and reading errors. Of course the direct and reversed angles should agree closely and be averaged.
6. The direct and reversed pointings and the average angle constitute one position. Observe several positions and average the results to improve the precision of the final average angle. Advance the horizontal circle approximately between positions in order to distribute the readings around the circle.

54.5 Plane Survey Computations

Plane surveys use a three-dimensional Cartesian coordinate system as shown in Fig. 54.22. The horizontal components of distance, angle, and direction are assumed to be in the plane defined by the X and Y axes. The vertical components are along the Z axis or “up axis” in a vertical plane perpendicular to the horizontal XY plane.

Plane surveys referenced to such an orthogonal Cartesian coordinate system ignore the effect of earth curvature, the fact that the actual level surface is perpendicular everywhere to the direction of gravity. Such a computation scheme is suitable only for local project surveys of limited extent. Plane survey computations can be used for horizontal control surveys over a large areal extent if all observations and positions are properly referenced to a grid system using an appropriate map projection. State plane coordinate systems, defined for each state, are an example of this technique.

Traverse

A traverse is an efficient and flexible method of field surveying used to connect points of interest and establish horizontal and/or vertical coordinate reference values for project control. A traverse consists of interconnected straight lines along a traverse route. The straight lines meet at angle points that must be permanently marked by a traverse station monument. The length of each line is measured by field survey and then reduced to the horizontal reference plane. At each traverse station a horizontal angle is measured that will relate the directions of each line to one another in the horizontal plane. The elevation of each traverse station may also be determined if required for the purpose of the survey. Elevation can be

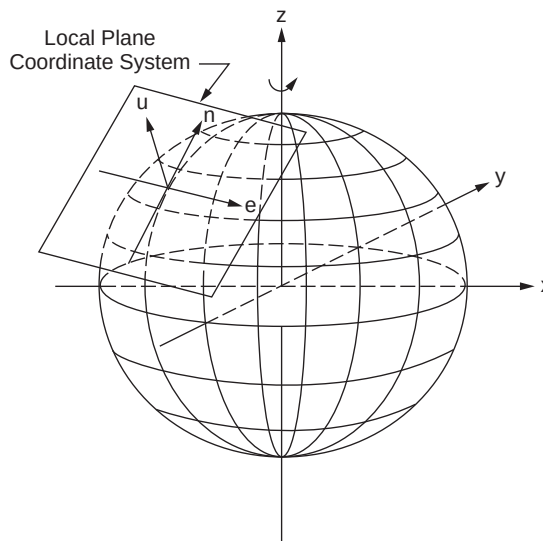


FIGURE 54.22 Local tangent plane coordinate system.

determined by trigonometric leveling as part of the traverse measurements or by differential leveling as a separate survey operation.

Traverses are characterized as open or closed. An open traverse has no check on computed direction or position at the end of the route. It is used only for preliminary and uncontrolled work. Whenever traverse stations are to be used to control subsequent engineering or surveying work, the route should be designed to close on the beginning traverse station or to close on another known point of equal or higher-order accuracy. The closed linear traverse between known control stations is preferred since it can best detect length measurement errors that affect the scale of a survey.

Traverse design is influenced by the purpose of the survey and the terrain through which the survey must progress. The overriding factor in determining where to locate traverse stations is the purpose of the station. In general, a traverse station becomes a reference system control point that is used for one or more of the following purposes:

1. Provide control for mapping by field survey or photogrammetric methods.
2. Provide control for construction layout.
3. Locate property boundaries.
4. Connect lines within the traverse and secondary traverses that may be added to the survey network.

The traverse station must be located so that it is accessible for its intended purpose. Secondary to the purpose of the survey, the traverse station should be located in an area where the monument will be stable and undisturbed for the intended useful life of the station.

The traverse survey route is flexible and generally follows the path of least resistance so that clear lines of sight from traverse station to traverse station are obtained. However, several guidelines should be kept in mind when planning a traverse route. First, avoid lines of sight that pass close to the intervening terrain between stations where atmospheric refraction will seriously degrade measurement accuracy. Second, avoid short lines of sight where setup errors and pointing errors can be the dominant measurement error source. Third, the traverse route should proceed along a generally straight path between terminal stations of a linear closed traverse. A useful rule of thumb is that the deviation of the route, as measured perpendicular to the straight-line path, should not exceed approximately one-third of the straight-line path length.

Accuracy standards for traversing are chosen to match the purpose of the survey. Specifications for the survey process are then developed to meet the required accuracy standards. For control surveys the accuracy standards and specifications developed by the Federal Geodetic Control Committee (FGCC) are often adopted. Surveys may be required to meet **first-**, **second-**, or **third-order accuracy**.

Surveying standards and specifications for surveys intended for other purposes may be available from a variety of sources, such as state statutes or licensing board rules regulating the type of survey, professional societies, or the agency contracting for the survey. For example, standards and specifications developed by ALTA/ACSM are often the requirements adopted for property surveys.

Traverse computations for local plane surveys are outlined in the following conventional procedure:

1. Draw a complete sketch of the traverse.
2. Compute the angular closure. If angle closure is equal to or less than allowable limit, adjust the angles; if angle closure is not acceptable, remeasure angles. The allowable closure is typically specified in the following form:

$$\text{Allowable closure} = c = k\sqrt{n}$$

where

c = allowable error in a series of measurements

k = a value specified for the accuracy order of the survey

n = number of angles measured

3. Compute the direction (azimuths or bearings) of all lines.

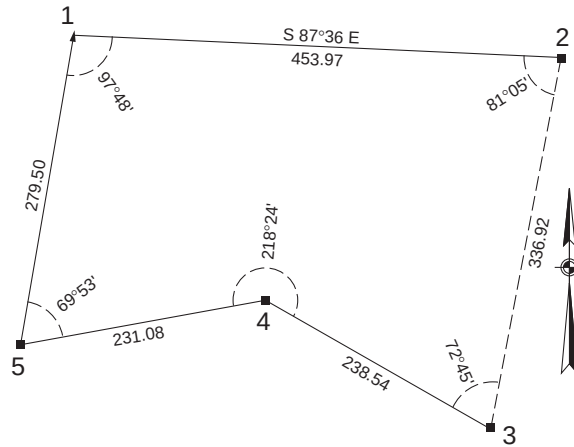


FIGURE 54.23 Sketch of sample traverse.

4. Compute the latitude and departure (relative northing and relative easting components) for each course. Set up the computations in tabular form.

$$\text{Latitude} = L \cos \alpha$$

$$\text{Departure} = L \sin \alpha$$

5. Compute the traverse misclosure.

$$c = \sqrt{(\text{Latitude error})^2 + (\text{Departure error})^2}$$

6. Compute the traverse precision ratio.

$$\text{Precision} = \frac{\text{Closure error}}{\text{Perimeter}} = \frac{1}{?}$$

7. If the precision ratio is equal to or better than the allowable precision specified for the accuracy order of the survey, then distribute the error of closure throughout the traverse by an appropriate rule or by a least-squares adjustment.
8. Compute the coordinates of the traverse stations using the balanced latitudes and departures.
9. Determine the adjusted traverse course lengths and directions using an inverse computation from the adjusted coordinates or latitudes and departures.
10. Compute the area of a closed loop traverse.

The traverse shown in Fig. 54.23 is used to illustrate conventional plane survey traverse computations. The measured horizontal distance of each line is given, and horizontal angles are measured as interior angles on the closed loop traverse. The orientation of the traverse is defined by the azimuth given for line 1-2. The coordinate system is defined by known coordinates at station of 1000.00N, 500.00E.

1. Draw a complete sketch.
2. Compute the angular closure.

$$\begin{aligned} \text{Field angle closure} &= (97^\circ 48' + 81^\circ 05' + 72^\circ 45' + 218^\circ 24' + 69^\circ 53') \\ &\quad - (5 - 2)180^\circ \\ &= 539^\circ 55' - 540^\circ 00' \\ &= .05' \end{aligned}$$

Assuming the angular error of closure is less than the maximum allowable, apply an equal correction to each angle.

$$\text{Correction per angle} = \frac{05'}{5 \text{ angles}} = 0^{\circ}01' \text{ to be added to each angle}$$

3. Compute bearings or azimuths. Use the adjusted angles. In this example the azimuth of each line is computed beginning from the given azimuth for line 1-2.

Azimuth 1 to 2 = given = 92°24'
 Azimuth 2 to 3 = 92°24' + 180° – 81°06' = 191°18'
 Azimuth 3 to 4 = 191°18' + 180° – 72°46' = 298°32'
 Azimuth 4 to 5 = 298°32' + 180° – 218°25' = 260°07'
 Azimuth 5 to 1 = 260°07' – 180° – 69°54' = 10°13'
 Azimuth 1 to 2 = 10°13' + 180° – 97°49' = 92°24'

Note that the azimuth of the first line is computed at the end to verify that the computations close on the given azimuth. The adjusted angles and azimuths are shown in Fig. 54.24.

4. Compute the latitude and departure of each course, as shown in Table 54.3.

TABLE 54.3 Computed Latitudes and Departures.

Course	Length (L)	Azimuth α	Latitude $= L \cos \alpha$	Departure $= L \sin \alpha$
1-2	453.97	92°24'	–19.010	453.572
2-3	336.92	191°18'	–330.389	–66.018
3-4	238.54	298°32'	113.943	–209.567
4-5	231.08	260°07'	–39.663	–227.651
5-1	279.50	10°13'	275.068	49.575
Sum =	1540.01	Error =	–0.051	–0.089
		Correction	+0.051	+0.089

5. Compute the traverse misclosure.

$$c = \sqrt{(-0.051)^2 + (-0.089)^2} = 0.102 \text{ ft}$$

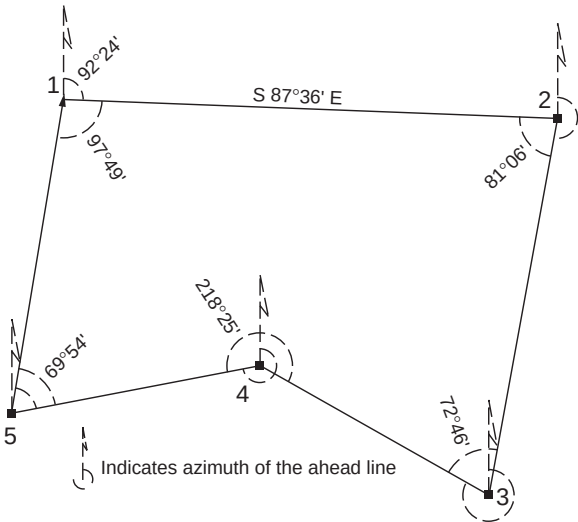


FIGURE 54.24 Adjusted angles and azimuths.

6. Compute the traverse precision.

$$\text{Precision} = \frac{0.102}{1540.01} = \frac{1}{15,100}$$

7. Distribute the error of closure. We will assume this precision is sufficient for the particular survey we are carrying out. Adjust the traverse by the compass rule:

$$C_{\text{lat}} = \frac{+0.051}{1540.01} (\text{Course length})$$

$$C_{\text{dep}} = \frac{+0.089}{1540.01} (\text{Course length})$$

8. Compute the balanced latitudes and departures by applying the corrections, and then calculate coordinates for all of the traverse stations, as shown in [Table 54.4](#).

TABLE 54.4 Compass Rule Adjustment

Station	Applying Corrections		Balanced		Coordinates	
	Latitude	Departure	Latitude	Departure	North (Y)	East (X)
1					1000.00	500.00
	− 19.010	453.572				
	+ .015	+ .026	− 18.995	+ 453.598	− 19.00	+ 453.60
2					981.00	953.60
	− 330.389	− 66.018				
	+ .011	+ .020	− 330.378	− 65.998	− 330.38	− 66.00
3					650.62	887.60
	113.943	− 209.567				
	+ .008	+ .014	+ 113.951	− 209.553	+ 113.95	− 209.55
4					764.57	678.05
	− 39.663	− 227.651				
	+ .008	+ .013	− 39.655	− 227.638	− 39.65	− 227.64
5					724.92	450.41
	275.068	49.575				
	+ .009	+ .016	+ 275.077	+ 49.591	+ 275.08	+ 49.59
1					1000.00	500.00
		Total =	0.000	0.000	Check	Check

9. Adjusted traverse line lengths and directions are computed by an inverse computation using the adjusted latitudes and departures or adjusted coordinates, as shown in [Table 54.5](#).

$$\text{Length} = \sqrt{Dep^2 + Lat^2}$$

$$= \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2}$$

$$\text{Azimuth}_{ij} = \tan^{-1} \left[\frac{Dep_{ij}}{Lat_{ij}} \right] = \tan^{-1} \left[\frac{X_j - X_i}{Y_j - Y_i} \right]$$

10. Compute the area. Traverse areas are typically computed using the coordinate method, as shown in [Table 54.6](#).

$$\text{Area} = \frac{\left| \sum_1^n X_i Y_{i-1} - \sum_1^n X_i Y_{i+1} \right|}{2}$$

TABLE 54.5 Adjusted Lengths and Azimuths

Station	Coordinates		Adjusted Length	Adjusted Azimuth
	North (Y)	East (X)		
1	1000.00	500.00	454.00	92°23'53"
2	981.00	953.60		
3	650.62	887.60	336.90	191°17'50"
4	764.57	678.05	238.53	298°32'12"
5	724.92	450.41	231.07	260°07'05"
			279.51	10°13'10"
1	1000.00	500.00		

TABLE 54.6 Coordinate Method for Area

Station	Coordinates		Double Area	
	North (Y)	East (X)	—	0
1	1000.00	500.00	490,500	
2	981.00	953.60	620,431	953,600
3	650.62	887.60	678,632	870,736
4	764.57	678.05	491,532	441,153
5	724.92	450.41	450,410	344,370
1	1000.00	500.00		362,460
Total =			2,731,505	2,972,319

The traverse area equals one-half of the difference between the totals of the double-area columns.

$$\text{Traverse area} = \left| \frac{2,731,505 - 2,972,319}{2} \right| = 120,407 \text{ ft}^2 = 2.764 \text{ acres}$$

Partitioning Land

Partitioning land is a problem that can usually be classified according to one of two types of dividing line — a line of known direction or a line through a known point. A preliminary line is often required that satisfies the given condition. Then the line is translated parallel to itself in the first condition or pivoted about the known point in the second condition to obtain the required area.

As an illustration of these methods, partition the adjusted traverse of the previous section so that 65,000 square feet lie west of a true north line. A preliminary cutoff line bearing true north can be constructed through station 4, as shown in Fig. 54.25. The area west of the preliminary line is 50,620 square feet. The line is translated true east a distance X so that 14,380 square feet will be added to the west parcel. The parcel added is a trapezoid as shown in Fig. 54.25, and the area can be expressed by

$$14,380 = 227.97X - \frac{1}{2}X(X \tan \theta_1) + \frac{1}{2}X(X \tan \theta_2)$$

where $\theta_1 = 2^\circ 23' 53''$ and $\theta_2 = 28^\circ 32' 12''$. Rearranging this expression results in a quadratic equation, and the solution for X is found to be 59.22 feet. Then the following distances can be determined for the final cutoff line:

$$6-7 = 59.27 \text{ feet}$$

$$4-8 = 67.41 \text{ feet}$$

$$7-8 = 257.69 \text{ feet}$$

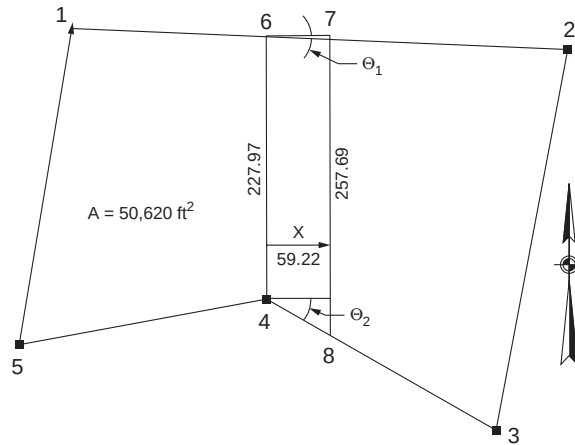


FIGURE 54.25 Partition by sliding a line.

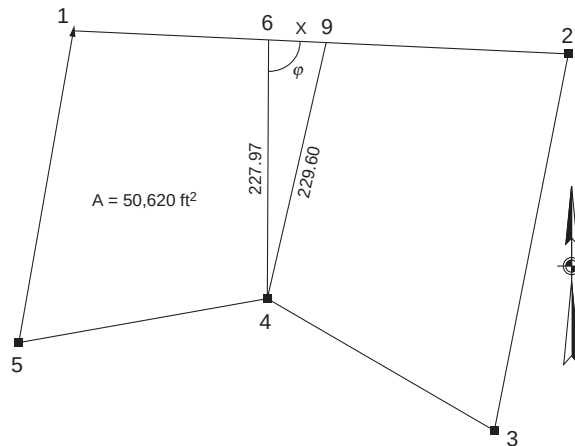


FIGURE 54.26 Partition by pivoting a line.

Next, partition the adjusted traverse so that 55,000 square feet lie west of a line through station 4. The same preliminary cutoff line can be used. The line is pivoted easterly about station 4 so that 4380 square feet is added to the west parcel. The parcel added is a triangle as shown in Fig. 54.26, and the area can be expressed by

$$4380 = \frac{1}{2} 227.97 X \sin \phi$$

where $\phi = 87^{\circ}36'07''$. The solution for X is found to be 38.46 feet, and the final cutoff line is

$$4-9 = 229.60 \text{ feet, } 9^{\circ}38'05'' \text{ azimuth}$$

54.6 Horizontal Curves

Horizontal curves are used in route projects to provide a smooth transition between straight-line tangent sections. These curves are simple circular curves. The components of a circular curve are illustrated in Fig. 54.27. The design of a circular curve requires that two curve elements be defined and then the

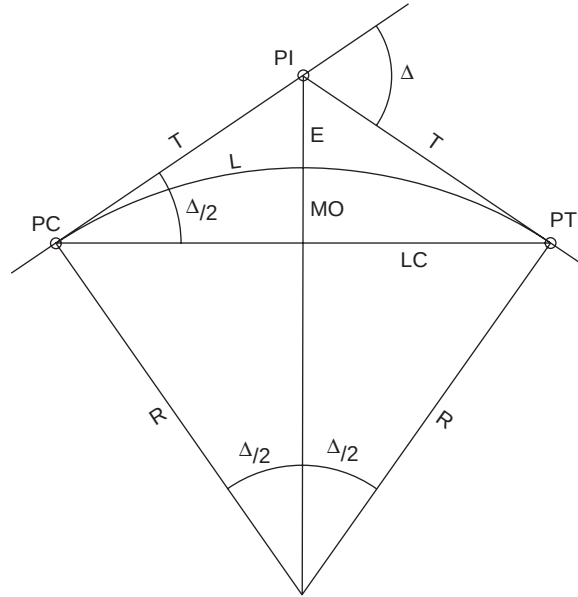


FIGURE 54.27 Circular horizontal curve components.

remaining elements are determined by computation. In the typical case the intersection angle of the tangents is determined by field survey and the curve radius is chosen to meet design specifications such as vehicle speed and minimum sight lengths. The principal relationships between curve elements are

$$T = R \tan \frac{\Delta}{2}$$

$$LC = \frac{1}{2}R \sin \frac{\Delta}{2}$$

$$E = R \left(\sec \frac{\Delta}{2} - 1 \right)$$

$$MO = R \left(1 - \cos \frac{\Delta}{2} \right)$$

$$L = R \frac{\pi}{180^\circ} \Delta$$

Important angle relationships used when solving circular curve problems are illustrated in [Fig. 54.28](#).

When a circular curve is staked out in the field, the most common method is to lay off deflection angles at the PC station. Field layout notes are prepared for a theodolite set up at the PC. A backsight is taken along the tangent line to the PI, and foresights are made to specific stations on the curve using the computed deflection angles. If a total station instrument is used the chord distance from the PC to the curve station can be used to set the station. If a tape is used the chord distance measured from the previous station on the curve is intersected with the line of sight to locate the curve station.

The following example illustrates both types of distances used to lay out a horizontal curve. Determine the field information necessary to stake out the horizontal curve shown in [Fig. 54.29](#). First determine the stationing of the PC and PT. The tangent distance is

$$T = R \tan \left(\frac{\Delta}{2} \right) = 572.96 \tan 40^\circ 00' = 503.46 \text{ ft}$$

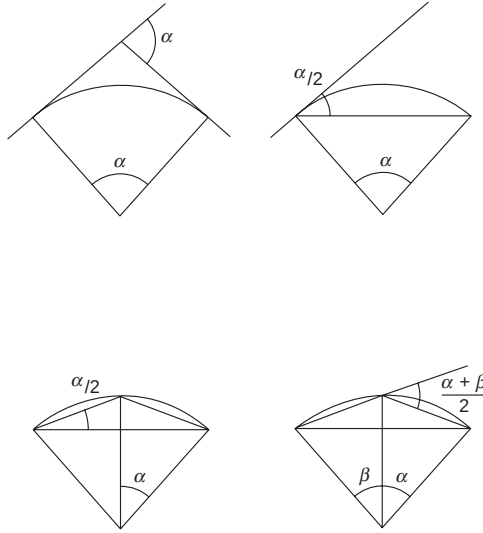


FIGURE 54.28 Angle relationships for circular curves.

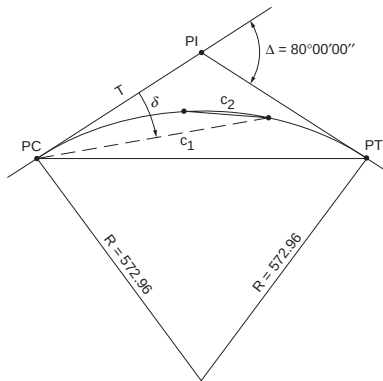


FIGURE 54.29 Horizontal curve layout.

so the PC station is

$$\text{PC sta.} = \text{PI sta.} - T = 100 + 50.00 - 5 + 03.46 = 95 + 46.54 \text{ ft}$$

The arc length of the curve is

$$L = R \left(\frac{\pi}{180^\circ} \right) \delta = 800.00 \text{ ft}$$

so the PT station is

$$\text{PT sta.} = \text{PC sta.} + L = 95 + 46.54 + 8 + 00.00 = 103 + 46.54 \text{ ft}$$

The central angle subtended per station of arc length is

$$D^\circ \text{ per sta} = \frac{\Delta}{L} = \frac{80^\circ 00'}{800.00} = 10^\circ 00'$$

TABLE 54.7 Deflection Angles and Chords for Layout of a Horizontal Curve

Station	Deflection Angle δ	Arc Length from PC	Chord Length C_1	Chord Length C_2
PC 95 + 46.54				
96 + 00.00	2°40'23"	53.46	53.44	53.44
97 + 00.00	7°40'23"	153.46	153.00	99.87
98 + 00.00	12°40'23"	253.46	251.40	99.87
99 + 00.00	17°40'23"	353.46	347.88	99.87
100 + 00.00	22°40'23"	453.46	441.72	99.87
101 + 00.00	27°40'23"	553.46	532.19	99.87
102 + 00.00	32°40'23"	653.46	618.62	99.87
103 + 00.00	37°40'23"	753.46	700.33	99.87
PT 103 + 46.54	$\Delta/240^\circ00'00''$	L 800.00	LC 736.58	46.53

The deflection angle is one-half of the central angle or

$$\delta \text{ per sta} = \frac{D}{2} = 5^\circ00'$$
$$\delta \text{ per ft} = \frac{D}{200} = 0^\circ03'$$

Recall the equation for a chord

$$C = 2R \sin\left(\frac{d}{2}\right)$$

where d is the central angle of the chord. The full stations falling on the curve are determined, and the values in Table 54.7 are computed for an instrument set up on the PC and oriented by a backsight on the PI. Note that C_1 denotes the chord measured from the PC and that C_2 denotes the chord measured from the previous station on the curve.

Alternative methods for staking out a curve include tangent offsets, chord offsets, middle ordinates, and radial staking out from the radius point of the curve or from the PI station.

54.7 Vertical Curves

Vertical curves are used in vertical alignments of route projects to provide a smooth transition between grade lines. These curves are usually equal-tangent parabolic curves. The point of vertical intersection, PVI, of the entrance and exit grade lines always occurs at the midpoint of the length of curve. The length of curve and all station distances are measured in the horizontal plane. Figures 54.30 and 54.31 illustrate the geometry of a sag and crest vertical curve, respectively. Note the following relationships between the curves in Figs. 54.30 and 54.31. The top curve is the elevation curve. The middle curve is the grade curve; it is the derivative of the elevation curve. The bottom curve is the rate of change of grade curve; it is the derivative of the grade curve. The rate of change of grade is always a constant, r , for a parabolic vertical curve.



Since the curves are related by the derivative/integration operation, the change in an ordinate value on one curve is equal to the area under the next lower curve. Therefore, in Fig. 54.30, the change in elevation on the curve from the PVC to a point at a distance X on the curve is equal to the trapezoidal area under the grade curve

$$Y_x - Y_0 = X \left(\frac{g_x + g_1}{2} \right)$$

The change in grade over the same interval X is the rectangular area under the rate of change of grade curve.

$$g_x - g_1 = Xr$$

These relationships can be used to calculate design values for vertical curves.

For example, suppose it is necessary to find the elevation and station of the high point for the crest curve in Fig. 54.31. Let X equal the distance to the high point that occurs at the point where the grade curve crosses the zero line. The change in grade must be equal to the area under the rate of change of grade curve.

$$\begin{aligned} g_{\text{high}} &= 0 \\ 0 - g_1 &= X(-r) \\ X &= \frac{-g_1}{-r} \end{aligned}$$

Note that r is negative for a crest value. Then the station of the high point is

$$P_{\text{high}} = \text{PC} + X$$

The elevation of the high point can be found by evaluating the elevation equation at the known value of X or by calculating the area under the grade curve:

$$\begin{aligned} Y_x - Y_0 &= X \left(\frac{g_1 + 0}{2} \right) \\ Y_x &= Y_0 + X \left(\frac{g_1}{2} \right) \end{aligned}$$

The design of a vertical curve involves choosing the length of the curve that will satisfy design speed considerations, earthwork considerations, and sometimes geometric constraints. As an example of a constrained design, suppose a PVC is located at station 150 + 40.00 and elevation 622.45 feet. The grade of the back tangent is -3.00% and the grade of the forward tangent is -7.00% . It is required that a vertical curve between these tangents must pass through station 152 + 10.00 at elevation 619.05 feet. Since this is a sag vertical curve, refer to Fig. 54.30. The value of r can be expressed as

$$r = \frac{(+7) - (-3)}{L} = + \frac{10}{L}$$

The distance X from the PVC to the known station on the curve is

$$X = (152 + 10.00) - (150 + 40.00) = 1.70000 \text{ stations}$$

The grade at the known station is then found from

$$g_x = (-3) + \left(\frac{10}{L} \right) 1.7 = \frac{17}{L} - 3$$

The change in elevation from the PVC to the known station can be set equal to the trapezoidal area under the grade curve.

$$619.05 - 622.45 = 1.70000 \left[\frac{\left(\frac{17}{L} - 3 \right) + (-3)}{2} \right]$$

Solving this expression for L , we obtain

$$L = 8.5000 \text{ stations} = 850.00 \text{ ft}$$

The design elevation at each full station along this curve can be evaluated from the parabolic equation. First the station and elevation of the PVI is

$$\text{PVI sta} = \text{PVC sta} + \frac{L}{2} = 150 + 40.00 + 4 + 25.00 = 154 + 65.00 \text{ ft}$$

$$\text{PVI elev} = \text{PVC elev} + g_1 \frac{L}{2} = 622.45 + (-3)(4.25) = 609.70 \text{ ft}$$

and the station and elevation of the PVT is

$$\text{PVT sta} = \text{PVI sta} + \frac{L}{2} = 154 + 65.00 + 4 + 25.00 = 158 + 90.00 \text{ ft}$$

$$\text{PVT elev} = \text{PVI elev} + g_2 \frac{L}{2} = 609.70 + (+7)(4.25) = 639.45 \text{ ft}$$

Then the elevation of any point on the curve is found from

$$Y = \left(\frac{r}{2} \right) X^2 + g_1 X + Y_0 = \left(\frac{10}{17} \right) X^2 - 3X + 622.45$$

A tabular solution for each full station along the curve is given in [Table 54.8](#).

54.8 Volume

The determination of volume is necessary before a project begins, throughout the project, and at the end of the project. In the planning stages, volumes are used to estimate project costs. After the project is started, volumes are determined so the contractor can receive partial payment for work completed. At the end, volumes are calculated to determine final quantities that have been removed or put in place to make final payment. The field engineer is often the person who performs the field measurements and calculations to determine these volumes. Discussed here are the fundamental methods used by field engineers.

General

To compute volumes, field measurements must be made. This typically involves determining the elevations of points in the field by using a systematic approach to collect the needed data. If the project is a roadway, cross-sectioning is used to collect the data that are needed to calculate volume.

If the project is an excavation for a building, borrow-pit leveling will be used to determine elevations of grid points to calculate the volume. Whatever the type of project, the elevation and the location of

TABLE 54.8 Vertical Curve Elevations

Station	X sta	$(r/2)X^2$	g_1X	Elevation Y ft
PVC 150 + 40.00				Y_0 622.45
	0.6000	0.21	− 1.80	
151 + 00.00	1.6000	1.50	− 4.80	620.86
152 + 00.00	2.6000	3.98	− 7.80	619.15
153 + 00.00	3.6000	7.62	− 10.80	618.83
154 + 00.00	4.6000	12.45	− 13.80	619.27
155 + 00.00	5.6000	18.45	− 16.80	621.10
156 + 00.00	6.6000	25.62	− 19.80	624.10
157 + 00.00	7.6000	33.98	− 22.80	628.27
158 + 00.00	8.5000	42.50	− 25.50	633.63
PVT 158 + 90.00				639.45

points will need to be determined. It is the responsibility of the field engineer to determine the most efficient method of field measurement to collect the data.

However, it should be mentioned that sometimes volumes can be determined using no field measurements at all. In some situations the contractor may be paid for the number of truckloads removed. Keeping track of the number of trucks leaving the site is all that may be necessary. However, this isn't a particularly accurate method since the soil that is removed expands or swells and takes up a larger amount of space than the undisturbed soil. Depending on how the project is bid, it is sometimes accurate enough.

Field Measurements for Volume Computations

Measurements for volume are nothing more than applying basic distance and elevation measurements to determine the locations and elevations of points where the volume is to be determined. It usually is not practical to take the time to collect data everywhere there is a slight change in elevation. Therefore, it must be understood that volume calculations do not give exact answers. Typically, approximations must be made and averages determined. The field engineer will analyze the data and make decisions that result in the best estimate of the volume.

Area

The key to volume calculation is the determination of area. Most volume calculation formulas contain within them the formula for an area, which is simply multiplied by the height to determine the volume. For instance, the area of a circle is pi times the radius squared. The volume of a cylinder is the area of the circle times the height of the cylinder. If an area can be determined, it is generally easy to determine the volume.

Counting Squares

Approximation is possible by plotting the figure to scale on cross-sectional paper and counting the squares. Each square represents x number of square feet. Incomplete squares along the edges of the cross section are visually combined and averaged.

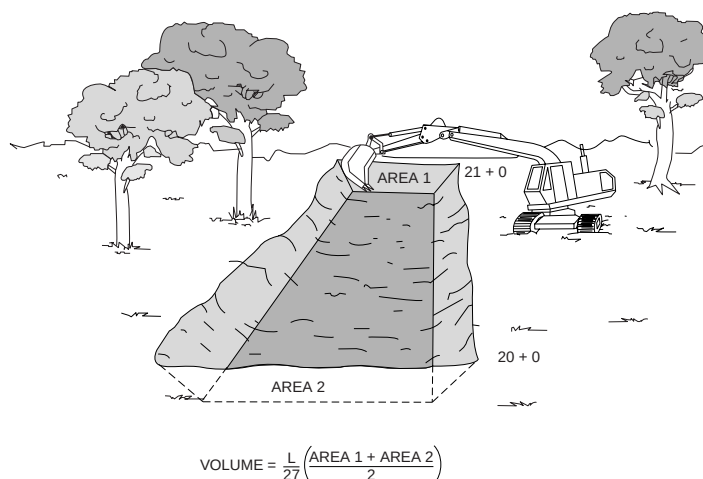


FIGURE 54.32 Average end area method.

Planimeter

The electromechanical digital planimeter is a quick method of determining the area of irregularly shaped figures. The irregular shape is drawn to scale and the planimeter is used to trace the outline of the shape. Inputting a scale factor into the planimeter results in a digital readout of the area.

Geometric Formula

Although a shape at first may seem irregular, it is often possible to break it into smaller regular shapes such as squares, rectangles, triangles, trapezoids, etc., that will allow the use of standard geometric formulas to determine the area. This method may be cumbersome because of all the shapes that may need to be calculated.

Cross-Section Coordinates

If cross-sectional field data are available, use of this data is the recommended method of calculating volume. Once understood, this process is fast and the most accurate way of determining area. Cross section data collected on a project represent elevation and location information for points on the ground. These points can be used as coordinates to determine area.

Volume Computations — Road Construction

In road construction the shape of the ground must be changed to remove the ups and downs of the hills and valleys for the planned roadway. Often mountains of dirt must be moved to create a gentle grade for the roadway. Payment for the removal and placement of dirt is typically on a unit cost basis. That is, the contractor will be paid per cubic yard of soil and will receive a separate price per cubic yard of rock. It can be seen that accurate determination of the volume moved is critical to the owner and to the contractor. Each wants an accurate volume so payment for the work is correct.

For road projects, cross sections of the ground elevations are measured at the beginning of the project, during the project, and at the end of the project. Comparisons between final cross sections and original cross sections are used to determine the volume moved. Areas of the cross sections are most easily determined by using the elevations of the points and their locations from the centerline (coordinates).

The average end area method uses the end areas of adjacent stations along a route and averages them. Refer to [Fig. 54.32](#). This average is then multiplied by the distance between the two end areas to obtain the volume between them. In formula form the process is as follows:

$$\text{Volume} = \frac{L}{27} \left(\frac{\text{Area 1} + \text{Area 2}}{2} \right)$$

where L represents the distance between the cross-sectional end areas being used in the formula, and 27 represents the number of cubic feet in 1 cubic yard. Dividing cubic feet by 27 converts to cubic yards.

Volume Computations — Building Excavation

A method known as *borrow-pit leveling* can be used effectively to determine volume on building projects. A grid is established by the field engineer and elevations on the grid points are determined both before the excavation begins and when the work is complete.

The borrow-pit method uses a grid and the average depth of the excavation to determine the volume. Before the excavating begins the field engineer creates a grid over the entire area where the excavation is to occur. Elevation data are collected at each of the grid points and recorded for future reference. At any time during the excavating, the field engineer can reestablish the grid and determine new elevations for each of the field points. Using the average height formula shown below, the volume of soil removed from each grid area can be determined. Refer to Fig. 54.33. The smaller the grid interval, the more accurate the volume.

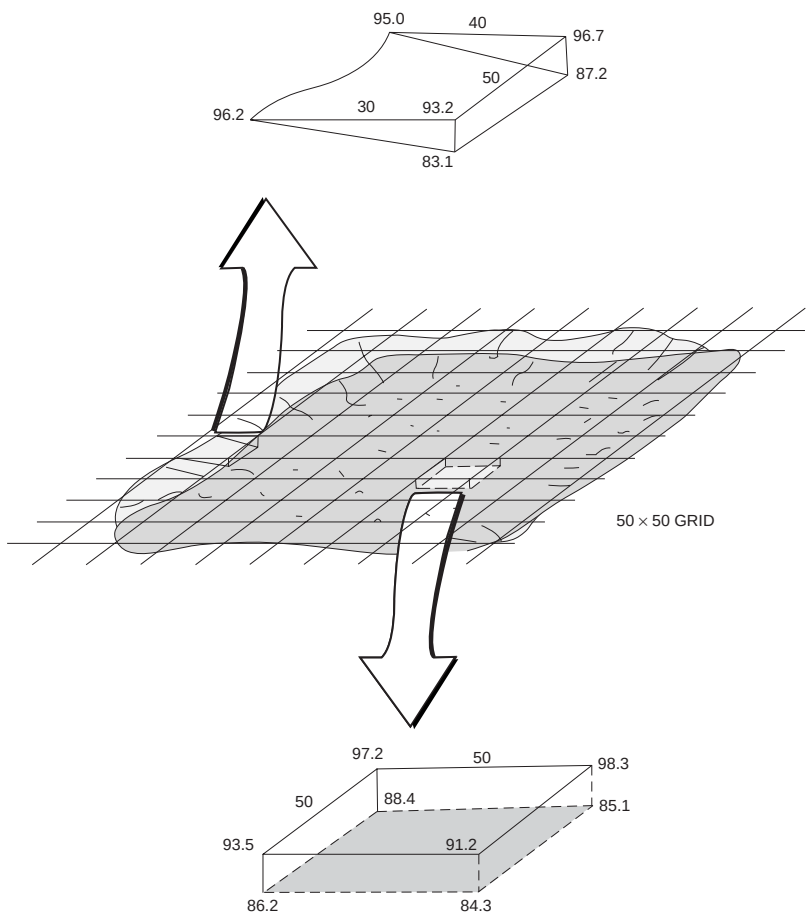


FIGURE 54.33 Borrow-pit method.

Summary

Only two general methods of calculating volumes have been presented here. There are many others that are very specific for the particular situation. For example, when determining volumes along a roadway, there is a constant transition from cut to fill to cut to fill, etc. To more accurately compute the volume, a prismoidal formula is used. The field engineer should check with textbooks that discuss in detail route surveying and earthwork for additional information.

Defining Terms

Accuracy — Refers to the degree of perfection obtained in measurements. It is a measure of the closeness to the true value.

Accuracy (first-order) — The highest accuracy required for engineering projects such as dams, tunnels, and high-speed rail system.

Accuracy (second-order) — The accuracy required for large engineering projects such as highways, interchanges, and short tunnels.

Accuracy (third-order) — The accuracy required for small engineering projects and topographic mapping control.

Accuracy ratio — The ratio of error of closure to the distance measured for one or a series of measurements.

Axis of level bubble — The line tangent to the top inner surface of a spirit level at the center of its graduated scale, and in the plane of the tube and its center of curvature.

Calibration — The process of comparing an instrument or chain with a standard.

Closure — The amount by which a value of a quantity obtained by surveying fails to agree with a value (of the same quantity) determined. It is also called *misclosure* or *error of closure*.

Datum — A reference elevation such as mean sea level or, in the case of some construction projects, a benchmark with elevation 100.00.

Horizontal axis — The axis about which the telescope rotates vertically.

Least squares — A mathematical method for the adjustment of observations based on the theory of probability.

Line of sight — The line extending from an instrument along which distinct objects can be seen. The straight line between two points.

Mean sea level — The average height of the surface of the sea measured over the complete cycle of high and low tides (a period of 18.6 years).

Monument — A physical structure that marks the location of a survey point.

Nadir — The point directly under the observer. The direction that a plumb bob points.

Precision — The closeness of one measurement to another. The degree of refinement in the measuring process. The repeatability of the measuring operation.

Random errors — Errors that are accidental in nature and always exist in all measurements. They follow the laws of probability and are equally high or low.

Refraction — The bending of light rays as they pass through the atmosphere.

Systematic error — Those errors that occur in the same magnitude and the same sign for each measurement of a distance, angle, or elevation. Can be eliminated by mechanical operation of the instrument or by mathematical formula.

Vertical — The direction in which gravity acts.

Zenith — The point directly above a given point on earth.

References

- Anderson, J.M., and Mikhail, E.M. 1985. *Introduction of Surveying*. McGraw-Hill, New York.
- Crawford, W.G. 1994. *Construction Surveying and Layout*, P.O.B. Publishing, Canton, MI.
- Davis, R.E. et al. 1981. *Surveying: Theory and Practice*, 6th ed. McGraw-Hill, New York.

Federal Geodetic Control Committee. 1984. *Standards and Specifications for Geodetic Control Networks*. National Geodetic Information Branch, NOAA, Silver Springs, MD.

Fronczek, C.J. 1980. *NOAA Technical Memorandum NOS NGS-10*. National Geodetic Information Branch, NOAA, Silver Springs, MD.

National Geodetic Survey. 1986. *Geodetic Glossary*. National Geodetic Information Branch, NOAA, Silver Springs, MD.

Professional Surveyor. American Surveyors Publishing Company, Suite 501, 2300 Ninth Street South, Arlington, VA.

Point of Beginning. Business News Publishing Company, Troy, MI.

Wolf, P.R., and Brinker, R.C. 1994. *Elementary Surveying*, 9th ed. HarperCollins, New York.

Further Information

The material here is intended only as an overview of plane surveying. There are many textbooks dedicated completely to the various aspects of surveying. The authors recommend the following books. For a more complete presentation of surveying theory, consult *Elementary Surveying*, 9th ed., by Wolf and Brinker, HarperCollins, 1992; or *Surveying: Theory and Practice*, 6th ed., by Davis et al., McGraw-Hill, 1981. For illustrated step-by-step descriptions of performing field work, consult *Construction Surveying and Layout*, by Crawford, P.O.B. Publishing, 1994.

To obtain detailed information on the capabilities of various instruments and software, P.O.B. Publishing prepares the trade magazine *P.O.B.*, and American Surveyors Publishing Company prepares the trade magazine *Professional Surveyor*. Each of these publications conducts annual reviews of theodolites, total stations, EDMs, data collectors, GPSs, and software. These listings allow the reader to keep up-to-date and compare “apples to apples” when analyzing equipment.

Survey control information, software, and many useful technical publications are available from the National Geodetic Survey (NGS). The address is

National Geodetic Survey Division
National Geodetic Information Branch, N/CG17
1315 East-West Highway, Room 9218
Silver Spring, MD 20910-3282

55

Geodesy

55.1 Introduction

55.2 Coordinate Representations

Two-Dimensional • Three-Dimensional • Coordinate Transformations • Curvilinear Coordinates and Transformations

55.3 Coordinate Frames Used in Geodesy and Some Additional Relationship

Earth-Fixed • Inertial and Quasi-Inertial • Relation between Earth-Fixed and Inertial

55.4 Mapping

Two Worlds • Conformal Mapping Using Cartesian Differential Coordinates • Conformal Mapping Using Polar Differential Coordinates • Coordinate Transformations and Conformal Mapping

55.5 Basic Concepts in Mechanics

Equations of Motion of a Point Mass in an Inertial Frame • Potential

55.6 Satellite Surveying

Numerical Solution of Three Second-Order Differential Equations • Analytical Solution of Three Second-Order Differential Equations • Orbit of a Satellite in a Noncentral Force Field • The Global Positioning System

55.7 Gravity Field and Related Issues

One-Dimensional Positioning: Heights and Vertical Control • Two-Dimensional Positioning: East–North and Horizontal Control • Three-Dimensional Positioning: Geocentric Positions and Full Three-dimensional Control

55.8 Reference Systems and Datum Transformations

Geodetic Reference Frames • Geodetic Reference System 1967 • Geodetic Reference System 1980 • 1983 Best Values • 1987 Best Values and Secular Changes • World Geodetic System 1984 • IERS Standards 1992 • Datum and Reference Frame Transformations • Textbooks and Reference Books • Journals and Organizations

B.H.W. van Gelder

Purdue University

55.1 Introduction

This chapter covers the basic mathematical and physical aspects of modeling the size and shape of the earth and its gravity field. Terrestrial and space geodetic measurement techniques are reviewed. Extra attention is paid to the relatively new technique of satellite surveying using the Global Positioning System (GPS). GPS surveying has not only revolutionized the art of navigation, but also brought about an efficient positioning technique for a variety of users, engineers not the least. It is safe to say that any

geometry-based data collecting scheme profits in some sense from the full constellation of 24 GPS satellites. Except for the obvious applications in geodesy, surveying, and photogrammetry, the use of GPS is applied in civil engineering areas such as transportation (truck and emergency vehicle monitoring, intelligent vehicle and highway systems, etc.) and structures (monitoring of deformation of such structures as water dams). Even in areas such as forestry and agriculture (crop yield management), GPS provides the geometric backbone to the (geographic) information systems.

Modern geodetic measurement techniques, using signals from satellites orbiting the earth, necessitate a new look at the science of geodesy. Classical measurement techniques divided the theoretical problem of mapping small or large parts of the earth into a horizontal issue and a vertical issue. Three-dimensional measurement techniques “solve” the geodetic problem at once. However, careful interpretation of these three-dimensional results is still warranted, probably even more so than before. This chapter will center around this three-dimensional approach. Less attention has been devoted to classical issues such as the computation of a geodesic on an ellipsoid of revolution. Although this issue still has some importance, the reader is referred to the textbooks listed at the end of this chapter.

More than in classical texts, three-dimensional polar (spherical) coordinate representations are used, because the fundamental issues pertaining to various geodetic models are easier to illustrate by spherical coordinates than by ellipsoidal coordinates. Moreover, the increased influence of the satellite techniques in everyday surveying revives the use of three-dimensional polar coordinate representations, because the three-dimensional location of a point is equally accurately represented by Cartesian, spherical, or ellipsoidal coordinates.

Throughout this chapter all coordinate frames are treated as right-handed orthogonal trihedrals. Because this also applies to curvilinear coordinates, the well-known geographic coordinates of latitude and longitude are presented in the following order:

1. Longitude (positive in east direction), λ
2. Latitude, ψ or ϕ
3. Height, h

In short, $\{\lambda, \psi, h\}$ or $\{\lambda, \phi, h\}$. Local Cartesian and curvilinear coordinates are treated in a similar fashion.

55.2 Coordinate Representations

For a detailed discussion on coordinate frames and transformations, the reader is referred to Chapter 53.

Two-Dimensional

In surveying and mapping, two-dimensional frames are widely used. The different representations are all dependent, because only two numbers suffice to define the location of a point in 2-space. Cartesian frames consist of two often perpendicular reference axes, denoted as x and y , or e (easting) and n (northing). Points in two-dimensional frames are equally well represented by polar coordinates r (distance from an origin) and α (polar angle, counted positive counterclockwise from a reference axis).

We have

$$\begin{aligned} x &= r \cos \alpha \\ y &= r \sin \alpha \end{aligned} \tag{55.1}$$

The polar coordinates $\{r, \alpha\}$ are expressed in terms of the Cartesian counterparts by

$$\begin{aligned} r &= \sqrt{x^2 + y^2} \\ \alpha &= \arctan(y/x) \end{aligned} \tag{55.2}$$

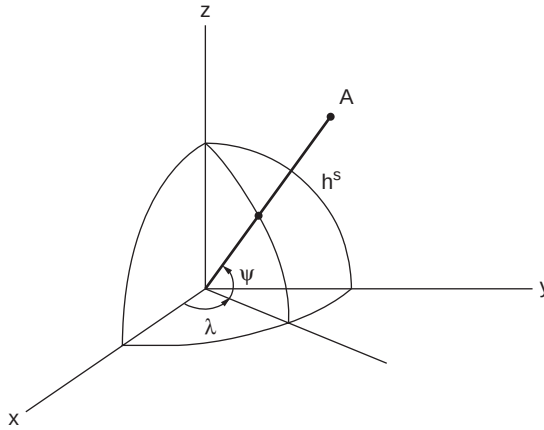


FIGURE 55.1 (Geographic) spherical coordinates: longitude λ , latitude ψ , and height h .

Three-Dimensional

Three-Dimensional Cartesian Coordinates

There are various ways to represent points in a three-dimensional space. One of the most well known is the representation by the so-called Cartesian coordinates x, y, z ; we represent the position of a point A through three distances x, y, z (coordinates) to three perpendicular planes, the yz -, xz -, xy -planes, respectively. The intersecting lines between the three planes are the perpendicular coordinate axes. The position of point A is thought to be represented by the vector $\mathbf{x}$ with elements $\{x, y, z\}$:

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad (55.3)$$

Three-Dimensional Polar Coordinates: Spherical

We may want to represent the position of these points with respect to a sphere with radius R . We make use of so-called spherical coordinates. The earth's radius is about $R = 6371.0$ km.

The sphere is intersected by two perpendicular planes, both of which pass through the center O of the sphere: a reference *equatorial plane* (perpendicular to the rotation axis) and a reference *meridian plane* (through the rotation axis). The angle between the vector and the reference equatorial plane is called *latitude*, ψ . The angle between the reference meridian plane (through Greenwich) and the local meridian plane (through A) is called *longitude*, λ (positive east). The distance to the surface of the sphere we call *height*, h . Consequently, the position of a point A is represented by $\{\lambda, \psi, h\}$ or $\{\lambda, \psi, r\}$ or $\{\lambda, \psi, R + h\}$; see Fig. 55.1.

Three-Dimensional Polar Coordinates: Ellipsoidal

The earth is flattened at the poles, and the average ocean surface has about the shape of an ellipsoid. For this reason, ellipsoidal coordinates are more often used in geodesy than spherical coordinates. We express the coordinates with respect to an ellipsoid of revolution with an equatorial semimajor axis a and a polar semiminor axis b . The semimajor axis thus represents the equatorial radius, and b is the distance between the ellipsoidal origin and the poles. The equation of such an ellipsoid of revolution is

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{b^2} = 1 \quad (55.4)$$

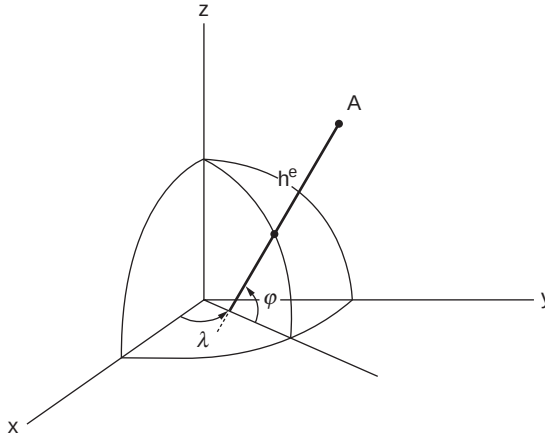


FIGURE 55.2 (Geodetic) ellipsoidal coordinates: longitude λ , latitude ϕ , and height h .

For the earth we have a semimajor axis $a = 6378.137$ km and a semiminor axis $b = 6356.752$ km. This means that the poles are about 21.4 km closer to the center of the earth than the equator.

The flattening of the earth is expressed by f and the (first) eccentricity by e :

$$f = \frac{a-b}{a} (\approx 1/298.257) \quad (55.5)$$

$$e^2 = \frac{a^2 - b^2}{a^2} (\approx 0.00669438) \quad (55.6)$$

See also [Fig. 55.2](#).

Coordinate Transformations

We have to distinguish between two classes of transformations:

- Transformations between dissimilar coordinate representations. An example would be the transformation between Cartesian coordinates and curvilinear coordinates, such as the ellipsoidal (geodetic) coordinates.
- Transformations between similar coordinate frames. An example is the relation between geocentric Cartesian coordinates and topocentric Cartesian coordinates.

The latter group is to be discussed subsequently in this section, after we consider transformations between dissimilar coordinate representations.

Transformations of Different Kind

If the xy -plane coincides with the equator plane and the xz -plane with the reference meridian plane, then we have the following:

From spherical to Cartesian:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = (R+h) \begin{pmatrix} \cos \psi \cos \lambda \\ \cos \psi \sin \lambda \\ \sin \psi \end{pmatrix} \quad (55.7)$$

From Cartesian to spherical:

$$\begin{pmatrix} \lambda \\ \psi \\ h \end{pmatrix} = \begin{pmatrix} \arctan(y/x) \\ \arctan\left(z/\sqrt{x^2 + y^2}\right) \\ \sqrt{x^2 + y^2 + z^2} - R \end{pmatrix} \quad (55.8)$$

From ellipsoidal to Cartesian:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} [N + h] \cos \phi \cos \lambda \\ [N + h] \cos \phi \sin \lambda \\ [N(1 - e^2) + h] \sin \phi \end{pmatrix} \quad (55.9)$$

with

$$N = \frac{a}{W} \quad (55.10)$$

and

$$W = \sqrt{1 - e^2 \sin^2 \phi} \quad (55.11)$$

In these equations the variable N has a distinct geometric significance: it is the radius of curvature in the prime vertical plane. This plane goes through the local normal and is perpendicular to the meridian plane. In other words, N describes the curvature of the curve obtained through the intersection of the prime vertical plane and the ellipsoid. The curve formed through the intersection of the meridian plane and the ellipsoid is given by M ; see Fig. 55.3.

The varying radius of curvature M of the elliptic meridian is given by

$$M = \frac{a(1 - e^2)}{W^3} \quad (55.12)$$

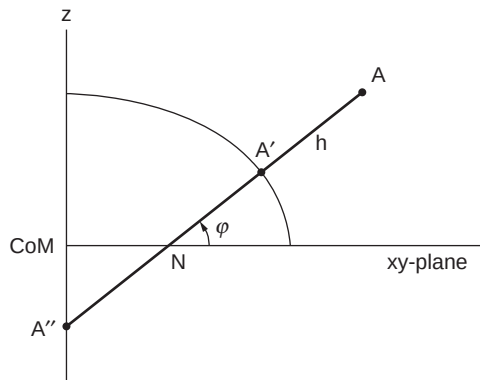


FIGURE 55.3 Meridian plane through point A.

From Cartesian to ellipsoidal:

$$\lambda = \arctan(y/x) \quad (55.13)$$

The geodetic latitude ϕ can be obtained by the following iteration scheme, starting with an approximate value for the geodetic latitude ϕ_0 :

$$\begin{aligned} \phi_0 &= \arctan \left[z / \sqrt{x^2 + y^2} \right] \\ N_0 &= a / \sqrt{1 - e^2 \sin^2 \phi_0} \\ \phi &= \arctan \left[(z + N_0 e^2 \sin \phi_0) / \sqrt{x^2 + y^2} \right] \end{aligned} \quad (55.14)$$

If $|\phi - \phi_0| > \epsilon$, then set ϕ_0 equal to ϕ and go back to the computation of N . After the iteration, h can be computed directly:

$$h = \sqrt{[x^2 + y^2 + (z + N e^2 \sin \phi)^2]} - N \quad (55.15)$$

Through more cumbersome expressions an analytical solution for the geodetic latitude ϕ as a function of the Cartesian coordinates $\{x, y, z\}$ is possible.

Transformations of Same Kind

Orthogonal Transformations: Translation and Rotation

When groups of points are known in their relative position with respect to each other, the use of Cartesian coordinates ($3n$ in total) becomes superfluous. As a matter of fact there are 6° of freedom, since the position of the origin with respect to the group is arbitrary, as is the orientation of the frame axes. Two groups of identical points or, for that matter, one and the same group of points expressed in two arbitrary but different coordinate frames may be represented by the following orthogonal transformation:

$$\mathbf{x}' = \mathbf{R}\mathbf{x} + \mathbf{t}' \quad (55.16)$$

or

$$\mathbf{x}' = \mathbf{R}(\mathbf{x} - \mathbf{t}) \quad (55.17)$$

The vector $\mathbf{t}$ represents a translation. In Eq. (55.16) $\mathbf{t}'$ represents the vector of the old origin in the new frame ($\mathbf{x}' - \mathbf{t}' = \mathbf{R}\mathbf{x}$); in Eq. (55.17) $\mathbf{t}$ represents the coordinates of the origin of the new frame in the old coordinate frame. The relation between the two translation vectors is represented by

$$\mathbf{t}' = -\mathbf{R}\mathbf{t} \quad (55.18)$$

The rotation matrix describes the rotations around the frame axes. In Eq. (55.16) $\mathbf{R}$ describes a rotation around axes through the origin of the $\mathbf{x}$ frame; in Eq. (55.17) $\mathbf{R}$ describes a rotation around axes through the origin of the $\mathbf{x}'$ frame. We define the sense of rotations as follows: the argument angle of the rotation matrix is taken positive if one views the rotation as counterclockwise from the positive end of the rotation axis looking back to the origin. For an application relating coordinates in a local frame to coordinates in a global frame, see Section 55.3.

In the above equations we assume three consecutive rotations, first around the z axis with an argument angle γ , then around the y axis around an argument angle β , and finally around the x axis with the argument angle α . So we have

$$\mathbf{R} = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma) \quad (55.19)$$

$$\mathbf{R} = \begin{pmatrix} \cos\beta & \cos\beta \sin\gamma & -\sin\beta \\ -\cos\alpha \sin\gamma + \sin\alpha \sin\beta \cos\gamma & \cos\alpha \cos\gamma + \sin\alpha \sin\beta \sin\gamma & \sin\alpha \cos\beta \\ \sin\alpha \sin\gamma + \cos\alpha \cos\beta \cos\gamma & -\sin\alpha \cos\gamma + \cos\alpha \sin\beta \sin\gamma & \cos\alpha \cos\beta \end{pmatrix} \quad (55.20)$$

One outcome of these orthogonal transformations is an inventory of variables that are invariant under these transformations. Without any proof, these include lengths, angles, sizes and shapes of figures, and volumes — important quantities for the civil or survey engineer.

Similarity Transformations: Translation, Rotation, Scale

In the previous section we saw that the relative location of n points can be described by fewer than $3n$ coordinates: $3n - 6$ quantities (for instance, an appropriate choice of distances and angles) are necessary but also sufficient. Exceptions have to be made for so-called critical configurations such as four points in a plane. The 6 is nothing else than the 6° of freedom supplied by the orthogonal transformation: three translations and three rotations.

A simple but different reasoning leads to the same result. Imagine a tetrahedron in a three-dimensional frame. The four corner points are connected by six distances. These are exactly the six necessary but sufficient quantities to describe the form and shape of the tetrahedron. These six sides determine this figure completely in size and shape. A fifth point will be positioned by another three distances to any three of the four previously mentioned points. Consequently, a field of n points (in three-dimensional) will be necessarily but sufficiently described by $3n - 6$ quantities. We need these types of reasoning in three-dimensional geometric satellite geodesy.

If we just consider the *shape* of a figure spanned by n points (we are not concerned any more about the *size* of the figure), then we need even one quantity fewer than $3n - 6$ (i.e., $3n - 7$); we are now ignoring the scale, in addition to the position and orientation of the figure. This constitutes just the addition of a seventh parameter to the six-parameter orthogonal transformation: the scale parameter σ . So we have

$$\mathbf{x}' = \sigma \mathbf{R} \mathbf{x} + \mathbf{t}' \quad (55.21)$$

or

$$\mathbf{x}' = \sigma \mathbf{R} (\mathbf{x} - \mathbf{t}) \quad (55.22)$$

Here also the vector $\mathbf{t}$ represents a translation. In Eq. (55.21) $\mathbf{t}'$ represents the vector of the old origin in the new scaled and rotated frame ($\mathbf{x}' - \mathbf{t}' = \sigma \mathbf{R} \mathbf{x}$); in Eq. (55.22) $\mathbf{t}$ represents the coordinates of the origin of the new frame in the old coordinate frame. The relation between the two translation vectors is represented by

$$\mathbf{t}' = -\sigma \mathbf{R} \mathbf{t} \quad (55.23)$$

One outcome of these similarity transformations is an inventory of invariant variables under these transformations. Without any proof, these include length ratios, angles, shapes of figures, and volume ratios, which are important quantities for the civil or survey engineer. The reader is referred to Leick and van Gelder [1975] for other important properties.

Curvilinear Coordinates and Transformations

One usually prefers to express coordinate differences in terms of the curvilinear coordinates on the sphere or ellipsoid or even locally, rather than in terms of the Cartesian coordinates. This approach also facilitates the study of effects due to changes in the adopted values for the reference ellipsoid (so-called *datum transformations*).

Curvilinear Coordinate Changes in Terms of Cartesian Coordinate Changes

Differentiating the transformation formulas in which the Cartesian coordinates are expressed in terms of the ellipsoidal coordinates (see Eq. (55.9)), we obtain a differential formula relating the Cartesian total differentials $\{dx, dy, dz\}$ as a function of the ellipsoidal total differentials $\{d\lambda, d\phi, dh\}$:

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \mathbf{J} \begin{pmatrix} d\lambda \\ d\phi \\ dh \end{pmatrix} \quad (55.24)$$

The projecting matrix $\mathbf{J}$ is nothing else than the Jacobian of partial derivatives:

$$\mathbf{J} = \frac{\partial(x, y, z)}{\partial(\lambda, \phi, h)} \quad (55.25)$$

Carrying out the differentiation, one finds

$$\mathbf{J} = \begin{bmatrix} -(N+h) \cos \phi \sin \lambda & -(M+h) \sin \phi \cos \lambda & \cos \phi \cos \lambda \\ (N+h) \cos \phi \cos \lambda & -(M+h) \sin \phi \sin \lambda & \cos \phi \sin \lambda \\ 0 & (M+h) \cos \phi & \sin \phi \end{bmatrix} \quad (55.26)$$

On inspection, this Jacobian $\mathbf{J}$ is simply a product of a rotation matrix $\mathbf{R}(\lambda, \phi)$ and a metric matrix $\mathbf{H}(\phi, h)$ [Soler, 1976]:

$$\mathbf{J} = \mathbf{R} \mathbf{H} \quad (55.27)$$

or, in full,

$$\mathbf{R} = \begin{bmatrix} -\sin \lambda & -\sin \phi \cos \lambda & \cos \phi \cos \lambda \\ \cos \lambda & -\sin \phi \sin \lambda & \cos \phi \sin \lambda \\ 0 & \cos \phi & \sin \phi \end{bmatrix} \quad (55.28)$$

$$\mathbf{H} = \begin{bmatrix} (N+h) \cos \phi & 0 & 0 \\ 0 & (M+h) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (55.29)$$

It turns out that the rotation matrix $\mathbf{R}(\lambda, \phi)$ relates the local $\{e, n, u\}$ frame to the geocentric $\{x, y, z\}$ frame; see further the discussion of earth-fixed coordinates in Section 55.3. The metric matrix $\mathbf{H}(\phi, h)$ relates the curvilinear coordinates' longitude, latitude, and height in radians and meters to the curvilinear coordinates, all expressed in meters.

The formulas just given are the simple expressions relating a small arc distance ds to the corresponding small angle $d\alpha$ through the radius of curvature. The radius of curvature for the longitude component is equal to the radius of the local parallel circle, which in turn

$$\begin{pmatrix} d\lambda_m \\ d\phi_m \\ dh_m \end{pmatrix} = \mathbf{H}(\phi, h) \begin{pmatrix} d\lambda_{\text{rad}} \\ d\phi_{\text{rad}} \\ dh_m \end{pmatrix} \quad (55.30)$$

equals the radius of curvature in the prime vertical plane times the cosine of the latitude.

The power of this evaluation is more apparent if one realizes that the inverse Jacobian, expressing the ellipsoidal total differentials $\{d\lambda, d\phi, dh\}$ as a function of the Cartesian total differentials $\{dx, dy, dz\}$, is easily obtained, whereas an analytic solution expressing the geodetic ellipsoidal coordinates in terms of the Cartesian coordinates is extremely difficult to obtain. So, we have

$$\begin{pmatrix} d\lambda \\ d\phi \\ dh \end{pmatrix} = \mathbf{J}^{-1} \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} \quad (55.31)$$

With the relationship in Eq. (55.27) $\mathbf{J}^{-1}$ becomes simply

$$\mathbf{J}^{-1} = (\mathbf{RH})^{-1} = \mathbf{H}^{-1}\mathbf{R}^T \quad (55.32)$$

or, in full,

$$\mathbf{J}^{-1} = \begin{bmatrix} \frac{\sin \lambda}{(N+h)\cos \phi} & \frac{\cos \lambda}{(N+h)\cos \phi} & 0 \\ \frac{-\sin \phi \cos \lambda}{M+h} & \frac{-\sin \phi \sin \lambda}{M+h} & \frac{\cos \phi}{M+h} \\ \cos \phi \cos \lambda & \cos \phi \sin \lambda & \sin \phi \end{bmatrix} \quad (55.33)$$

This equation gives a simple analytic expression for the inverse Jacobian, whereas the analytic expression for the original function is virtually impossible.

Curvilinear Coordinate Changes Due to a Similarity Transformation

Differentiating Eq. (55.21) with respect to the similarity transformation parameters $\alpha, \beta, \gamma, t'_x, t'_y, t'_z$, and σ , one obtains

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix}_7 = \mathbf{J}_7 \begin{pmatrix} d\alpha \\ d\beta \\ d\gamma \\ dt'_x \\ dt'_y \\ dt'_z \\ d\sigma \end{pmatrix} \quad (55.34)$$

with

$$\mathbf{J}_7 = \frac{\partial(x, y, z)}{\partial(\alpha, \beta, \gamma, t'_x, t'_y, t'_z, \sigma)} \quad (55.35)$$

The Jacobian $\mathbf{J}_7$ is a matrix that consists of seven column vectors

$$\mathbf{J}_7 = [\mathbf{j}_1 | \mathbf{j}_2 | \mathbf{j}_3 | \mathbf{j}_4 | \mathbf{j}_5 | \mathbf{j}_6 | \mathbf{j}_7] \quad (55.36)$$

with

$$\begin{aligned}\mathbf{j}_1 &= \sigma \mathbf{L}_1 \mathbf{R}_1(\alpha) \mathbf{R}_2(\beta) \mathbf{R}_3(\gamma) \mathbf{x} \\ &= \sigma \mathbf{L}_1 \mathbf{R} \mathbf{x} \\ &= \mathbf{L}_1 \mathbf{x}'\end{aligned}\tag{55.37}$$

$$\mathbf{j}_2 = \sigma \mathbf{R}_1(\alpha) \mathbf{L}_2 \mathbf{R}_2(\beta) \mathbf{R}_3(\gamma) \mathbf{x}\tag{55.38}$$

$$\begin{aligned}\mathbf{j}_3 &= \sigma \mathbf{R}_1(\alpha) \mathbf{R}_2(\beta) \mathbf{R}_3(\gamma) \mathbf{L}_3 \mathbf{x} \\ &= \sigma \mathbf{R} \mathbf{L}_3 \mathbf{x}\end{aligned}\tag{55.39}$$

$$[\mathbf{j}_4 | \mathbf{j}_5 | \mathbf{j}_6] = \mathbf{I} \quad (3 \times 3) \text{ identity matrix}\tag{55.40}$$

$$\begin{aligned}\mathbf{j}_7 &= \mathbf{R}_1(\alpha) \mathbf{R}_2(\beta) \mathbf{R}_3(\gamma) \mathbf{x} \\ &= \mathbf{R} \mathbf{x} \\ &= (\mathbf{x}' - \mathbf{t}')/\sigma\end{aligned}\tag{55.41}$$

since

$$\partial \mathbf{R}_1 / \partial \alpha = \mathbf{L}_1 \mathbf{R}_1(\alpha) = \mathbf{R}_1(\alpha) \mathbf{L}_1\tag{55.42}$$

$$\partial \mathbf{R}_2 / \partial \beta = \mathbf{L}_2 \mathbf{R}_2(\beta) = \mathbf{R}_2(\beta) \mathbf{L}_2\tag{55.43}$$

$$\partial \mathbf{R}_3 / \partial \gamma = \mathbf{L}_3 \mathbf{R}_3(\gamma) = \mathbf{R}_3(\gamma) \mathbf{L}_3\tag{55.44}$$

and

$$\mathbf{L}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}\tag{55.45}$$

$$\mathbf{L}_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}\tag{55.46}$$

$$\mathbf{L}_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\tag{55.47}$$

The advantage of these $\mathbf{L}$ matrices is that in many instances the derivative matrix (product) can be written as the original matrix pre- or postmultiplied by the corresponding $\mathbf{L}$ matrix [Lucas, 1963].

Curvilinear Coordinate Changes Due to a Datum Transformation

Differentiating Eq. (55.9) with respect to the semimajor axis a and the flattening f , one obtains

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix}_{a,f} = \mathbf{J}_{a,f} \begin{pmatrix} da \\ df \end{pmatrix} \quad (55.48)$$

with (see Soler and van Gelder [1987])

$$\mathbf{J}_{a,f} = \frac{\partial(x, y, z)}{\partial(a, f)} \quad (55.49)$$

and

$$\mathbf{J}_{a,f} = \begin{bmatrix} \cos\phi \cos\lambda/W & a(1-f)\sin^2\phi \cos\phi \cos\lambda/W^3 \\ \cos\phi \sin\lambda/W & a(1-f)\sin^2\phi \cos\phi \sin\lambda/W^3 \\ (1-e^2)\sin\phi/W & (M\sin^2\phi - 2N)(1-f)\sin\phi \end{bmatrix} \quad (55.50)$$

Also see Soler and van Gelder [1987] for the second-order derivatives.

Curvilinear Coordinate Changes Due to a Similarity and a Datum Transformation

The curvilinear effects of a redefinition of the coordinate frame due to a similarity transformation and a datum transformation are computed by adding Eqs. (55.34) and (55.48) and substituting them into

$$\begin{pmatrix} d\lambda \\ d\phi \\ dh \end{pmatrix} = \mathbf{J}^{-1} \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} \quad (55.51)$$

with

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix}_7 + \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix}_{a,f} \quad (55.52)$$

55.3 Coordinate Frames Used in Geodesy and Some Additional Relationships

Earth-Fixed

Earth-Fixed Geocentric

From the moment satellites were used to study geodetic aspects of the earth, one had to deal with modeling the motion of the satellite (a point mass) around the earth's center of mass (CoM). The formulation of the equations of motion is easiest when referred to the CoM. This point became almost naturally the origin of the coordinate frame in which the earthbound observers were situated. For the orientation of

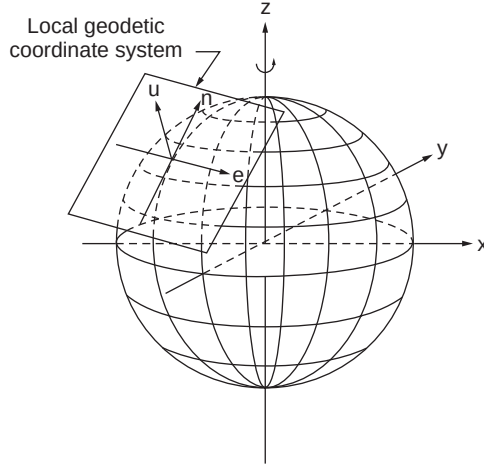


FIGURE 55.4 A geocentric and a local Cartesian coordinate frame.

the x and z axes, see the discussion of spherical three-dimensional polar coordinates in Section 55.2 and the discussion of polar motion in this section.

Earth-Fixed Topocentric Cartesian

An often used local frame is the earth-fixed topocentric coordinate frame. The origin resides at the position of the observer's instrument. Although in principle arbitrary, one often chooses the x axis pointing east, the y axis pointing north, and the z axis pointing up. This e, n, u frame is again a right-handed frame. With respect to the direction of the local z or u axis, various choices are possible: the u axis coincides with the negative direction of the local gravity vector (the first axis of a leveled theodolite) or along the normal perpendicular to the surface of the ellipsoid.

An Important Relationship Using an Orthogonal Transformation

The transformation formulas between a geocentric coordinate frame and a local coordinate frame are (see Fig. 55.4):

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{R}_3\left(-\lambda_a - \frac{\pi}{2}\right) \mathbf{R}_1\left(+\phi_a - \frac{\pi}{2}\right) \begin{pmatrix} e \\ n \\ u \end{pmatrix} + \begin{pmatrix} [N_a + h_a] \cos \phi_a \cos \lambda_a \\ [N_a + h_a] \cos \phi_a \sin \lambda_a \\ [N_a (1 - e^2) + h_a] \sin \phi_a \end{pmatrix} \quad (55.53)$$

One should realize that this transformation formula is of the orthogonal type shown in Eq. (55.16):

$$\mathbf{x}' = \mathbf{R}\mathbf{x} + \mathbf{t}' \quad (55.54)$$

whereby $\mathbf{x}'$ = the geocentric Cartesian vector

$$\mathbf{R} = \mathbf{R}_3(-\lambda_a - \pi/2) \mathbf{R}_1(\phi_a - \pi/2)$$

$\mathbf{t}'_a$ = the location of a in the (new) geocentric frame and is equal to:

$$\mathbf{t}'_a = \begin{pmatrix} [N_a + h_a] \cos \phi_a \cos \lambda_a \\ [N_a + h_a] \cos \phi_a \sin \lambda_a \\ [N_a (1 - e^2) + h_a] \sin \phi_a \end{pmatrix} \quad (55.55)$$

The rotation matrix $\mathbf{R} = \mathbf{R}_3(-\lambda_a - \pi/2)\mathbf{R}_1(\phi_a - \pi/2)$ is given in Section 55.2 as Eq. (55.28).

Given the geocentric coordinates $\{x, y, z\}$ of an arbitrary point (e.g., a satellite), if one wants to compute the local coordinates $\{e, n, u\}$ of that point, the local frame being centered at a , then one obtains for the inverse relationship

$$\begin{pmatrix} e \\ n \\ u \end{pmatrix} = \mathbf{R}_1\left(-\phi_a + \frac{\pi}{2}\right)\mathbf{R}_3\left(+\lambda_a + \frac{\pi}{2}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} - \begin{pmatrix} [N_a + h_a] \cos \phi_a \cos \lambda_a \\ [N_a + h_a] \cos \phi_a \sin \lambda_a \\ [N_a (1 - e^2) + h_a] \sin \phi_a \end{pmatrix} \quad (55.56)$$

The rotation matrix $\mathbf{R}_1 = (-\phi_a + \pi/2)\mathbf{R}_3(+\lambda_a + \pi/2)$ is the transpose of the matrix given in Section 55.2 as Eq. (55.28).

Earth-Fixed Topocentric Spherical

Satellites orbit the earth at finite distances. For such purposes as visibility calculations, one relates the local e, n, u coordinates to local spherical coordinates El (elevation or altitude angle), Az (azimuth, clockwise positive from the north), and Sr (slant range to the object):

$$\begin{pmatrix} e \\ n \\ u \end{pmatrix} = \text{Sr} \begin{pmatrix} \cos \text{El} \sin \text{Az} \\ \cos \text{El} \cos \text{Az} \\ \sin \text{El} \end{pmatrix} \quad (55.57)$$

The inverse relationships are

$$\begin{pmatrix} \text{El} \\ \text{Az} \\ \text{Sr} \end{pmatrix} = \begin{pmatrix} \arctan\left[u/\sqrt{e^2 + n^2}\right] \\ \arctan(e/n) \\ \sqrt{e^2 + n^2 + u^2} \end{pmatrix} \quad (55.58)$$

Note again that El, Az, and Sr form themselves a right-handed (curvilinear) frame.

Some Important Relationships Using Similarity and Datum Transformations

Increasing measurement accuracies and improved insights in the physics of the earth often cause reference frames to be reviewed. For instance, if coordinates of a station are given in an old frame, then with current knowledge of similarity transformation parameters relating the old $\mathbf{x}$ frame to the new $\mathbf{x}'$ frame, the new coordinates can be computed according to

$$\mathbf{x}' = \sigma \mathbf{R} \mathbf{x} + \mathbf{t}' \quad (55.59)$$

In many instances the translation and rotation transformation parameters are small, and the scale parameter σ deviates little from 1, so we introduce the following new symbols:

$$\begin{aligned} \sigma &= 1 + \delta\sigma \\ \alpha &= \delta\epsilon \\ \beta &= \delta\psi \end{aligned} \quad (55.60)$$

$$\gamma = \delta\omega$$

$$t'_x = \Delta x$$

$$t'_y = \Delta y$$

$$t'_z = \Delta z$$

Neglecting second-order effects, the rotation matrix $\mathbf{R}$ can be written as the sum of an identity matrix $\mathbf{I}$ and a skew-symmetric matrix $\delta\mathbf{R}$:

$$\mathbf{R} = \mathbf{R}_1(\delta\epsilon)\mathbf{R}_2(\delta\psi)\mathbf{R}_3(\delta\omega) \quad (55.61)$$

$$= \mathbf{I} + \delta\mathbf{R} \quad (55.62)$$

with

$$\delta\mathbf{R} = \begin{pmatrix} 0 & \delta\omega & -\delta\psi \\ -\delta\omega & 0 & \delta\epsilon \\ \delta\psi & -\delta\epsilon & 0 \end{pmatrix} \quad (55.63)$$

Equation (55.59) becomes

$$\mathbf{x}' = (1 + \delta\sigma)(\mathbf{I} + \delta\mathbf{R})\mathbf{x} + \Delta\mathbf{x} \quad (55.64)$$

or, neglecting second-order effects,

$$\mathbf{x}' = \mathbf{x} + \mathbf{dx} \quad (55.65)$$

with

$$\mathbf{dx} = \delta\sigma\mathbf{x} + \delta\mathbf{R}\mathbf{x} + \Delta\mathbf{x} \quad (55.66)$$

See Section 55.8 for a variety of parameter sets relating the various reference frames and datum values.

Inertial and Quasi-Inertial

Inertial Geocentric Coordinate Frame

For the derivation of the equations of motion of point masses in space we need so-called inertial frames. These are frames where Newton's laws apply. These frames are nonrotating, where point masses either have uniform velocity or are at rest. Popularly speaking, in these frames the stars or, better, extragalactic points or quasars, are “fixed” (i.e., not moving in a rotational sense). Since the stars are at such large distances from the earth, it is often sufficient in geodetic astronomy to consider the inertial directions. Instead of the inertial coordinates of the stars we consider the vector $\mathbf{d}$, consisting of the three direction cosines. One has to realize that these direction cosines are dependent on only two angles. Consequently, only two direction cosines contain independent information because the three direction cosines squared sum up to 1.

The two angles are (see [Fig. 55.5](#))

α right ascension

δ declination

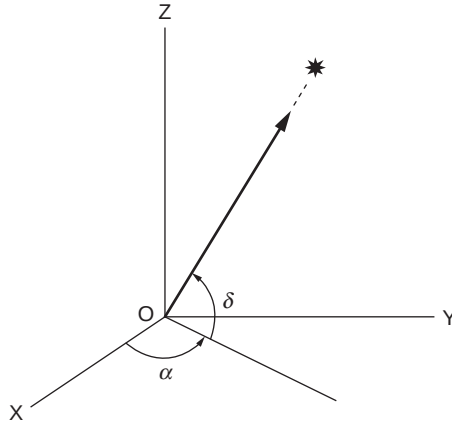


FIGURE 55.5 Direction to a satellite or star: right ascension, α , and declination, δ .

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = l\mathbf{d} \quad (55.67)$$

with

$$\mathbf{d} = \begin{pmatrix} \cos\delta & \cos\alpha \\ \cos\delta & \sin\alpha \\ \sin\delta \end{pmatrix} \quad (55.68)$$

The right ascension is counted counterclockwise positive from the X axis and is defined as the intersection of the earth's equatorial plane and the plane of the earth's orbit around the sun. One of the points of intersection is called the *vernal equinox*: it is that point in the sky among the stars where the sun appears as viewed from earth at the beginning of spring in the northern hemisphere. The declination is counted from the equatorial plane in the same manner as the latitude; see [Fig. 55.6](#).

Quasi-Inertial Coordinate Frame

In the previous section the position of the origin was not defined yet: the origin of the inertial frame is not to coincide with the center of mass of the earth; since the earth itself orbits around the sun, the center of mass of the earth is subject to accelerations. Similarly, the center of mass of the sun and all its planets is rotating around the center of our galaxy, and the galaxy experiences gravitational forces from other galaxies. A continuation of this reasoning will improve the quality of “inertiality” of the coordinate frame, but the practical application for the description of the motion of earth-orbiting satellites has been completely lost.

In Section 55.5 a practical solution is presented: in orientation the frame is as inertial as possible, but the origin has been chosen to coincide with the earth's center of mass. Such frames are called *quasi-inertial* frames. The apparent forces caused by the (small) accelerations of the origin have to be accounted for later.

Relation between Earth-Fixed and Inertial

Satellite equations of motion are easily dealt with in an inertial frame, but we observers are likely to model our positions and relatively slow velocities in an earth-fixed frame. The relationship between these two frames has to be dealt with.

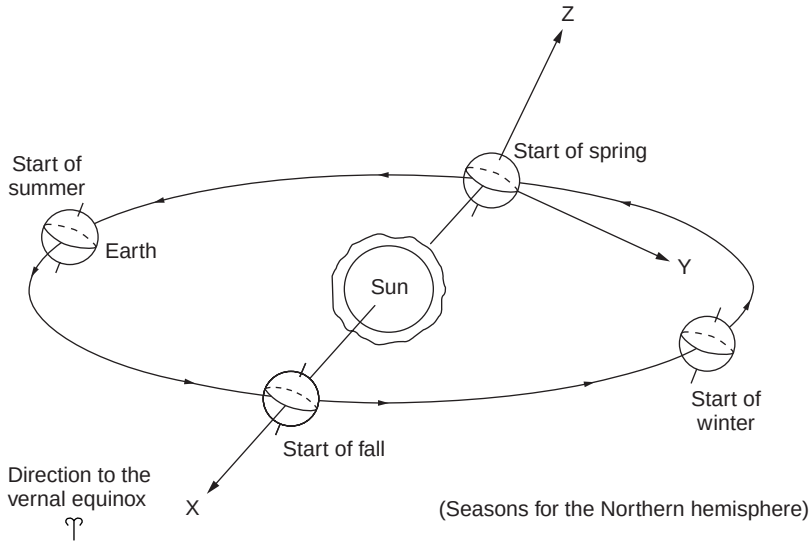


FIGURE 55.6 The (quasi-)inertial reference frame with respect to sun and earth.

Time and Sidereal Time

The diurnal rotation of the earth is given by its average angular velocity:

$$\omega_e = 7.292115 \times 10^{-5} \text{ rad/s} \quad (55.69)$$

This inertial angular velocity results in an average day length of

$$T = \frac{2\pi}{\omega_e} = 86164.1 \text{ sec} \quad (55.70)$$

This *sidereal day*, based on the earth's spin rotation with respect to the fixed stars, deviates from our $24 \times 60 \times 60 = 86,400$ -sec day by 3 min, 55.9 sec. That is why we see an arbitrary star constellation in the same position in the sky each day about 4 min earlier. Our daily lives are based on the earth's spin with respect to the sun, the *solar day*. Since the earth advances about 1° per day in its orbit around the sun, the earth has not completed a full spin with respect to the sun when it completes one full turn with respect to the stars; see Fig. 55.7.

For practical purposes the angular velocity must include the effect of precession; see the next section. We have

$$\omega_c^* = 7.2921158553 \times 10^{-5} \text{ rad/s} \quad (55.71)$$

The angle between the vernal equinox and the Greenwich meridian as measured along the equator is *Greenwich apparent sidereal time* (GAST). This angle increases in time by ω_c^* per second. With the help of the formula of Newcomb, we are able to compute GAST [IERS, 1992]:

$$\text{GAST}(t) = a + bT_u + cT_u^2 + dT_u^3 + ee \quad (55.72)$$

with

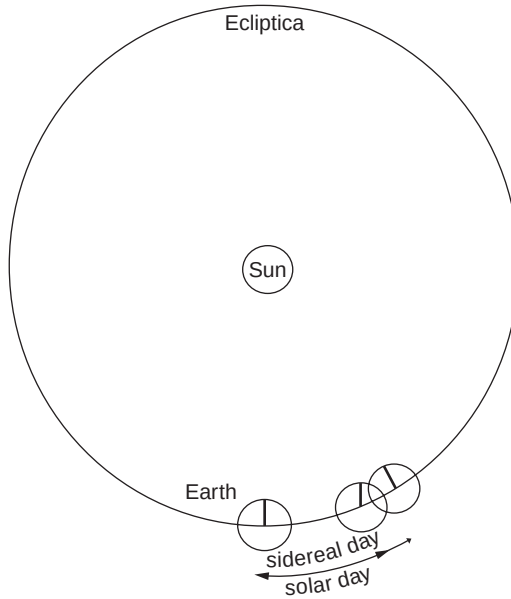


FIGURE 55.7 The earth's orbit around the sun: sidereal day and solar day.

$$\begin{aligned}
 a &= 18\text{ h } 41\text{ min } 50.54841\text{ s} \\
 b &= 8,640,184.812866\text{ s/century}^3 \\
 c &= 0.093104\text{ s/century}^3 \\
 d &= -0.0000062\text{ s/century}^3 \\
 ee &= \text{the equation of the equinoxes}
 \end{aligned}
 \tag{55.73}$$

T_u is measured in Julian centuries of 36,525 universal days, since 1.5 January 2000 ($JD_0 = 2,451,545.0$). This means that T_u , until the year 2000, is negative. T_u can be computed from

$$T_u = \frac{(JD - 2,451,545.0)}{36,525}
 \tag{55.74}$$

when the Julian day number, JD , is given.

Polar Motion

Polar motion, or on a geological time scale “polar wandering,” represents the motion of the earth's spin axis with respect to an earth-fixed frame. Polar motion changes our latitudes, since if the z axis were chosen to coincide with the instantaneous position of the spin axis, our latitudes would change continuously.

Despite the earth's nonelastic characteristics, excitation forces keep polar motion alive. Polar motion is the motion of the instantaneous rotation axis, or celestial ephemeris pole (CEP), with respect to an adopted reference position, the conventional terrestrial pole (CTP) of the conventional terrestrial reference frame (CTRF). The adopted reference position, or conventional international origin (CIO), was the main position of the CEP between 1900 and 1905. Since then the mean CEP has drifted about 10 m away from the CIO in a direction of longitude 280° .

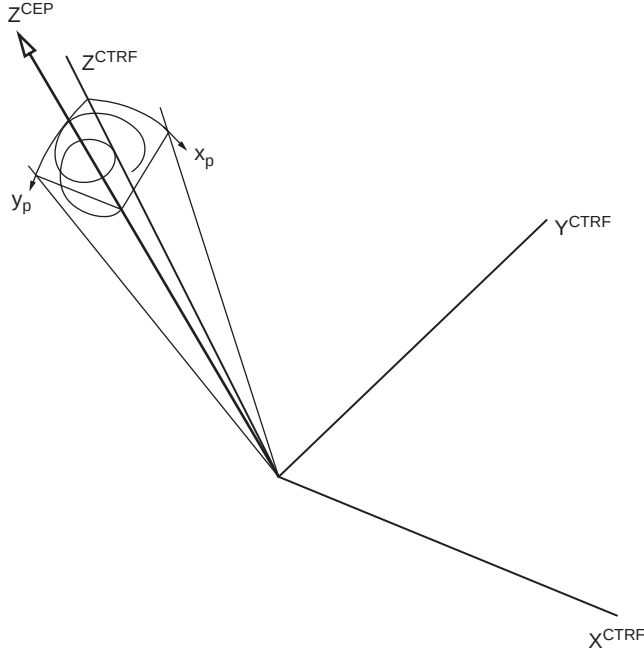


FIGURE 55.8 Local curvilinear pole coordinates x_p, y_p with respect to the conventional terrestrial reference frame.

The transformation from the somewhat earth-fixed frame $\mathbf{x}_{\text{CEP}}$ to $\mathbf{x}_{\text{CTRF}}$ is

$$\mathbf{x}_{\text{CTRF}} = \mathbf{R}_2(-x_p) \mathbf{R}_1(-y_p) \mathbf{x}_{\text{CEP}} \quad (55.75)$$

The position of the pole is expressed in local curvilinear coordinates that have their origin in the CIO. The angle x_p (radians) increases along the Greenwich meridian south, whereas the angle y_p increases along the meridian $\lambda = 270^\circ$ south; see Fig. 55.8. As we saw in the previous subsection, the earth rotates daily around its (moving) CEP axis. Expanding the transformation of Eq. (55.75) but now also including the sidereal rotation of the earth, we obtain the following relationship between an inertial reference frame and the CTRF:

$$\mathbf{x}_{\text{CEP}} = \mathbf{R}_3(\text{GAST}) \mathbf{x}_{\text{in}} \quad (55.76)$$

This relationship is shown in Fig. 55.9. Combining Eqs. (55.75) and (55.76), we have

$$\mathbf{x}_{\text{CTRF}} = \mathbf{R}_2(-x_p) \mathbf{R}_1(-y_p) \mathbf{R}_3(\text{GAST}) \mathbf{x}_{\text{in}} \quad (55.77)$$

or, in short,

$$\mathbf{x}_{\text{CTRF}} = \mathbf{R}_s \mathbf{x}_{\text{in}} \quad (55.78)$$

where $\mathbf{R}_s$ represents the combined earth rotation due to polar motion and diurnal rotation (length of day).

Two main frequencies make up the polar motion: the Chandler wobble of 435 days (14 months, more or less) and the annual wobble of 365.25 days. A prediction model for polar motion is

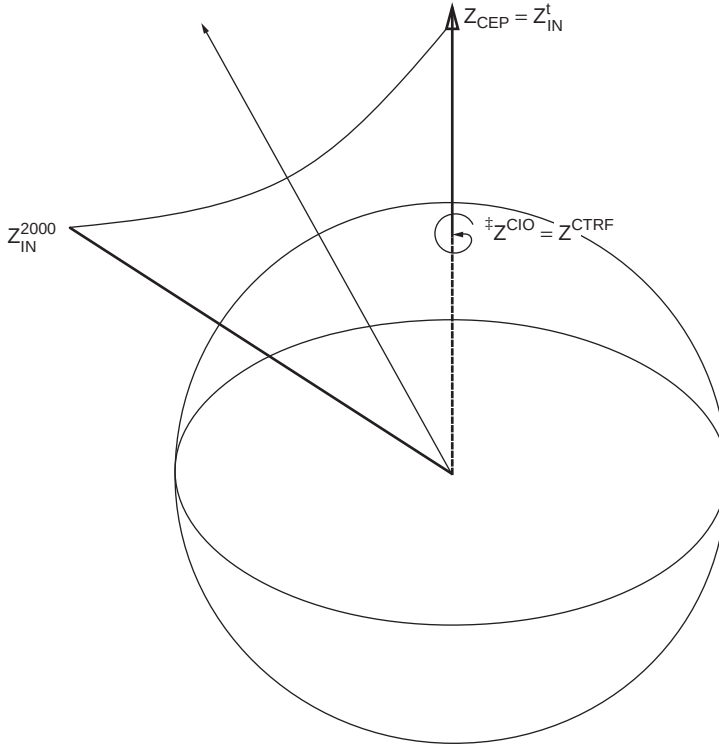


FIGURE 55.9 Relationship between the quasi-earth-fixed frame $\mathbf{X}_{\text{CEP}}$ and the inertial frame $\mathbf{X}_{\text{in}}$.

$$\begin{aligned}
 x_p(t) = & x_p(t_0) + \dot{x}_p(t-t_0) + A_s^x \sin \left[\frac{2\pi}{T_1}(t-t_0) \right] + A_c^x \cos \left[\frac{2\pi}{T_1}(t-t_0) \right] \\
 & + C_s^x \sin \left[\frac{2\pi}{T_2}(t-t_0) \right] + C_c^x \cos \left[\frac{2\pi}{T_2}(t-t_0) \right]
 \end{aligned} \tag{55.79}$$

$$\begin{aligned}
 y_p(t) = & y_p(t_0) + \dot{y}_p(t-t_0) + A_s^y \sin \left[\frac{2\pi}{T_1}(t-t_0) \right] + A_c^y \cos \left[\frac{2\pi}{T_1}(t-t_0) \right] \\
 & + C_s^y \sin \left[\frac{2\pi}{T_2}(t-t_0) \right] + C_c^y \cos \left[\frac{2\pi}{T_2}(t-t_0) \right]
 \end{aligned} \tag{55.80}$$

where $T_1 = 365.25$ days
 $T_2 = 435$ days

Similarly, the ever-increasing angle GAST has to be corrected for seasonal variations in ω . These variations are in the order of ± 1 msec. A prediction model for length-of-day variations is

$$\delta \text{UT1} = a \sin 2\pi t + b \cos 2\pi t + c \sin 4\pi t + d \cos 4\pi t \tag{55.81}$$

with $a = +0.0220$ sec
 $b = -0.0120$ sec
 $c = -0.0060$ sec
 $d = +0.0070$ sec
 $t = 2000.000 + [(\text{MJD} - 51544.03)/365.2422]$.

With Eq. (55.72), we have

$$\text{GAST} = \text{GAST}_{(t)} + \delta\text{UT1} \quad (55.82)$$

Due to the flattening, f , $\mathbf{x}_{\text{in}}$ in Eq. (55.77) is not star-fixed but would be considered a proper (quasi-) inertial frame if we froze the frame at epoch t . So, Eq. (55.78) is more properly expressed as

$$\mathbf{x}_{\text{CTRF}} = \mathbf{R}_S \mathbf{x}_{\text{in}}^t \quad (55.83)$$

The motions of the CEP with respect to the stars are called *precession* and *nutation*.

Precession and Nutation

Precession and nutation represent the motion of the earth's spin axis with respect to an inertial frame.

To facilitate comparisons between observations of spatial objects (satellites, quasars, stars) that may be made at different epochs, a transformation is carried out between the inertial frame at epoch t (CEP at t) and the mean position of the inertial frame at an agreed-upon reference epoch t_0 . The reference epoch is again 1.5 January 2000, for which $JD_0 = 2,451,545.0$; see the discussion on time and sidereal time earlier in this section. At this epoch we define the conventional inertial reference frame (CIRF). So we have

$$\mathbf{x}_{\text{in}}^t = \mathbf{R}_{\text{NP}} \mathbf{x}_{\text{in}}^{2000} \quad (55.84)$$

All masses in the ecliptic plane (earth and sun) and close to it (planets) exert a torque on the tilted equatorial bulge of the earth. The result is that the CEP describes a cone with its half top angle equal to the obliquity ϵ : precession. The period is about 25,800 years. The individual orbits of the planets and moon cause deviations with respect to this cone: nutation. The largest effect is about 9 sec of arc, with a period of 18.6 years, caused by the inclined orbit of the moon.

The transformation $\mathbf{R}_{\text{NP}}$ is carried out in two steps. The mean position of the CEP is first updated for precession to a mean position at epoch t (now):

$$\mathbf{x}_{\text{in}}^{t,\text{mean}} = \mathbf{R}_p \mathbf{x}_{\text{in}}^{2000,\text{mean}} \quad (55.85)$$

Subsequently, the mean position of the CEP at epoch t is transformed to the true position of the CEP at epoch t due to nutation:

$$\mathbf{x}_{\text{in}}^{t,\text{true}} = \mathbf{R}_N \mathbf{x}_{\text{in}}^{t,\text{mean}} \quad (55.86)$$

Combining all transformations, we have

$$\mathbf{x}_{\text{CTRF}} = \mathbf{R}_S \mathbf{R}_N \mathbf{R}_p \mathbf{x}_{\text{CIRF}} \quad (55.87)$$

where $\mathbf{x}_{\text{CTRF}}$ is identical to $\mathbf{x}_{\text{earth-fixed}}^{\text{CIO}}$ and $\mathbf{x}_{\text{CIRF}}$ is identical to $\mathbf{x}_{\text{star-fixed}}^{2000}$

The rotation matrices $\mathbf{R}_p$ and $\mathbf{R}_N$ depend on the obliquity ϵ , longitude of sun, moon, etc. We have

$$\mathbf{R}_p = \mathbf{R}_3(-z) \mathbf{R}_2(\theta) \mathbf{R}_3(-\zeta) \quad (55.88)$$

with (see, e.g., IERS [1992])

$$\begin{aligned} \zeta &= 2 \ 306''.218 \ 1 T_u + 0''.301 \ 88 T_u^2 + 0''.017 \ 998 T_u^3 \\ \theta &= 2 \ 004''.310 \ 9 T_u - 0''.426 \ 65 T_u^2 - 0''.041 \ 833 T_u^3 \\ z &= 2 \ 306''.218 \ 1 T_u + 1''.094 \ 68 T_u^2 + 0''.018 \ 203 T_u^3 \end{aligned}$$

and

$$\mathbf{R}_N = \mathbf{R}_1(-\epsilon - \Delta\epsilon)\mathbf{R}_3(-\Delta\psi)\mathbf{R}_1(\epsilon) \quad (55.89)$$

with $\epsilon = 84381''.448 - 46''.8150T_u - 0''.018203T_u^2 + 0''.001813T_u^3$.

For the nutation in longitude, $\Delta\psi$, and the nutation in obliquity, $\Delta\epsilon$, a trigonometric series expansion is available consisting of $106 \times 2 \times 2$ parameters and five arguments: mean anomaly of the moon, the mean anomaly of the sun, mean elongation of the moon from the sun, the mean longitude of the ascending node of the moon, and the difference between the mean longitude of the moon and the mean longitude of the ascending node of the moon. There are 2×2 constants in each term: a sine coefficient, a cosine coefficient, a time-invariant coefficient, and a time-variant coefficient.

For more detail on these transformations, refer to earth orientation literature, such as Mueller [1969], Moritz and Mueller [1988], and IERS [1992].

55.4 Mapping

The art of mapping is referred to a technique that maps information from an n -dimensional space $\mathbf{R}_n$ to an m -dimensional space $\mathbf{R}_m$. Often information belonging to a high-dimensional space is mapped to a low-dimensional space. In other words, one has

$$n > m \quad (55.90)$$

In geodesy and surveying one may want to map a three-dimensional world onto a two-dimensional world. In photogrammetry, an aerial photograph can be viewed as a mapping procedure as well: a two-dimensional photo of the three-dimensional terrain. In least-squares adjustment we map an n -dimensional observation space onto a u -dimensional parameter space. In this section we restrict our discussion to the mapping:

$$\mathbf{R}_n \rightarrow \mathbf{R}_m \quad (55.91)$$

with

$$n = 3 \quad (55.92)$$

$$m = 2 \quad (55.93)$$

A three-dimensional earth, approximated by a sphere or, better, by an ellipsoid of revolution, cannot be mapped onto a two-dimensional surface, which is flat at the start (plane) or can be made flat (the surface of a cylinder or a cone), without distorting the original relative positions in $\mathbf{R}_3$. Any figure or, better, the relative positions between an arbitrary number of points on a sphere or ellipsoid, will also be distorted when mapped onto a plane, cylinder, or cone. The distortions of the figure (or part thereof) will increase with the area. Likewise, the mapping will introduce distortions that will become larger as the extent of the area to be mapped increases.

If one approximates (maps) a sphere of the size of the earth, the radius R being

$$R \approx 6371.000 \text{ km} \quad (55.94)$$

onto a plane tangent in the center of one's engineering project of diameter D km, one finds increased errors in lengths, angles, and heights the further one gets away from the center of the project. [Table 55.1](#) lists these errors in distance dS , angle $d\alpha$, and height dh , if one assumes the following case: one measures in the center of the project one angle of 60° and two equal distances of S km. In the plane assumption we would find in the two terminal points of both lines two equal angles of 60° and a distance between them of exactly S km. Basically, we have an equilateral triangle. In reality, on the curved, spherical earth we would measure angles larger than 60° and a distance between them shorter than S km. The error dh shows how the earth curves away from underneath the tangent plane in the center of the project.

TABLE 55.1 Errors in Length dS (km), Angle $d\alpha$ (arcseconds), and Height dh (km)^a

Diameter, D (km)	Length, S (km)	dS		$d\alpha$		dh (m)
		(m)	(ppm)	(arcseconds)	(ppm)	
0.100	0.050	−0.000	−0.000	0.000	0.000	0.000
0.200	0.100	−0.000	−0.000	0.000	0.000	0.001
0.500	0.250	−0.000	−0.000	0.000	0.000	0.005
2.000	1.000	−0.000	−0.003	0.001	0.005	0.078
5.000	2.500	−0.000	−0.019	0.007	0.032	0.491
20.000	10.000	−0.003	−0.308	0.110	0.509	7.848
50.000	25.000	−0.048	−1.925	0.688	3.184	49.050
200.000	100.000	−3.080	−30.797	11.002	50.937	784.790
500.000	250.000	−48.123	−192.494	68.768	318.372	4904.409
2000.000	1000.000	−3084.329	−3084.329	1101.340	5098.796	78,319.621

^a Depending on the diameter D (km) of a project or distance S ($=D/2$) from the center of the project for an equilateral triangle with sides of S km.

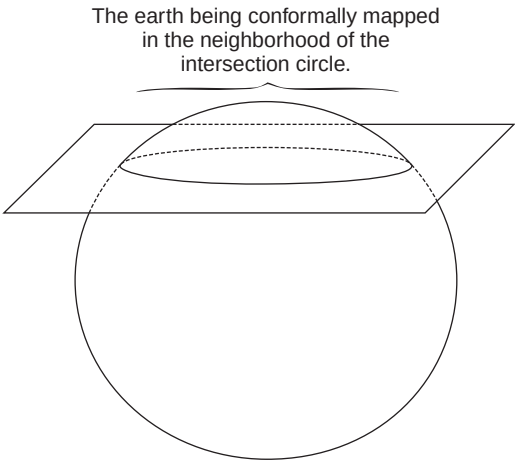
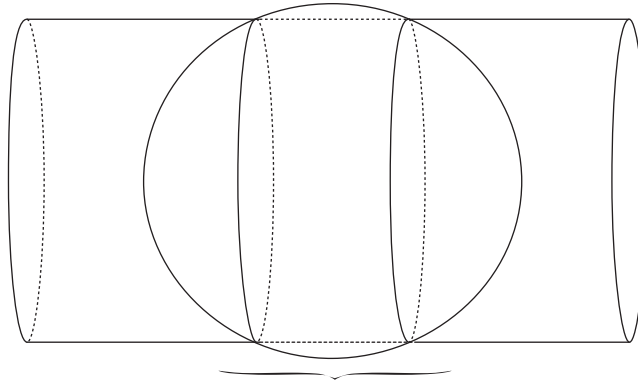


FIGURE 55.10 Mapping a sphere to a plane or lowered plane.

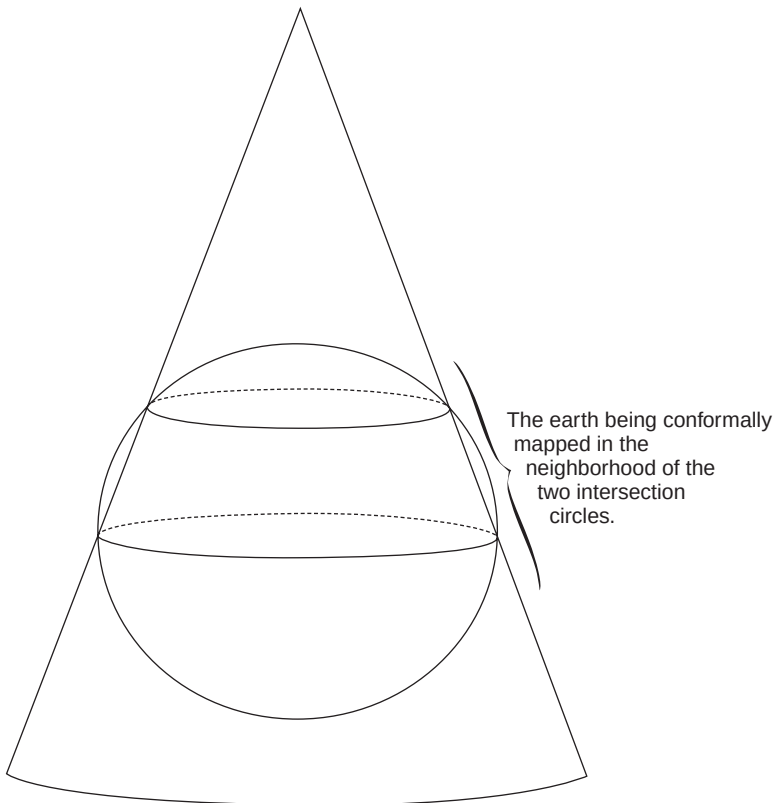
The table shows that errors in length and angle of larger than 1 ppm start to occur for project diameters larger than 20 km. Height differences obtained through leveling would be accurate enough, but vertical angles would start deviating from 90° by 1 arcsecond per 30 m. Within an area of 20 km fancy mapping procedures would not be needed to avoid errors of 1 ppm. The trouble starts if one wants to map an area of the size of the state of Indiana. A uniform strict mapping procedure has to be adhered to if one wants to work in one consistent system of mapping coordinates. In practice, a state of the size of Indiana is actually divided into two regions to keep the distortions within bounds. One may easily reduce the errors by a factor of 2 by making the plane not tangent to the sphere, but by lowering the plane from the center of the project by such an amount that the errors of dS in the center of the area are equal, but of opposite sign, to the errors dS at the border of the area (see Fig. 55.10). The U.S. State Plane Coordinate Systems are based on this practice.

As an alternative one may choose the mapping plane to be cylindrical or conical so that the mapping plane “follows” the earth’s curvature at least in one direction (see Figs. 55.11 and 55.12). Also, with these alternatives distortions are reduced even further by having the cylindrical or conical surface not tangent to the sphere or ellipsoid, but intersecting the surface to be mapped just below the tangent point. After mapping, the cylinder or cone can be “cut” and made into a two-dimensional map. A cylinder and a cone are called *developable surfaces*.



The earth being conformally mapped
in the neighborhood of two
intersection circles.

FIGURE 55.11 A cylinder trying to follow the earth's curvature.



The earth being conformally mapped in the
neighborhood of the
two intersection
circles.

FIGURE 55.12 A cone trying to follow the earth's curvature.

The notion of distortion of a figure is applied to different elements of a figure. If a (spherical) triangle is mapped, one may investigate how the length of a side or the angle between two sides is distorted in the mapping plane.

Two Worlds

We live in R_3 , which is to be mapped into R_2 . In R_3 we need three quantities to position ourselves: $\{x, y, z\}$, $\{\lambda, \psi, h\}_R$, or $\{\lambda, \phi, h\}_{a,f}$; see Section 55.2. In R_2 we need only two quantities, such as two Cartesian mapping coordinates $\{X, Y\}$ or two polar mapping coordinates $\{r, \alpha\}$.

Real world $\rightarrow$ Mapped world

$$\left. \begin{array}{l} \{x, y, z\} \\ \text{or} \\ \{\lambda, \psi, h\}_R \\ \text{or} \\ \{\lambda, \phi, h\}_{a,f} \end{array} \right\} \rightarrow \left\{ \begin{array}{l} \{X, Y\} \\ \text{or} \\ \{r, \alpha\} \end{array} \right.$$

The mapping M may be written symbolically as

$$\{r, \alpha\} = M\{x, y, z\} \quad (55.95)$$

or

$$\{X, Y\} = M'\{\lambda, \phi, h\}_{a,f} \quad (55.96)$$

M represents a mere mapping “prescription” of how the R_3 world is condensed into R_2 information. Note that the computer era made it possible to “store” the R_3 world digitally in R_3 (a file with three numbers per point). Only at the very end, if that information is to be presented on a map or computer screen, do we map to R_2 .

The analysis of distortion is, in this view, the mere comparison of corresponding geometrical elements in the real world and in the mapped world. A distance $s(x, y, z)$ between points i and j in the real world is compared to a distance $S = S(X, Y)$ in the mapped world. A scale distortion may be defined as the ratio

$$\sigma = \frac{S(X_i, Y_i, X_j, Y_j)}{s(x_i, y_i, z_i, x_j, y_j, z_j)} \quad (55.97)$$

or, for infinitely small distances,

$$\sigma = \frac{dS(dX_{ij}, dY_{ij})}{ds(dx_{ij}, dy_{ij}, dz_{ij})} \quad (55.98)$$

where

$$x_{ij} = x_j - x_i \quad \text{and so forth} \quad (55.99)$$

Similarly, an angle θ_{jik} in point i to points j and k in the real world is being compared to an angle Θ_{jik} in the mapped world, the angular distortion $d\theta$:

$$d\theta = \Theta_{jik} - \theta_{jik} \quad (55.100)$$

Other features may be investigated to assess the distortion of a certain mapping; for example, one may want to compare the area in the real world to the area in the mapped world. Various mapping prescriptions (mapping equations) exist that minimize scale distortion, angular distortion, area distortion, or combinations of these. In surveying engineering, and for that matter in civil engineering, the most widely applied map projection is the one that minimizes angular distortion. Moreover, there exists a class of map projections that do not show any angular distortion throughout the map; in other words,

$$d\theta = 0 \quad (55.101)$$

This class of map projections is known as *conformal map projections*. One word of caution is needed: the angles are only preserved in an infinitely small area. In other words, the points i, j, k have to be infinitely close together.

The surveyor or civil engineer often works in a relatively small area (compared to the dimensions of the earth). Therefore, it is extremely handy for his or her angular measurements from theodolite or total station, made in the real world, to be preserved in the mapped world. As a matter of fact, any of a variety of conformal map projections are used throughout the world by national mapping agencies. The most widely used map projection, also used by the military, is a conformal map projection, the so-called *Universal Transverse Mercator* (UTM) projection. In the U.S. all states have adopted some sort of a state plane coordinate system. This system is basically a local reference frame based on a certain type of conformal mapping. See the following three subsections.

Without proof, the necessary and sufficient conditions for conformality are that the real-world coordinates p and q and the mapping coordinates X and Y fulfill the Cauchy–Riemann equations or conditions:

$$\begin{aligned} \frac{\partial X}{\partial p} &= + \frac{\partial Y}{\partial q} \\ \frac{\partial X}{\partial q} &= - \frac{\partial Y}{\partial p} \end{aligned} \quad (55.102)$$

Purposely, the real-world variables p and q have not been identified. First we want to enforce the natural restriction on p, q, X, Y : they have to be isometric coordinates. The mapping coordinates are often isometric by definition; however, the real-world coordinates, as the longitude λ and the spherical latitude ψ (or geodetic latitude ϕ), are not isometric.

For instance, one arcsecond in longitude expressed in meters is very latitude dependent and, moreover, is not equal to one arcsecond in latitude in the very same point; see [Table 55.2](#).

Conformal Mapping Using Cartesian Differential Coordinates

In principle, we have four choices to map from R_3 to R_2 :

- A1: three-dimensional Cartesian $\rightarrow$ two-dimensional Cartesian
- A2: three-dimensional Cartesian $\rightarrow$ two-dimensional curvilinear
- B1: three-dimensional curvilinear $\rightarrow$ two-dimensional Cartesian
- B2: three-dimensional curvilinear $\rightarrow$ two-dimensional curvilinear

Although all four modes have known applications, we treat an example in this section with the B1 mode of mapping, whereas the following subsection deals with an example from the B2 mode.

From Eqs. (55.29) and (55.30) we have

$$\begin{aligned} d\lambda_m &= (N + h) \cos \phi d\lambda_{\text{rad}} \\ d\phi_m &= (M + h) d\phi_{\text{rad}} \end{aligned} \quad (55.103)$$

TABLE 55.2 Radius of Curvature in the Meridian M , Radius of Curvature in the Prime Vertical N , and Metric Equivalence of 1 Arcsecond in Ellipsoidal or Spherical Longitude λ (m) and in Ellipsoidal or Spherical Latitude ϕ/ψ (m) as a Function of Geodetic Latitude ϕ and Spherical Latitude ψ

ϕ/ψ	M	N	Ellipsoid		Sphere	
			1 in. λ	1 in. ϕ	1 in. λ	1 in. ψ
00.0	6,335,439	6,378,137	30.922	30.715	30.887	30.887
10.0	6,337,358	6,378,781	30.455	30.724	30.418	30.887
20.0	6,342,888	6,380,636	29.069	30.751	29.025	30.887
30.0	6,351,377	6,383,481	26.802	30.792	26.749	30.887
40.0	6,361,816	6,386,976	23.721	30.843	23.661	30.887
50.0	6,372,956	6,390,702	19.915	30.897	19.854	30.887
60.0	6,383,454	6,394,209	15.500	30.948	15.444	30.887
70.0	6,392,033	6,397,072	10.607	30.989	10.564	30.887
80.0	6,397,643	6,398,943	05.387	31.017	05.364	30.887
90.0	6,399,594	6,399,594	00.000	31.026	00.000	30.887

Note: Ellipsoidal values for WGS84: $a = 6,378,137$ m, $1/f = 298.257\ 223\ 563$.
Spherical values: $R = 6,371,000$ m.

A line element (small distance) in the real world (on the ellipsoid, $h = 0$) is

$$ds^2 = d\lambda_m^2 + d\phi_m^2 \quad (55.104)$$

A line element (small distance) in the mapped world (on paper) is

$$dS^2 = dX^2 + dY^2 \quad (55.105)$$

Equation (55.104) leads to

$$ds^2 = N^2 \cos^2 \phi d\lambda_{\text{rad}}^2 + M^2 d\phi_{\text{rad}}^2 \quad (55.106)$$

or

$$ds^2 = N^2 \cos^2 \phi \left(d\lambda_{\text{rad}}^2 = \frac{M^2}{N^2 \cos^2 \phi} d\phi_{\text{rad}}^2 \right) \quad (55.107)$$

Since we want to work with isometric coordinates in the real world (note that the mapping coordinates $\{X, Y\}$ are already isometric), we introduce the new variable dq . In other words, Eq. (55.107) becomes

$$ds^2 = N^2 \cos^2 \phi (d\lambda^2 + dq^2) \quad (55.108)$$

So, we have

$$dq = \frac{M}{N \cos \phi} d\phi_{\text{rad}} \quad (55.109)$$

Upon integration of Eq. (55.109) we obtain the isometric latitude q :

$$q = \ln \left[\tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right) \left(\frac{1 - e \sin \phi}{1 + e \sin \phi} \right)^{e/2} \right] \quad (55.110)$$

The isometric latitude for a sphere ($e = 0$) becomes simply

$$q = \ln \tan \left(\frac{\pi}{4} + \frac{\psi}{2} \right) \quad (55.111)$$

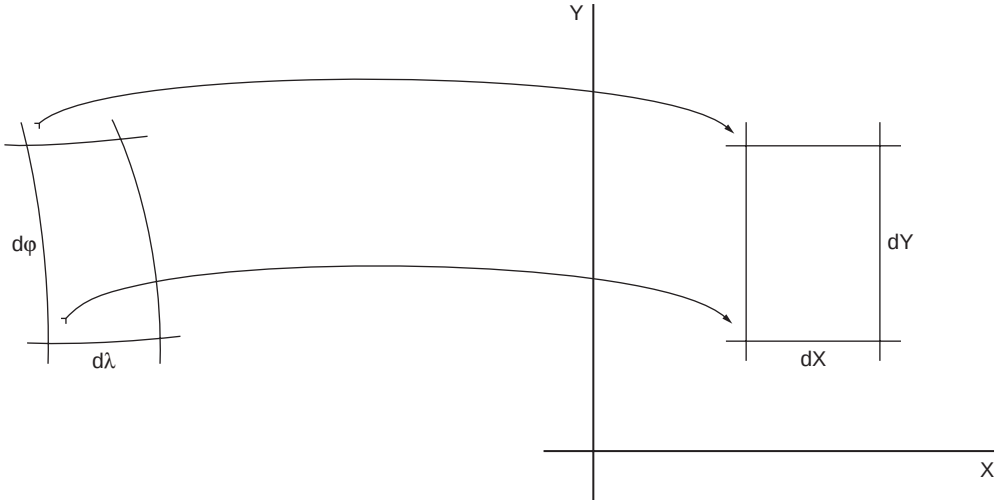


FIGURE 55.13 Spherical/ellipsoidal quadrangle mapped onto a planar Cartesian quadrangle.

Equating the variables from Eqs. (55.105) and (55.108) we have the mapping M between Cartesian mapping coordinates $\{X, Y\}$ and curvilinear coordinates $\{\lambda, q\}$ for both the sphere and the ellipsoid. Note that $\{\lambda, q\}$ play the role of the isometric coordinates $\{p, q\}$ in the Cauchy–Riemann equations.

So the mapping equations are simply

$$\begin{aligned} X &= \sigma \lambda \\ Y &= \sigma q \end{aligned} \quad (55.112)$$

or, for the sphere,

$$\begin{aligned} X &= \sigma \lambda \\ Y &= \sigma \ln \tan \left(\frac{\pi}{4} + \frac{\psi}{2} \right) \end{aligned} \quad (55.113)$$

Equation (55.113) represents the conformal mapping equations from the sphere to R_2 . These are the formulas of the well-known Mercator projection (cylindrical type). Inspection of the linear scale σ reveals that this factor depends on the term $M/(N \cos \phi)$ for the ellipsoid or $1/\cos \phi$ for the sphere. This means that the linear distortion is only latitude dependent. In order to minimize this distortion, we simply apply this mapping to regions that are elongated in the longitudinal direction, where the linear distortion is constant. In case we want to map an arbitrarily oriented elongated region, we simply apply a coordinate transformation. In the subsection on coordinate transformations and conformal mapping we will perform such a coordinate transformation on these mapping equations.

So far, we have mapped a spherical quadrangle $d\lambda, d\psi$ or ellipsoidal “quadrangle” $d\lambda, d\phi$ to a planar quadrangle dX, dY ; see Fig. 55.13.

Conformal Mapping Using Polar Differential Coordinates

The alternative is to map an ellipsoidal or spherical quadrangle onto a polar quadrangle. Figure 55.14 shows that one option is to look for a mapping between the polar mapping coordinates $\{r, \alpha\}$ and the real-world coordinates $\{\lambda, \phi\}$ or $\{\lambda, \psi\}$.

The similarity of roles played by the radius r and the latitude ϕ or ψ is more apparent if we view the *colatitude* θ , since this real-world coordinate radiates from one point as the radius r does:



Considering now the line element ds on a sphere, we have

or

Introducing the new variable dq' , we have

Integration of Eq. (55.117) gives the isometric colatitude q' :

or

A line element on the sphere in terms of isometric coordinates is

A line element dS (small distance) in the mapped world (on paper) is

© 2003 by CRC Press LLC

Now we want to derive isometric coordinates in the mapped world as well (we do not have a Cartesian, but a polar, representation). Among various options we choose

$$dS^2 = r^2 \left(d\alpha^2 + \frac{dr^2}{r^2} \right) \quad (55.122)$$

Introducing the new variable $d\rho$,

$$d\rho = \frac{dr}{r} \quad (55.123)$$

Upon integration of Eq. (55.123), we obtain the isometric radius ρ :

$$\rho = \int d\rho = \int \frac{dr}{r} = \ln r + c_r \quad (55.124)$$

or

$$r = e^\rho \quad r^2 = e^{2\rho} \quad (55.125)$$

The line element dS becomes with the new variable

$$dS^2 = e^{2\rho} (d\alpha^2 + d\rho^2) \quad (55.126)$$

The line element ds was

$$ds^2 = R^2 \sin^2 \theta (d\lambda^2 + dq'^2) \quad (55.127)$$

So the mapping equations are simply

$$\begin{aligned} \alpha &= \sigma \lambda \\ \rho &= \sigma q' \end{aligned} \quad (55.128)$$

or

$$\begin{aligned} \alpha &= \sigma \lambda \\ \ln r &= \sigma \left[-\ln \left(\cot \frac{\theta}{2} \right) \right] \end{aligned} \quad (55.129)$$

If one appropriately chooses the integration constant c_r , Eq. (55.126), to be

$$c_r = \ln 2 \quad (55.130)$$

the mapping Eq. (55.129) becomes

$$\begin{aligned} \alpha &= \sigma \lambda \\ \ln r &= \sigma \left[\ln 2 - \ln \left(\cot \frac{\theta}{2} \right) \right] = \ln \left(2 \tan \frac{\theta}{2} \right) \end{aligned} \quad (55.131)$$

or

$$\begin{aligned} \alpha &= \sigma \lambda \\ r &= 2\sigma \tan \frac{\theta}{2} \end{aligned} \quad (55.132)$$

Equation (55.132) also represents conformal mapping equations from the sphere to R_2 . They are the formulas of the well-known stereographic projection (planar type). Other choices of integration constants and integration interval would have led to the Lambert conformal projection (conical type).

The approaches laid out in this subsection and the preceding one are the theoretical basis of the U.S. State Plane Coordinate Systems. Refer to Stem [1991] for the formulas of the ellipsoidal equivalents.

Coordinate Transformations and Conformal Mapping

The two examples treated in the preceding two subsections can be treated for any arbitrary curvilinear coordinates. The widely used Transverse Mercator projection for the sphere is easily derived using a simple coordinate transformation. Rather than having the origin of the (co)latitude variable at the pole, we define a similar pole at the equator, and the new equator will be perpendicular to the old equator.

Having mapping poles at the equator leads to transverse types of conformal mapping. If the pole is neither at the North Pole nor at the equator, we obtain oblique variants of conformal mapping.

For the Transverse Mercator, the new equator may pass through a certain (old) longitude λ_0 . For a UTM projection this λ_0 has specified values; for a state plane coordinate system the longitude λ may define the central meridian in a particular (part of the) state.

Two successive rotations will bring the old $\mathbf{x}$ frame to the new $\mathbf{x}'$ frame (see Section 55.2):

$$\mathbf{x}' = \mathbf{R}_1(\pi/2)\mathbf{R}_3(\lambda_0)\mathbf{x} \quad (55.133)$$

The original $\mathbf{x}$ frame expressed in curvilinear coordinates is

$$\mathbf{x} = R \begin{pmatrix} \cos \psi \cos \lambda \\ \cos \psi \sin \lambda \\ \sin \psi \end{pmatrix} \quad (55.134)$$

The new $\mathbf{x}'$ -frame expressed in curvilinear coordinates is

$$\mathbf{x}' = R \begin{pmatrix} \cos \psi' \cos \lambda' \\ \cos \psi' \sin \lambda' \\ \sin \psi' \end{pmatrix} \quad (55.135)$$

Multiplying out the rotations in Eq. (55.135) we get

$$\mathbf{x}' = R \begin{pmatrix} \cos \psi' \cos \lambda' \\ \cos \psi' \sin \lambda' \\ \sin \psi' \end{pmatrix} = R \begin{pmatrix} x \cos \lambda_0 + y \sin \lambda_0 \\ z \\ x \sin \lambda_0 + y \cos \lambda_0 \end{pmatrix} \quad (55.136)$$

Substituting Eq. (55.134) into Eq. (55.136) and dividing by R we obtain

$$\begin{pmatrix} \cos \psi' \cos \lambda' \\ \cos \psi' \sin \lambda' \\ \sin \psi' \end{pmatrix} = \begin{pmatrix} \cos \psi \cos \lambda \cos \lambda_0 + \cos \psi \sin \lambda \sin \lambda_0 \\ \sin \psi \\ \cos \psi \cos \lambda \sin \lambda_0 - \cos \psi \sin \lambda \cos \lambda_0 \end{pmatrix} \quad (55.137)$$

which directly leads to

$$\begin{aligned}\tan \lambda' &= \frac{\sin \psi}{\cos \psi \cos(\lambda - \lambda_0)} = \frac{\tan \psi}{\cos(\lambda - \lambda_0)} \\ \tan \psi' &= \frac{\cos \psi \sin(\lambda - \lambda_0)}{\sqrt{\cos^2 \psi \cos^2(\lambda - \lambda_0) + \sin^2 \psi}} = \frac{\sin(\lambda - \lambda_0)}{\sqrt{\cos^2(\lambda - \lambda_0) + \tan^2 \psi}}\end{aligned}\quad (55.138)$$

The new longitudes λ' and latitudes ψ' are subjected to a (normal) Mercator projection according to

$$\begin{aligned}X' &= \sigma \lambda' \\ Y' &= \sigma q' = \sigma \ln \left[\tan \left(\frac{\pi}{4} + \frac{\psi'}{2} \right) \right]\end{aligned}\quad (55.139)$$

The Transverse Mercator projection with respect to the central meridian λ_0 is obtained by a simple rotation, about -90° . The final mapping equations are

$$\begin{pmatrix} X \\ Y \\ 0 \end{pmatrix} = \mathbf{R}_3 \left(-\frac{\pi}{2} \right) \begin{pmatrix} X' \\ Y' \\ 0 \end{pmatrix} = \begin{pmatrix} -Y' \\ X' \\ 0 \end{pmatrix}\quad (55.140)$$

When we start with ellipsoidal curvilinear coordinates, we cannot apply this procedure directly. However, when we follow a two-step procedure — mapping the ellipsoid conformal to the sphere, and then using the “rotated conformal” mapping procedure as just described — the treatise in this section will have a more general validity.

For more details on conformal projections using ellipsoidal coordinates, consult Bugayevskiy and Snyder [1998], Maling [1993], Stem [1991], and others.

55.5 Basic Concepts in Mechanics

Equations of Motion of a Point Mass in an Inertial Frame

To understand the motion of a satellite around the earth, we resort to two fundamental laws of physics: Isaac Newton’s second law (the law of inertia) and Newton’s law of gravitation.

Law of Inertia

The second law of Newton (the law of inertia) is as follows:

$$\mathbf{F} = m\mathbf{a}\quad (55.141)$$

The mass of a point mass m is the constant ratio that experimentally exists between the force $\mathbf{F}$, acting on that point mass, and the acceleration that is the result of that force.

The acceleration $\mathbf{a}$, the velocity $\mathbf{v}$, and the distance $\mathbf{s}$ are related as follows:

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2\mathbf{s}}{dt^2}\quad (55.142)$$

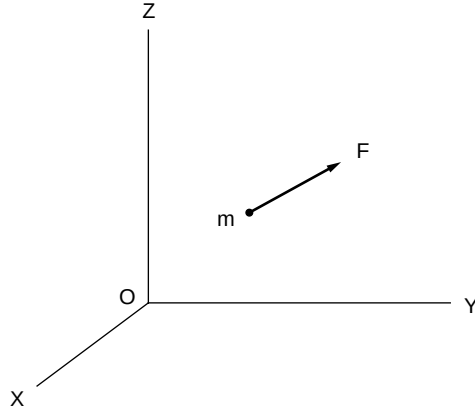


FIGURE 55.15 The acceleration of a point mass m .

Equation (55.141) can be written in vector form (see Fig. 55.15):

$$\mathbf{F} = m\mathbf{a} \quad \text{or} \quad \begin{pmatrix} F_x \\ F_y \\ F_z \end{pmatrix} = m \begin{pmatrix} \mathbf{a}_x \\ \mathbf{a}_y \\ \mathbf{a}_z \end{pmatrix} = m \begin{pmatrix} \ddot{X} \\ \ddot{Y} \\ \ddot{Z} \end{pmatrix} = m\ddot{\mathbf{X}} \quad (55.143)$$

with

$$\ddot{X} = \frac{d^2 X}{dt^2} \quad \text{and similarly for } \ddot{Y} \text{ and } \ddot{Z} \quad (55.144)$$

Law of Gravitation

Until now we did not mention the cause of the force $\mathbf{F}$ acting on point mass m . If this force is caused by the presence of a second point mass, then Newton's law of gravitation says

$$F = G \frac{m_1 m_2}{|\mathbf{X}_{12}|^2} \quad (55.145)$$

Two point masses m_1 and m_2 attract each other with a force that is proportional to the masses of each point mass and inversely proportional to the square of the distance between them, $|\mathbf{X}_{12}|$.

$$G = 6.67259 \pm 0.00085 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2} \quad (55.146)$$

is the gravitation constant (e.g., Cohen and Taylor [1988]), and

$$|\mathbf{X}_{12}|^2 = (X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2 \quad (55.147)$$

The force $\mathbf{F}$ will be written in vector form (see Fig. 55.16):

$$\mathbf{F}_{12} = \begin{pmatrix} F_{12X} \\ F_{12Y} \\ F_{12Z} \end{pmatrix} = |\mathbf{F}_{12}| \begin{pmatrix} \sin \alpha \\ \sin \beta \\ \sin \gamma \end{pmatrix} = |\mathbf{F}_{12}| \begin{pmatrix} X_{12}/|\mathbf{X}_{12}| \\ Y_{12}/|\mathbf{X}_{12}| \\ Z_{12}/|\mathbf{X}_{12}| \end{pmatrix} \quad (55.148)$$

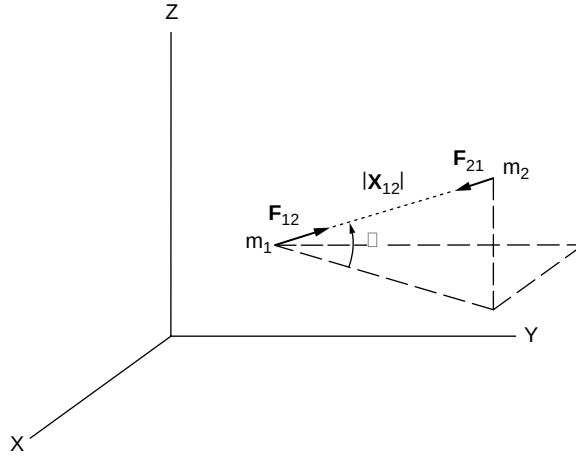


FIGURE 55.16 The attracting force $\mathbf{F}_{12}$ between two point masses m_1 and m_2 .

with

$$X_{12} = X_2 - X_1 \quad \text{and so forth} \quad (55.149)$$

Substituting Eq. (55.148) into Eq. (55.145),

$$\mathbf{F}_{12} = \frac{Gm_1m_2}{|\mathbf{X}_{12}|^2} \begin{pmatrix} X_{12}/|\mathbf{X}_{12}| \\ Y_{12}/|\mathbf{X}_{12}| \\ Z_{12}/|\mathbf{X}_{12}| \end{pmatrix} = \frac{Gm_1m_2}{|\mathbf{X}_{12}|^3} \begin{pmatrix} X_{12} \\ Y_{12} \\ Z_{12} \end{pmatrix} \quad (55.150)$$

or

$$\mathbf{F}_{12} = \frac{Gm_1m_2}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.151)$$

Equations (55.143) and (55.151) applied subsequently to two point masses m_1 and m_2 yield:

For m_1 ,

$$\mathbf{F}_1 = m_1 \ddot{\mathbf{X}}_1 = \mathbf{F}_{12} = \frac{Gm_1m_2}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.152)$$

or

$$\ddot{\mathbf{X}}_1 = \frac{Gm_2}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.153)$$

For m_2 ,

$$\mathbf{F}_2 = m_2 \ddot{\mathbf{X}}_2 = \mathbf{F}_{21} = -\mathbf{F}_{12} = \frac{-Gm_1m_2}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.154)$$

or

$$\ddot{\mathbf{X}}_2 = \frac{-Gm_1}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.155)$$

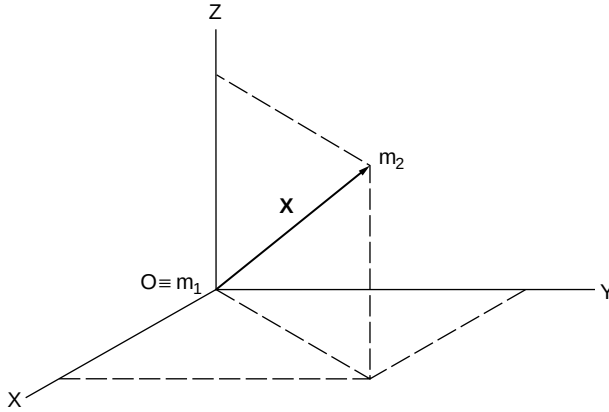


FIGURE 55.17 The inertial frame centered in m_1 .

By subtracting Eq. (55.153) from Eq. (55.155) we get

$$\ddot{\mathbf{X}}_{12} = \ddot{\mathbf{X}}_2 - \ddot{\mathbf{X}}_1 = \frac{-G(m_1 + m_2)}{|\mathbf{X}_{12}|^3} \mathbf{X}_{12} \quad (55.156)$$

If m_1 resides in the origin of the inertial frame, then the subindices may be omitted:

$$\ddot{\mathbf{X}} = \frac{-G(m_1 + m_2)}{|\mathbf{X}|^3} \mathbf{X} \quad (55.157)$$

Equation (55.157) represents the equations of motion of m_2 in an inertial frame centered at m_1 ; see Fig. 55.17.

It will be clear that in m_1 the origin of an inertial frame can be defined if and only if $m_1 + m_2$. For the equations of motion of a satellite $m = m_2$ orbiting the earth $M = m_1$, Eq. (55.157) simplifies to

$$\ddot{\mathbf{X}} = \frac{-GM}{|\mathbf{X}|^3} \mathbf{X} \quad (55.158)$$

In Eq. (55.158) the product of the gravitational constant G and the mass of the earth M appears. For this product, the *geocentric gravitational constant*, a new equation is introduced:

$$\mu = GM \quad (55.159)$$

Potential

The equations of motion around a mass M ,

$$\ddot{\mathbf{X}} = \frac{-\mu}{|\mathbf{X}|^3} \mathbf{X} \quad (55.160)$$

can be obtained through the definition of a scalar function V :

$$V = V(X, Y, Z) = \frac{\mu}{|\mathbf{X}|} = \frac{\mu}{(X^2 + Y^2 + Z^2)^{1/2}} \quad (55.161)$$

The partial derivatives of V with respect to $\mathbf{X}$ are

$$\frac{\partial V}{\partial X} = -\frac{\mu}{|\mathbf{X}|^2} \frac{\partial |\mathbf{X}|}{\partial X} = -\frac{\mu}{|\mathbf{X}|^3} X \quad (55.162)$$

$$\frac{\partial V}{\partial Y} = -\frac{\mu}{|\mathbf{X}|^2} \frac{\partial |\mathbf{X}|}{\partial Y} = -\frac{\mu}{|\mathbf{X}|^3} Y \quad (55.163)$$

$$\frac{\partial V}{\partial Z} = -\frac{\mu}{|\mathbf{X}|^2} \frac{\partial |\mathbf{X}|}{\partial Z} = -\frac{\mu}{|\mathbf{X}|^3} Z \quad (55.164)$$

Consequently, the equations of motion may be written as

$$\ddot{\mathbf{X}} = \begin{pmatrix} \frac{\partial V}{\partial X} \\ \frac{\partial V}{\partial Y} \\ \frac{\partial V}{\partial Z} \end{pmatrix} \equiv \text{grad } V \equiv \nabla V \quad (55.165)$$

V has physical significance: it is the potential of a point mass of negligible mass in a gravity field of a point mass with sizable mass M at a distance $|\mathbf{X}|$.

55.6 Satellite Surveying

The Global Positioning System has become a tool used in a variety of fields within and outside engineering.

Positioning has become possible with accuracies from the subcentimeter level for high-accuracy geodetic applications — as used in state, national, and global geodetic networks and for deformation analysis in engineering and geophysics — and to the hectometer level in navigation applications. As in the space domain, a variety of accuracy classes may be assigned to the time domain: GPS provides a means to obtain position and velocity determinations averaged over time spans from subseconds (instantaneous) to 1 or 2 days. Stationary applications of the observatory type are used in GPS tracking for orbit improvement.

First the physics and mathematics of the space segment will be given (without derivations).

Numerical Solution of Three Second-Order Differential Equations

Equation (55.158) represents the equations of motion of a satellite expressed in vector form. Written in the three Cartesian components, we have

$$\begin{pmatrix} \frac{d^2 X}{dt^2} \\ \frac{d^2 Y}{dt^2} \\ \frac{d^2 Z}{dt^2} \end{pmatrix} = \frac{-GM}{(X^2 + Y^2 + Z^2)^{3/2}} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (55.166)$$

or

$$\begin{aligned}\ddot{X} &= C \left[X / (X^2 + Y^2 + Z^2)^{3/2} \right] \\ \ddot{Y} &= C \left[Y / (X^2 + Y^2 + Z^2)^{3/2} \right] \\ \ddot{Z} &= C \left[Z / (X^2 + Y^2 + Z^2)^{3/2} \right]\end{aligned}\tag{55.167}$$

with $C = GM$. Rather than solving for three second-order differential equations (DEs), we make a transformation to six first-order DEs. Introduce three new variables U, V, W :

$$U = \dot{X} = \frac{dX}{dt}; \quad V = \dot{Y} = \frac{dY}{dt}; \quad W = \dot{Z} = \frac{dZ}{dt}\tag{55.168}$$

Equations (55.162) through (55.165) yield the following six DEs:

$$\begin{aligned}U &= \dot{X} \\ V &= \dot{Y} \\ W &= \dot{Z} \\ \dot{U} &= C \left[X / (X^2 + Y^2 + Z^2)^{3/2} \right] \\ \dot{V} &= C \left[Y / (X^2 + Y^2 + Z^2)^{3/2} \right] \\ \dot{W} &= C \left[Z / (X^2 + Y^2 + Z^2)^{3/2} \right]\end{aligned}\tag{55.169}$$

Integration of Eq. (55.169) results in six constants of integration. One is free to choose at an epoch t_0 six variables, $\{X_0, Y_0, Z_0, U_0, V_0, W_0\}$ or $\{X_0, Y_0, Z_0, \dot{X}_0, \dot{Y}_0, \dot{Z}_0\}$. These six starting values determine uniquely the orbit of m around M . In other words, if we know the position $\{X_0, Y_0, Z_0\}$ and its velocity $\{\dot{X}_0, \dot{Y}_0, \dot{Z}_0\}$ at an epoch t_0 , then we are able to determine the position and velocity of m at any other epoch t by numerical integration of Eq. (55.169).

Analytical Solution of Three Second-Order Differential Equations

The differential equations of an earth-orbiting satellite can also be solved analytically. Without derivation, the solution is presented in computational steps in terms of transformation formulas.

In history the solution to the motion of planets around the sun was found before its explanation. Through the analysis of his own observations and those made by Tycho Brahe, Johannes Kepler discovered certain regularities in the motions of planets around the sun and formulated the following three laws:

1. Formulated 1609: The orbit of each planet around the sun is an ellipse. The sun is in one of the two focal points.
2. Formulated 1609: The sun–planet line sweeps out equal areas in equal time periods.
3. Formulated 1611: The ratio between the square of a planet's orbital period and the third power of its average distance from the sun is constant.

Kepler's third law leads to the famous equation

$$n^2 a^3 = GM\tag{55.170}$$

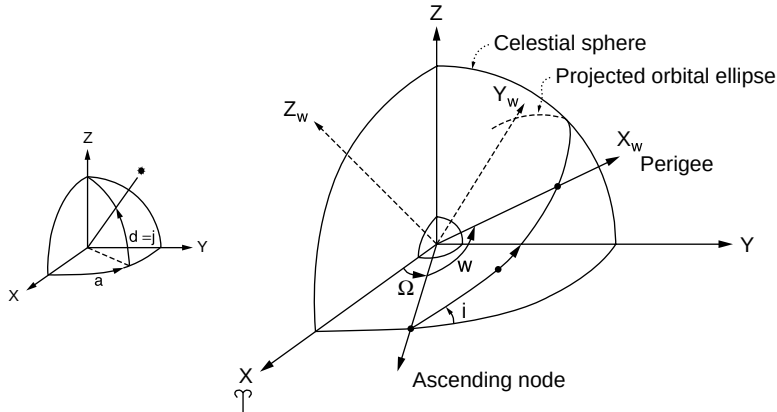


FIGURE 55.18 Celestial sphere with projected orbit ellipse and equator.

in which n is the average angular rate and a the semimajor axis of the orbital ellipse.

In 1665–1666 Newton formulated his more fundamental laws of nature (which were only published after 1687) and showed that Kepler's laws follow from them.

Orientation of the Orbital Ellipse

In a (quasi-)inertial frame the ellipse of an earth-orbiting satellite has to be positioned: the focal point will coincide with the center of mass of the earth. Instead of picturing the ellipse itself we project the ellipse on a celestial sphere centered at the CoM. On the celestial sphere we also project the earth's equator (see Fig. 55.18).

The orientation of the orbit ellipse requires three orientation angles with respect to the inertial frame XYZ : two for the orientation of the plane of the orbit (Ω and I) and one for the orientation of the ellipse in the orbital plane in terms of the point of closest approach, the perigee (ω).

Ω represents the right ascension (α) of the ascending node. The ascending node is the (projected) point where the satellite rises above the equator plane. I represents the inclination of the orbital plane with respect to the equator plane. ω represents the *argument of perigee*: the angle from the ascending node (in the plane of the orbit) to the perigee (for planets, the perihelion), which is that point where the satellite (planet) approaches closest to the earth (sun) or, more precisely, the CoM of the earth (sun).

We define now another reference frame X_ω , of which the $X_\omega Y_\omega$ plane coincides with the orbit plane. The X_ω axis points to the perigee, and the origin coincides with the earth's CoM ($\equiv$ focal point ellipse $\equiv$ origin of X frame). The relationship between the inertial frames X_I and X_ω is

$$X_I = R_{I?} X_\omega \quad (55.171)$$

in which

$$R_{I?} = R_3(-\Omega) R_1(-I) R_3(-\omega) \quad (55.172)$$

and

$$X_\omega = \begin{pmatrix} X_\omega \\ Y_\omega \\ Z_\omega = 0 \end{pmatrix} \quad (55.173)$$

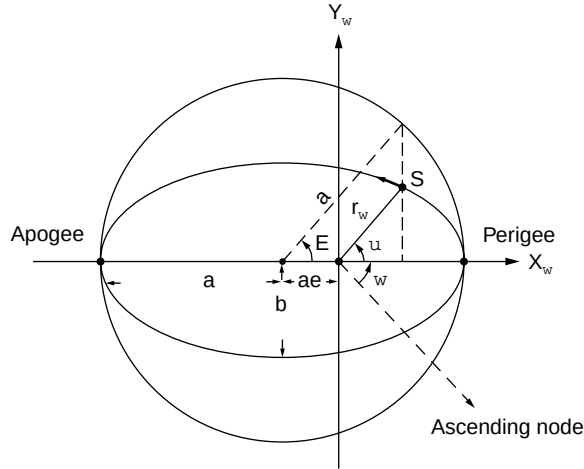


FIGURE 55.19 The position of the satellite *S* in the orbital plane.

Reference Frame in the Plane of the Orbit

Now that we know the orientation of the orbital ellipse we have to define the size and shape of the ellipse and the position of the satellite along the ellipse at a certain epoch t_0 .

Similarly to the earth's ellipsoid, discussed in Section 55.2, we define the ellipse by a semimajor axis a and eccentricity e . In orbital mechanics it is unusual to describe the shape of the orbital ellipse by its flattening.

The position of the satellite in the orbital $X_w Y_w$ plane is depicted in Fig. 55.19. In the figure the auxiliary circle enclosing the orbital ellipse reveals the following relationships:

$$X_w = r_w \cos v = a(\cos E - e) \quad (55.174)$$

$$Y_w = r_w \sin v = a\sqrt{1-e^2} \sin E \quad (55.175)$$

in which

$$r_w = a(1 - e \cos E) \quad (55.176)$$

In Fig. 55.19 and the Eqs. (55.174) through (55.176), a is the semimajor axis of the orbital ellipse, b is the semiminor axis of the orbital ellipse, e is the eccentricity of the orbital ellipse, with

$$e^2 = \frac{a^2 - b^2}{a^2} \quad (55.177)$$

v is the true anomaly, sometimes denoted by an f , and E is the eccentric anomaly.

The relation between the true and the eccentric anomalies can be derived to be

$$\tan\left(\frac{E}{2}\right) = \sqrt{\frac{1-e}{1+e}} \tan\left(\frac{v}{2}\right) \quad (55.178)$$

Substitution of Eqs. (55.174) through (55.176) into (55.171) gives

$$\mathbf{X}_I = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \mathbf{R}_3(-\Omega)\mathbf{R}_1(-I)\mathbf{R}_3(-\omega) \begin{pmatrix} a(\cos E - e) \\ a\sqrt{1-e^2} \sin E \\ 0 \end{pmatrix} \quad (55.179)$$

In Eq. (55.179) the Cartesian coordinates are expressed in the six so-called Keplerian elements: a , e , I , Ω , ω , and E . Paraphrasing an earlier remark: if we know the position of the satellite at an epoch t_0 through $\{a, e, I, \Omega, \omega, \text{ and } E_0\}$, we are capable of computing the position of the satellite at an arbitrary epoch t through the Eq. (55.179) if we know the relationship in time between E and E_0 . In other words, how does the angle E increase with time?

We define an auxiliary variable (angle) M , which increases linearly in time with the mean motion $n (= (GM/a^3)^{1/2})$ according to Kepler's third law. The angle M , the *mean anomaly*, may be expressed as function of time by

$$M = M_0 + n(t - t_0) \quad (55.180)$$

Through Kepler's equation

$$M = E - e \sin E \quad (55.181)$$

the (time) relationship between M and E is given. Kepler's equation is the direct result of the enforcement of Kepler's second law (equal area law).

Combining Eqs. (55.180) and (55.181) gives an equation that expresses the relationship between a given eccentric anomaly E_0 (or M_0 or v_0) at an epoch t_0 and the eccentric anomaly E at an arbitrary epoch t :

$$E - E_0 = e(\sin E - \sin E_0) + n(t - t_0) \quad (55.182)$$

Transformation from Keplerian to Cartesian Orbital Elements

So far, the position vector $\{X, Y, Z\}$ of the satellite has been expressed in terms of the Keplerian elements. The transformation is complete when we express the velocity vector $\{\dot{X}_0, \dot{Y}_0, \dot{Z}_0\}$ in terms of those Keplerian elements. Differentiating Eq. (55.171) with respect to time we get

$$\dot{\mathbf{X}}_I = \mathbf{R}_{I?} \dot{\mathbf{X}}_? + \dot{\mathbf{R}}_{I?} \mathbf{X}_? \quad (55.183)$$

Since we consider the two-body problem with $m_1 = M$? $m = m_2$, the orientation of the orbital ellipse is time independent in the inertial frame. This means that the orientation angles I , Ω , ω are time independent as well:

$$\dot{\mathbf{R}}_{I\omega} = [0] \quad (55.184)$$

Equation (55.183) simplifies to

$$\dot{\mathbf{X}}_I = \mathbf{R}_{I?} \dot{\mathbf{X}}_? \quad (55.185)$$

Differentiating Eqs. (55.174) through (55.176) with respect to time we have

$$\dot{X}_\omega = -a\dot{E} \sin E \quad (55.186)$$

$$\dot{Y}_\omega = -a\dot{E}\sqrt{1-e^2}\cos E \quad (55.187)$$

$$\dot{r}_\omega = a\dot{E}e\sin E \quad (55.188)$$

The remaining variable $\dot{E}$ is obtained through differentiation of Eq. (55.181):

$$\dot{E} = \frac{n}{1-e\cos E} \quad (55.189)$$

Now all transformation formulas express the Cartesian orbital elements (state vector elements) in terms of the six Keplerian elements:

$$[\mathbf{X}_I \cdot \dot{\mathbf{X}}_I] = \mathbf{R}_{I\omega} [\mathbf{X}_\omega \cdot \dot{\mathbf{X}}_\omega] \quad (55.190)$$

or

$$[\mathbf{X}_I \cdot \dot{\mathbf{X}}_I] = \begin{bmatrix} X & \cdot & \dot{X} \\ Y & \cdot & \dot{Y} \\ Z & \cdot & \dot{Z} \end{bmatrix} = \quad (55.191)$$

$$= \mathbf{R}_3(-\Omega)\mathbf{R}_1(-I)\mathbf{R}_3(-\omega) \begin{bmatrix} a(\cos E - e) & \cdot & -a\dot{E}\sin E \\ a\sqrt{1-e^2}\sin E & \cdot & a\dot{E}\sqrt{1-e^2}\cos E \\ 0 & \cdot & 0 \end{bmatrix} \quad (55.192)$$

Transformation from Cartesian to Keplerian Orbital Elements

To compute the inertial position of a satellite in a central force field, it is simpler to perform a time update in the Keplerian elements than in the Cartesian elements. The time update takes place in Eqs. (55.180) through (55.182).

Schematically the following procedure is to be followed:

$$\begin{array}{ll} t_0 : \{X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}\} & \\ \downarrow & \text{Conversion to Keplerian elements, this subsection} \\ t_0 : \{a, e, I, \Omega, \omega, E_0\} & \\ \downarrow & \text{Conversion to Keplerian, Eq. (55.182)} \\ t_1 : \{a, e, I, \Omega, \omega, E_1\} & \\ \downarrow & \text{Conversion to Cartesian elements, previous subsection} \\ t_1 : \{X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}\} & \end{array}$$

The conversion from Keplerian elements to state vector elements has been treated in the previous subsection. In this section the somewhat more complicated conversion from position and velocity vector to Keplerian representation will be described. Basically we “invert” Eq. (55.192) by solving for the six elements $\{a, e, I, \Omega, \omega, E_0\}$ in terms of the six state vector elements.

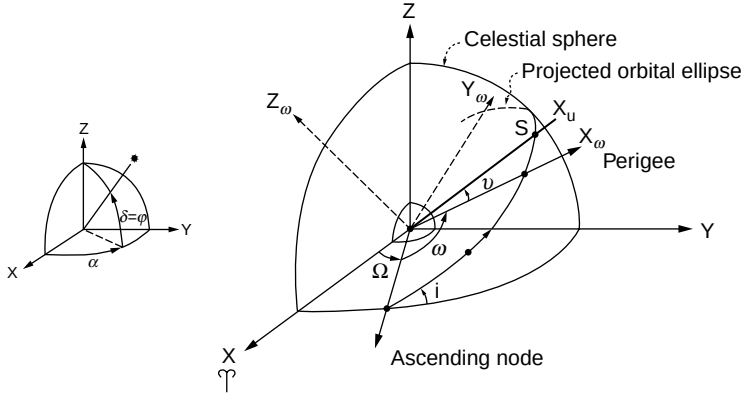


FIGURE 55.20 The orbital reference frame X_u .

First, we introduce another reference frame X_u :

X_I, Y_I, Z_I : inertial reference frame (X_I axis $\rightarrow$ vernal equinox)

$X_\omega, Y_\omega, Z_\omega$: orbital reference frame (X_ω axis $\rightarrow$ perigee)

X_u, Y_u, Z_u : orbital reference frame (X_u axis $\rightarrow$ satellite)

The X_u frame is defined similarly to the X_ω frame, except that the X_u axis continuously points to the satellites (see Fig. 55.20). Thus,

$$\mathbf{X}_u = \begin{pmatrix} \mathbf{X}_u \\ \mathbf{Y}_u \\ \mathbf{Z}_u \end{pmatrix} = \begin{pmatrix} \mathbf{X}_\omega \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad (55.193)$$

The angle in the orbital plane enclosed by the X_u axis and the direction to the ascending node is called u , the *argument of latitude*, with

$$u = \omega + \nu \quad (55.194)$$

As in the discussion of the orientation of the orbit ellipse, the following relationships hold:

$$\mathbf{X}_\omega = \mathbf{R}_{\omega I} \mathbf{X}_I \quad (55.195)$$

$$\mathbf{X}_u = \mathbf{R}_{uI} \mathbf{X}_I \quad (55.196)$$

$$\mathbf{X}_u = \mathbf{R}_{uI} \mathbf{X}_I = \mathbf{R}_{uI} \mathbf{R}_{\omega I} \mathbf{X}_I \quad (55.197)$$

Figure 55.20 reveals that

$$\mathbf{R}_{\omega I} = \mathbf{R}_3(\omega) \mathbf{R}_I(I) \mathbf{R}_3(\Omega) \quad (55.198)$$

$$\mathbf{R}_{uI} = \mathbf{R}_3(u) \mathbf{R}_I(I) \mathbf{R}_3(\Omega) \quad (55.199)$$

$$= \mathbf{R}_3(\nu + \omega) \mathbf{R}_I(I) \mathbf{R}_3(\Omega) \quad (55.200)$$

$$= \mathbf{R}_3(\nu) \mathbf{R}_3(\omega) \mathbf{R}_I(I) \mathbf{R}_3(\Omega) \quad (55.201)$$

$$= \mathbf{R}_3(\nu) \mathbf{R}_{\omega I} = \mathbf{R}_{u\omega} \mathbf{R}_{\omega I} \quad (55.202)$$

We define a vector $\mathbf{h}$ perpendicular to the orbital plane according to

$$\mathbf{h} \equiv \mathbf{X} \times \dot{\mathbf{X}} = |\mathbf{h}| \mathbf{w} = |\mathbf{h}| \begin{pmatrix} \sin \Omega \sin I \\ -\cos \Omega \sin I \\ \cos I \end{pmatrix} \quad (55.203)$$

in which $\mathbf{w}$ is the unit vector along $\mathbf{h}$. Consequently,

$$\mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} = \begin{pmatrix} Y\dot{Z} - Z\dot{Y} \\ Z\dot{X} - X\dot{Z} \\ X\dot{Y} - Y\dot{X} \end{pmatrix} \quad (55.204)$$

$\mathbf{h}$ represents the angular momentum vector (vector product of position vector and velocity vector). The Keplerian elements Ω and I follow directly from Eqs. (55.203) and (55.204):

$$\tan \Omega = \frac{h_1}{-h_2} \quad (55.205)$$

$$\tan I = \frac{\sqrt{h_1^2 + h_2^2}}{h_3} \quad (55.206)$$

From

$$\mathbf{R}_3(-u) \mathbf{X}_u = \mathbf{R}_I(I) \mathbf{R}_3(\Omega) \mathbf{X}_I \quad (55.207)$$

it follows that

$$X_u \cos u = X \cos \Omega + Y \sin \Omega \quad (55.208)$$

$$X_u \sin u = -X \cos I \sin \Omega + Y \cos I \cos \Omega + Z \sin I \quad (55.209)$$

and

$$\tan u = \frac{-X \cos I \sin \Omega + Y \cos I \cos \Omega + Z \sin I}{X \cos \Omega + Y \sin \Omega} \quad (55.210)$$

Before determining the third Keplerian element defining the orientation of the orbit (ω) from the argument of latitude (u), we define the following quantities:

- Length r of the radius vector $\mathbf{X}$:

$$r \equiv |\mathbf{X}| = (X^2 + Y^2 + Z^2)^{1/2} \quad (55.211)$$

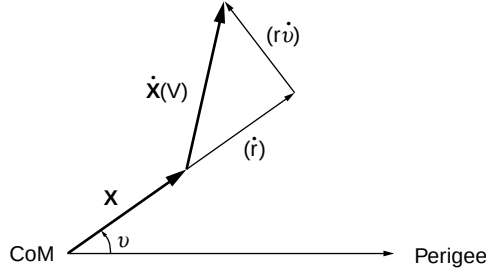


FIGURE 55.21 Illustration of the Vis-Viva equation.

- Length $\dot{r}$ of the radial component of the velocity vector:

$$\dot{r} \equiv |\mathbf{X} \cdot \dot{\mathbf{X}}|/r = |X \dot{X} + Y \dot{Y} + Z \dot{Z}|/r \quad (55.212)$$

The *Vis-Viva equation* gives the relationship between the length (magnitude) V of the velocity vector $\dot{\mathbf{X}}$ and the length r of the position vector $\mathbf{X}$ through the radial component $(\dot{r})$ and the tangential component $(r\dot{\nu})$ (see Fig. 55.21):

$$V^2 \equiv |\dot{\mathbf{X}}|^2 = \dot{r}^2 + (r\dot{\nu})^2 = \dot{r}^2 + |\mathbf{h}|^2/r^2 = GM \left(\frac{2}{r} - \frac{1}{a} \right) \quad (55.213)$$

From Eq. (55.213) it follows that

$$a = \frac{|\mathbf{X}|GM}{2GM - |\mathbf{X}|V^2} = \frac{rGM}{2GM - rV^2} \quad (55.214)$$

In a similar manner one arrives at

$$1 - e^2 = \frac{|\mathbf{h}|^2}{aGM} \quad (55.215)$$

From Eqs. (55.188), (55.189), (55.176), and (55.170) the eccentric anomaly E may be computed:

$$\tan E = \frac{\sin E}{\cos E} = \left(\frac{r\dot{r}}{e\sqrt{GMa}} \right) / \left(\frac{a-r}{ae} \right) \quad (55.216)$$

The mean anomaly M is given by

$$M = E - e \sin E \quad (55.217)$$

The true anomaly ν follows from Eqs. (55.174) and (55.175):

$$\tan \nu = \frac{\sqrt{1-e^2} \sin E}{\cos E - e} \quad (55.218)$$

after which, finally, the last Keplerian element, ω , is determined through Eq. (55.194):

$$\omega = u - \nu \quad (55.219)$$

Orbit of a Satellite in a Noncentral Force Field

The equations of motion for a real satellite are more difficult than reflected by Eq. (55.165). First of all, we do not deal with a central force field: the earth is not a sphere, and it does not have a radial symmetric density. Secondly, we deal with other forces, chiefly the gravity of the moon and the sun, atmospheric drag, and solar radiation pressure. Equation (55.165) gets a more general meaning if we suppose that the potential function is generated by the sum of the forces acting on the satellite:

$$V = V_c + V_{nc}^t + V_{sun}^t + V_{moon}^t + \dots \quad (55.220)$$

with the central part of the earth's gravitational potential

$$V_c = \mu / |\mathbf{X}| \quad (55.221)$$

and the noncentral and time-dependent part of the earth's gravitational field

$$V_{nc}^t = [\text{see Eq. (52.224)}] \quad (55.222)$$

and so forth.

The superscript t has been added to various potentials to reflect their time variance with respect to the inertial frame.

The equations of motion to be solved are

$$\begin{aligned} \ddot{\mathbf{X}} &= \nabla(V_c + V_{nc}^t + V_{sun}^t + V_{moon}^t + \dots) \\ &= \nabla V_c + \nabla V_{nc}^t + \nabla V_{sun}^t + \nabla V_{moon}^t + \dots \end{aligned} \quad (55.223)$$

For the earth's gravitational field we have (in an earth-fixed frame)

$$V_c + V_{nc} = \frac{\mu}{r} \left[1 + \sum_{l=1}^{\infty} \sum_{m=0}^l \left(\frac{a_e}{r} \right)^l (C_{lm} \cos m\lambda + S_{lm} \sin m\lambda) P_{lm}(\sin \phi) \right] \quad (55.224)$$

With Eq. (55.224) one is able to compute the potential at each point $\{\lambda, \phi, r\}$ necessary for the integration of the satellite's orbit. The coefficients C_{lm} and S_{lm} of the spherical harmonic expansion are in the order of 10^{-6} , except for C_{20} ($l = 2, m = 0$), which is about 10^{-3} . This has to do with the fact that the earth's equipotential surface at mean sea level can be best approximated by an ellipsoid of revolution. One has to realize that the coefficients C_{lm}, S_{lm} describe the shape of the potential field and not the shape of the physical earth, despite a high correlation between the two. $P_{lm} \sin \phi$ are the associated Legendre functions of the first kind, of degree l and order m ; a_e is some adopted value for the semimajor axis (equatorial radius) of the earth. See Section 55.8 for values of $a_e, \mu (= GM)$, and $C_{20} (= -J_2)$.

The equatorial radius a_e , the geocentric gravitational constant GM , and the dynamic form factor J_2 characterize the earth by an ellipsoid of revolution of which the surface is an equipotential surface.

Restricting ourselves to the central part ($\mu = GM$) and the dynamic flattening ($C_{20} = -J_2$), Eq. (55.224) becomes

$$V_c + V_{nc} = \frac{\mu}{r} \left[1 + \frac{J_2 a_e^2}{2r^2} (1 - 3 \sin^2 \phi) \right] \quad (55.225)$$

with

$$\sin \phi = (\sin \delta) \frac{z}{r} \quad (55.226)$$

in which ϕ is the latitude and δ the declination; see [Fig. 55.20](#).

A solution of the DE (by substitution of Eq. (55.225) in Eq. (55.165))

$$\ddot{\mathbf{X}} = \nabla(V_c + V_{20}) \quad (55.227)$$

in a closed analytical expression is not possible. The solution expressed in Keplerian elements shows periodic perturbations and some dominant secular effects. An approximate solution using only the latter effects is (position only)

$$\mathbf{X}_I = \mathbf{R}_3[-(\Omega_0 + \dot{\Omega}\Delta t)]\mathbf{R}_1(-I)\mathbf{R}_3[-(\omega_0 + \dot{\omega}\Delta t)]\mathbf{X}_\omega \quad (55.228)$$

with

$$\Delta t = t - t_0 \quad (55.229)$$

with

$$\dot{\Omega} = -\frac{3}{2} \frac{J_2 a_e^2}{a^2 (1-e^2)^2} n \cos I \quad (55.230)$$

$$\dot{\omega} = \frac{3}{2} \frac{J_2 a_e^2}{a^2 (1-e^2)^2} n \left(2 - \frac{5}{2} \sin^2 I \right) \quad (55.231)$$

$$n = n_0 \left[1 + \frac{3}{2} \frac{J_2 a_e^2 \sqrt{1-e^2}}{a^2 (1-e^2)^2} \left(1 - \frac{3}{2} \sin^2 I \right) \right] \quad (55.232)$$

$$n_0 = \sqrt{\frac{GM}{a^3}} \quad (55.233)$$

whenever we have:

$I = 0^\circ$	equatorial orbit
0°	$< I < 90^\circ$ direct orbit
$I = 90^\circ$	polar orbit
$90^\circ < I < 180^\circ$	retrograde orbit
$I = 180^\circ$	retrograde equatorial orbit

Equation (55.230) shows that the ascending node of a direct orbit slowly drifts to the west. For a satellite at a height of about 150 km above the earth's surface the right ascension of the ascending node decreases about 9° per day. For satellites used in geodesy and geodynamics, such as STARLETTE ($a = 7340$ km, $I = 50^\circ$) and LAGEOS-I ($a = 12,270$ km, $I = 110^\circ$), these values are -4° per day and $+1/3^\circ$ per day, respectively.

The satellites belonging to the Global Positioning System have an inclination of about 55° . Their nodal regression rate is about -0.04187° per day.

The Global Positioning System

Introduction

The Navstar GPS space segment consists of 24 satellites, plus a few spare ones. This means that the full satellite constellation, in six orbital planes at a height of about 20,000 km, is completed. With this number of satellites, three-dimensional positioning is possible every hour of the day. However, care must be exercised, since an optimum configuration for three-dimensional positioning is not available on a full day's basis.

In the meantime, GPS receivers, ranging in cost between \$300 and \$20,000, are readily available. Over 100 manufacturers are marketing receivers, and the prices are still dropping. Magazines such as *G.I.M.*,

GPS World, P.O.B., and Professional Surveyor publish regularly on the latest models; see, for example, the recent GPS equipment surveys in GPS World [2002].

GPS consumer markets have been rapidly expanded. In the areas of land, marine, and aviation navigation, of precise surveying, of electronic charting, and of time transfer the deployment of GPS equipment seems to have become indispensable. This holds for military as well as civilian users.

Positioning

Two classes of positioning are recognized; standard positioning service (SPS) and precise positioning service (PPS):

In terms of positional accuracies one has to distinguish between SPS, with and without selective availability (SA), on the one hand and PPS on the other. Selective availability deliberately introduces clock errors and ephemeris errors in the data broadcast by the satellite. On May 1, 2000, selective availability was suspended. The current positional accuracy using the (civil) signal without SA is in the order of 10 m or better. With SA implemented, SPS accuracy is degraded to 50 to 100 m.

In several applications, GPS receivers are interfaced with other positioning systems, such as inertial navigation systems (INSs), hyperbolic systems, or even automatic braking systems (ABSs) in cars.

GPS receivers in combination with various equipment are able to answer such general questions as [Wells and Kleusberg, 1990]

- Absolute positioning: Where am I?
Where are you?
- Relative positioning: Where am I with respect to you?
Where are you with respect to me?
- Orientation: Which way am I heading?
- Timing: What time is it?

All of these questions may refer to either an observer at rest (*static* positioning) or one in motion (*kinematic* positioning). The questions may be answered immediately (*real-time* processing, often mis-named *DGPS*, for differential GPS) or after the fact (*batch* processing). The various options are summarized in Table 55.3 [Wells and Kleusberg, 1990]. Similarly, the accuracies for time dissemination are summarized in Table 55.4 [Wells and Kleusberg, 1990].

TABLE 55.3 Accuracy of Various GPS Positioning Modes

Absolute Positioning:	
SPS with SA	100 m
SPS without SA	40 m
PPS	20 m
Relative/Differential Positioning:	
Differential SPS	10 m
Carrier-smoothed code	2 m
Ambiguity-resolved carrier	10 cm
Surveying between fixed points	1 mm to 10 cm

Note: Accuracy of differential models depends on inter-receiver distance.

TABLE 55.4 Accuracy of Various GPS Time Dissemination Modes

Time and Time Interval:	
With SA	500 ns
Without SA, correct position	100 ns
Common mode, common view	<25 ns

Limiting Factors

Physics of the environment, instruments, broadcast ephemeris, and the relative geometry between orbits and network all form limiting factors on the final accuracy of the results. *Dilution of precision* (DOP) is used as a scaling factor between the observational accuracy and positioning accuracy.

The atmosphere of the earth changes the speed and the geometrical path of the electromagnetic signals broadcast by the GPS satellites. In the uppermost part of the atmosphere (the ionosphere) charged particles vary in number spatially as well as temporally. The so-called ionospheric refraction errors may amount to several tens of meters. Since this effect is frequency dependent, the first-order effect can be largely eliminated by the use of dual-frequency receivers. The lower part of the atmosphere (the troposphere) causes refraction errors of several meters. Fortunately, the effect can be modeled rather well by measuring the atmospheric conditions at the measuring site.

GPS instruments are capable of measuring one or a combination of the following signals:

- C/A code: with an accuracy of a few meters
- P code: with an accuracy of a few decimeters
- Carrier phase: with an accuracy of a few millimeters

In addition to this measurement noise, receiver clock errors are present and have to be modeled as to-be-solved-for parameters. It is this synchronization parameter between satellite time and receiver time that makes it necessary to have at least four satellites in view in order to get a three-dimensional fix.

Because of the high frequency of the GPS signals, multipath effects may hamper the final accuracy; the signal arriving at the receiver through a reflected path may be stronger than the direct signal. By careful antenna design and positioning of the antenna, multipath effects are reduced. The phase center of the antenna needs to be carefully calibrated with respect to a geometric reference point on the antenna assembly. However, because of the varying inclination angle of the incoming electromagnetic signals, effects of a moving phase center may be present at all times.

Information on the orbit of the satellite, as well as the orbital geometry relative to the network and receiver geometry, influences the overall positioning accuracy. The information the satellite broadcasts on its position and velocity is necessarily the result of a process of prediction. This causes the broadcast ephemeris to be contaminated with extrapolation errors. Typical values are

- Radial error: about 5 m
- Across-track error: about 10 m
- Along-track error: about 15 m

Also, the onboard satellite clock is not free of errors. Orbital and satellite clock errors can be largely taken care of by careful design of the functional model.

As mentioned before, deliberate contamination of the broadcast ephemeris and satellite time degrades the system to an accuracy of several tens of meters. PPS users are able to use the P code on two frequencies and have access to the SA code. Consequently, they are capable of eliminating the ionospheric effects and of removing deliberately introduced orbital and satellite clock errors. The resulting measurement error will be about 5 m. SPS users are able to use the C/A code (on one frequency only). They do not have access to the SA code. Consequently, the ionospheric error can only be roughly modeled, and these users are stuck with the deliberate errors. The resulting measurement error may be as large as 50 m.

Translocation techniques, such as having a stationary receiver continuously supporting the other (roving) receivers, will reduce the measurement error to well below the 10 m for SPS users.

Differencing techniques applied to the carrier phase measurements are successfully used to eliminate a wide variety of errors, provided the receivers are not too far apart. In essence, two close-by receivers are influenced almost equally by (deliberate) orbital errors and by part of the atmosphere error. Differencing of the measurements of both receivers will cancel a large portion of the first-order effects of these errors.

Modeling and the GPS Observables

Developing well-chosen functional models $\mathbf{F}$, relating the GPS measurements $\mathbf{L}$ to the modeled parameters $\mathbf{X}$, enables users to fit GPS perfectly to their needs. A wide class of applications, from monitoring the subsidence of oil rigs in the open sea to real-time navigating vehicles collecting geoinformation, belong to the range of possibilities that have been opened up by the introduction of GPS.

In satellite geodesy, one traditionally modeled the state of the satellite: a vector combining the positional ($\mathbf{X}$) and velocity ($\dot{\mathbf{X}}$) information. Although a known fact in the area of navigation, nowadays GPS provides geodesists, or geoscientists in general, with a tool by which the state of the observer, also in terms of position ($\mathbf{x}$) and velocity ($\dot{\mathbf{x}}$) can be determined with high accuracy and often in real time.

The GPS satellite geodetic model has evolved to

$$\mathbf{L} = \mathbf{F}(\mathbf{X}, \dot{\mathbf{X}}, \mathbf{x}, \dot{\mathbf{x}}, \mathbf{p}, t) \quad (55.234)$$

where L = the C/A code, P code, or carrier phase observations
 $i = 1, \dots, n$
 $\mathbf{X}$ = the three-dimensional position of the satellite at epoch t
 $\dot{\mathbf{X}}$ = the three-dimensional velocity of the satellite at epoch t
 $\mathbf{x}$ = the one-dimensional, two-dimensional, or three-dimensional position of the observer at epoch t
 $\dot{\mathbf{x}}$ = the one-dimensional, two-dimensional, or three-dimensional velocity of the observer at epoch t
 $\mathbf{p}$ = the vector of modeled (known or unknown) parameters
 $j = 1, \dots, u$
 t = the epoch of measurement taking

Various differencing operators $\mathbf{D}^k$, up to order 3, are applied to the original observations in order to take full advantage of the GPS measurements. The difference operator $\mathbf{D}^k$ may be applied in the observation space spanned by the vector $\mathbf{L}$, Eq. (55.235), or the $\mathbf{D}^k$ operator may be applied in the parameter space $\mathbf{x}$, Eq. (55.236):

$$\mathbf{D}^k[\mathbf{L}] = \mathbf{D}^k[\mathbf{F}(\mathbf{X}, \dot{\mathbf{X}}, \mathbf{x}, \dot{\mathbf{x}}, \mathbf{p}, t)] \quad (55.235)$$

$$\mathbf{L} = \mathbf{F}[\mathbf{X}, \dot{\mathbf{X}}, \mathbf{D}^1(\mathbf{x}, \dot{\mathbf{x}}), \mathbf{p}, t] \quad (55.236)$$

The latter method is sometimes referred to as *delta positioning*.

This is a difficult way of saying that one may either construct so-called derived observations from the original observations by differencing techniques or model the original observations, compute parameters (e.g., coordinates) in this way, and subsequently start a differencing technique on the results obtained from the roving receiver and the base receiver.

Pseudorangeing

We restrict the discussion to the C/A-based *pseudorange observables*. The ranges are called *pseudo* because this technique is basically a one-way ranging technique with two independent clocks: the offset δt_E^S between the satellite clock S and the receiver clock E , which yields one additional parameter to solve for. Writing the observation equation in the earth-fixed reference frame, we have

$$\text{pr} = \sqrt{(x^S - x_E)^2 + (y^S - y_E)^2 + (z^S - z_E)^2} - c\delta t_E^S \quad (55.237)$$

Inspection of the partials

$$\frac{\partial \text{pr}}{\partial x^S} = \frac{x^S - x_E}{\text{pr}} = -\frac{\partial \text{pr}}{\partial x_E} \quad (55.238)$$

$$\frac{\partial \text{pr}}{\partial y^s} = \frac{y^s - y_E}{\text{pr}} = -\frac{\partial \text{pr}}{\partial y_E} \quad (55.239)$$

$$\frac{\partial \text{pr}}{\partial z^s} = \frac{z^s - z_E}{\text{pr}} = -\frac{\partial \text{pr}}{\partial z_E} \quad (55.240)$$

$$\frac{\partial \text{pr}}{\partial (\delta t_E^s)} = -c \quad (55.241)$$

reveals that:

- The coordinates of the stations are primarily obtained in a frame determined by the satellites or, better, by their broadcast ephemeris.
- Partials evaluated for neighboring stations are practically identical, so the coordinates of one station need to be adopted.

Phase (Carrier Wave) Differencing

For precise engineering applications the phase of the carrier wave is measured. Two wavelengths are available in principle:

$$\begin{aligned} L_1: \quad \lambda_1 &= \frac{c}{f_1} \quad \text{with} \quad f_1 = 1.57542 \text{ GHz} \\ &= 19.0 \text{ cm} \end{aligned} \quad (55.242)$$

and

$$\begin{aligned} L_2: \quad \lambda_2 &= \frac{c}{f_2} \quad \text{with} \quad f_2 = 1.22760 \text{ GHz} \\ &= 24.4 \text{ cm} \end{aligned} \quad (55.243)$$

For phase measurements the following observation equation can be set up:

$$\text{Range} = \Phi + N \lambda_l; \quad l = 1, \dots, 2 \quad (55.244)$$

or

$$\Phi_E^s = \sqrt{(x^s - x_E)^2 + (y^s - y_E)^2 + (z^s - z_E)^2} - N_E^s \lambda_l \quad (55.245)$$

where Φ_E^s is the phase observable in a particular S - E combination and N_E^s is the integer multiple of wavelengths in the range: the *ambiguity*. Phase measurements can be done with probably 1% accuracy. This yields an observational accuracy — in case the ambiguity N can be properly determined — in the millimeter range.

Using the various differencing operators on the phase measurements:

- $\mathbf{D}^k=1$ yields single differences:
 - Between receiver differences, $\Delta\Phi$, eliminating or reducing satellite-related errors
 - Between satellite differences, $\nabla\Phi$ eliminating or reducing receiver-related errors
 - Between epoch differences, $\delta\Phi$, eliminating phase ambiguities per satellite and receiver combination

- $D^k=2$ yields double differences:
 - Between receiver and satellite differences, $\nabla\Delta\Phi$ eliminating or reducing satellite- and receiver-related errors, and so forth
- $D^k=3$ yields triple differences:
 - Between epoch, receiver, and satellite differences, $\delta\nabla\Delta\Phi$ eliminating or reducing satellite- and receiver-related errors and ambiguities

Receivers that use carrier wave observations have, in addition to the electronic components that do the phase measurements, a counter, which counts the complete cycles between selected epochs. GPS analysis software uses the triple differences to detect and possibly repair cycle slips occurring during loss of lock.

Design specifications and receiver selection depend on the specific project accuracy requirements. In the United States the Federal Geodetic Control Committee has adopted the specifications [FGCC, 1989] given in [Tables 55.5](#) and [55.6](#).

GPS Receivers

A variety of receivers are on the market. Basically, they can be grouped in four classes, listed in [Table 55.7](#). The types of observations of the first three types of receivers are subjected to models that can be characterized as *geometric* models. The position of the satellite is considered to be known, mostly the information taken from the broadcast ephemeris. The known positions are of course not errorless. First of all, the positions are predicted; thus they contain errors because of an extrapolation process in time. Second, the positions being broadcast may be corrupted by intentional errors (due to SA, see previous paragraphs). Differencing techniques are capable of eliminating most of the error if the separation between base station and roving receiver is not too large.

Millimeter-accurate observations from geodesy-grade receivers are often subjected to analysis through models of the *dynamic* type. Software packages containing dynamic models are very elaborate and allow for some kind of orbit improvement estimation process.

GPS World [2002] lists an overview of recent GPS receivers.

GPS Base Station

GPS, like most other classical survey techniques, has to be applied in a differential mode if one wants to obtain reliable relative positional information. This implies that for most applications of GPS in geodesy — surveying and mapping, photogrammetry, Geographic Information Systems (GIS), and so forth — one has to have at least two GPS receivers at one's disposal. If one of the receivers occupies a known location during an acceptable minimum period, than one may obtain accurate coordinates for the second receiver *in the same frame*. In surveying and geodesy applications one should preferably include three stations with known horizontal coordinates and at least four with known vertical (orthometric) heights. In most GIS applications one receiver is left at one particular site. This station serves as a so-called *base station*.

GIS, Heights, and High-Accuracy Reference Networks

In order to reduce influences from satellite-related errors and atmospheric conditions in geodesy, surveying, and Geographic Information Systems applications, GPS receivers are operated in a differential mode. Whenever the roving receiver is not too far from the base station receiver, “errors at high altitudes (satellite and atmosphere)” are more or less canceled if the “fix from the field” is differenced with the “fix from the base.”

Washington editor of *GPS World*, Hale Montgomery, writes, “As a peripheral industry, the reference station business has grown almost into an embarrassment of riches, with stations proliferating nationwide and sometimes duplicating services.” William Strange, chief geodesist at the National Geodetic Survey (NGS), is quoted as saying, “Only about 25 full-service, fixed stations would be needed to cover the entire United States” [Montgomery, 1993]. A group of the interagency Federal Geodetic Control subcommittee has compiled a list of about 90 operating base stations on a more or less permanent basis. If all GIS and GPS base stations being planned or in operation are included, the feared proliferation will be even larger.

From the point of view of the U.S. taxpaying citizen, “duplication of services” of work may be wasteful, although decentralization of services may often be more cost-effective than all-encompassing projects run by even more all-encompassing agencies. At the time of this writing (January 2002) the website of the National Geodetic Survey (<http://www.ngs.noaa.gov/>) lists a network of almost 300 continuous operating reference stations (CORS) in the U.S.

From the geodetic point of view, the duplication of services (base station-generated fixes in the field) will proliferate the coordinate fields and the reference frames they supposedly are tied to. The loss of money and effort in the years to come while trying to make sense out of these most likely not-matching point fields may be far larger than the money lost in “service duplicating” base stations. If we are not careful, the “coordinate duplicating” base stations will create chaos among GIS-applying agencies.

Everyone is convinced about the necessity to collect GIS data in one *common frame*. Formerly, the GIS community was satisfied with positions of the 3- to 5-m accuracy. Manufacturers are now aggressively marketing GPS and GIS equipment with 0.5-m accuracy (\$10,000 per receiver, and remember, you need at least two). The increased demand for accuracy requires that a reference frame be in place that lasts at least two decades. This calls for a consistent reference of one, probably two orders of magnitude more accurate than presently available.

The accuracy of the classical horizontal control was in the order of 1 part in 100,000 (1 cm over 1 km). GPS is a survey tool with an accuracy of 1 part per 1,000,000 (1 cm over 10 km). All U.S. states put new High-Accuracy Reference Networks (HARNs) in place to accommodate the accuracy of GPS surveys. Even for GIS applications where 0.5-m accuracies are claimed for the roving receivers, one may speak of 1-ppm surveys whenever those rovers operate at a distance of 500 km from their base station.

It should not be forgotten that GPS is a geometric survey tool yielding results in terms of earth-fixed coordinate differences x_{ij} . From these coordinate differences expressed in curvilinear coordinates we obtain, at best, somewhat reproducible ellipsoidal height differences. These height differences are *not* easily converted to orthometric height differences of equal accuracy. The latter height differences are of interest in engineering and GIS applications; see Section 55.7.

55.7 Gravity Field and Related Issues

One-Dimensional Positioning: Heights and Vertical Control

One of the most accurate measurements surveyors are able to make are the determinations of height differences by spirit leveling. Since a leveling instrument’s line of sight is tangent to the potential surface, one may say that leveling actually determines the height differences with respect to equipotential surfaces. If one singles out one particular equipotential surface at mean sea level (the so-called *geoid*), then the heights a surveyor determines are actually orthometric heights; see Fig. 55.22. Leveling in a closed loop is not a check on the actual height differences in a metrical sense, but in a potential sense: the distance between equipotential surfaces varies due to local gravity variations.

In spherical approximation the potential at a point A is

$$V = -\frac{GM}{r} = -\frac{GM}{R+h} \quad (55.246)$$

The gravity is locally dependent on the change in potential per height unit, or

$$\frac{dV}{dr} = g = \frac{GM}{r^2} \quad (55.247)$$

The potential dV difference between two equipotential surfaces is

$$dV = g \, dr \quad (55.248)$$

Consequently, if one levels in a loop, one has

$$\sum dV = \sum g dr = 0 \quad (55.249)$$

$$\oint dV = \oint g dr \quad (55.250)$$

This implies that for each metrically leveled height difference dr , one has to multiply this difference by the local gravity. Depending on the behavior of the potential surfaces in a certain area and the diameter of one's project, one has to "carry along a gravimeter" while leveling. The variations of local gravity vary depending on the geology of the area. Variations in the order of $10^{-7} g$ may yield errors as large as 10 mm for height differences in the order of several hundred meters. For precise leveling surveys (≤ 0.1 mm/km) gravity observations must be made with an interval of:

- 2 to 3 km in relatively flat areas
- 1 to 2 km in hilly terrain
- 0.5 to 1.5 km in mountainous regions

For more design criteria on leveling and gravity surveys, see FGCC [1989] and [Table 55.8](#).

GPS surveys yield, at best, ellipsoidal height differences. These are rather meaningless from the engineering point of view. Therefore, extreme caution should be exercised when GPS height information, even after correction for geoidal undulations, is to be merged with height information from leveling. For two different points i and j ,

$$h_i = H_i + N_i \quad (55.251)$$

$$h_j = H_j + N_j \quad (55.252)$$

Subtracting Eq. (55.251) from Eq. (55.252), we find the ellipsoidal height differences h_{ij} (from GPS) in terms of the orthometric height differences H_{ij} (from leveling) and the geoidal height differences N_{ij} (from gravity surveys):

$$h_{ij} = H_{ij} + N_{ij} \quad (55.253)$$

where

$$\begin{aligned} h_{ij} &= h_j - h_i \\ H_{ij} &= H_j - H_i \\ N_{ij} &= N_j - N_i \end{aligned} \quad (55.254)$$

For instance, with the National Geodetic Survey's software program geoid99 package, geoidal height differences are as accurate as 5 cm over 100 km for the conterminous U.S. For GPS leveling this means that GPS competes with third-order leveling as long as the stations are more than 5 km apart.

In principle any equipotential surface can act as a vertical datum such as the National Geodetic Vertical Datum of 1929 (NGVD29). Problems may arise merging GPS heights, gravity surveys, and orthometric heights referring to NGVD29. Heights referring to the NGVD88 datum will be more suitable for use with GPS surveys. In the United States about 600,000 vertical control stations are in existence.

Two-Dimensional Positioning: East–North and Horizontal Control

In classical geodesy the measurements in height had to be separated (leveling) from the horizontal measurements (directions, angles, azimuths, and distances). To allow for the curvature of the earth and the varying gravity field, the horizontal observations were reduced first to the geoid, taking into account the orthometric heights. Subsequently, if it was desired to take advantage of geometrical properties between the once-reduced horizontal observations, the observations had to be reduced once more, but now from the geoid to the ellipsoid. An ellipsoid approximates the geoid up to 0.01%; the variations of the geoid are nowhere larger than 150 m. On the ellipsoid, which is a precise mathematical figure, one could check, for instance, whether the sum of the three angles equaled a prescribed value.

So far, geodesists have relied on a biaxial ellipsoid of revolution. A semimajor axis a and a semiminor axis b define the dimensions of the ellipsoid. Rather than this semiminor axis, one specifies the *flattening* of the ellipsoid

$$f = \frac{a-b}{a} \approx \frac{1}{298.257} \quad (55.255)$$

For a semimajor axis of about 6378.137 km, it implies that the semiminor axis is $6378.137/298.257 \approx 22$ km shorter than a .

Distance measurements need to be reduced to the ellipsoid. Angular measurements made with theodolites, total stations, or other instruments need to be corrected for several effects:

- The direction of local gravity does not coincide with the normal to the ellipsoid. The direction of the first axis of the instrument coincides with the direction of the local gravity vector. Notwithstanding this effect, the earth's curvature causes nonparallelism of first axes of 1 arcsecond for each 30 m.
- The targets aimed at generally do not reside on the ellipsoid.

The noncoincidence of the gravity vector and the normal is called *deflection of the vertical*. Proper knowledge of the behavior of the local geopotential surfaces is needed for proper distance and angle reductions. Consult Vaniček and Krakiwsky [1982], for example, for the mathematical background of these reductions. The FGCC adopted accuracy standards for horizontal control using classical geodetic measurement techniques; see [Table 55.9](#). In the U.S. over 270,000 horizontal control stations exist.

Three-Dimensional Positioning: Geocentric Positions and Full Three-dimensional Control

Modern three-dimensional survey techniques, most noticeably GPS, allow for immediate three-dimensional relative positioning. Three-dimensional coordinates are equally accurately expressed in ellipsoidal, spherical, or Cartesian coordinates. Care should be exercised in properly labeling curvilinear coordinates as spherical (*geographic*) or ellipsoidal (*geodetic*). [Table 55.10](#) shows the large discrepancies between the two. At the midlatitudes they may differ by more than 11'. This could result in a north–south error of 20 km. While merging GIS data sets, one should be aware of the meaning LAT/LON in any instance.

Despite their three-dimensional characteristics, networks generated by GPS are the weakest in the vertical component, not only because of the lack of physical significance of GPS determined heights, as described in the preceding subsection, but also because of the geometrical distribution of satellites with respect to the vertical: no satellite signals are received from “below the network.” This lopsidedness makes the vertical the worst determined component in three-dimensional characteristics.

Because of the inclination of the GPS satellites, there are places on earth, most notoriously the midlatitudes, where there is not an even distribution of satellites in the azimuth sense. For instance, in the northern midlatitudes we never have as many satellites to the north as we have to the south; see, for

example, Santerre [1991]. This makes the latitude the second-best-determined curvilinear coordinate. For space techniques the FGCC has proposed the classification in [Table 55.11](#).

55.8 Reference Systems and Datum Transformations

Geodetic Reference Frames

There has always been a necessity, in a variety of fields but especially in engineering, to rely on a set of adopted parameter values that describe various quantities related to the earth to the current state of the art of accuracy. Questions asked of geodesists include:

- What is the best estimate of the equatorial radius of the earth?
- What is the mass of the earth?
- How fast does the earth rotate?
- How flattened is the earth?

The values adopted at the General Assemblies of the International Association of Geodesy (IAG) of Madrid (1924) and Stockholm (1930) were valid for a long time:

$$a = 6,378.388 \text{ km} \quad \text{and} \quad f = 1/297.0$$

This was the so-called international ellipsoid of the American geodesist J.F. Hayford (1909). The appropriate international gravity formula was

$$\gamma = 978.0490(1 + 0.005\,288\,4 \sin^2 \phi - 0.000\,0059 \sin 2\phi) \text{ cm sec}^{-2} \text{ (gal)}$$

This formula for γ describes the gravity on the surface of the international ellipsoid as function of the geodetic (ellipsoidal) latitude c . After some years these values turned out to be not accurate enough, and the Soviet geodesist F.N. Krassovsky proposed a new ellipsoid (1943):

$$a = 6378.245 \text{ km} \quad \text{and} \quad f = 1/298.3$$

Geodetic Reference System 1967

In Hamburg (1964) the International Astronomical Union (IAU), after the advice of the International Union of Geodesy and Geophysics (IUGG), adopted values of three variables:

- a the equatorial radius of the earth
- GM the geocentric gravitational constant of the earth, including the atmosphere
- J_2 the dynamical form factor of the earth

The General Assembly of the IAG in Lucerne (1967) followed this proposal, resulting in the so-called Geodetic Reference System (GRS) 1967:

$$a = 6,378,160 \text{ m}$$

$$GM = 398,603 \times 10^9 \text{ m}^3 \text{ sec}^{-2}$$

$$J_2 = 10,827 \times 10^{-7}$$

It was agreed upon that the semiminor axis of the reference ellipsoid would be parallel to the direction of the rotation axis as defined by the conventional international origin. In addition, the reference meridian would be parallel to the zero meridian, as resulted from the adopted values for the longitudes of various observatories by the Bureau International de l'Heure (BIH) in Paris.

TABLE 55.5 Guidelines for GPS Field Survey Procedures

	AA	A	B	C
	AA	A	B	1, 2-I&II, 3
Geometric Relative Positioning Standards	0.01	0.1	1.0	10, 20, 50, 100
Two-frequency observations (I and L2) required ^a : daylight observations ^b	Y	Y	Y	op
Recommended number of receivers observing simultaneously, not less than	5	5	4	3
Satellite observations: GDOP values during observing session (m/cycle) ^d				
Period of observing session (observing span), not less than (min) (4 or more simultaneous satellite observations) ^c				
Triple difference processing ^f	na	na	240	60–120
Other processing techniques ^g				
General requirements ^h	240	240	120	30–60
Continuous and simultaneous between all receivers, period not less than ⁱ	180	120	60	20–30
Data sampling rate — maximum time interval between observations (sec)	15	30	30	15–30
Minimum number of quadrants from which satellite signals are observed	4	4	3	3 or 2 ^j
Maximum angle above horizon for obstructions ^u (degrees)	10	15	20	20–40
Independent occupations per station ^k				
Three or more (percent of all stations, not less than)	80	40	20	10
Two or more (percent of stations, not less than)				
New stations	100	80	50	30
Vertical control stations	100	100	100	100
Horizontal control stations	100	75	50	25
Two or more for each station of “station-pairs” ^l	Y	Y	Y	Y
Master or fiducial stations ^m				
Required, yes or no ⁿ	Y	Y	Y	op
If yes, minimum number	4	3	2	—
Repeat baseline measurements, about equal number in N–S and E–W directions, minimum not less than (percent of total independently (nontrivial) determined base lines)	25	15	5	5
Loop closure, requirements when forming loops for postanalyses				
Baselines from independent observing sessions, not less than	3	3	2	2
Baselines in each loop, total not more than	6	8	10	10
Loop length, generally not more than (km)	2000	300	100	100
Baselines not meeting criteria for inclusion in any loop, not more than (percent of all independent nontrivial lines ^o)	0	5	20	30
Stations not meeting criteria for inclusion in any loop, not more than (percent of all stations)	0	5	10	15
Direct connections are required; between any adjacent (NGRS or GPS) stations (new or old, GPS or non-GPS) located near or within project area, when spacing is less than (km)	30	10	5	3
Antenna setup				
Number of antenna phase center height measurements per session, not less than	3 ^p	3 ^p	2	2
Independent plumb point check required ^q	Y	Y	Y	op
Photograph (closeup) or pencil rubbing required for each mark occupied	Y	Y	Y	Y
Meteorological observations				
Per observing session, not less than	3 ^r	3 ^r	2 ^s	2 ^s or op
Sampling rate (measurement interval), not more than (min)	30	30	60	60
Water vapor radiometer measurements required at selected stations?	op	op	N	N
Frequency standard warm-up time (h) ^t				
Crystal	12	12	<i>u</i>	<i>u</i>
Atomic	—	1	<i>t</i>	<i>t</i>

Note: na = not applicable; op = optional.

^a If two-frequency observations cannot be obtained, it is possible that an alternative method for estimating the ionospheric refraction correction would be acceptable, such as modeling the ionosphere using two-frequency data obtained from other sources. Or, if observations are during darkness, single-frequency observations may be acceptable, depending on the expected magnitude of the ionospheric refraction error.

^b When spacing between any two stations occupied during an observing session is more than 50 km, two-frequency observations may need to be considered for accuracy standards of order 2 or higher.

^c Multiple baseline processing techniques.

TABLE 55.5 (continued) Guidelines for GPS Field Survey Procedures

- ^d Studies are under way to investigate the relationship of geometric dilution of precision (GDOP) values to the accuracy of the baseline determinations. Initial results of these studies indicate a possible correlation. It appears that the best results may be achieved when the GDOP values are changing during the observing session.
- ^e The number of satellites that are observed simultaneously cannot be less than the number specified for more than 25% of the specified period for each observing session.
- ^f Absolute minimum criterion is 100% of specified period.
- ^g "Other" includes processing carrier phase data using single, double, nondifferencing, or other comparable precise relative positioning processing techniques.
- ^h The times for the observing span are conservative estimates to ensure that the data quantity and quality will give results that will meet the desired accuracy standard.
- ⁱ Absolute minimum criterion for the data collection observing span is that period specified for an observing session that includes continuous and simultaneous observations. Continuous observations are data collected that do not have any breaks involving all satellites; occasional breaks for individual satellites caused by obstructions are acceptable, but must be minimized. A set of observations for each measurement epoch is considered simultaneous when it includes data from at least 75% of the receivers participating in the observing session.
- ^j Satellites should pass through quadrants diagonally opposite each other.
- ^k Two or more independent occupations for the stations of a network are specified to help detect instrument and operator errors. Operator errors include those caused by antenna centering and height offset blunders. When a station is occupied during two or more sessions, back-to-back, the antenna/tripod will be reset and replumbed between sessions to meet the criteria for an independent occupation. To separate biases caused by receiver and/or antenna equipment problems from operator-induced blunders, a calibration test may need to be performed.
- ^l Redundant occupations are required when pairs of intervisible stations are established to meet azimuth requirements, when the distance between the station pair is less than 2 km, and when the order is 2 or higher.
- ^m Master or fiducial stations are those that are continuously monitored during a sequence of sessions, perhaps for the complete project. These could be sites with permanently tracking equipment in operation where the data are available for use in processing with data collected with the mobile units.
- ⁿ If simultaneous observations are to be processed in the session or network for baseline determinations while adjusting one or more components of the orbit, then two or more master stations shall be established.
- ^o For each observing session there are $r - 1$ independent baselines, where r is the number of receivers collecting data simultaneously during a session; e.g., if there were 10 sessions and 4 receivers used in each session, 30 independent baselines would be observed.
- ^p A measurement will be made both in meters and feet, at the beginning, midpoint, and end of each station occupation.
- ^q To ensure that the antenna was centered accurately with the optical plummet over the reference point on the marker, when specified, a heavyweight plumb bob will be used to check that the plumb point is within specifications.
- ^r Measurements of station pressure (in millibars), relative humidity, and air temperature (in °C) will be recorded at the beginning, midpoint, and end, depending on the period of the observing session.
- ^s Report only unusual weather conditions, such as major storm fronts passing over the sites during the data collection period. This report will include station pressure, relative humidity, and air temperature.
- ^t The amount of warm-up time required is very instrument dependent. It is important to follow the manufacturer's specifications.
- ^u An obstruction is any object that would effectively block the signal arriving from the satellite. Included are buildings, trees, fences, humans, vehicles, etc.

Source: FGCC, *Geometric Geodetic Accuracy Standards and Specifications for Using GPS Relative Positioning Techniques*, Version 5.0, Federal Geodetic Control Committee, Rockville, MD, 1989.

A fourth parameter completed the Geodetic Reference System 1967: the (inertial or sidereal) angular velocity of the earth:

$$\omega = 0.000\,072\,921\,151\,467 \text{ rad/sec}$$

The duration of one turn of the earth around its spin axis is then

$$\text{Length of day} = \frac{2\pi}{\omega} \approx 86,164.1 \text{ sec} \quad (55.256)$$

This is not equal to $24 \times 60 \times 60 = 86,400$ sec, as explained in Section 55.3.

Officially, the XVth General Assembly of the IAG in Moscow (1971) approved and adopted these values [IAG, 1971].

TABLE 55.6 Office Procedures for Classifying GPS Relative Positioning Networks Independent of Connections to Existing Control

	AA	A	B	1	2-I	2-II	3
Geometric Relative Positioning Standards	0.01	0.1	1.0	10	20	50	100
Ephemerides							
Orbit accuracy, minimum (ppm)	0.008	0.05	0.5	5	10	25	50
Precise ephemerides required?	Y ^a	Y ^a	Y	op	op	N	N
Loop Closure Analyses^b — when forming loops, the following are minimum criteria:							
Baselines in loop from independent observations not less than	4	3	2	2	2	2	2
Baselines in each loop, total not more than	6	8	10	10	10	15	15
Loop length, not more than (km)	2000	300	100	100	100	100	100
Baselines not meeting criteria for inclusion in any loop, not more than (percent of all independent lines)	0	0	5	20	30	30	30
In any component (X, Y, Z), maximum misclosure not to exceed (cm)	10	10	15	25	30	50	100
In any component (X, Y, Z), maximum misclosure, in terms of loop length, not to exceed (ppm)	0.2	0.2	1.25	12.5	25	60	125
In any component (X, Y, Z), average misclosure, in terms of loop length, not to exceed (ppm)	0.09	0.09	0.9	8	16	40	80
Repeat Baseline Differences							
Baseline length, not more than (km)	2000	2000	500	250	250	100	50
In any component (X, Y, Z), maximum not to exceed (ppm)	0.01	0.1	1.0	10	20	50	100

Note: op = optional.

^a The precise ephemerides is presently limited to an accuracy of about 1 ppm. The accuracy has been improved to about 0.1 ppm. It is unlikely orbital coordinate accuracies of 0.01 ppm will be achieved in the near future. Thus to achieve precisions approaching 0.01 ppm, it will be necessary to collect data simultaneously with continuous trackers or fiducial stations. Then all the data are processed in a session or network solution mode where the initial orbital coordinates are adjusted while solving for the baselines. In this method of processing the carrier phase data, the coordinates at the continuous trackers, are held fixed.

^b Between any combination of stations, it must be possible to form a loop through three or more stations that never passes through the same station more than once.

Source: FGCC, *Geometric Geodetic Accuracy Standards and Specifications for Using GPS Relative Positioning Techniques*, Version 5.0, Federal Geodetic Control Committee, Rockville, MD, 1989.

TABLE 55.7 Accuracy Grades of (Civilian/Commercial) GPS Receivers

Grade	Accuracy	Code
Navigation grade	40–100 m	C/A code in stand-alone mode
Mapping (GIS) grade	2–5 m	C/A code in differenced mode
Surveying grade	1–2 cm (within 10 km)	C/A code plus phase, differenced
Geodesy grade	5–15 mm over any distance	C/A plus P codes plus phase, differenced

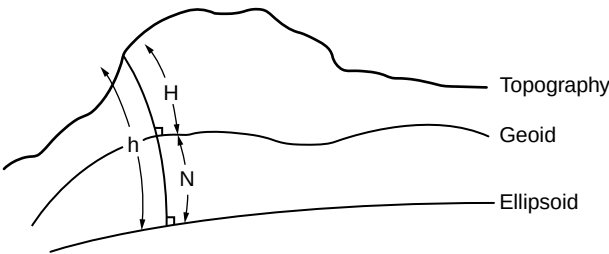


FIGURE 55.22 Orthometric heights.

TABLE 55.8 FGCC Vertical Control Accuracy Standards (Differential Leveling)

Class	b^a
First Order	
I	< 0.5
II	0.7
Second Order	
I	1.0
II	1.3
Third Order	2.0

^a $b = S/\sqrt{d}$ [mm km^{-1/2}], where S = standard deviation of elevation difference between control points (in mm) and d = approximate horizontal distance along leveled route (in km).

TABLE 55.9 FGCC Horizontal Control Accuracy Standards (Classical Techniques)

Class	Accuracy
First Order	1:100,000 (10 mm/km)
Second Order	
Class I	1:50,000 (20 mm/km)
Class II	1:20,000 (50 mm/km)
Third Order	
Class I	1:10,000 (100 mm/km)
Class II	1:5,000 (200 mm/km)

TABLE 55.10 Geographic (Spherical) Latitude as a Function of Geodetic Latitude

Geodetic Latitude			Geographic Latitude			Geodetic Minus Geographic Latitude		
°	'	"	°	'	"	°	'	"
00	0	0.000	00	00	00.000	00	00	00.000
10	0	0.000	09	56	03.819	00	03	56.181
20	0	0.000	19	52	35.868	00	07	24.132
30	0	0.000	29	50	01.089	00	09	58.911
40	0	0.000	39	48	38.198	00	11	21.802
50	0	0.000	49	48	37.402	00	11	22.598
60	0	0.000	59	49	59.074	00	10	00.926
70	0	0.000	69	52	33.576	00	07	26.424
80	0	0.000	79	56	02.324	00	03	57.676
90	0	0.000	90	00	00.000	00	00	00.000

TABLE 55.11 FGCC Three-Dimensional Accuracy Standards (Space System Techniques)

Order	Accuracy
AA order (global)	3 mm + 1:100,000,000 (1 mm/100 km)
A order (primary)	5 mm + 1:10,000,000 (1 mm/10 km)
B order (secondary)	8 mm + 1:1,000,000 (1 mm/km)
C order (dependent)	10 mm + 1:100,000 (10 mm/km)

Geodetic Reference System 1980

Accurate results of modern geodetic and astronomic measurement techniques required the replacement of GRS 1967 within 13 years by the Geodetic Reference System 1980:

$$a = 6378\,137 \text{ m}$$

$$GM = 3986\,05 \times 10^8 \text{ m}^3 \text{ sec}^{-2}$$

$$J_2 = 108\,263 \times 10^{-8}$$

$$\omega = 7292\,115 \times 10^{-11} \text{ rad sec}^{-1}$$

GRS 1980 was adopted at the XVIIth General Assembly of the IUGG in Canberra, December 1979 [IAG, 1980]. Agreements with respect to the orientation of the reference ellipsoid were not altered.

1983 Best Values

During the XVIIIth General Assembly of the IUGG in Hamburg, August 1983, the IAG adopted a resolution that stated that the GRS 1980 did not need to be replaced, but that the following values were (at the time) the most representative [IAG, 1984]:

1983 Best Values	
Velocity of light in vacuum	$c = (299,792,458 \pm 1.2) \text{ m sec}^{-1}$
Gravitational constant (Newton)	$G = (6,673 \pm 1) \times 10^{-14} \text{ m}^3 \text{ sec}^{-2} \text{ kg}^{-1}$
Angular velocity of the earth	$\omega = 7,292,115 \pm 10^{-11} \text{ rad sec}^{-1}$
Geocentric gravitational constant of the earth, including the atmosphere	$GM = (39,860,044 \pm 1) \times 10^7 \text{ m}^3 \text{ sec}^{-2}$
Geocentric gravitational constant of atmosphere	$GM_A = (35 \pm 0.3) \times 10^7 \text{ m}^3 \text{ sec}^{-2}$
Second-degree harmonic coefficient without permanent deformation due to tides	$J_2 = (1,082,629 \pm 1) \times 10^9$
Equatorial radius of the earth	$a = (6,378,136 \pm 1) \text{ m}$
Equatorial gravity	$\gamma_e = (978,032 \pm 1) \times 10^{-5} \text{ m sec}^{-2}$
Flattening	$1/f = (298,257 \pm 1) \times 10^{-3}$
Potential of the geoid	$W_0 = (6,263,686 \pm 2) \times 10 \text{ m}^2 \text{ sec}^{-2}$
Triaxial ellipsoid (rounded values):	
Equatorial flattening	$1/f_1 = 90,000$
Longitude semimajor axis	$\lambda_1 = 345^\circ \text{ (east)}$

1987 Best Values and Secular Changes

In Hamburg a special study group (SSG 5.100), whose chairman was B.H. Chovitz, got the task to evaluate the status of the GRS 1980 and to prepare possible recommendations to the XIXth General Assembly of the IUGG in Vancouver, August, 1987. The result of the four-year study was that the GRS 1980 still did not need replacement [IAG, 1988a], but that a series of 1987 best values and secular changes could be forwarded. The interesting conclusion was that, from this moment on, two values had to be recognized for many parameters: a time-invariant component and the first derivative with respect to time [IAG, 1988b].

In the meantime the velocity of light had been raised to a physical constant, adopting the value as listed in the previous section [Cohen and Taylor, 1988]:

$$c = 299,792,458 \text{ m sec}^{-1} \text{ exactly}$$

This means that the meter is now a derived parameter through the adopted length of a second.

1987 Best Values	
Angular velocity of the earth (rounded)	$\omega = 7.292115 \times 10^{-5} \text{ rad sec}^{-1}$
Geocentric gravitational constant of the earth, including the atmosphere	$GM = (3,986,004.40 \pm 0.03) \times 10^8 \text{ m}^3 \text{ sec}^{-2}$
Equatorial radius of the earth	$a = (6,378,136 \pm 1) \text{ m}$
Second-degree spherical harmonic coefficient without permanent effect due to tides	$J_2 = (1,082,626 \pm 2) \times 10^{-9}$

1987 Best Secular Changes

Decrease of angular velocity due to tides	$d\omega_T/dt = (-6.0 \pm 0.3) \times 10^{-22} \text{ rad/sec}^{-2}$
Decrease of angular velocity due to other causes	$d\omega_{NT}/dt = (+1.4 \pm 0.3) \times 10^{-22} \text{ rad/sec}^{-2}$
Total decrease in angular velocity	$d\omega/dt = (-4.6 \pm 0.4) \times 10^{-22} \text{ rad/sec}^{-2}$
Relative change of the gravitational constant	$(dG/dt)/G = (0 \pm 1) \times 10^{-11} \text{ year}^{-1}$
Change of the earth's mass	$dM/dt = 0 \text{ kg year}^{-1}$
Change of the earth's radius	$d\alpha/dt = 0 \text{ mm year}^{-1}$
Decrease in the second-degree harmonic coefficient	$dJ_2/dt = (-2.8 \pm 0.3) \times 10^{-11} \text{ year}^{-1}$

It is interesting to note that we do not zero in on the earth's semimajor axis in an arbitrary manner. The earth seems to “shrink” under our increasingly accurate estimates of its equatorial radius, as shown in [Table 55.12](#).

World Geodetic System 1984

The World Geodetic System 1984 (WGS 84) and its three predecessors (WGS 60, WGS 66, and WGS 72), developed by the American Department of Defense/Defense Mapping Agency (DoD/DMA), are also reference *systems* in the real sense of the word. The WGS 84 consists of a reference coordinate frame, an ellipsoidal gravity formula, an ellipsoid, an earth gravity model, a geoid, and a series of transformation parameters and formulas between various geodetic reference frames and datum values [DMA, 1988].

WGS 84 Coordinate Frame

The WGS 84 coordinate frame is based on the frames that were originally developed for the U.S. Navy's Doppler/Transit System. One of its most recent reference frames, known as NSWC 9Z-2, had the following known problems:

- Its geocentricity or, better, nongeocentricity
- The misorientation of the zero meridian, compared to the zero meridian adopted by the BIH
- A scale error of almost 1 ppm

These problems can be expressed in terms of specific values:

- The origin with respect to the center of mass of the earth, as accurately determined by satellite laser ranging (SLR), was 4.5 m too high.
- The x axis of the NSWC frame was 0.814 arcseconds to the east of the BIH zero meridian — an error of about 24 m at the equator.
- The scale of NSWC 9Z-2 was 0.6×10^6 too large.

The WGS 84 corrected these problems. The transformation formulas (see the discussion of earth-fixed frames in Section 55.3) from $\mathbf{X}_{\text{NSWC 9Z-2}}$ to $\mathbf{X}_{\text{WGS 84}}$ become

$$\mathbf{x}_{\text{WGS 84}} = \sigma \mathbf{R}_3(\gamma) \mathbf{x}_{\text{NSWC 9Z-2}} + \mathbf{t} \quad (55.257)$$

TABLE 55.12 Decrease of the Semimajor Axis
in Recent Decades

Reference	Estimate
International ellipsoid 1924/Hayford 1909	6,378,388 m
Krassovsky 1943	6,378,245 m
GRS 1967	6,378,160 m
GRS 1980	6,378,137 m
Best values 1983/1987	6,378,136 m

with

$$\sigma = 1. + \delta\sigma = 1. - 0.000\,000\,6 \quad (55.258)$$

$$\gamma = 0''.814 \quad (55.259)$$

$$\mathbf{t} = \begin{pmatrix} 0 \\ 0 \\ 4.5 \text{ m} \end{pmatrix} \quad (55.260)$$

WGS 84 Ellipsoid

The origin of the WGS 84 coordinate frame is also the center of the WGS 84 ellipsoid, so that the symmetry axis of the ellipsoid of revolution coincides with the z axis of the WGS 84 frame. The values for the standard WGS 84 earth are as follows:

WGS 84 Ellipsoid

Semimajor axis	$a = (6,378,137 \pm 2) \text{ m}$
Geocentric gravitational constant of the earth, including the atmosphere	$GM = (3,986,005 \pm 0.6) \times 10^8 \text{ m}^3 \text{ sec}^{-2}$
Geocentric gravitational constant of atmosphere	$GM_A = (3.5 \pm 0.1) \times 10^8 \text{ m}^3 \text{ sec}^{-2}$
Geocentric gravitational constant of the earth without atmosphere	$GM_{-A} = (3,986,001.5 \pm 0.6) \times 10^8 \text{ m}^3 \text{ sec}^{-2}$
Normalized second-degree zonal gravitational coefficient without permanent deformation due to tides	$\bar{C}_{2,0}^{-T} = (-484.166\,85 \pm 0.00130) \times 10^{-6}$
Normalized second-degree zonal gravitational coefficient with permanent deformation due to tides ($J_2 = -\sqrt{5}\bar{C}_{2,0}$)	$\bar{C}_{2,0}^{+T} = (-484.171\,01 \pm 0.00130) \times 10^{-6}$
Angular velocity of the earth	$\omega = (7,292,115 \pm 0.1500) \times 10^{-11} \text{ rad sec}^{-1}$
Angular velocity of the earth in a precessing frame	$\omega^* = (7,292,115.8553 \times 10^{-11} + 4.3 \times 10^{-15} T_u) \text{ rad sec}^{-1}$ with T_u in Julian centuries, since January 1.5, 2000; see Eq. (55.74)

The velocity of light is the same as in GRS 1980; see earlier in this section.

Since the WGS 84 ellipsoid is a geocentric equipotential ellipsoid, the flattening f of this reference ellipsoid is a derived constant. It means that f can be computed from the listed “fundamental” parameters; see Heiskanen and Moritz [1967, sections 2-7 through 2-10] or IAG [1971].

The flattening f has the value:

$$\text{flattening (ellipticity)} \quad f = 1/298.257\,223\,563$$

WGS 84 Ellipsoidal Gravity Formula

The normal gravity γ along the surface of the WGS 84 ellipsoid is given by

$$\gamma = 978,032.677\,14 \times \left(1 + 0.001\,931\,851\,386\,39 \times \sin^2 \phi\right) / \left(1 - 0.006\,694\,379\,990\,13 \times \sin^2 \phi\right)^{1/2} \text{ milligal}$$

where 1 milligal = $0.001 \text{ cm sec}^{-2}$.

WGS 84 Earth Gravitational Model

The WGS 84 Earth Gravitational Model (EGM) is a gravity model with harmonic coefficients up to degree and order 180. For civilian users, the coefficients above degree and order 18 are classified.

WGS 84 Geoid

As with the EGM, for civilian users the geoid based on coefficients above degree and order 18 is classified.

In general, the geoidal undulations (geoidal heights) are accurate to about 2 to 6 m. For 55% of the earth's coverage the accuracy is 2 to 3 m; for 93%, 2 to 4 m.

WGS 84 Refinements since 1994

Regarding the WGS 84 Reference Frame, two adjustments have been implemented since January 1, 1994; the reference frames WGS 84 (G730), between January 2, 1994 and September 28, 1996, and the WGS 84 (G873), since September 29, 1996. Also a modified WGS 84 Earth Gravitational Model as well as a modified WGS 94 Geoid Model have been adopted since October 1, 1996: the WGS 84 EGM 96, see NIMA [1997]. These modifications led to a refined value for the WGS 84 GM parameter:

Geocentric gravitational constant of the earth including the atmosphere

$$GM = (3,986,004.418 \pm 0.008) \times 10^8 \text{m}^3 \text{s}^{-2}$$

With the geocentric gravitational constant of the atmosphere unchanged, a consistent value for GM excluding the atmosphere has become:

Geocentric gravitational constant of the earth without atmosphere

$$GM_A = (3,986,000.9 \pm 0.6) \times 10^8 \text{m}^3 \text{s}^{-2}$$

The WGS 84 EGM 96 is a gravity model with harmonic coefficients up to degree and order 360, consisting of 130,317 coefficients. The WGS 84 EGM96 Geoid has an error range between 0.5 and 1.0 m (1 sigma) anywhere on Earth. Geoid undulations are available on a 15 ft × 15 ft (arcminutes) grid.

IERS Standards 1992

The International Earth Rotation Service (IERS) adopted a standard reference system to be used in earth rotation analysis. The system includes not only numerical values for a set of variables, but also sets of adopted equations reflecting certain physical models. The reader is referred to IERS [1992] for more details.

Datum and Reference Frame Transformations

In Section 55.3 the transformation formulas are described relating coordinates given in two different reference frames. The adoption of new ellipsoidal parameters (so-called datum transformations) causes the geodetic coordinates to change. Tables 55.13 to 55.15, from such sources as Soler and Hothem [1988], DMA [1988], and NIMA [1997], list the values for various ellipsoids and reference frame transformations.

TABLE 55.13 Transformation Parameters among Several Global Reference Frames

Coordinate System (Datum) (1)	Δx (m) (2)	Δy (m) (3)	Δr (m) (4)	$\delta \epsilon$ (arc sec) (5)	$\delta \psi$ (arc sec) (6)	$\delta \omega$ (arc sec) (7)	δs (ppm) (8)
NWL-9D→WGS-72	0	0	0	0	0	−0.26	−0.827
NWL-9D→WGS-84	0	0	+4.5	0	0	−0.814	−0.6
WGS-72→WGS-84	0	0	+4.5	0	0	−0.554	+0.227
BTS87→NWL-9D	+0.071	−0.509	−4.666	−0.0179	+0.005	+0.8073	+0.583
BTS87→WGS-84	+0.071	−0.509	−0.166	−0.0179	+0.005	−0.0067	−0.107
BTS87→VLBI (NGS)	−0.089	+0.143	−0.016	+0.0043	−0.0093	+0.0033	+0.009
BTS87→SLR (GSFC)	0.000 ^a	0.000 ^a	0.000 ^a	+0.0018	−0.0062	+0.0075	0.000 ^a
WGS-84→WGS-84 (GPS)	+0.026	−0.006	+0.093	+0.001	0.000	+0.002	−0.128

^a These values were held fixed (i.e., the BTS87 frame origin and scale are assumed to be defined through satellite laser ranging (GSFC)).

Source: Solar, T. and Hothem, L.D., *J. Surveying Eng.*, 114, 84, 1988.

TABLE 55.14 Parameters of Some Adopted Reference Ellipsoids

Coordinate System (Datum) (1)	Reference Ellipsoid Used (2)	a (m) (3)	$1/f$ (4)
AGD	AN (or SA-69)	6,378,160	298.25
ED-79	International	6,378,388	297
GEM-8	GEM-8	6,378,145	298.255
GEM-9 (or GEM-10)	GEM-9 (or GEM-10)	6,378,140	298.255
GEM-10B	GEM-10B	6,378,138	298.257
GEM-T1	GEM-T1	6,378,137	298.257
NAD-27	Clarke 1866	6,378,206.4	294.9786982
NAD-83	GRS-80	6,378,137	298.257222101
NWL-9D = NSWC-9Z2	WGS-66	6,378,145	298.25
SA-69	SA-69 (or AN)	6,378,160	298.25
WGS-72	WGS-72	6,378,135	298.26
WGS-84	WGS-84	6,378,137	298.257223563

Note: AGD = Australian geodetic datum; AN = Australian national; ED = European datum; GEM = Goddard earth model; GRS = Geodetic Reference System; NAD = North American datum; NSWC = Naval Surface Warfare Center; NWL = Naval Weapons Laboratory; SA = South American; WGS = World Geodetic System.

Source: Solar, T. and Hothem, L.D., *J. Surveying Eng.*, 114, 84, 1988.

TABLE 55.15 Transformation Parameters, Local Geodetic Systems to WGS 84

Local Geodetic Systems	Reference Ellipsoids and Parameter Differences			Transformation Parameters		
	Name	Δa (m)	$\Delta f \times 10^4$	ΔX (m)	ΔY (m)	ΔZ (m)
Adindan	Clark 1880	-112.145	-0.54750714	-162	-12	206
Afgooye	Krassovsky	-108	0.00480795	-43	-163	45
Ain El ABD 1970	International	-251	-0.14192702	-150	-251	-2
Anna 1 Astro 1965	Australian National	-23	-0.00081204	-491	-22	435
ARC 1950	Clarke 1880	-112.145	-0.54750714	-143	-90	-294
ARC 1960	Clarke 1880	-112.145	-0.54750714	-160	-8	-300
Ascension Island 1958	International	-251	-0.14192702	-207	107	52
Astro Beacon "E"	International	-251	-0.14192702	145	75	-272
Astro B4 Sorol Atoll	International	-251	-0.14192702	114	-116	-333
Astro Dos 71/4	International	-251	-0.14192702	-320	550	-494
Astronomic Station 1952	International	-251	-0.14192702	124	-234	-25
Australian Geodetic 1966	Australian National	-23	-0.00081204	-133	-48	148
Australian Geodetic 1984	Australian National	-23	-0.00081204	-134	-48	149
Bellevue (IGN)	International	-251	-0.14192702	-127	-769	472
Bermuda 1957	Clarke 1866	-69.4	-0.37264639	-73	213	296
Bogota Observatory	International	-251	-0.14192702	307	304	-318
Campo Inchauspe	International	-251	-0.14192702	-148	136	90
Canton Astro 1966	International	-251	-0.14192702	298	-304	-375
Cape	Clarke 1880	-112.45	-0.54750714	-136	-108	-292
Cape Canaveral	Clarke 1866	-69.4	-0.37264639	-2	150	181
Carthage	Clarke 1880	-112.145	-0.54750714	-263	6	431
Chatham 1971	International	-251	-0.14192702	175	-38	113
Chua Astro	International	-251	-0.14192702	-134	229	-29
Corrego Alegre	International	-251	-0.14192702	-206	172	-6
Djakarta (Batavia)	Bessel 1841	739.845	0.10037483	-377	681	-50
DOS 1968	International	-251	-0.14192702	230	-199	-752
Easter Island 1967	International	-251	-0.14192702	211	147	111
European 1950	International	-251	-0.14192702	-87	-98	-121
European 1979	International	-251	-0.14192702	-86	-98	-119
Gandajika Base	International	-251	-0.14192702	-133	-321	50

TABLE 55.15 Transformation Parameters, Local Geodetic Systems to WGS 84

Local Geodetic Systems	Reference Ellipsoids and Parameter Differences			Transformation Parameters		
	Name	$\Delta a(\text{m})$	$\Delta f \times 10^4$	$\Delta X(\text{m})$	$\Delta Y(\text{m})$	$\Delta Z(\text{m})$
Geodetic Datum 1949	International	−251	−0.14192702	84	−22	209
Guam 1963	Clarke 1866	−69.4	−0.37264639	−100	248	259
GUX 1 Astro	International	−251	−0.14192702	252	−209	−751
Hjorsey 1955	International	−251	−0.14192702	−73	46	−86
Hong Kong 1963	International	−251	−0.14192702	−156	−271	−189
India	Everest	860.655	0.28361368	214	836	303
				289	734	257
Ireland 1965	Modified Airy	796.811	0.11960023	506	−122	611
ISTS 073 Astro 1969	International	−251	−0.14192702	208	−435	−229
Johnson Island 1961	International	−251	−0.14192702	191	−77	−204
Kandawala	Everest	860.655	0.28361368	−97	787	86
Kerguelen Island	International	−251	−0.14192702	145	−187	103
Kertau 1948	Modified Everest	832.937	0.2861368	−11	851	5
L.C. 5 Astro	Clarke 1866	−69.4	−0.37264639	42	124	147
Liberia 1964	Clarke 1880	−112.145	−0.54750714	−90	40	88
Luzon	Clarke 1866	−69.4	−0.37264639			
Philippines				−133	−77	−51
Mindanao Island				−133	−79	−72
Mahe 1971	Clarke 1880	−112.145	−0.54750714	41	−220	−134
Macro Astro	International	−251	−0.14192702	−289	−124	60
Massawa	Bessel 1841	739.845	0.10037483	639	405	60
Merchich	Clarke 1880	−112.145	−0.54750714	31	146	47
Midway Astro 1961	International	−251	−0.14192702	912	−58	1227
Minna	Clarke 1880	−112.145	−0.54750714	−92	−93	122
Nahrwan	Clarke 1880	−112.145	−0.54750714			
Masirah Island				−247	−148	369
United Arab Emirates				−249	−156	381
Saudi Arabia				−231	−196	482
Naparima, BWI	International	−251	−0.14192702	−2	374	172
North American 1927	Clarke 1866	−69.4	−0.37264639			
Mean Value (Conus)				−8	160	176
Alaska				−5	135	172
Bahamas				−4	154	178
San Salvador Island				1	140	165
Canada				−10	158	187
Canal Zone				0	125	201
Caribbean				−7	152	178
Central America				0	125	194
Cuba				−9	152	178
Greenland				11	114	195
Mexico				−12	130	190
North American 1983	GRS	0	−0.00000016	0	0	0
Observatorio 1966	International	−251	−0.14192702	−425	−169	81
Old Egyptian	Helmert 1906	−63	0.00480795	−130	110	−13
Old Hawaiian	Clarke 1866	−69.4	−0.37264639	61	−285	−181
Oman	Clarke 1880	−112.145	−0.54750714	−346	−1	224
Ordnance Survey of Great Britain 1936	Airy	573.604	0.11960023	375	−111	431
Pico De Las Nieves	International	−251	−0.14192702	−307	−92	127
Pitcairn Astro 1967	International	−251	−0.14192702	185	165	42
Provisional South Chilean 1963	International	−251	−0.14192702	16	196	93
Provisional South American 1956	International	−251	−0.14192702	−288	175	−376
Puerto Rico	Clarke 1866	−69.4	−0.37264639	11	72	−101
Qatar National	International	−251	−0.14192702	−128	−283	22
Qornoq	International	−251	−0.14192702	164	138	−189
Reunion	International	−251	−0.14192702	94	−948	−1262

TABLE 55.15 Transformation Parameters, Local Geodetic Systems to WGS 84

Local Geodetic Systems	Reference Ellipsoids and Parameter Differences			Transformation Parameters		
	Name	$\Delta a(\text{m})$	$\Delta f \times 10^4$	$\Delta X(\text{m})$	$\Delta Y(\text{m})$	$\Delta Z(\text{m})$
Rome 1940	International	−251	−0.14192702	−225	−65	9
Santo (DOS)	International	−251	−0.14192702	170	42	84
Sao Braz	International	−251	−0.14192702	−203	141	53
Sapper Hill 1943	International	−251	−0.14192702	−355	16	74
Schwarazeck	Bessel 1841	653.135	0.10037483	616	97	−251
South American 1969	South American 1969	−23	−0.00081204	−57	1	−41
South Asia	Modified Fischer 1960	−18	0.00480795	7	−10	−26
Southeast Base	International	−251	−0.14192702	−499	−249	314
Southwest Base	International	−251	−0.14192702	−104	167	−38
Timbalai 1948	Everest	860.655	0.28361368	−689	691	−46
Tokyo	Bessel 1841	739.845	0.10037483	−128	481	664
Tristan Astro 1968	International	−251	−0.14192702	−632	438	−609
Viti Levu 1916	Clarke 1880	−112.145	−0.54750714	51	391	−36
Wake–Eniwetok 1960	Hough	−133	−0.14192702	101	52	−39
Zanderij	International	−251	−0.14192702	−265	120	−358

Source: DMA, DMA Technical Report 8350.2, DMA, March 1, 1988.

References

Bugayevskiy, L.M. and Snyder, J.P., *Map Projections: A Reference Manual*, Taylor & Francis, Philadelphia, 1998.

Cohen, E.R. and Taylor, B.N., The fundamental physical constants, *Physics Today*, 41, 9, 1988.

DMA, Department of Defense World Geodetic System: Its Definition and Relationships with Local Geodetic Systems, DMA Technical Report 8350.2, Defense Mapping Agency, 1988 (revised March 1, 1988).

FGCC, Geometric Geodetic Accuracy Standards and Specifications for Using GPS Relative Positioning Techniques, Version 5.0, Federal Geodetic Control Committee, Rockville, MD, 1989.

GPS World, GPS world receiver survey, *GPS World*, 13, 28, 2002.

Heiskanen, W.A. and Moritz, H., 1967. *Physical Geodesy*, W.H. Freeman and Company, San Francisco, 1967.

IAG, Geodetic Reference System 1967, Publication Special 3, International Association of Geodesy, Paris, 1971.

IAG, Geodetic Reference System 1980, in *The Geodesist's Handbook 1980*, Moritz, H., Compiler, *Bull. G  od  sique*, 54, 395, 1980.

IAG, Geodetic Reference System 1980, in *The Geodesist's Handbook 1980*, Moritz, H., Compiler, *Bull. G  od  sique*, 58, 388, 1984.

IAG, Geodetic Reference System 1980, in *The Geodesist's Handbook 1980*, Moritz, H., Compiler, *Bull. G  od  sique*, 62, 348, 1988a.

IAG, Parameters of common relevance of astronomy, geodesy, and geodynamics, in *The Geodesist's Handbook 1980*, Chovitz, B.H., Compiler, *Bull. G  od  sique*, 62, 359, 1988b.

IERS, IERS Technical Note 12, in *IERS Standards*, McCarthy, D.D., Ed., Central Bureau of the International Earth Rotation Service, Observatoire de Paris, 1992.

Leick, A. and van Gelder, B.H.W., On Similarity Transformations and Geodetic Network Distortions Based on Doppler Satellite Coordinates, Reports of the Department of Geodetic Science, No. 235, Ohio State University, Columbus, 1975.

Lucas, J., Differentiation of the orientation matrix by matrix multipliers, *Photogrammetric Eng.*, 29, 708, 1963.

- Maling, D.H., *Coordinate Systems and Map Projections*, Pergamon Press, Oxford, 1993.
- Montgomery, H., City streets, airports, and a station roundup, *GPS World*, 4, 16, 1993.
- Moritz, H. and Mueller, I.I., *Earth Rotation: Theory and Observation*, Frederick Ungar Publishing Co., New York, 1988.
- Mueller, I.I., *Spherical and Practical Astronomy, As Applied to Geodesy*, Frederick Ungar Publishing Co., New York, 1969.
- NIMA, Department of Defense World Geodetic System: Its Definition and Relationships with Local Geodetic Systems, NIMA Technical Report 8350.2, National Imagery and Mapping Agency, 1997 (revised January 3, 2000).
- Santerre, R., Impact of GPS satellite sky distribution, *Manuscripta Geodaetica*, 61, 28, 1991.
- Soler, T., On Differential Transformations between Cartesian and Curvilinear (Geodetic) Coordinates, Reports of the Department of Geodetic Science, No. 236, Ohio State University, Columbus, 1976.
- Soler, T. and Hothem, L.D., Coordinate systems used in geodesy: basic definitions and concepts, *J. Surveying Eng.*, 114, 84, 1988.
- Soler, T. and van Gelder, B.H.W., On differential scale changes and the satellite Doppler z-shift, *Geophys. J. Roy. Astron. Soc.*, 91, 639, 1987.
- Stem, J.E., State Plane Coordinate System of 1983, NOAA Manual NOS NGS 5, NOAA, Rockville, MD, 1991.
- Vanič, P. and Krakiwsky, E.J., *Geodesy: The Concepts*, North-Holland Publishing Company, Amsterdam, 1982.
- Wells, D. and Kleusberg, A., GPS: a multipurpose system, *GPS World*, 1, 60, 1990.

Further Information

Textbooks and Reference Books

For additional reading and more background — from the very basic to the advanced level — in geodesy, satellite geodesy, physical geodesy, mechanics, orbital mechanics, and relativity, refer to the following textbooks (in English):

- Bomford, G., *Geodesy*, Clarendon Press, Oxford, 1980.
- Escobal, P.R., *Methods of Orbit Determination*, John Wiley & Sons, New York, 1976.
- FGCC, *Standards and Specifications for Geodetic Control Networks*, Federal Geodetic Control Committee, Rockville, MD, 1984 (reprint version, 1991).
- Goldstein, H., *Classical Mechanics*, Addison-Wesley, Reading, MA, 1965.
- Grewal, M.S., Weill, L.R., and Andrews, A.P., *Global Positioning Systems, Inertial Navigation, and Integration*, John Wiley & Sons, New York, 2001.
- Heitz, S., *Coordinates in Geodesy*, Springer-Verlag, Berlin, 1985.
- Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J., *GPS: Theory and Practice*, Springer-Verlag, New York, 2001.
- Jeffreys, Sir H., *The Earth: Its Origin, History and Physical Constitution*, Cambridge University Press, New York, 1970.
- Jekeli, C., *Inertial Navigation Systems with Geodetic Applications*, Walter de Gruyter, Berlin, 2000.
- Kaplan, E.D., Ed., *Understanding GPS: Principles and Applications*, Artech House Publishers, Boston, 1966.
- Kaula, W.M., *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*, Blaisdell Publishing Company, Waltham, MA, 1966.
- Kennedy, M., *The Global Positioning System and GIS: An Introduction*, Ann Arbor Press, Chelsea, MI, 1996.
- Lambeck, K., *Geophysical Geodesy: The Slow Deformations of the Earth*, Clarendon Press, Oxford, 1988.
- Leick, A., *GPS: Satellite Surveying*, John Wiley & Sons, New York, 1995.
- McElroy, S., *Getting Started with GPS Surveying*, GPS Consortium (GPSCO), Bathhurst, Australia, 1996.
- Melchior, P., *The Tides of the Planet Earth*, Pergamon Press, Oxford, 1978.

Moritz, H., *The Figure of the Earth: Theoretical Geodesy and the Earth's Interior*, Wichmann, Karlsruhe, Germany, 1990.

Munk, W.H. and MacDonald, G.J.F., *The Rotation of the Earth: A Geophysical Discussion*, Cambridge University Press, New York, 1975.

NATO, Standardization Agreement on NAVSTAR Global Positioning System (GPS), System Characteristics: Preliminary Draft, STANAG 4294, North Atlantic Treaty Organization, 1988 (revised April 15, 1988).

Parkinson, B.W. and Spilker, J.J., Eds., *Global Positioning System: Theory and Applications Volumes I and II*, Volumes 163 and 164 in Progress in Astronautics and Aeronautics, AIAA, Washington, D.C., 1996.

Seeber, G., *Satellite Geodesy: Foundations, Methods, and Applications*, Walter de Gruyter, Berlin, 1993.

Soffel, M.H., *Relativity in Astrometry, Celestial Mechanics and Geodesy*, Springer-Verlag, Berlin, 1989.

Strang, G. and Borre, K., *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press, Wellesley, MA, 1997.

Teunissen, P.J.G. and Kleusberg, A., Eds., *GPS for Geodesy*, Springer-Verlag, Berlin, 1998.

Torge, W., *Geodesy*, Walter de Gruyter, Berlin, 2001.

Van Sickle, J., *GPS for Land Surveyors*, Ann Arbor Press, Inc., Chelsea, MI, 1996.

Wells, D., Ed., *Guide to GPS Positioning*, Canadian GPS Associates, Fredericton, New Brunswick, Canada, 1986.

Journals and Organizations

The latest results from research in geodesy is published in:

- *Journal of Geodesy*: The *Journal of Geodesy* (<http://link.springer.de/link/service/journals/00190/>) is the recent merger of two international magazines, all published by Springer-Verlag, Berlin, under the auspices of the International Association of Geodesy:
 - Bulletin G  od  sique
 - Manuscripta Geodaetica

Geodesy- and geophysics-related articles can be found in:

- American Geophysical Union: EOS and *Journal of Geophysical Research*, Washington, D.C.
- Royal Astronomical Society: *Geophysical Journal International*, London

GPS-related articles can be found in:

- *GPS Solutions*, published by Wiley (<http://www.interscience.wiley.com/jpages/1080-5370/info.html>)
- Institute of Navigation: *Navigation*
- American Society of Photogrammetry and Remote Sensing: *Photogrammetric Engineering & Remote Sensing*

Many national mapping organizations publish journals in which recent results in geodesy, surveying, and mapping are documented:

- American Congress of Surveying and Mapping: *Surveying and Land Information Systems, Cartography and Geographic Information Systems* (<http://www.acsm.net/publist.html>)
- American Society of Civil Engineers: *Journal of Surveying Engineering* (<http://www.pubs.asce.org/journals/su.html>)
- Deutscher Verein f  r Vermessungswesen: *Zeitschrift f  r Vermessungswesen*, Konrad Wittwer Verlag, Stuttgart
- The Canadian Institute of Geomatics: *Geomatica* (<http://www.cig-acsg.ca/page.asp?intNodeID=15>)
- The Royal Society of Chartered Surveyors: *Survey Review* (<http://www.surveyreview.org.uk/>)
- Institution of Surveyors of Australia: *Australian Surveyor*

Worth special mention are the following trade magazines:

- *GPS World*, published by Advanstar Communications, Eugene, OR (<http://www.gpsworld.com>)
- *P.O.B.* (Point of Beginning), P.O.B. Publishing Company, Canton, MI (<http://www.pobonline.com>)
- *Professional Surveyor*, published by American Surveyors Publishing Company, Inc., Arlington, VA (<http://www.profsurv.com>)
- *G.I.M. International (Geomatics Info Magazine)*, published by Geodetical Information & Trading Centre B.V., Lemmer, the Netherlands (<http://www.gim-international.com>)

National mapping organizations, such as the U.S. National Geodetic Survey (<http://www.ngs.noaa.gov/>), regularly make geodetic software available (free and at cost). Information can be obtained from:

National Geodetic Survey
Geodetic Services Branch
National Ocean Service, NOAA
1315 East–West Highway, Station 8620
Silver Spring, MD 20910–3282

56

Photogrammetry and Remote Sensing

56.1 Basic Concepts in Photogrammetry

Scale and Coverage • Relief and Tilt Displacement • Parallax and Stereo

56.2 Sensors and Platforms

Cameras • Scanners • Pushbroom Linear Sensors

56.3 Mathematics of Photogrammetry

Condition Equations • Block Adjustment • Object Space Coordinate Systems

56.4 Instruments and Equipment

Stereoscopes • Monocomparator, Stereocomparator, Point Marker • Stereo Restitution: Analogue, Analytical, Softcopy • Scanners • Plotters

56.5 Photogrammetric Products

Topographic Maps • Image Products • Digital Elevation Models • Geographic Information Systems and Photogrammetry

56.6 Digital Photogrammetry

Data Sources • Digital Image Processing Fundamentals • Matching Techniques

56.7 Photogrammetric Project Planning

Flight Planning • Control Points

56.8 Close-Range Metrology

Equipment • Applications

56.9 Remote Sensing

Data Sources • Geometric Modeling • Interpretive Remote Sensing

J.S. Bethel

Purdue University

56.1 Basic Concepts in Photogrammetry

The term *photogrammetry* refers to the measurement of photographs and images for the purpose of making inferences about the size, shape, and spatial attributes of the objects appearing in the images. The term *remote sensing* refers to the analysis of photographs and images for the purpose of extracting the best interpretation of the image content. Thus the two terms are by no means mutually exclusive and each one includes some aspects of the other. However, the usual connotations designate geometric inferences as photogrammetry and radiometric inferences as remote sensing. Classically, both photogrammetry and remote sensing relied on photographs, that is, silver halide emulsion products, as the imaging medium. In recent years digital images or computer resident images have taken on an increasingly important role in both photogrammetry and remote sensing. Thus many of the statements to be found herein will refer to the general term *images* rather than to the more restrictive term *photographs*. The

characteristic of photogrammetry and remote sensing that distinguishes them from casual photography is the insistence on a thorough understanding of the sensor geometry and its radiometric response.

The predominant type of imaging used for civil engineering applications is the traditional 23-cm format frame aerial photograph. Photogrammetric and remote sensing techniques can be equally applied to satellite images and to close-range images acquired from small format cameras. The main contributions of photogrammetry and remote sensing to civil engineering include topographic mapping, orthophoto production, planning, environmental monitoring, database development for geographic information systems (GIS), resource inventory and monitoring, and deformation analysis.

Scale and Coverage

A typical aerial camera geometry is shown in Fig. 56.1. The perspective geometry of an ideal camera would dictate that each object point, A , would lie along a line containing the corresponding image point, a , and the perspective center, L . H represents the flying height above the terrain; f is the focal length or principal distance of the camera or sensor. In the ideal case that the camera axis is strictly vertical and the terrain is a horizontal plane surface, the *scale* of the image can be expressed as a fraction:

$$\text{Scale} = \frac{\text{Image distance}}{\text{Object distance}} \quad (56.1)$$

The units of measure should be the same for the numerator and the denominator so the fraction becomes a unitless ratio. Usually one forces the numerator to be 1, and this is often called the representative fraction or scale. In practice, units are sometimes introduced, but this should be discouraged. For instance, “one inch equals one hundred feet” should be converted as

$$\frac{1 \text{ inch}}{100 \text{ feet}} = \frac{1 \text{ inch}}{1200 \text{ inches}} = \frac{1}{1200} \quad (56.2)$$

This scale could also be represented in the form 1:1200. By the geometry shown in Fig. 56.1, the scale may also be expressed as a ratio of focal length and flying height above the terrain:

$$\text{Scale} = \frac{f}{H} \quad (56.3)$$

In practice, images never fulfill the ideal conditions stated above. In the general case of tilted photographs and terrain which is not horizontal, such a simple scale determination is only approximate. In fact the scale may be different at every point, and further, the scale at each point may be a function of direction. Because of this one often speaks of a *nominal scale*, which may apply to a single image or to a strip or block of images.

From Eq. (56.3) it is clear that high altitude imagery, having large H , yields a small scale ratio compared with lower altitude imagery, for which H is smaller. Thus one speaks of small scale imagery from high altitude and large scale imagery from low altitude. The area covered by a fixed size image will of course be inversely related to scale. Small scale images would cover a larger area than large scale images.

Relief and Tilt Displacement

If the ideal conditions of horizontal terrain and vertical imagery were fulfilled, a scale could be determined for the image and it could be used as a map. Terrain relief and image tilt are always present to some extent, however, and this prevents the use of raw images or photographs as maps. It is common practice to

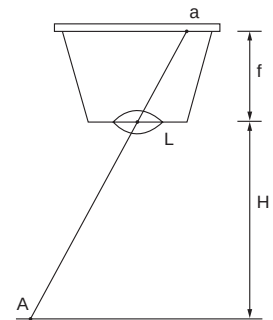


FIGURE 56.1 Typical aerial camera geometry.

modify images to remove tilt displacement (this is referred to as *rectification*); such images can be further modified to remove relief displacement (this is referred to as *differential rectification*). Such a differentially rectified image is referred to as an *orthophoto*.

The concept of relief displacement is shown schematically in Fig. 56.2. If the flagpole were viewed from infinity its image would be a point and there would be no relief displacement. If it is viewed from a finite altitude its image will appear to “lay back” on the adjacent terrain. The displacement vector in the image is aa' . The magnitude of this displacement vector will depend on the height of the object, the flying height, and its location in the image. Such displacements in the image are always radial from the image of the *nadir point*. The nadir point is the point exactly beneath the perspective center. In a vertical photograph this would also coincide with the image *principal point*, or the foot of the perpendicular from the perspective center. Assuming vertical imagery, the height of an object can be determined by relief displacement to be

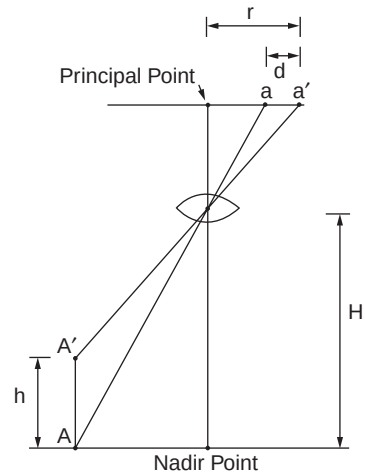
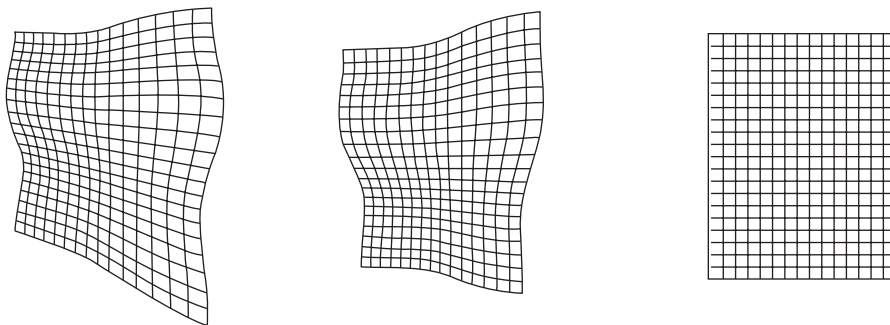


FIGURE 56.2 Relief displacement.

$$h = \frac{dH}{r} \quad (56.4)$$

d and r should be measured in the same units in the image, and h and H should be in the same object space units. Equation (56.4) could be used to obtain approximate heights for discrete objects with respect to the surrounding terrain. The same displacement occurs with the terrain itself, if it is higher than the reference plane. However, because the terrain is a continuous surface this displacement is not obvious as in the case of the flagpole.

Image tilt also creates image point displacements that would not be present in a vertical image. Extreme image tilts are sometimes introduced on purpose for *oblique photography*. For nominally vertical images, tilts are usually kept smaller than three degrees. Figure 56.3 illustrates some of these concepts in a sequence of sketches. Figure 56.3(a) depicts a tilted image of a planimetrically orthogonal grid draped over a terrain surface, Fig. 56.3(b) shows the image with tilt effects removed, and Fig. 56.3(c) shows the image with relief displacement effects removed. Only after the steps to produce image (c) can one use the image as a map. Prior to this there are severe systematic image displacements.



(a) Tilted Perspective Image

(b) Rectified Perspective Image
(Equivalent Vertical Photograph)

(c) Differentially Rectified Image
(Orthophotograph)

FIGURE 56.3 Image rectification sequence.

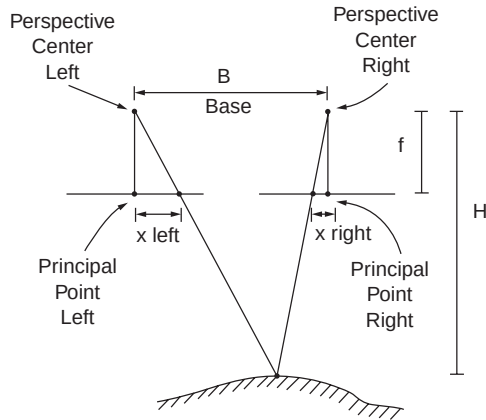


FIGURE 56.4 Parallax geometry.

Parallax and Stereo

Parallax is defined as the apparent shift in the position of an object, caused by a shift in the position of the viewer. Alternately closing one eye and then the other will demonstrate the concept of parallax, as near objects appear to shift, whereas far objects will appear stationary. This effect is the primary mechanism by which we achieve binocular depth perception, or stereo vision. This effect is exploited in photogrammetry when two overlapping photographs are taken. Within the overlap area, objects are imaged from two different exposure positions and parallax is evidenced in the resulting images. Parallax measurements can be used for approximate height computations in the following way using the geometry in Fig. 56.4. Two vertical overlapping photographs are taken at the same altitude, and an image coordinate system is set up within each image with the x axis parallel to the flight line, and with origin at the principal point in each image. For a given point, the parallax is defined as

$$p = x_{\text{left}} - x_{\text{right}} \quad (56.5)$$

The dimension H can be computed from

$$H = \frac{fB}{p} \quad (56.6)$$

B represents the base, or distance between the exposure stations. As is evident from Eq. (56.6) the distance H is inversely related to parallax, so that large parallax yields small H . Equation (56.6) is most often used with a pair of nearby points to determine a difference in elevation rather than the absolute elevation itself. This is given by the following equation, in which b is the image dimension of the base, and Δp is the difference in parallax.

$$\Delta H = \frac{H \Delta p}{b} \quad (56.7)$$

Figure 56.4 also illustrates the concept of B/H or *base–height ratio*. For a given focal length, format size, and overlap there is a corresponding B/H value for some average elevation in the area of interest. Large B/H yields strong geometry for elevation determination and small B/H yields weak geometry. For typical aerial photography for engineering mapping one would have $f = 152$ mm, $H = 732$ m, $B = 439$ m, overlap = 60%, and $B/H = 0.6$.

The viewing geometry as opposed to the taking geometry for this imagery yields a B/H of 0.3. The difference between the taking geometry and the viewing geometry causes *vertical exaggeration*.

That is, features appear stretched in the z -dimension compared to the planimetric dimension. This does not affect measurements, but only the viewer's perception of depth.

56.2 Sensors and Platforms

Cameras

Aerial cameras for engineering mapping use standard aerial film with 23-cm width. The emulsion type is usually panchromatic (black and white), but color film is becoming more popular. For some interpretation tasks, particularly involving vegetation, color infrared film is also used. This should not be confused with thermal infrared imaging, which requires a special sensor and cannot be directly captured on film. Resolution of camera components can be, individually, in the range of 140 line pairs per millimeter. But taken as a system the resolution would be in the range of 30 line pairs per millimeter. To maintain high resolution at low altitude, modern cameras employ image motion compensation to translate the film in synchronism with the apparent ground motion while the shutter is open. The ratio of image velocity in the image plane, v , to camera velocity over the ground, V , for a nominally vertical photograph is given by

$$\frac{v}{V} = \text{Scale} = \frac{f}{H} \quad (56.8)$$

A vacuum system holds the film firmly in contact with the platen during exposure, releasing after exposure for the film advance. During a photo flight, the camera is typically rotated to be parallel with the flight direction, rather than the aircraft body, in case there is a crosswind. This prevents *crab* between the adjacent photographs.

Calibration of aerial cameras should be executed every two to three years. Such a calibration determines the resolution, shutter characteristics, fiducial mark coordinates, principal point location, and a combination of focal length and lens distortion parameters that permits the modeling of the system to have central perspective geometry. Specifications for engineering mapping should always include a requirement for up-to-date camera calibration. A typical radial lens distortion curve is given in Fig. 56.5. The horizontal axis is radial distance from the principal point in millimeters, and the vertical axis is distortion at the image plane in micrometers.

Terrestrial cameras for photogrammetry usually employ a smaller film format than the standard aerial case; 70-mm film is a popular size for terrestrial photogrammetry. Whereas aerial cameras are at fixed, infinity focus, terrestrial cameras are typically used at a variety of focus positions. Thus calibration is more difficult since the lens components, and hence their image-forming attributes, are changing from one focus position to another. Often a camera manufacturer will provide principal distance and lens distortion data for each of several focus positions. Alternatively, the user may wish to model such behavior with additional parameters in a block adjustment with self-calibration.

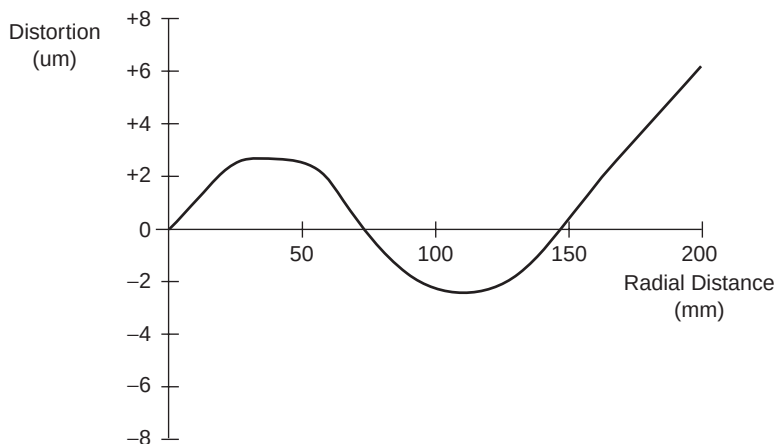


FIGURE 56.5 Lens distortion curve.

Before leaving the subject of photogrammetric cameras, mention should be made of the possibility of using an electronic image sensor in place of hardcopy film. In the case of large scale, aerial, engineering mapping photography, sensors of a size and resolution to duplicate the functionality of film are not available, and will not be for many years into the future. In the case of small format cameras, it is a different matter. Area sensors with 2048×2048 picture elements are available, with an element size of 10 micrometers. The total area of such an array would be therefore about 20 by 20 mm. This begins to approach the format size and resolution that could be usable for terrestrial photogrammetry. The direct capture of a digital image has advantages when feature extraction or enhancement by image processing is to be employed.

Scanners

Satellite imagers necessarily employ electronic sensors rather than film-based image capture because of the difficulty of retrieving film packets versus the relative ease of transmitting digital data by telemetry. A variety of sensor technologies is employed in satellite imagers, and the first of these, mechanical scanners, is discussed here. The MSS, multispectral scanner, on Landsats 1 to 5 is shown schematically in Fig. 56.6. For this instrument the mirror scans across the track of the satellite. The focal plane elements consist of light-gathering fiber-optic bundles which carry the radiation to PMTs, photomultiplier tubes, or photodiode detectors. The *IFOV*, *instantaneous field of view*, determined by detector size, telescope magnification, and altitude, is 83 m. The scan rate and the sampling rate determine the *pixel size*. For the MSS it is 83 m in the along-track dimension and 68 m in the across-track dimension. The MSS has detectors in 4 bands covering the visible and near infrared portions of the spectrum. The TM, thematic mapper, on Landsats 4 and 5 has a similar design but enhanced performance compared to the MSS. Data recording takes place on both swings of the mirror, and there is a wider spectral range among the seven bands. The pixel size is 30 m at ground scale.

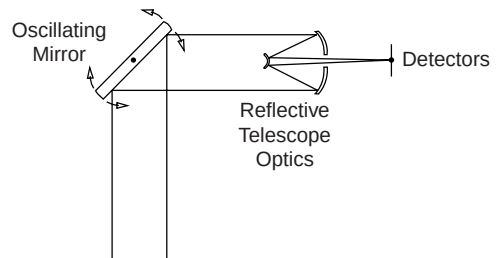


FIGURE 56.6 Schematic of mechanical scanner.

Pushbroom Linear Sensors

Linear CCDs, charge-coupled devices, are gaining wide acceptance for satellite imagers. A typical linear array is 2000 to 8000 elements in length. Longer effective pixel widths may be obtained by combining multiple arrays end to end (or equivalently using optical beam splitters). The individual elements in the sensor collect incident illumination during the integration period. The resulting charges are then transferred toward one end of the chip, and emerge as an analogue signal. This signal must be digitized for transmission or storage. Pixel sizes are usually in the 5- to 15-micrometer range. Since telescope optics have far superior resolution at the image plane, a purposeful defocusing of the optical system is performed to prevent aliasing. (See Fig. 56.7.)

56.3 Mathematics of Photogrammetry

Condition Equations

Using photogrammetry to solve spatial position problems inevitably leads to the formation of equations which link the observables to the quantities of interest. It is extremely rare that one would directly observe these quantities of interest. The form of the condition equations will reflect the nature of the observations, such as 2-D image coordinates or 3-D model coordinates, and the particular problem that one wishes to solve, such as space resection or relative orientation.

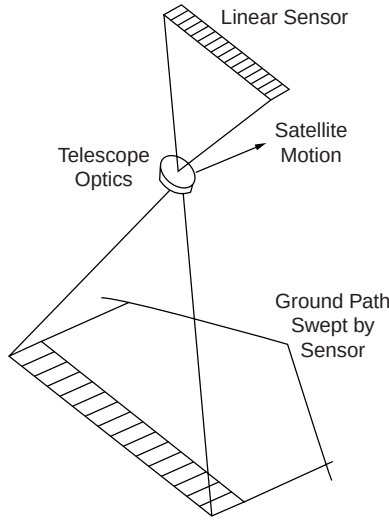


FIGURE 56.7 Pushbroom linear sensor.

Preliminaries

The 3×3 rotation that is often employed in developing photogrammetric condition equations is a function of three independent quantities. These quantities are usually the sequential rotations ω , ϕ , and κ , about the x , y , and z axes, respectively. The usual order of application is

$$M = M_{\kappa} M_{\phi} M_{\omega} \quad (56.9)$$

The elements are given by

$$M = \begin{bmatrix} \cos \phi \cos \kappa & \cos \omega \sin \kappa + \sin \omega \sin \phi \cos \kappa & \sin \omega \sin \kappa - \cos \omega \sin \phi \cos \kappa \\ -\cos \phi \sin \kappa & \cos \omega \cos \kappa - \sin \omega \sin \phi \sin \kappa & \sin \omega \cos \kappa + \cos \omega \sin \phi \sin \kappa \\ \sin \phi & -\sin \omega \cos \phi & \cos \omega \cos \phi \end{bmatrix} \quad (56.10)$$

An occasionally useful approximation to this matrix, in the case of small angles (i.e., near vertical imagery) is given by

$$M \approx \begin{bmatrix} 1 & \kappa & -\phi \\ -\kappa & 1 & \omega \\ \phi & -\omega & 1 \end{bmatrix} \quad (56.11)$$

in which the assumption is made that all cosines are 1, and the product of sines is zero. Other rotations and other parameters may also be used to define this matrix.

Collinearity Equations

The fundamental imaging characteristic of an ideal camera is that each object point, its corresponding image point, and the lens perspective center all lie along a line in space. This can be expressed in the following way, referring to [Fig. 56.8](#):

$$\vec{a} = k \vec{A} \quad (56.12)$$

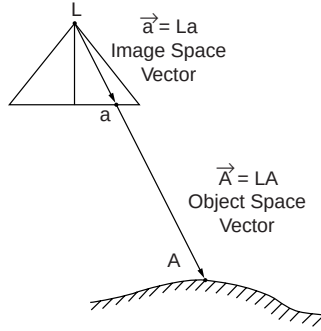


FIGURE 56.8 Collinearity geometry.

This equation is valid only if the two vectors are expressed in the same coordinate system. The image space vector is inevitably expressed in an image coordinate system:

$$\vec{a} = \begin{bmatrix} x - x_0 \\ y - y_0 \\ -f \end{bmatrix} \quad (56.13)$$

while the object space vector is expressed in an object space coordinate system (shifted to the camera perspective center):

$$\vec{A} = \begin{bmatrix} X - X_L \\ Y - Y_L \\ Z - Z_L \end{bmatrix} \quad (56.14)$$

One thus needs to scale, rotate, and translate one of these two coordinate systems until they are coincident. This transformation is usually applied to the object space vector and is expressed as follows:

$$\begin{bmatrix} x - x_0 \\ y - y_0 \\ -f \end{bmatrix} = kM \begin{bmatrix} X - X_L \\ Y - Y_L \\ Z - Z_L \end{bmatrix} \quad (56.15)$$

Eliminating the scale parameter k yields the classical form of the collinearity equations:

$$\begin{aligned} x - x_0 &= -f \frac{m_{11}(X - X_L) + m_{12}(Y - Y_L) + m_{13}(Z - Z_L)}{m_{31}(X - X_L) + m_{32}(Y - Y_L) + m_{33}(Z - Z_L)} \\ y - y_0 &= -f \frac{m_{21}(X - X_L) + m_{22}(Y - Y_L) + m_{23}(Z - Z_L)}{m_{31}(X - X_L) + m_{32}(Y - Y_L) + m_{33}(Z - Z_L)} \end{aligned} \quad (56.16)$$

Examples of particular problems for which the collinearity equations are useful include *space resection* (camera exterior orientation unknown, object points known, observed image coordinates given, usually implying a single image), *space intersection* (camera exterior orientations known, object point unknown, observed image coordinates given, usually implying a single object point), and *bundle block adjustment* (simultaneous resection and intersection, multiple images, and multiple points). This equation is non-linear and a linear approximation is usually made if we attempt to solve for any of the variables as unknowns. This dictates an iterative solution.

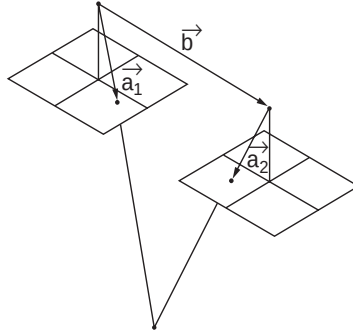


FIGURE 56.9 Coplanarity geometry.

Coplanarity Equation

The phrase *conjugate image points* refers to multiple image instances of the same object point. If we consider a pair of properly oriented images and the pair of rays defined by two conjugate image points, then this pair of rays together with the base vector between the perspective centers should define a plane in space. The *coplanarity* condition enforces this geometrical configuration. This is done by forcing these three vectors to be coplanar, which is in turn guaranteed by setting the triple scalar product to zero. An alternative explanation is that the parallelepiped defined by the three vectors as edges has zero volume. Figure 56.9 illustrates this geometry. The left vector is given by

$$a_1 = \begin{bmatrix} u_1 \\ v_1 \\ w_1 \end{bmatrix} = M_1^t \begin{bmatrix} x - x_0 \\ y - y_0 \\ -f \end{bmatrix}_1 \quad (56.17)$$

The right vector is given by

$$a_2 = \begin{bmatrix} u_2 \\ v_2 \\ w_2 \end{bmatrix} = M_2^t \begin{bmatrix} x - x_0 \\ y - y_0 \\ -f \end{bmatrix}_2 \quad (56.18)$$

and the base vector is given by

$$b = \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} = \begin{bmatrix} X_{L_2} - X_{L_1} \\ Y_{L_2} - Y_{L_1} \\ Z_{L_2} - Z_{L_1} \end{bmatrix} \quad (56.19)$$

The coplanarity condition equation is the above-stated triple scalar product

$$F = \begin{vmatrix} b_x & b_y & b_z \\ u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \end{vmatrix} = 0 \quad (56.20)$$

The most prominent application for which the coplanarity equation is used is relative orientation. The equation is nonlinear and a linear approximation is usually made in order to solve for any of the variables as unknowns. This dictates an iterative solution.

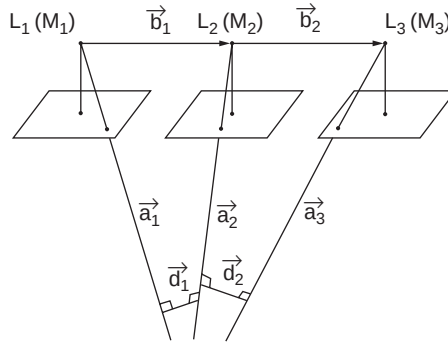


FIGURE 56.10 Scale restraint geometry.

Scale Restraint Equation

If photograph one is relatively oriented to photograph two, and photograph two is relatively oriented to photograph three, there is no guarantee that photograph one and photograph three are also relatively oriented. There are several methods to enforce this condition, and among them the most robust would be the *scale restraint* condition. This states that the intersection of conjugate rays from the three photographs should in fact occur at a single point. From Fig. 56.10 we see the three rays, $\mathbf{a}_i$, and two “mismatch” vectors, $\mathbf{d}_i$:

$$\mathbf{a}_1 = \begin{bmatrix} a_{1_x} \\ a_{1_y} \\ a_{1_z} \end{bmatrix} = \mathbf{M}_1^t \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \\ -f \end{bmatrix} \quad (56.21)$$

$$\mathbf{a}_2 = \begin{bmatrix} a_{2_x} \\ a_{2_y} \\ a_{2_z} \end{bmatrix} = \mathbf{M}_2^t \begin{bmatrix} x_2 - x_0 \\ y_2 - y_0 \\ -f \end{bmatrix} \quad (56.22)$$

$$\mathbf{a}_3 = \begin{bmatrix} a_{3_x} \\ a_{3_y} \\ a_{3_z} \end{bmatrix} = \mathbf{M}_3^t \begin{bmatrix} x_3 - x_0 \\ y_3 - y_0 \\ -f \end{bmatrix} \quad (56.23)$$

$$\mathbf{d}_1 = \mathbf{a}_1 \times \mathbf{a}_2 \quad (56.24)$$

$$\mathbf{d}_2 = \mathbf{a}_2 \times \mathbf{a}_3 \quad (56.25)$$

The scale restraint equation itself forces the independent scale factors for the common ray to be equal:

$$F = \frac{\begin{vmatrix} a_{1_x} & d_{1_x} & b_{1_x} \\ a_{1_y} & d_{1_y} & b_{1_y} \\ a_{1_z} & d_{1_z} & b_{1_z} \end{vmatrix}}{\begin{vmatrix} a_{1_x} & d_{1_x} & a_{2_x} \\ a_{1_y} & d_{1_y} & a_{2_y} \\ a_{1_z} & d_{1_z} & a_{2_z} \end{vmatrix}} + \frac{\begin{vmatrix} b_{2_x} & d_{2_x} & a_{3_x} \\ b_{2_y} & d_{2_y} & a_{3_y} \\ b_{2_z} & d_{2_z} & a_{3_z} \end{vmatrix}}{\begin{vmatrix} a_{2_x} & d_{2_x} & a_{3_x} \\ a_{2_y} & d_{2_y} & a_{3_y} \\ a_{2_z} & d_{2_z} & a_{3_z} \end{vmatrix}} = 0 \quad (56.26)$$

This equation is used primarily in the analytical formation of strips by successive relative orientation of image pairs in the strip. Common points in any photo triplet (that is, two adjacent models) would be subjected to the scale restraint condition. This equation is nonlinear and would require linear approximation for practical use in the given application.

Linear Feature Equations

It can happen, particularly in close-range photogrammetry, that the object space parameters of a straight line feature are to be determined. If at the same time stereo observation is either unavailable or difficult because of convergence or scale, then it becomes helpful if one can observe the feature monoscopically on each image without the need for conjugate image points. This can be elegantly accomplished with a condition equation which forces the ray associated with an observed image coordinate to pass through a straight line in object space. For each point in each photograph, a condition equation of the following kind may be written:

$$\begin{vmatrix} \rho_x & \rho_y & \rho_z \\ \beta_x & \beta_y & \beta_z \\ LC_x & LC_y & LC_z \end{vmatrix} = 0 \quad (56.27)$$

In this equation the vector ρ is the object space vector from the observed image point, the vector β is the vector along the straight line in object space, and LC is the vector from the perspective center to the point on the line closest to the origin. The six linear feature parameters must be augmented by two constraints which fix the magnitude of β to 1, and guarantee that β and C are orthogonal. A variation on this technique is the case where the object space feature is a circle in space. In this case each point on each image contributes an equation of the form

$$\left\| (L - C) - \frac{(L - C)\eta}{\rho\eta} \rho \right\| = r \quad (56.28)$$

where ρ and L have the same meaning as before. η represents the normal vector to the circle plane, r represents the circle radius, and C represents the circle center. In this case the normal vector must be constrained to unit magnitude. As in every case described here these equations are nonlinear in the variables of interest, and when we solve for them, the equations must be approximated using the Taylor series.

Block Adjustment

The internal geometry of a block of overlapping photographs or images may be sufficient to determine relative point positions, but for topographic mapping and feature extraction, one needs to tie this block to a terrestrial coordinate system. It would be possible to provide field survey determined coordinates for every point, but this would be prohibitively expensive. Thus arises the need to simultaneously tie all the photographs to each other, as well as to a sparse network of terrestrial control points. This process is referred to as *block adjustment*. The minimum amount of control necessary would be seven coordinate components, i.e., two horizontal (X, Y) points and three vertical (Z) points, or two complete control points (X, Y, Z) and one point with only vertical (Z). In practice, of course, one usually provides control in excess of the minimum, the redundancy providing increased confidence in the results.

Block Adjustment by Bundles

Block adjustment by bundles is the most mathematically rigorous way to perform this task. Observations consist of 2-D photograph image coordinates, usually read from a comparator or analytical plotter, transformed to the principal point origin, and refined for all known systematic errors. These systematic errors, described below, consist of at least lens distortion and atmospheric refraction, although at low altitude refraction may be considered negligible. Each image point, i , on each image, j , contributes two collinearity condition equations of the form

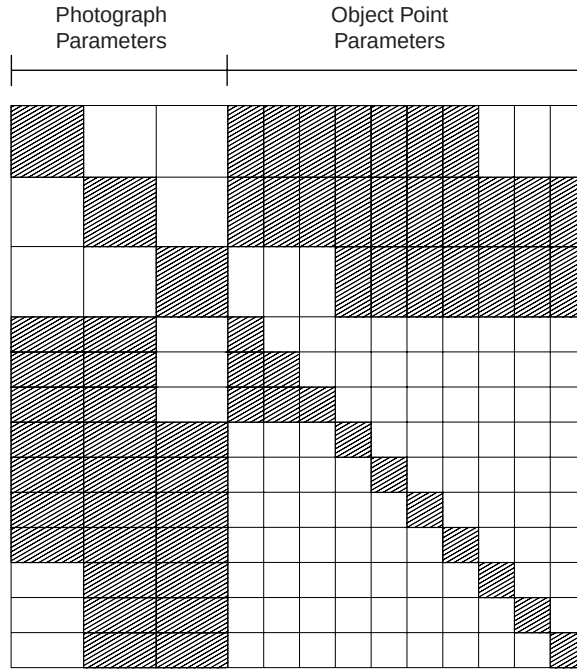


FIGURE 56.11 Organization of normal equations for bundle adjustment.

$$\begin{aligned}
 x_i - x_0 &= -f \frac{m_{11}(X_i - X_{L_j}) + m_{12}(Y_i - Y_{L_j}) + m_{13}(Z_i - Z_{L_j})}{m_{31}(X_i - X_{L_j}) + m_{32}(Y_i - Y_{L_j}) + m_{33}(Z_i - Z_{L_j})} \\
 y_i - y_0 &= -f \frac{m_{21}(X_i - X_{L_j}) + m_{22}(Y_i - Y_{L_j}) + m_{23}(Z_i - Z_{L_j})}{m_{31}(X_i - X_{L_j}) + m_{32}(Y_i - Y_{L_j}) + m_{33}(Z_i - Z_{L_j})}
 \end{aligned}
 \tag{56.29}$$

in which (x_i, y_i) are the observed image coordinates, transformed into a coordinate system defined by the camera fiducial coordinates. The variables (x_0, y_0) represent the position of the principal point and f represents the focal length or principal distance. The last three variables would often be considered as fixed constants from a camera calibration report, or they may be carried as unknowns in the adjustment. The variables (X_i, Y_i, Z_i) represent the object coordinates of the point i . They may be known or partially known if point i is a control point, or they may be unknown if point i is a pass point. The variables $(X_{L_j}, Y_{L_j}, Z_{L_j})$ represent the coordinates of the exposure station or perspective center of image j . Each image, j , also has an associated orientation matrix, M_j , whose elements are shown in the equation. If there are n points observed on m images, the total number of condition equations will be $2nm$. The total number of unknowns will be $3n + 6m - (\text{number of fixed coordinate components})$. With the advent of GPS in the photo aircraft, control may be introduced not only at the object points but also at the exposure stations. If the solution to the overdetermined problem is carried out by normal equations, the form of these equations is shown in Fig. 56.11.

Block Adjustment by Models

Block adjustment by models is necessary if the original observations of the photographs are made in stereo with an observed (x, y, z) model coordinate for each control point and pass point. Together with the model coordinates of the perspective centers, all model points are simultaneously transformed to the object (or control) coordinate system. There is usually redundant information to specify this transformation, so a

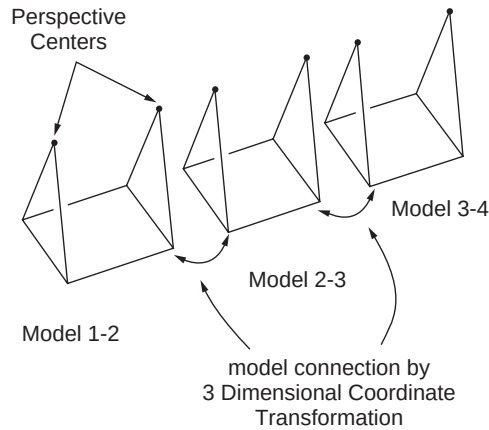


FIGURE 56.12 Model connection and strip formation.

least squares estimation is necessary. Practitioners have used comparator-derived image coordinates to compute model coordinates, and then subjected these derived model coordinates to an independent model block adjustment. This practice should be discouraged in favor of using the image coordinates directly in a bundle adjustment as described above. For the independent model block adjustment, each point, i , in model (stereo pair), j , will be related to the object space coordinates by a seven-parameter transformation unique to each model. The seven parameters include scale, s , rotations, Ω , Φ , and K , and translations T_x , T_y , T_z . For each point, i , in each model, j , the following three equations can be written:

$$\begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} = - \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_i + s_j M_j \begin{bmatrix} x \\ y \\ z \end{bmatrix}_i + \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}_j = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (56.30)$$

The uppercase coordinate vector represents the coordinate system of the control points, the lowercase vector represents the model coordinates, and the matrix M contains the rotation parameters. Only for control points will the (X, Y, Z) values be known; for all other points these will be unknown parameters solved for in the block adjustment. For n points and m models the total number of condition equations would be $3nm$. The total number of unknown parameters would be $3n + 7m - (\text{number of fixed coordinate components})$.

Strip Formation and Block Adjustment by Polynomials

Strip formation and block adjustment by polynomials assumes that the input data are xyz model coordinates. These would usually come directly from a relatively oriented model in an analogue stereoplotter. They could also come from analytical computation from image coordinates. If this is the case it would be preferable to use the image coordinates directly in a bundle block adjustment. A similar comment was made with regard to the independent model block adjustment. The strip formation consists of linking successive models (including perspective center coordinates) by seven-parameter transformations, and then transforming each new model into the strip system based on the first model. This process is illustrated in Fig. 56.12. If a single strip is sufficiently short, say five models or less, the strip can be fitted to the control points by a global seven-parameter transformation. This is given in the following equation:

$$\begin{bmatrix} E \\ N \\ h \end{bmatrix} = sM \begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \quad (56.31)$$

If the strip is longer than this, then because of adverse error propagation, artificial bends and bows will be present in the strip coordinates and polynomials present a way to model such effects. This technique is primarily of historical interest since its computational efficiency is no longer a compelling attribute. Both model coordinates and ground coordinates are transformed into an “axis of flight” coordinate system centered within the strip. Then conformal polynomials are used to transform the strip planimetric coordinates into the control system:

$$x' = x + a_1 + a_3x - a_4y + a_5(x^2 - y^2) - 2a_6xy + a_7(x^3 - 3xy^2) - a_8(3x^2y - y^3) + \dots \quad (56.32)$$

$$y' = y + a_2 + a_4x + a_3y + a_6(x^2 - y^2) + 2a_5xy + a_7(3x^2y - y^3) + a_8(x^3 - 3x^2y) + \dots \quad (56.33)$$

The vertical control points are used in the following polynomial:

$$z' = z + b_0 - 2b_2x + 2b_1y + c_1x^2 + c_2x^3 + c_3x^4 + d_1xy + d_2x^2y + d_3x^3y + d_4x^4y + e_1y^2 + e_2xy^2 \quad (56.34)$$

The number of terms is selected based on the quantity of control points and the length of the strip. Following the polynomial estimation, the points are transformed back into the original control coordinate system.

Image Coordinate Refinement

Raw stage coordinates from a comparator or analytical plotter must undergo a number of transformations and refinements before being used in further photogrammetric processing such as relative orientation or bundle block adjustment. Firstly the stage coordinates are transformed into the coordinate system defined by the camera *fiducial marks* or registration marks. This is usually done with a four- or six-parameter transformation. They are then shifted to the *principal point of autocollimation* by the principal point offsets (x_0, y_0). Following this they are corrected for radial lens distortion based on their position with respect to the *principal point of best symmetry*. The radial lens distortion is provided as part of the calibration of the camera either in the form of a table, a graph, or a polynomial function. A sample radial lens distortion graph is shown in Fig. 56.5. The usual form for polynomial lens distortion functions is given by

$$\Delta r = k_0r + k_1r^3 + k_2r^5 + k_3r^7 \quad (56.35)$$

in which the radial distance r is the distance from the symmetry point mentioned above. If a distortion table or function is given, the correction should be applied with the opposite sign. Conventionally, “+” indicates radial distortion outward from the principal point. Thus the correction equations would be

$$\begin{aligned} x_c &= x \left(1 - \frac{\Delta r}{r} \right) \\ y_c &= y \left(1 - \frac{\Delta r}{r} \right) \end{aligned} \quad (56.36)$$

Following lens distortion correction the image coordinates should be corrected for atmospheric refraction (if it is significant, i.e., on the order of a micrometer or larger). The expression for radial image displacement due to atmospheric refraction is given by

$$d_r = K \left(r + \frac{r^3}{f^2} \right) \quad (56.37)$$

where r is the radial distance from the principal point of autocollimation and f is the focal length. The value for K is a function of the camera altitude and terrain elevation, and is given according to the ARDC Model Atmosphere (Air Research and Development Command of the U.S. Air Force):

$$K = \left[\frac{2410H}{H^2 - 6H + 250} - \frac{2410h}{h^2 - 6h + 250} \left(\frac{h}{H} \right) \right] \times 10^{-6} \quad (56.38)$$

where H is the flying height in kilometers above sea level and h is the terrain height, also in kilometers above sea level. The displacement due to atmospheric refraction is always radially outward, therefore the correction is always radially inward. The same correction formulas may be used as in the case of lens distortion, replacing Δr by dr . Some practitioners have advocated handling earth curvature effects by modifying the image coordinates. This is to be discouraged. A better solution is to ensure that the object space coordinate system is truly Cartesian, and then the “problem” disappears. See the following section for a discussion of this.

Object Space Coordinate Systems

Geodetic Coordinates ϕ, λ, h

Geodetic coordinates, that is, latitude, longitude, and height, are the most fundamental way to represent the position of a point in space with respect to a terrestrial ellipsoid. However, photogrammetric condition equations are usually expressed in terms of rectangular, Cartesian coordinates. Thus, for the purpose of providing a reference frame for photogrammetric computations, one would usually transform ϕ, λ , and h into a rectangular system.

Space Rectangular Coordinates

Geocentric space rectangular coordinates may be derived from geodetic coordinates in such a way that the Z axis is parallel with the axis of rotation of the ellipsoid, and the X axis passes through the meridian of Greenwich in the plane of the equator. The Y axis is constructed so that the system is right-handed. In the following, h is assumed to be the ellipsoid height; if geoid height is given, it must be modified by the local geoid separation. The equations transforming geodetic coordinates into geocentric space rectangular coordinates are given by

$$\begin{aligned} X &= (N + h) \cos \phi \cos \lambda \\ Y &= (N + h) \cos \phi \sin \lambda \\ Z &= [N(1 - e^2) + h] \sin \phi \end{aligned} \quad (56.39)$$

where N , the radius of curvature in the prime vertical is given by

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}} \quad (56.40)$$

where a = the semimajor axis of the ellipsoid
 b = the semiminor axis
 e = the eccentricity given by

$$e = \sqrt{\frac{(a^2 - b^2)}{a^2}} \quad (56.41)$$

The inverse transformation cannot be given in a closed form. One can solve for ϕ, λ , and h by choosing an initial approximation and proceeding iteratively by the conventional Newton method,

$$X_{i+1} = X_i - J^{-1}F(X_i) \quad (56.42)$$

or

$$\begin{bmatrix} \phi_{i+1} \\ \lambda_{i+1} \\ h_{i+1} \end{bmatrix} = \begin{bmatrix} \phi_i \\ \lambda_i \\ h_i \end{bmatrix} - \begin{bmatrix} \frac{\partial F_1}{\partial \phi} & \frac{\partial F_1}{\partial \lambda} & \frac{\partial F_1}{\partial h} \\ \frac{\partial F_2}{\partial \phi} & \frac{\partial F_2}{\partial \lambda} & \frac{\partial F_2}{\partial h} \\ \frac{\partial F_3}{\partial \phi} & \frac{\partial F_3}{\partial \lambda} & \frac{\partial F_3}{\partial h} \end{bmatrix}^{-1} \begin{bmatrix} F_1(\phi_i, \lambda_i, h_i) \\ F_2(\phi_i, \lambda_i, h_i) \\ F_3(\phi_i, \lambda_i, h_i) \end{bmatrix} \quad (56.43)$$

and the three functions are the ones given in Eq. (56.39). One possible difficulty with the use of geocentric space rectangular coordinates is the large magnitude of the coordinate values. If one wished to maintain point precision to the nearest millimeter, 10 significant digits would have to be carried in the coordinates and single precision floating point computations would be insufficient. An alternative is the local space rectangular system, which is just the geocentric space rectangular coordinates, rotated so that Z' passes through a local point, Y' is in the meridian plane, and X' is constructed for a right-handed system. The LSR, or local space rectangular coordinates, are given by

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix}_{\text{LSR}} = M \left[\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{GSR}} - \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix} \right] \quad (56.44)$$

where GSR refers to the geocentric space rectangular coordinate vector, the T vector is the translation to the local origin, and the rotation matrix M is given by

$$M = \begin{bmatrix} -\sin \lambda & \cos \lambda & 0 \\ -\sin \Phi \cos \lambda & -\sin \Phi \sin \lambda & \cos \Phi \\ \cos \Phi \cos \lambda & \cos \Phi \sin \lambda & \sin \Phi \end{bmatrix} \quad (56.45)$$

Map Projections Coordinates

The common map projections used to express terrestrial control points are the Lambert conformal conic and the transverse mercator. In the U.S., each state has a state plane coordinate system utilizing possibly multiple zones of these projections. Globally, there is the UTM, or universal transverse mercator system, in which the globe is divided into 60 zones of width 6 degrees. Zone 1 is from 180 degrees west to 174 degrees west, and the zone numbering proceeds eastward until the globe is covered at zone 60. The zones are limited to ± 80 degrees latitude, and the scale factor at the central meridian is 0.9996. All of the above map projection coordinates have a common deficiency when used in photogrammetry. The XY coordinate is with respect to the developed projection surface, but the height coordinate is usually with respect to a sea level datum. Thus the system is not Cartesian. Over a very small region one could neglect the curved Z -reference, but over any substantial project area the nonorthogonality of the coordinate system will present itself in photogrammetric computations as a so-called earth curvature effect. The best approach to handling this situation is to either transform all control points into an LSR system described above or construct a local tangent plane system from the map projection coordinates, modifying the height component as follows:

$$h_{tp} = h_{sl} - \frac{D^2}{2} R \quad (56.46)$$

where the subscript tp refers to the tangent plane system, sl refers to the sea level system, D is the distance from a project centered tangent point, and R is the nominal earth radius. Following all photogrammetric computations, that is, block adjustment, the heights can be corrected back into the sea level system for use by compilers and engineers.

56.4 Instruments and Equipment

Stereoscopes

Stereo viewing is possible with the unaided eyes if conjugate imagery is placed at a spacing approximately equal to the eye base, at a comfortable distance in front of the eyes. Prolonged viewing at such a distance may produce eye fatigue and therein lies the value of a stereoscope. A simple *lens stereoscope* allows the eyes to focus comfortably at infinity, thus permitting longer working sessions. For frame photographs, the overlapping pair should be laid out with the flight lines coincident and the spacing adjusted for comfortable viewing. Only a small portion of a standard 23-cm photograph overlap area can be viewed in this way, and some bending of the paper prints may be necessary to access the full model area. A *mirror stereoscope*, being larger, permits viewing of almost an entire overlap area, at a necessarily smaller scale. Approximate elevations can be read via a *parallax bar* and the associated 3-D measuring mark. Some modern softcopy stereo viewing systems employ nothing more than a simple mirror stereoscope to view conjugate imagery presented in split-screen mode on a video monitor.

Monocomparator, Stereocomparator, Point Marker

Both of the comparator instruments have been largely superseded by the analytical plotter, which is really nothing more than a computer-controlled stereocomparator. In any case, a monocomparator is a single two-axis stage with a measuring microscope and a coordinate readout, preferably with an accuracy of 1 or 2 micrometers. A stereocomparator is a pair of two-axis stages which permit stereo viewing by a pair of measuring microscopes, and simultaneous coordinate readout of two pairs of (XY) coordinates. Accuracy levels should be comparable to that mentioned for the monocomparator. Both of these comparator instruments are used chiefly for *aerial triangulation*, *bridging*, or *control extension*. In this process, all *control points* and *pass points* are read for all photographs in a *strip* or *block*. The photos are then linked by geometric condition equations and tied to the ground coordinate system, thus producing ground coordinates for all observed pass points. These pass points may then be used for individual model setups in a stereo restitution instrument. If pass points are desired in an area of the photograph without identifiable detail points, artificial emulsion marks or “pug points” are introduced by a *point marker* or “pug.” These marks are typically 40- to 80-micrometer-diameter drill holes in the photograph emulsion, sized to be compatible with the stereo measuring device.

Stereo Restitution: Analogue, Analytical, Softcopy

Early instruments for map compilation consisted of optical projectors and a small viewing screen with a means to direct the image from one projector to the left eye and from the other projector to the right eye. This binocular separation was effected by *anaglyph* (red and blue filters), by mechanical shutter, and by polarization. Analogue instruments in use today employ exclusively mechanical projection in which a collection of gimbals, space rods, and cardan joints emulate the optical light paths. All analogue instruments must provide a way to re-create the inner camera geometry by positioning the principal point (via the fiducial marks) and setting the principal distance or focal length. These steps constitute the interior orientation. A procedure is also necessary to reestablish the *relative orientation* of the photographs at instant of exposure. This is accomplished by clearing y -*parallax*, or y displacement in model space between the projected images, in at least five points spaced throughout the model. For the point layout in Fig. 56.13 the sequence of steps for two-projector relative orientation is as follows:

1. Clear at point 1 with kappa-right.
2. Clear at point 2 with kappa-left.
3. Clear at point 3 with phi-right.
4. Clear at point 4 with phi-left.
5. Clear at point 5 with omega-left or omega-right.
6. Check for no parallax at point 6.

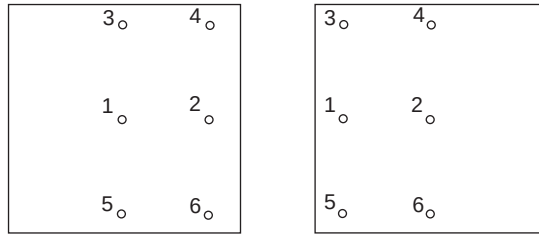


FIGURE 56.13 Point layout for relative orientation.

If there is parallax at point 6, the procedure is repeated until no parallax is seen at point 6. Convergence can be speeded up by overcorrecting by about half at step 5. When this is complete, the entire model should be free of γ -parallax. If there is visible parallax at other points, it could be due to uncompensated lens distortion, excessive film deformation, or other factors. Following relative orientation comes the *absolute orientation*, in which the relation is established between the model coordinates and the ground coordinates, defined by control points in the model. In the past this was done by physically orienting a map manuscript to a mechanical tracing device. This physical procedure would involve scaling, by adjusting the base components, and leveling, by adjusting either common rotation elements or combinations of projector rotations and corresponding base components. Now it is done analytically by computing the parameters of the three-dimensional similarity transformation between the model and ground coordinates. The computed rotations would then be introduced into the instrument as before. This computationally assisted absolute orientation requires that the instrument be fitted with position encoders for coordinate readout of xyz model coordinates. Accuracies on the order of 5 micrometers are typically seen for this task. Schematic depictions of an optical and a mechanical stereo restitution instrument are shown in Fig. 56.14(a) and (b). In addition to map compilation of planimetry and elevation data, an analogue stereo instrument can also be used to collect model coordinates for independent model aerial triangulation. This requires an additional step of determining the model coordinates of the perspective center, which is necessary to link adjacent models in a strip.

All of the functions of an analogue instrument can be duplicated and usually exceeded in an *analytical plotter*. Such a device, shown schematically in Fig. 56.14(c), consists of two computer-controlled stages,

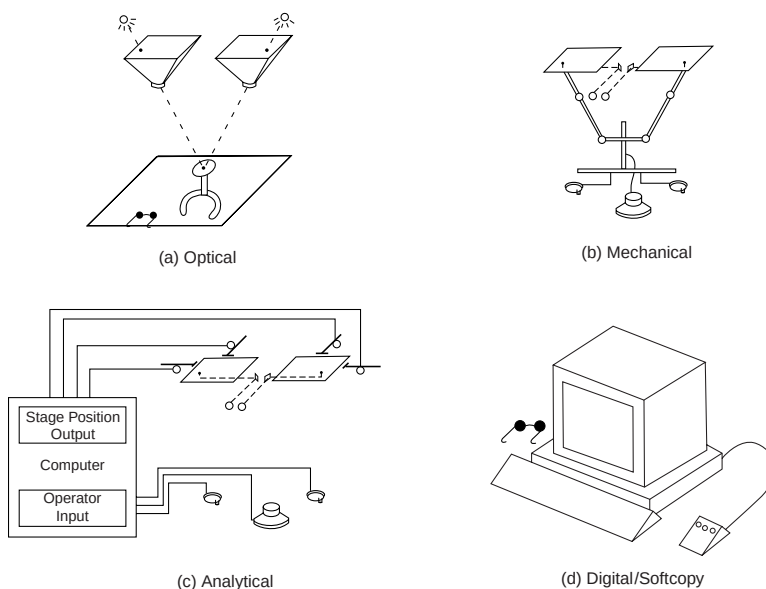


FIGURE 56.14 Schematic diagrams of stereo restitution instruments.

a viewing stereomicroscope, operator controls for three-axis motion, and a suite of computer software to automate and assist in all of the desired operations. Interior orientation consists of measuring the fiducial marks and introducing the calibrated camera parameters. Relative orientation consists of measuring conjugate points and computing the five orientation parameters. Absolute orientation consists of measuring the control points and computing the seven-parameter transformation as above. Of course, the two steps of relative and absolute orientation can be combined in a two-photo bundle solution using the collinearity equations as described previously. In addition to conventional map compilation, analytical plotters are well suited to aerial triangulation and block adjustment, digital elevation model collection, cross section and profile collection, and terrestrial or close-range applications. Stage accuracies are typically 1 or 2 micrometers. Today one would always have a CAD system connected to the instrument for direct digitizing of features into the topographic or GIS database.

The most recent variant on the stereo viewer/plotter is the *softcopy* stereo system. Here the stereo images are presented to the operator on a computer video monitor. This is shown schematically in Fig. 56.14(d). In the two previous cases the input materials were hardcopy film transparencies. In this case the input material is a pair of digital image files. These usually come from digitized photographs, but can also come from sensors which provide digital image data directly, such as SPOT. Softcopy stereo systems present interesting comparisons with hardcopy-based instruments. Spatial resolution may be inferior to that visible in the original hardcopy image, depending on the resolution and performance of the scanner, but possibilities for automation and operator assistance by digital image processing are abundant and are being realized today. In addition, the complicated task of overlaying vector graphics onto the stereo images in a hardcopy instrument becomes a simple task in a softcopy environment. This can be enormously beneficial for editing and completeness checking. The orientation aspect of a softcopy system is very similar to the analytical plotter in that all computations for orientation parameters are done via computer from image coordinate input. The dramatic impact of softcopy systems will not be apparent until specialized image processing tools for feature extraction and height determination are improved to the point that they can reliably replace manual operation for substantial portions of the map compilation task. A few definitions are now presented to encourage standardized terminology. *Digital mapping* refers to the collection of digital map data into a CAD or GIS system. This can be done from any type of stereo device: analogue, analytical, or softcopy. *Digital photogrammetry* refers to any photogrammetric operations performed on digital images. *Softcopy photogrammetry* is really synonymous with digital photogrammetry, with the added connotation of softcopy stereo viewing.

Scanners

With the coming importance of digital photogrammetry, scanners will play an important role in the conversion of hardcopy photograph transparencies into digital form. For aircraft platforms, and therefore for the majority of large scale mapping applications, film-based imaging is still preferred because of the high resolution and the straightforward geometry of non-time-dependent imagery. Thus arises the need for scanning equipment to make this hardcopy imagery available to digital photogrammetric workstations. To really capture all of the information in high-performance film cameras, a pixel size of about 5 micrometers would be needed. Because of large resulting file sizes, many users are settling for pixel sizes of 12 to 30 micrometers. For digital orthophotos, sometimes an even larger size is used. Table 56.1 shows the relation between pixel size and file size for a 230-mm square image assuming no compression and assuming that each pixel is quantized to one of 256 gray levels (8 bits).

There are three main scanner architectures: (1) drum with a point sensor, usually a PMT (photomultiplier tube); (2) flatbed with area sensor; and (3) flatbed with linear sensor. Radiometric response of a scanner is usually set so that the imagery uses as much of the 256 level gray scale as possible. The relation between gray level and image density should be known from system calibration. In some cases, gray values can be remapped so that they are linear with density or transmittance. Most photogrammetric scanners produce color by three passes over the imagery with appropriate color filters. These can be recorded in a band sequential or band interleaved manner as desired. There are a large number of image file formats in use.

TABLE 56.1 File Sizes for Given Pixel Sizes

Pixel Size	File Size, 230-mm Image
5 μm	2.1 Gb
10 μm	530 Mb
15 μm	235 Mb
20 μm	130 Mb
25 μm	85 Mb
50 μm	21 Mb
100 μm	5 Mb

Plotters

Until recently the majority of engineering mapping (as opposed to mass production mapping) has been produced on vector plotters. These produce vector line work on stable base material. They are usually based on a rotating drum for one axis and a moving pen carriage for the other axis. Flatbed designs also exist with a two-axis cross-slide for the pen carriage. These have the additional possibility to handle scribing directly in addition to ink. Electrostatic plotters are essentially raster plotters which may emulate a vector plotter by vector to raster conversion. Digital photogrammetry, with the integration of images and vectors, requires a raster-oriented device. Likewise GIS, which often calls for graphic presentations with area fills or raster layers, may require a raster device.

56.5 Photogrammetric Products

Topographic Maps

The classical product of photogrammetric compilation is a *topographic map*. A topographic map consists typically of planimetric features such as roads, buildings, and waterways, as well as terrain elevation information usually in the form of contours. In the past these were manually drafted in ink or scribed onto scribe coat material. Today they are recorded directly into a CAD system or GIS. U.S. National Map Accuracy Standards (NMAS) dictate an accuracy of planimetric features as well as contour lines, which is tied to hard copy scale. In the digital environment, such standards may need to be revised to reflect the increasingly prominent role of the (scaleless) digital map representation. The following is a summary of the Office of Management and Budget standards:

1. Horizontal accuracy. For maps with publication scale greater than 1:20,000, not more than 10% of the “well-defined” points tested shall be in error by more than 1/30th of an inch at publication scale. For maps with publication scale less than 1:20,000, the corresponding tolerance is 1/50th of an inch.
2. Vertical accuracy. Not more than 10% of the elevations tested shall be in error by more than one-half contour interval. Allowances are made for contour line position errors as above.
3. Any testing of map accuracy should be done by survey systems of a higher order of accuracy than that used for the map compilation.
4. Published maps meeting these accuracy requirements shall note this fact in their legends, as follows: “This map complies with national map accuracy standards.”
5. Published maps whose errors exceed these limits shall omit from their legends any mention of compliance with accuracy standards.
6. When a published map is a considerable enlargement of a map designed for smaller-scale publication, this fact shall be stated in the legend. For example, “This map is an enlargement of a 1:20,000 scale map.”

Other commonly accepted accuracy standards are as follows. Reference grid lines and control point positions should be within 1/100th of an inch of their true position. Ninety percent of spot elevations

should be accurate to within one-fourth contour interval, and the remaining 10% shall not be in error by more than one-half contour interval.

Image Products

Image products from photogrammetry include uncontrolled mosaics, controlled mosaics, rectified enlargements, and orthophotos. Mosaics are collections of adjoining photograph enlargements mated in a way to render the join lines as invisible as possible. In the case of controlled mosaics, the photographs are enlarged in a rectifier, using control points to remove the effects of camera tilt, and to bring all enlargements for the mosaic to a common scale. Uncontrolled mosaics are similar except that no tilt removal is done, enlargement scales are less accurately produced, and continuity is attempted by the careful matching of image features. In the past all mosaicking has been done with paper prints, glue, and considerable manual dexterity. If the photographs are scanned, or if the imagery is originally digital, then the mosaicking process can be entirely digital. Digital techniques allow great flexibility for such tasks as tone matching, scaling, rectifying, and vector/annotation addition. Orthophotos are photographs which have been differentially rectified to remove both tilt displacements as well as relief displacements. A well-produced orthophoto can meet horizontal map accuracy standards, and can function as a planimetric map. Individual orthophotos can be further merged into orthophoto mosaics. Digital techniques for orthophoto production are also becoming very popular because of the flexibility mentioned above. People who are not mapping specialists seem to have a particularly easy time interpreting an orthophoto, compared to an abstract map with point, line, and area symbology that may be unfamiliar. With digital orthophoto generation, it is particularly effective to overlay contour lines on the imagery.

Digital Elevation Models

The concept of digitally recording discrete height points to characterize the topographic surface has been in practice for a number of years. Names and acronyms used to describe this concept include *DTM*, *digital terrain model*; *DEM*, *digital elevation model*; *DHM*, *digital height model*; and *DTED*, *digital terrain elevation data*. The philosophy behind this concept is that one obtains sufficient digital data to describe the terrain, and then generates graphic products such as contours or profiles only as a means for visualizing the terrain. This is in contrast to the conventional practice of recording a contour map and having the contours be the archival record which represents the landforms. The advantage of the DEM approach is that the height database can be used for several different applications such as contour generation, profile/cross section generation, automated road design, and orthophoto interpolation control. Potential pitfalls in this approach mostly revolve around decisions to balance the conflicting demands of accurate terrain representation versus fast data collection and reasonable file sizes. There are basically two alternatives to consider when collecting such data: random data points selected to describe the terrain, or a regular grid of points, with interval selected to describe the terrain.

Random Data Points

Random data points in a DEM may be used directly in a *TIN*, *triangulated irregular network*, or they may be used to interpolate a regular grid via a variety of interpolation methods. The TIN may be created by a number of algorithms which are producing the equivalent *Dirichlet tessellation*, *Thiessen Polygons*, or the *Delauney triangulation*. One of the simpler methods is the basic Watson algorithm, described by the following steps:

1. Create three fictitious points such that the defined triangle includes all of the data points.
2. Pick a point.
3. Find all of the triangles whose “circumcircle” (the circle passing through triangle vertices) contains the point.
4. The union of all triangles in step 3 forms an “insertion polygon.”
5. Destroy all internal edges in the insertion polygon, and connect the current point with all vertices of the polygon.

6. Go to step 2, until no more points are left.
7. When done, eliminate any triangle with a vertex consisting of one of the initial three fictitious points.

To enforce a breakline, one can overlay the breakline on the preliminary TIN, and introduce new triangles as required by the breakline. Interpolation within a TIN usually means locating the required triangular plane facet, and evaluating the plane for Z as a function of X and Y . Contour line generation is particularly easy within the triangular plane facets, with all lines being straight and parallel. Connecting contour lines between facets requires a searching and concatenation operation.

Interpolation in a random point DEM not organized as a TIN can be carried out by various *moving surface* methods, or by *linear prediction*. An example of a moving surface model would be a moving “tilted” plane:

$$z = a_0 + a_1x + a_2y \quad (56.47)$$

One equation is written for each point within a certain radius, possibly with weighting inversely related to distance from desired position, and the three parameters are solved for, thereby allowing an estimate of the interpolated height. Such moving surface models can be higher-order polynomials in two dimensions, as dictated by point density and terrain character. With linear prediction, an elevation is interpolated as follows:

$$z_0 = \sigma_{zz_0}^t \Sigma_{zz}^{-1} z \quad (56.48)$$

where z is an $n \times 1$ vector representing height points in the vicinity of the point to be interpolated, Σ_{zz} is an $n \times n$ matrix of covariances between the reference points, usually based on distance, and σ_{zz_0} is an $n \times 1$ vector representing the covariances between the point to be interpolated and the reference points, again based on distance. Breaklines can be enforced in these methods by not allowing reference points on one side of the break to influence interpolations on the other side of the break.

Gridded Data Points

Points can be easily collected directly in a regular grid by programming an analytical stereo instrument to move in this pattern. Grids may also be “created” by interpolating at grid “posts” using random data as outlined above. Interpolation within a regular grid could be done by the methods outlined for random points, but is more often done by bilinear interpolation. Bilinear interpolation can be done by making two linear interpolations along one axis, followed by a single linear interpolation along the other axis. Alternatively, one could solve uniquely for the following four parameters a, b, c, d at the grid cell corners, and evaluate the equation at the unknown point. x and y can be local grid cell coordinates.

$$z = a + bx + cy + dxy \quad (56.49)$$

Contouring is also relatively straightforward in a gridded DEM by linear interpolation along the edges of each grid cell. An ambiguity may arise when there is the same elevation on all four sides of the grid cell.

Geographic Information Systems and Photogrammetry

Historically, map data collected over the same region for different purposes was typically stored separately with no attempts at registration and integration. Likewise, textual attribute data describing the map features were likely kept in yet another storage location. Increasingly, the trend, particularly for municipalities, is to coregister all of this diverse map and attribute data within a *geographic information system*, or GIS. The photogrammetric process typically plays a very important role in creating the *land base*, or *land fabric*, based on well-defined geodetic control points, to which all other GIS data layers are registered. Photogrammetry also plays a role in making periodic updates to the GIS land base in order to keep it current with new land development, subdivision, or construction. The photogrammetric compiler may

also be involved in collecting facilities features and tagging them with attributes or linking them to existing attribute records in a facilities database.

56.6 Digital Photogrammetry

The term *digital photogrammetry* refers to photogrammetric techniques applied to digital imagery. The current trend in favor of digital photogrammetry has been driven by the enormous increase in computer power and availability, and the advances in image capture and display techniques.

Data Sources

Digital image data for photogrammetry can come from a number of sources. Satellite imagers transmit data directly in digital form. Primary among the satellite sources would be SPOT and Landsat, each with a series of spacecraft and sensors to provide continuous availability of imagery. Digital cameras for airborne or terrestrial use have up to now been a minor contributor as source imagery for photogrammetry. This is principally due to the limited size and density of sensor element packing for area sensors, and to the adverse restitution capabilities with linear sensors using motion-induced scanning. The predominant source of digital imagery for engineering photogrammetry is the scan conversion of conventional film transparencies on high-accuracy scanning systems. Such image data is typically collected as 8-bit (256 level) gray values for monochrome imagery, or as 3×8 -bit red, green, and blue components for color imagery. Monochrome digital images are stored as a *raster*, or rectangular grid format. For fast access they may also be *tiled*, or subdivided into subgrids. Color images may be stored in *band sequential* or *band interleaved* order.

Digital Image Processing Fundamentals

Sampling

Digital image data is generated by sampling from a continuous source of incident radiation. In the case of direct imaging, the source is truly continuous; in the case of sampling from a silver-halide emulsion image, the source is usually of sufficiently higher resolution as to be effectively continuous. The *sampling theorem* states that for a band-limited spatial signal with period T corresponding to the highest frequency present, we must sample it with a period no greater than $T/2$ in order to be able to reconstruct it without error. If there are frequencies present which are higher than can be reconstructed, *aliasing* will occur and may be visible in the reconstructed image as *moiré patterns* or other effects. If this is the case, one should purposely blur or defocus the image until the maximum frequencies present are consistent with the sampling interval. This creates an antialiasing filter. The sampled image will be referred to as $F(x, y)$. The gray values recorded in the digital image may represent estimates of well-defined photometric quantities such as irradiance, density, or transmittance. They may also be a rather arbitrary quantization of the range of irradiance levels reaching the sensor.

Histogram Analysis and Modification

The *histogram* of a digital image is a series of counts of frequency of occurrence of individual gray levels or ranges of gray levels. It corresponds to the probability function or probability density function of a random variable. Any operation on an image, A ,

$$B(x, y) = f[A(x, y)] \quad (56.50)$$

in which the output gray level at a point is a function strictly of the input gray level is referred to as a *point operation* and will result in a modification of the histogram. If an image operation is a function of the position as well as the input gray level,

$$B(x, y) = f[A(x, y), x, y] \quad (56.51)$$

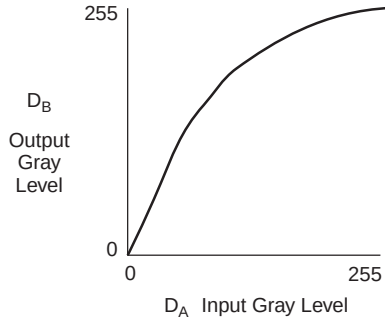


FIGURE 56.15 Gray level remapping.

This may be referred to as a *spatially variant point operation*. It will result in a modification of the histogram that is position dependent. This would be used to correct for a sensor which has a position-dependent response. To illustrate a point operation, function $f(D)$ in Fig. 56.15 maps an input gray level D_A to an output gray level D_B . We assume for simplicity that f is monotone increasing or decreasing, and therefore has an inverse. The output histogram, $H_B(D)$, as a function of the input histogram, $H_A(D)$, and of the function, f , is

$$H_B(D) = \frac{H_A[f^{-1}(D)]}{|df/dD|} \quad (56.52)$$

This is analogous to the transformation of distributions based on a functional relation between random variables. Other common applications of histogram modification include contrast stretching and brightness shifting, that is, gain and offset.

Resampling, Geometric Image Manipulation

Whenever the geometric form of an image requires modification, *resampling* is necessary. This is really just two-dimensional interpolation. Examples of such geometric modifications are rotation, rectification, and epipolar line extraction. If a simple change of scale is desired, magnification can be accomplished by *pixel replication*, and minification can be accomplished by *pixel aggregation*, or more quickly (with aliasing) by *subsampling*. For the resampling task, the fastest method is *nearest neighbor*, which assigns the gray level of the nearest pixel center to the point being interpolated. The most common method would be *bilinear* interpolation. This math model was described under digital elevation models. Less common would be interpolation by higher-order polynomials, such as cubic polynomials. As an example, the process to generate a digital orthophoto can be easily summarized in this context. A grid is defined in object space, which represents the locations of the “output” pixels. The elevation for these points is observed directly or interpolated from a digital terrain model. Each of these (XYZ) object points is passed through the collinearity equations using the known orientation of an image, and image coordinates are obtained. These image coordinates are transformed into the row/column coordinate space of the digital image. If the output spacing is on the order of the digital image pixel spacing, bilinear interpolation can be used to interpolate the orthophoto pixel gray level. If the output spacing is much greater than the image spacing, the output gray value should be obtained by averaging over all image pixels within the larger pixel coverage, thus avoiding aliasing problems.

Filtering

For continuous functions in one dimension, *convolution* is expressed as

$$y(t) = f * g = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau \quad (56.53)$$

For continuous functions in two dimensions, convolution is given by

$$h(x, y) = f * g = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u, v) g(x - u, y - v) du dv \quad (56.54)$$

For discrete two-dimensional functions, that is, images, convolution is given by

$$H(i, j) = F * G = \sum_m \sum_n F(m, n) G(i - m, j - n) \quad (56.55)$$

Image enhancements and feature exaggeration can be introduced by convolving the following 3×3 kernels with an image. For an edge parallel to azimuth 90 degrees,

$$G = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{bmatrix} \quad (56.56)$$

For an edge parallel to azimuth 135 degrees,

$$G = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & 1 \\ -1 & -1 & 1 \end{bmatrix} \quad (56.57)$$

For general edge sharpening via a Laplacian,

$$G = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (56.58)$$

For an edge parallel to azimuth 0 degrees using a Sobel operator,

$$G = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (56.59)$$

For low pass filtering,

$$G = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (56.60)$$

Matching Techniques

Digital image matching represents one of the means by which digital photogrammetry can potentially far exceed the productivity of conventional photogrammetry. Point mensuration and stereo compilation are tasks requiring skilled operators if done manually, and few means are available to speed up the manual process. Point mensuration and stereo “perception” by image matching will increase in speed with each advance in computer technology, including parallel processing. Point mensuration occurs at numerous stages in the photogrammetric restitution process, such as interior orientation (fiducial marks), relative

orientation (parallax points), absolute orientation (signalized control points), and aerial triangulation (pass points and signalized control points). In the case of signalized points, a strategy would be as follows:

1. Derive a rotation-independent detection template for the target type. Convolve this template with the image(s) under study, and, by thresholding, obtain the approximate locations of the points.
2. With a fine-pointing template estimate the target position in the image, while simultaneously modeling rotation, scale, affinity, and radiometry.

The search criterion can be the maximization of a correlation coefficient, or the minimization of a sum of squares of residuals. The correlation coefficient is given by

$$C_{uv} = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{[\sum(u_i - \bar{u})^2 \sum(v_i - \bar{v})^2]^{1/2}} \quad (56.61)$$

where u and v represent image and template, or vice versa, and

$$\bar{u} = \frac{\sum u_i}{N} \quad \text{and} \quad \bar{v} = \frac{\sum v_i}{N} \quad (56.62)$$

The other significant application area of digital image matching is digital elevation model extraction. Three techniques will be described for this task: (1) vertical line locus, (2) least squares matching, and (3) epipolar matching.

Vertical Line Locus

Vertical line locus (VLL) is used to estimate the elevation at a single point, appearing in two photographs. An initial estimate is required of the object space elevation of the point to be estimated. A “search range” is then established in the Z -dimension, extending above and below the initial estimate. This search range is then subdivided into “test levels.” At each test level, a matrix of points is defined surrounding the point to be estimated, all in a horizontal plane. All points in this matrix are then projected back into each of the two photographs, and via interpolation in the digital image, a corresponding matrix of gray levels is determined for each photograph, at each level. When the test level most nearly coincides with the actual terrain elevation at the point, the match between the pair of gray level matrices should be the maximum. Some measure of this match is computed, often the correlation coefficient, and the elevation corresponding to the peak in the match function is the estimated elevation. Variants on this procedure involving an iterative strategy of progressively finer elevation intervals, variable sized matrix, and “matrix shaping” based on estimated terrain slope can all be implemented, yielding more accurate results at the expense of more computing effort.

Least Squares Matching

In the usual least squares matching (LSM) approach, one assumes that the two images are the same except for an affine geometric relationship and a radiometric relationship with a gain and offset. One takes the gray values from one image of the pair to be observations and then computes geometric and radiometric transformation parameters in order to minimize the sum of squares of the discrepancies (residuals) with the second image. In the following equations, g will represent the first image and h will represent the second image. The affine geometric relationship is defined by

$$x_h = a_1 x_g + a_2 y_g + a_3 \quad (56.63)$$

$$y_h = b_1 x_g + b_2 y_g + b_3$$

The linearized equation relating the gray levels is

$$g(x, y) = h(x, y) + \frac{\partial h}{\partial x} dx + \frac{\partial h}{\partial y} dy + h(x, y) dk_1 + dk_2 \quad (56.64)$$

Taking differentials of Eq. (56.63),

$$dx_h = x_g da_1 + y_g da_2 + da_3 \quad (56.65)$$

$$dy_h = x_g db_1 + y_g db_2 + db_3$$

making the following substitution for compact notation:

$$h_x = \frac{\partial h(x, y)}{\partial x} \quad \text{and} \quad h_y = \frac{\partial h(x, y)}{\partial y} \quad (56.66)$$

adding a residual to g , and substituting the differentials into Eq. (56.64), we obtain the condition equation to be used for the least squares estimation,

$$g + v_g - h - h_x x_g da_1 - h_x y_g da_2 - h_x da_3 - h_y x_g db_1 - h_y y_g db_2 - h_y db_3 - h dk_1 - dk_2 = 0 \quad (56.67)$$

In matrix form,

$$v + B\Delta = f \quad (56.68)$$

the condition equation becomes

$$v_g + \begin{bmatrix} -h_x x_g & -h_x y_g & -h_x & -h_y x_g & -h_y y_g & -h_y & -h & -1 \end{bmatrix} \begin{bmatrix} da_1 \\ da_2 \\ da_3 \\ db_1 \\ db_2 \\ db_3 \\ dk_1 \\ dk_2 \end{bmatrix} = h - g \quad (56.69)$$

One such equation may be written for each pixel in the area surrounding the point to be matched. The usual solution of the least squares problem yields the parameter estimates, in particular the shift parameters to yield the “conjugate” image point. If the parameters are not small, it may be necessary to resample for a new “shaped” image, h , and solve repeated iterations until the parameter estimates are sufficiently small.

Epipolar Matching

Both of the previous techniques involved matching areas, albeit small ones. In epipolar matching, the two matched signals are one-dimensional. Also of interest is the degree to which these methods make use of a priori knowledge of the image orientation. In the VLL technique we make use of this information. In LSM we do not. In epipolar matching this information is used.

In theory, the use of this information should further restrict the solution space and thereby yield faster solutions. In the epipolar technique, one takes the two planes defined by the two images to be matched, and a third plane defined by the two perspective centers and intersecting the two photograph planes. This third plane is referred to as an epipolar plane (there is an infinite number of them). The intersection of this epipolar plane with each of the photograph planes defines two lines, one in each of the photographs. If the orientation is correct, it is guaranteed that each point on one of the lines has a conjugate point on the other line. Thus to search for this conjugate point one needs only to search in one dimension. The practical steps necessary to implement this technique would be as follows:

1. For each photograph, determine two points in object space which lie in the epipolar plane.
2. For each photograph, project the two corresponding points into the photograph plane and solve for the parameters defining the resulting line in image space.
3. Resample each digital image at an appropriate interval along these lines.
4. Determine a match point interval in image space. Select points in one image at this spacing along the epipolar line.
5. For each match point in one image, bracket it by some pixels on either side of it. Then search the epipolar line in the other image for the best match “element.” The match criterion is usually the above-mentioned cross-correlation function, but it can be any objective function which measures the degree of “sameness” between the elements.
6. Create in this way an irregularly spaced line of points in object space.
7. Rotate the epipolar plane about the photograph baseline by a small amount, and repeat the process. This will create an irregular grid of XYZ points in object space.
8. Use these points directly to form a TIN, or interpolate a regular grid as described earlier.

Up to now, digital image matching has been most effective and reliable when used on small and medium scale imagery. Large scale images, showing the fine structure of the terrain along with individual trees, buildings, and other man-made objects, may contain steep slopes and vertical planes (i.e., not a smooth, continuous surface). This generally interferes with the simple matching algorithms described here. For large scale engineering mapping, more research needs to be done to develop more robust matching methods.

56.7 Photogrammetric Project Planning

Project planning usually starts with analysis of the specifications for the maps or image products to be produced. Requirements for these final products generally place fairly rigid constraints on the choices available to the planner. The constraints arise because the specified accuracies must be achieved and on the other hand the process must be economical.

Flight Planning

For conventional topographic mapping in the U.S., a compilation system (camera, field survey system, stereoplotter, operators, etc.) is often characterized by a *C-factor*, used as follows:

$$\text{Flying height} = C\text{-factor} \times \text{Contour interval} \quad (56.70)$$

This *C-factor* can be in the range of 150 to 250 depending on the quality of the equipment and the skill of the personnel doing the compilation. Thus if one knows from the specifications that a 1-m contour interval is required and the *C-factor* is 2000, the maximum flying height (above the terrain) would be 2000 m. One could always be conservative and fly at a lower height, resulting in more photographs at a larger scale. Table 56.2 gives a series of representative parameters for some of the traditional mapping scales used in the U.S. Conventional forward overlap between successive photographs is 60%. Conventional side overlap is 30%. In Table 56.2 the enlargement factor is the factor from negative scale to map scale. Figure 56.16 shows the geometry for determining the base, *B*, or forward gain, from the ground dimension of the photograph coverage, *W*, and the forward overlap as a fraction of *W* (60% is a common value). The following equation gives *B* as described (*OL* = 0.6, the forward overlap fraction):

$$B = (1 - OL) \times W \quad (56.71)$$

The photograph coverage, *W*, is related to the actual photograph dimension, *w*, by the scale

$$W = \text{Scale} \times w \quad (56.72)$$

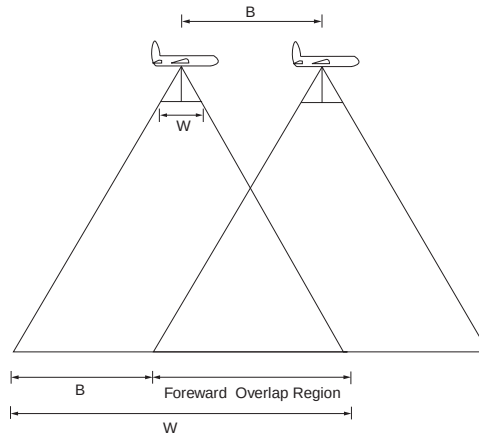


FIGURE 56.16 Forward overlap geometry.

TABLE 56.2 Typical Flight Parameters (Length Units: Feet)

Map Scale	Ratio	Cont. Intvl.	Neg. Scale	Engl. Factor	H	C-Factor	W	Fwd. Gain	Flt. Line Spc.
1" = 50'	600	1	4,200	7 ×	2,100	2100	3,150	1,260	2,205
1" = 100'	1,200	2	7,800	6.5 ×	3,900	1950	5,850	2,340	4,095
1" = 200'	2,400	5	12,000	5 ×	6,000	1200	9,000	3,600	6,300
1" = 400'	4,800	5	16,800	3.5 ×	8,400	1680	12,600	5,040	8,820
1" = 1000'	12,000	10	30,000	2.5 ×	15,000	1500	22,500	9,000	15,750
1" = 2000'	24,000	10	384,000	1.6 ×	19,200	1920	28,800	11,520	20,160
1" = 4000'	48,000	20	57,600	1.2 ×	28,800	1440	43,200	17,280	30,240

The photograph dimension, w , is approximately 23 cm for standard aerial film. An analogous situation is shown in Fig. 56.17 for determining S , the distance between flight lines, from W and from the side overlap as a fraction of W (30% is a common value). The following equation gives S ($SL = 0.3$, the side overlap fraction):

$$S = (1 - SL) \times W \quad (56.73)$$

Control Points

Horizontal control points (XY or NE) allow proper scaling of the model data collected from the stereo photographs. Vertical control points (Z or h) allow the proper terrain slopes to be preserved. Classical surveying techniques determined these two classes of control separately and the distinction is still made. Even with GPS-derived control points, which are inherently 3-D, the elevation or Z -dimension is often disregarded because we lack an adequate geoid undulation map to convert ellipsoid heights to orthometric (sea level) referenced heights. Control points may be targeted or signalized before the photo flight so that they are visible in the images, or one may use *photo-ID* or natural points such as manhole covers or sidewalk corners. Coordinates are determined by field survey either before or after the flight, though usually after in the case of *photo-ID* points. If artificial targets are used, the center panel (at image scale) should be modestly larger than the measuring mark in the stereocomparator or plotter. There should also be two or more prominent legs radiating from the target center panel to allow easy identification in the photographs. Each stereo model requires an absolute minimum of two horizontal control points and three vertical control points. These could all be established by field techniques but this would usually be too expensive. More commonly a sparser network of control points is established, and the needed control in between the field control is determined by bridging or aerial triangulation. When aerial triangulation

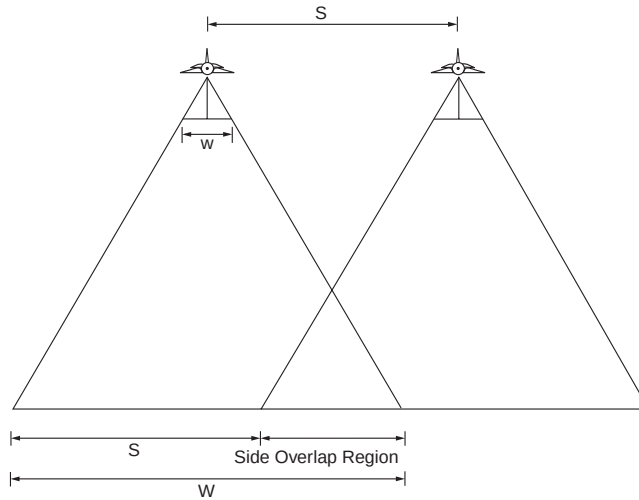


FIGURE 56.17 Side overlap geometry.

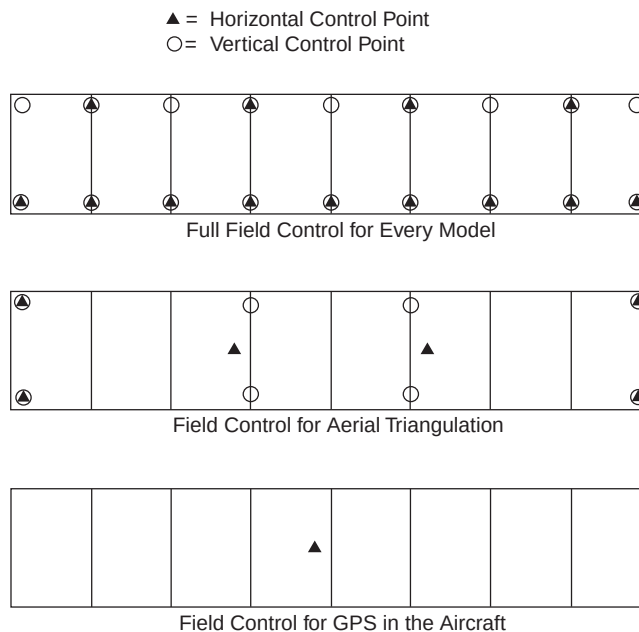


FIGURE 56.18 Control point distributions.

is used, artificial marking or pugging of the film diapositives is necessary to create points for the operator to use for absolute orientation. When GPS is used in the aircraft to determine exposure station coordinates independently, then in theory, assuming one has photo strips crossing at near right angles, no ground control is needed. Without crossing strips, at least one control point is necessary. In practice the best number to use will probably fall between 1 and the larger number necessary for ground control-based aerial triangulation. These control point requirements are shown in [Fig. 56.18](#).

56.8 Close-Range Metrology

Close range metrology using photogrammetry offers some unique advantages and presents some equally unique challenges. The advantages are (1) it is a noncontact measurement method, (2) photography can be acquired relatively quickly in dangerous or contaminated areas, (3) the photographs represent an archival record in case any dimensions are disputed, and (4) the camera can be brought to the object, rather than the other way around (as with a coordinate measuring machine). Accuracies obtainable using close-range photogrammetry (as in aerial photogrammetry) would fall in the range of 1/5,000 to 1/100,000 of the object distance.

Equipment

Cameras are typically small format, usually with 70-mm film size. Stereometric cameras are mounted rigidly in pairs, so that fewer parameters are required in solving for the orientations. The cameras used would preferably be metric, although with very special handling, and much extra effort, nonmetric cameras can be used for some applications. There should be at least some sort of fiducial marks in the focal plane, preferably a reseau grid covering the image area. Film flatness can be a severe problem in nonmetric cameras. Ideally there should be a vacuum back permitting the film to be flattened during exposure. There should be detents in the focus settings to prevent accidental movement or slippage in the setting. Calibration will be carried out at each of these focus settings, resulting in a set of principal distances and lens distortion curves for each setting. These are often computed using added parameters in the bundle adjustment, performed by the user, rather than sending the camera to a testing laboratory. Lighting can be a problem in close-range photogrammetry. For some objects and surfaces which have very little detail and texture, some sort of “structured lighting” is used to create a texture. To obtain strong geometry for the best coordinate accuracy, highly convergent photography is often used. This may prevent conventional stereo viewing, leading to all photo observations being made in monoscopic view. This introduces an additional requirement for good targetting, and can make advantageous the use of feature equations rather than strictly point equations.

Applications

Close-range photogrammetry has been successfully used for tasks such as mapping of complex piping systems, shape determination for parabolic antennas, mating verification for ship hull sections, architectural/restoration work, accident reconstruction, and numerous medical/dental applications.

56.9 Remote Sensing

Remote sensing is considered here in its broad sense, including the photogrammetric aspects of using nonframe imagery from spaceborne or airborne platforms. A thorough treatment must include the metric aspects of the image geometry and the interpretive and statistical aspects of the data available from these sources.

Data Sources

Following is a partial listing of sensors, with associated platform and image descriptions. These sensors provide imagery which could be used to support projects in civil engineering.

1. MSS, multispectral scanner, Landsat 1-5, altitude 920 km, rotating mirror, telescope focal length 0.826 m, IFOV (instantaneous field of view) 83 m on the ground, gray levels 64, image width 2700 pixels, 4 spectral bands: 0.4–1.0 micrometers
2. TM, thematic mapper, Landsat 4-5, altitude 705 km, rotating mirror, telescope focal length 1.22 m, IFOV 30 m on the ground, gray levels 256, image width 6000 pixels, 6 spectral bands: 0.4–0.9, 1.5–1.7, 2.1–2.4, 10.4–12.5 micrometers

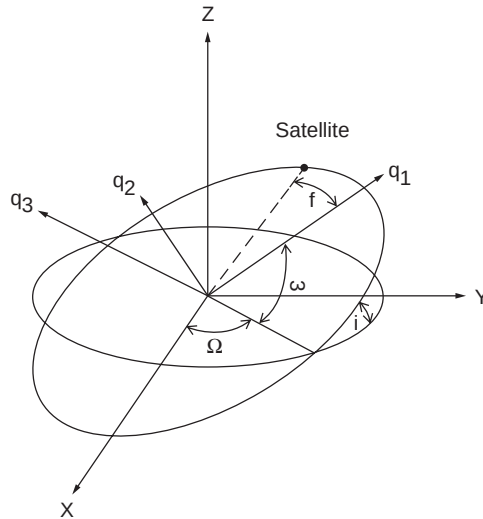


FIGURE 56.19 Orbital parameters.

3. SPOT, SPOT 1-3, panchromatic (multispectral not described), altitude 822 km, pushbroom, telescope focal length 1.082, IFOV 10 m on the ground, gray levels 256, image width 6000 pixels, 1 panchromatic band 0.55–0.75 micrometers plus 3 multispectral bands, off-nadir viewing capability for stereo

Geometric Modeling

Coordinate Systems

The two primary coordinate systems needed for describing elliptical orbits as occupied by imaging sensor platforms are the XYZ system and the $q_1q_2q_3$ system shown in Fig. 56.19. The XYZ system has the XY plane defined by the earth equatorial plane, Z oriented with the spin axis of the earth, and with X through the vernal equinox or the first point of Aries. The second system has q_1q_2 in the orbit plane with origin at one focus of the ellipse, as shown in Fig. 56.19. q_1 is along the semimajor axis; q_2q_3 are as described above to define a right-handed coordinate system.

Orbital Mechanics

The gravitational attraction between the earth and an orbiting satellite is

$$F = \frac{GMm}{r^2} \quad (56.74)$$

where m = the mass of the satellite
 r = the geocentric distance to the satellite
 GM = the earth's gravitational constant:

$$GM = 3,986,004 \times 10^8 \text{ m}^3 \text{ s}^{-2}$$

For a small mass satellite,

$$\mu = GMm \approx GM \quad (56.75)$$

Expressing the position vector in 3-D coordinates and the position in the orbital plane in polar coordinates, extracting an expression for the acceleration, and further manipulating the expressions yields a differential equation whose solution is a general conic section in polar coordinates:

$$r = \frac{p}{1 + e \cos(f)} \quad (56.76)$$

where f is the true anomaly as shown in Fig. 56.19, p is the semilatus rectum, and e is the eccentricity. For $e > 1$ the orbit is a hyperbola, for $e = 1$ the orbit is a parabola, and for $e < 1$ the orbit is an ellipse. The use of the term *anomaly* as a synonym for angle is a vestige of Ptolemaic misconceptions, originally implying an angle which did not vary linearly with time. The true anomaly, f , the eccentric anomaly, E , and the mean anomaly, M , are related by Kepler's equation:

$$M = E - e \sin E \quad (56.77)$$

and

$$\tan\left(\frac{f}{2}\right) = \sqrt{\frac{1+e}{1-e}} \tan\left(\frac{E}{2}\right) \quad (56.78)$$

The point on an elliptical orbit furthest from the earth is the *apogee*; the point closest is the *perigee*. At perigee the true anomaly is zero; at apogee it is 180 degrees. An orbit is specified when the satellite position and velocity vectors, $\mathbf{r}$ and $\dot{\mathbf{r}}$ are given at a particular time. A more common set of six parameters to specify an orbit are the classical Keplerian elements $\{a, e, T_0, \Omega, i, \text{and } \omega\}$, where a is the semimajor axis, e is the eccentricity, T_0 is the time of perigee passage, Ω is the right ascension of the ascending node (shown in Fig. 56.19), i is the inclination, and ω is the argument of the perigee.

In order to obtain the Kepler elements from the position and velocity vectors at a given time, the following equations can be used. The angular momentum vector is

$$\mathbf{h} = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} = \mathbf{r} \times \dot{\mathbf{r}} \quad (56.79)$$

This yields

$$\Omega = \tan^{-1}\left(\frac{h_x}{-h_y}\right) \quad (56.80)$$

$$i = \tan^{-1}\left(\frac{\sqrt{h_x^2 + h_y^2}}{h_z}\right) \quad (56.81)$$

With v representing the velocity,

$$a = \frac{r}{2 - (rv^2/\mu)} \quad (56.82)$$

$$e = \sqrt{1 - \frac{h^2}{\mu a}} \quad (56.83)$$

Obtaining the eccentric anomaly from

$$\cos E = \frac{a-r}{ae} \quad (56.84)$$

the true anomaly is found from

$$f = \tan^{-1} \left(\frac{\sqrt{1-e^2} \sin E}{\cos E - e} \right) \quad (56.85)$$

Defining an intermediate coordinate system in the orbit plane such that p_1 is along the nodal line,

$$p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = R_1(i)R_3(\Omega)r \quad (56.86)$$

$$\omega + f = \tan^{-1} \left(\frac{p_2}{p_1} \right) \quad (56.87)$$

T_0 can be obtained from

$$M = \sqrt{\frac{\mu}{a^3}}(t - T_0) \quad (56.88)$$

To go in the other direction, from the Kepler elements to the position and velocity vectors, begin with Eq. (56.88) to obtain the mean anomaly, then solve Kepler's equation, Eq. (56.77), numerically for the eccentric anomaly, E . The true anomaly, f , is found by Eq. (56.78). The magnitude of the position vector in the orbit plane is

$$r = \frac{a(1-e^2)}{1+e \cos(f)} \quad (56.89)$$

the position vector is

$$r = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} = \begin{pmatrix} r \cos(f) \\ r \sin(f) \\ 0 \end{pmatrix} \quad (56.90)$$

and the velocity vector is

$$v = \sqrt{\frac{\mu}{p}} \begin{pmatrix} -\sin(f) \\ e + \cos(f) \\ 0 \end{pmatrix} \quad (56.91)$$

where p , the semilatus rectum, is given by

$$p = a(1-e^2) \quad (56.92)$$

The rotation matrix relating the orbit plane system and the vernal equinox, spin-axis system is given by

$$r_{XYZ} = R r_{q_1 q_2 q_3} \quad (56.93)$$

where R is

$$R = R_3(-\Omega) R_1(-i) R_3(-\omega) \quad (56.94)$$

This R can be used to transform both the position and velocity vectors above into the XYZ “right ascension” coordinate system defined above.

Platform and Sensor Modeling

If we consider a linear sensor on an orbiting platform, with the sensor oriented such that the long dimension is perpendicular to the direction of motion (i.e., pushbroom), we can construct the imaging equations in the following way. Each line of imagery, corresponding to one integration period of the linear sensor, can be considered a separate perspective image, having only one dimension. Each line would have its own perspective center and orientation parameters. Thus, what may appear to be a static image frame is, in fact, a mosaic of many tiny “framelets.” These time dependencies within the image “frame” are the result of normal platform translational velocity, angular velocity, and additional small, unpredictable velocity components. The instantaneous position and orientation of a platform coordinate system is shown in Fig. 56.20.

The equations which relate an object point, $(X_m, Y_m, Z_m)^t$, and the time-varying perspective center, $(X_s(t), Y_s(t), Z_s(t))^t$ are of the form

$$\begin{bmatrix} 0 \\ y \\ -f \end{bmatrix} = k \times R_{III} \times R_{II} \times R_I \times \begin{bmatrix} X_m - X_s(t) \\ Y_m - Y_s(t) \\ Z_m - Z_s(t) \end{bmatrix} \quad (56.95)$$

R_{III} is a platform to sensor rotation matrix which should be fixed during a normal frame. This could accommodate off-nadir view angle options, for instance. R_{II} transforms the ideal platform to the actual

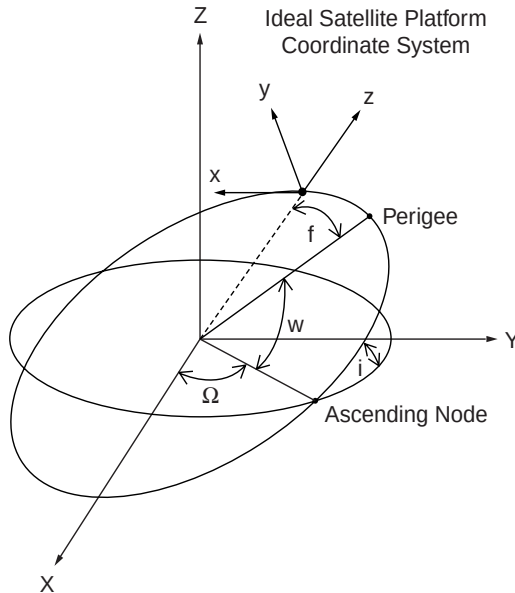


FIGURE 56.20 Imaging satellite orientation.

platform and the time dependency in its angles would probably take the form of low degree polynomial functions. R_I represents the time-dependent rotation from a space-fixed geocentric coordinate system to the instantaneous platform coordinate system. In a manner similar to the previous section, this matrix would be the composition of the rotations,

$$R_I = R_z(\pi) \times R_y\left(\frac{\pi}{2} - \omega - f(t)\right) \times R_x\left(i - \frac{\pi}{2}\right) \times R_z(\Omega) \quad (56.96)$$

with the orbit elements as described in the previous section. The Cartesian coordinates on the right side of Eq. (56.95) are with respect to a space-fixed coordinate system (i.e., not rotating with the earth). In order to transform a point in a terrestrial, earth-fixed coordinate system, a coordinate rotation, R_0 , is necessary to account for the earth's rotation, polar motion, precession, and nutation. For $(X_e, Y_e, Z_e)^t$ in an earth-fixed system,

$$\begin{bmatrix} X_m \\ Y_m \\ Z_m \end{bmatrix} = R_0 \times \begin{bmatrix} X_e \\ Y_e \\ Z_e \end{bmatrix} \quad (56.97)$$

Analogous to the decomposition of conventional orientation in photogrammetry into exterior and interior components, we could construct here a decomposition of the imaging Eq. (56.95) into a platform component and a sensor component. Eliminating the nuisance scale parameter in Eq. (56.95) would yield two equations per object point, per image “frame.” Unknown parameters would consist of a user-selected subset of the platform and sensor orientation parameters, some time varying, the orbit parameters, and the object point parameters. These equation systems could be formed for a single image (resection), a set of sequential images along the orbit path (strip), or an arbitrary collection of adjacent frames (block).

The imminent arrival of commercially available satellite data in the 1- to 3-meter pixel range means that this kind of image modeling will become even more important in the future as digital sensors continue to slowly supplant film-based photography.

Interpretive Remote Sensing

Remote sensing, as the term is usually employed, implies the study of images to identify the objects and features in them, rather than merely to determine their size or position. In this way, remote sensing is a natural successor to the activity described by the slightly outdated and more restrictive term *photo interpretation*. Remote sensing almost always implies digital data sources and processing techniques. It often implies multiple sensors in distinct spectral bands, all simultaneously imaging the same field of view. Along with the multispectral concept, there is no restriction that sensors respond only within the range of visible radiation. Indeed, many remote sensing tasks rely on the availability of a wide range of spectral data. Witness the recent move toward “hyperspectral” data, which may have hundreds of distinct, though narrow, spectral bands.

Systems have been deployed with sensitivities in the ultraviolet, visible, near infrared, thermal infrared, and microwave frequencies. Microwave systems are unique in that they can be *active* (providing source radiation as well as detection), whereas all of the others listed are strictly *passive* (detection only). Active microwave systems include both SLAR, side-looking airborne radar, and synthetic aperture radar.

Multispectral Analysis

Multispectral image data can be thought of as a series of coregistered image planes, each representing the scene reflectance in a discrete waveband. At each pixel location, the values from the spectral bands constitute an n -dimensional data vector for that location. Consider an example with two spectral bands in which the scene has been classified a priori into three categories: water (W), agricultural land (A),

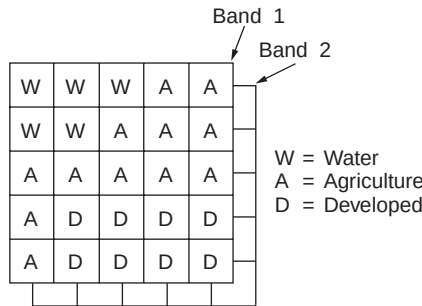


FIGURE 56.21 Training sample for supervised classification.

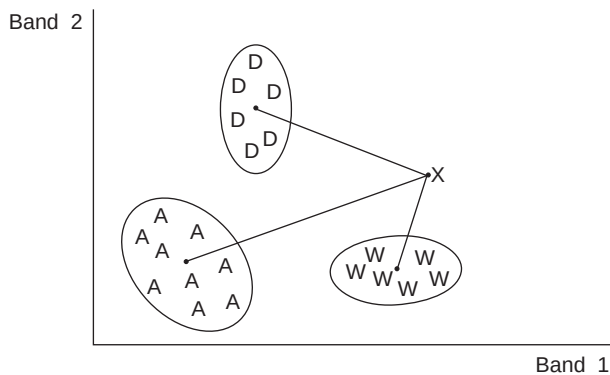


FIGURE 56.22 Clustering of 2 band data in scatter diagram.

and developed land (D), as shown in Fig. 56.21. From this *training sample* one can (assuming normality) construct a mean vector and a variance/covariance matrix for each class. New observation vectors could then be assigned to one of the defined classes by selecting the minimum of the Mahalanobis function,

$$D^2 = (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) + \ln |\Sigma_i| \quad (56.98)$$

where μ_i and Σ_i are evaluated for each category. Clustering of feature classes is shown in Fig. 56.22. This is an example of a maximum likelihood classifier, using *supervised* classification. Other techniques can be used such as discriminant functions and Bayesian classification. In addition to statistical and heuristic approaches to pattern recognition, one can also employ syntactic methods, wherein a feature is identified by its context or relationship to adjacent or surrounding features.

Change Detection

In order to detect changes on a portion of the earth's surface it is necessary to have imagery at different times. In addition, when possible, it is desirable to minimize apparent, but spurious, differences due to season, view altitude, and time of day. In order to make a proper pixel-by-pixel comparison, both images should be resampled and brought to the same map projection surface. This is best done by using identifiable ground control points (road intersections, etc.) and a modeling of the sensor and platform positions and orientations. Less effective registration is sometimes done using two-dimensional polynomial functions and a "rubber sheet" transformation to the control points. Once the scenes have been brought to a common coordinate base, a classification is made individually on each scene. A pixel-by-pixel comparison is made on the classification results, and a "change" layer can be generated where differences are encountered.

Microwave Remote Sensing

Microwave radiation has the useful characteristic of penetrating clouds and other weather conditions which are opaque to visible wavelengths. Aircraft-based imaging radars are thus able to acquire imagery under less restrictive flight conditions than, for example, conventional photography. The geometry of the radar image is fundamentally different from the near perspective geometry of other sensor systems. In both real aperture side-looking airborne radar and synthetic aperture radar, imagery is presented as a succession of scan lines perpendicular to the flight direction. The features displayed along the scan lines have positions proportional to either the slant range or the ground range, and the image density or gray level is related to the strength of the return. Radar imagery can be acquired of the same area from two flight trajectories, inducing height-related parallax into the resulting stereo pair. With the proper imaging equations, stereo restitution and feature compilation can be carried out, giving rise to the term *radargrammetry*.

Until recently, remote sensing activities were the exclusive domain of a few highly industrialized countries. Now a number of countries and even some commercial ventures are becoming involved in systems to provide remote sensing imagery. For the civil engineering community, this increased availability of data can be very beneficial.

References

- Burnside, C.D. 1985. *Mapping from Aerial Photographs*. John Wiley & Sons, New York.
- Colwell, R.N., ed. 1983. *Manual of Remote Sensing*. American Society of Photogrammetry and Remote Sensing, Bethesda, MD.
- Ebner, H., Fritsch, D., and Heipke, C. 1991. *Digital Photogrammetric Systems*. Wichman, Karlsruhe.
- Faugeras, O. 1993. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, MA.
- Karara, H.M., ed. 1989. *Non-Topographic Photogrammetry*, 2nd ed. American Society of Photogrammetry and Remote Sensing, Bethesda, MD.
- Kraus, K. 1993. *Photogrammetry*. Dummler Verlag, Bonn.
- Leick, A. 1990. *GPS Satellite Surveying*. John Wiley & Sons, New York.
- Moffitt, F.H., and Mikhail, E.M. 1980. *Photogrammetry*, 3rd ed. Harper & Row, New York.
- Pease, C.B. 1991. *Satellite Imaging Instruments*. Ellis Horwood, Chichester.
- Richards, J.A. 1993. *Remote Sensing Digital Image Analysis: An Introduction*, 2nd ed. Springer-Verlag, New York.
- Slama, C.C., ed. 1980. *Manual of Photogrammetry*, 4th ed. American Society of Photogrammetry and Remote Sensing, Bethesda, MD.
- Wolf, P.R. 1983. *Elements of Photogrammetry*. McGraw-Hill, New York.

Further Information

The following journals are useful sources of information:

- Photogrammetric Engineering and Remote Sensing*, American Society of Photogrammetry and Remote Sensing, Bethesda, MD.
- The Photogrammetric Record*, The Photogrammetric Society, London.
- ISPRS Journal of Photogrammetry and Remote Sensing* (formerly *Photogrammetria*), Elsevier Science Publishers B.V., Amsterdam, The Netherlands.
- CISM Journal* (formerly *Canadian Surveyor*), Canadian Institute of Surveying and Mapping, Ottawa, Canada.
- Journal of Surveying Engineering*, American Society of Civil Engineers, New York.

The following organizations provide valuable technical and reference data:

- American Society of Photogrammetry and Remote Sensing, 5410 Grosvenor Lane, Suite 210, Bethesda, MD 20814
- American Congress on Surveying and Mapping, 5410 Grosvenor Lane, Bethesda, MD 20814

AM/FM International, 14456 E. Evans Ave., Aurora, CO 80014

U.S. Geological Survey, EROS Data Center, Sioux Falls, SD 57198

U.S. Geological Survey, Earth Science Information Center, 53. National Center, Reston, VA 22092

SPIE, The International Society for Optical Engineering, P.O. Box 10, Bellingham, WA 98227

American Society of Civil Engineers, 345 East 47th Street, New York, NY

57

Geographic Information Systems

57.1 Introduction

Background • Applications

57.2 Geographic Information Components

Geometry (Graphics) • Attributes (Nongraphics) • Vector • Raster • Topology

57.3 Modeling Geographic Information

Layer-Based Approaches • Relational Approaches • Object-Oriented Approaches • Extended Relational Model • Security and Information Sharing in a GIS

57.4 Building and Maintaining a GIS

Reference Coordinate Systems • Consideration of Scale • Data Sources • Data Entry and Processing • Structure/Topology • Maintenance Operations

57.5 Spatial Analysis

Database Operations • Coupling to External Analyses/Applications • Data Interchange Standards and Formats

57.6 Information Extraction

Displays and Reporting • Spatial Query Languages

57.7 Applications

Basemap and Infrastructure in Government • Facilities Management • Development Tools (Means to Customize/Build New Applications)

57.8 Summary

Jolyon D. Thurgood

Leica, Inc.

J.S. Bethel

Purdue University

57.1 Introduction

Background

Information about our world has been depicted on maps of various forms for many centuries. During the golden age of exploration maps showed critical paths of navigation in the known world, as well as strategic political boundaries and information about settlements and natural resources. Over the past 300 years the art of cartography has been complemented by the development of scientific methods of surveying and related technologies, which have enabled increasingly accurate and complete representation of both physical and cultural features.

Most recently, computer-related advances have led to a revolution in the handling of geographic information.

First of all, raw spatial information can be gathered and processed much more efficiently and quickly, based on technologies such as analytical and digital photogrammetry, Global Positioning System (GPS), and satellite remote sensing.

Second, it has become possible to automate drafting and map production techniques to replace manual drafting procedures.

Third, instead of providing simply a graphical representation through paper maps, it has become possible to model the real world in a much more structured fashion, and to use that spatial model as the basis for comprehensive and timely analyses. Based on this model, it is possible to geographically reference critical data generated by government agencies and private enterprises, and using information modeling and management techniques, to query large amounts of spatially integrated data, and to do so across multiple departments and users, thereby sharing common elements.

The result of this revolution is that there has been a very rapid growth in the production and manipulation of geographic information, to the extent that many organizations can fulfill their production and operational goals only through the use of a geographic information system (GIS).

A geographic information system may be defined as an integrated system designed to collect, manage, and manipulate information in a spatial context. The geographic component, the various technologies involved, and the approach to information modeling set a GIS apart from other types of information systems. A geographic information system provides an abstract model of the real world, stored and maintained in a computerized system of files and databases in such a way as to facilitate recording, management, analysis, and reporting of information. It can be more broadly stated that a geographic information system consists of a set of software, hardware, processes, and organization that integrates the value of spatial data. Various authors provide more detailed definitions of geographic information systems [Antenucci et al., 1991; Dueker, 1987; Parker, 1988].

Early automated mapping systems used interactive computer graphics to generate, display, and edit cartographic elements using computer-aided drafting (CAD) techniques, more or less emulating the manual processes previously used. Over the past decade, more advanced techniques designed to more comprehensively integrate geometric (graphic) elements with associated nongraphic elements (attributes) and designed specifically for map-based and geographic data have resulted in more powerful and flexible implementation of GIS. Continuous mapping, in which a seamless geographic database system replaces map sheets or arbitrary facets earlier used, and the manipulation of geographic information as spatial objects or features are two aspects of most recent geographic information systems that provide users with more intuitive and realistic models of the real world. Also, the integration of vector-based graphics, imagery, and other cell-based information has provided increasingly powerful visualization and analysis capabilities.

Applications

At a broad range of scales, maps have become increasingly important as legal documents that convey land ownership and jurisdictional boundaries, as tools to support decision making (for example, in urban planning), and as a means of visualizing multiple levels of information on political, social, and ecological issues, for example, in thematic mapping of demographic data.

It is estimated that typically 70 to 80% of information maintained by government agencies may be geographically referenced. In addition to directly specifying spatial location on basemap information, such elements as taxpayer identifier, home-owner address, phone numbers, and parcel numbers may be used as the spatial key. Perhaps for this reason, GIS is often seen as the means to promote information sharing and more efficient information management and maintenance, and as a key to providing better and more timely services in a competitive environment. In addition, GIS applications are often both graphics- and database-intensive and provide strong visualization capabilities. The GIS offers the power

to process large amounts of various types of information, but also to present results in a powerful graphical medium: The most common standard product of a GIS is for the time being still the printed map, but it is likely to be a cartographic product customized for a specific task or analysis, as opposed to a standard map series product.

In general, a GIS can provide the following information on geographic elements or features: location, characteristics, logical and geometric relationships with other features, and dependencies on other features. This information can generally be used as the basis for tabular reports, standard and custom map output plots, spatial decision support, trend analysis, as well as output to other potential users and analyses. A geographic information system may be accessed from a single PC, a local area network (LAN) of UNIX workstations, or through a virtual, wide area network (WAN) of distributed information.

Standard or common components of a GIS that enable full implementation of such tasks include drafting, data entry, polygon processing and network analysis, spatial querying, and application development tools (macro language, programming libraries).

From earliest times, maps have been used to establish land ownership. One of the first application areas for modern GIS has been in the area of property ownership and records. Within a municipality, the assessor's office or appraisal district is normally responsible for the identification, listing, and appraisal of parcels of real estate and personal property. A GIS provides real benefits to such an office by allowing accurate and complete appraisals, based on access not just to property attributes such as lot size and building square footage, but also to spatial information, such as the comparison of similar properties within a neighborhood. Once the complete map base has been established in digital form and linked to the nongraphic attribute database system, such tasks as property transactions and applications for building permits can be performed efficiently and without a lengthy manual, and often bureaucratic, delay.

Such a parcel-based land information system can provide the basis for a much more sophisticated GIS. For example, within an urban environment, various boundaries define school, library, fire department, sewer and water supply districts, special business zones, and other special tax assessment districts. The allocation of real estate taxes for a given property may be determined by overlaying all of these special districts with property boundaries. Done manually, it is a cumbersome process, and one that makes redistricting — that is, changing the boundaries of any of the constituent districts — a complex process. Polygon processing within a GIS provides the means to perform such an overlay and to determine very quickly how the various tax components apply to one or many properties. The same function can be used to provide answers to discussions regarding proposed changes to these districts, for example, to examine the impact on a city's tax base by annexing an adjacent unincorporated business region. The GIS therefore offers benefits in two areas — first in new capabilities, and second in its ability to produce results in a timely manner: two months of visual inspection and transcription can be replaced by one hour of computer time.

Another key application area is one based on linear networks, such as those defining transportation routes, or an electricity distribution network. In the area of transportation the GIS provides the ability to model individual road elements and intersections and to analyze routes between any two points within an urban street network. Such a network trace can be used in conjunction with emergency services planning to identify the shortest path to a hospital or to examine the average response time to a call to the fire department. By extending the GIS data structure to incorporate one-way streets, turn restrictions in a downtown area, and rush-hour speed statistics, a sophisticated, multipurpose model of the transportation network may be derived. This model can be designed and optimized exclusively for emergency response activities or for planning purposes only — for example, to examine commuting patterns and traffic congestion projections.

The geographic information system provides the ability to completely model utility networks, such as those supplying water, power, and telecommunications to large numbers of consumers. Such a system may operate at a variety of scales, modeling service connections to consumers, service districts, as well as detailed facilities inventories and layouts, such as transformers, valves, conduits, and schematic diagrams.

The GIS then becomes a key element at many levels: in customer support (to respond to service failure), in maintenance and daily operations (to identify work requirements and assess inventories), and in planning (to respond to projected needs). It provides the link between many information systems, including engineering, planning, and customer billing, which can increase overall performance and operational efficiency.

These simple examples identify the key elements of a geographic information system: a base model that identifies spatial features and spatial relationships, a set of descriptors that can be used to discriminate and identify individual elements, and a set of functional processes and tools that operate against all information components. This structure is also shown in Fig. 57.1. Typical bases for application areas are shown in Fig. 57.2.

GIS Overview

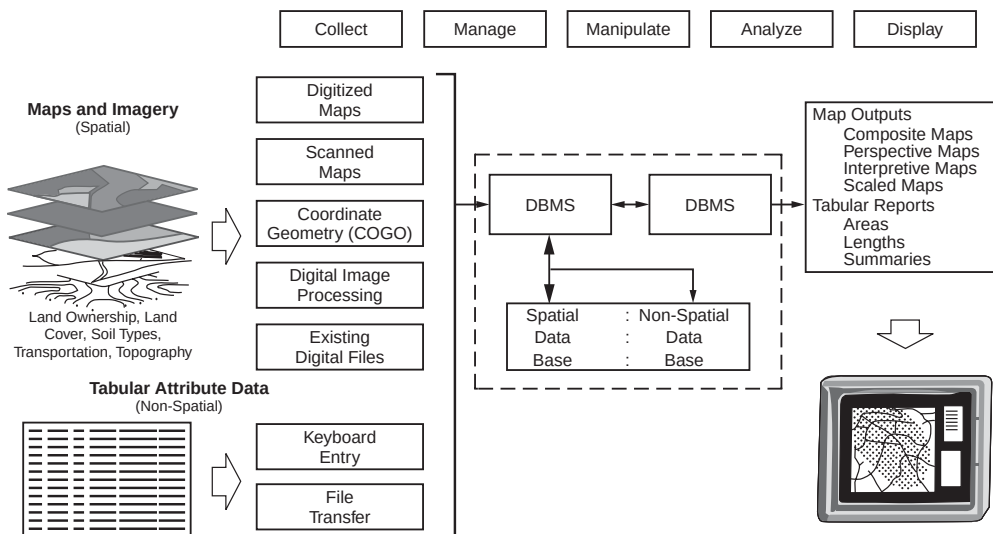


FIGURE 57.1 Overview of a geographic information system (GIS).

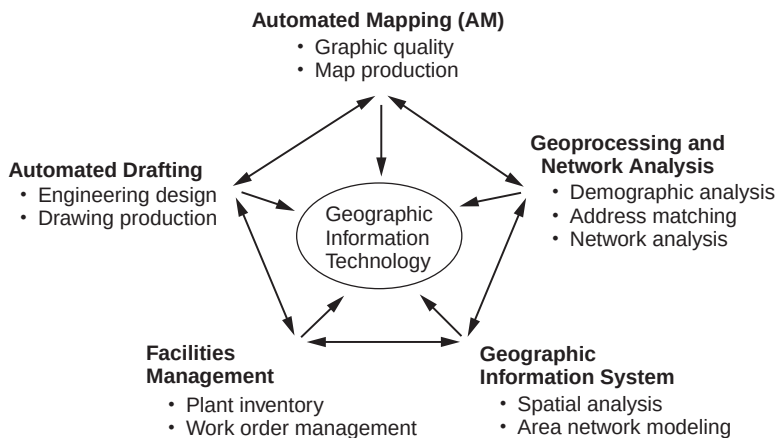


FIGURE 57.2 Application areas based on geographic information technology. (Source: Antenucci et al. 1991. *Geographic Information Systems: A Guide to the Technology*. Van Nostrand Reinhold, New York.)

Key	Type	Usage
absolute location	X, Y(Z) coordinates	survey monuments
relative location	distance, bearing	property boundaries
parcel identifier	alphanumeric	land ownership transactions
street address	alphanumeric	ownership
street segment ID	alphanumeric	traffic engineering
manhole ID	alphanumeric	water/wastewater utilities
zip code	numeric	demographics
telephone number	numeric	pizza delivery

FIGURE 57.3 Keys to geographic information.

57.2 Geographic Information Components

Geography provides a reference for a very large amount of information commonly stored and maintained in information systems. In many cases this reference or key is not necessarily absolute position or a set of geographic coordinates, but an indirect key to location such as street address, parcel or property identifier, phone number, or taxpayer number. One of the distinguishing properties of a geographic information system is the ability to tie such keys to a common geographic base, such as a basemap containing streets, property information, and so on. Some common keys for geographic information are identified in [Fig. 57.3](#).

In fact, the geographic information system is often seen as a focal point for various types of graphical and nongraphical data collections.

Initially, we can look at these components separately.

Geometry (Graphics)

As previously discussed, GIS has generally evolved from computer graphics systems that allowed the graphical representation conventionally depicted on map products to be modeled as layers that can be displayed, edited, and otherwise manipulated by means of specialized software. Today's CAD systems still provide the same type of structure. In such a GIS, graphical elements forming a logical grouping or association are stored on distinct layers or even in separate files. A final graphical display or map output is formed by switching on or off the appropriate layers of information and assigning to each layer a predefined cartographic representation designed for the scale of map or specific application. The symbology or line style to be used is traditionally stored with the layer definition, although it is also normally possible to define special representations for specific graphical elements.

Spatial location is typically stored directly or indirectly within the graphical component of a geographic information system. In earlier systems a complete GIS project stretching across many map sheet boundaries would be stored still in the form of distinct tiles or facets, each representing a drawing with its original spatial extent. Absolute spatial location in a reference coordinate system, such as a state plane coordinate system, would be obtained by interpreting for each drawing part a local transformation (a two-dimensional conformal transformation typically) applied to drawing coordinates. A librarian system would allow such transformations to be applied transparently to the human operator or indeed to applications interested only in absolute spatial location.

More recent systems provide a more seamless continuous map base, where geometric components are stored in a single reference coordinate system that models the real-world representation of geographic features. Any trimming or splitting of the database to map sheet boundaries or other artificial tiling systems is more typically completely hidden from anyone but the project manager or system administrator, as depicted in [Fig. 57.4](#).

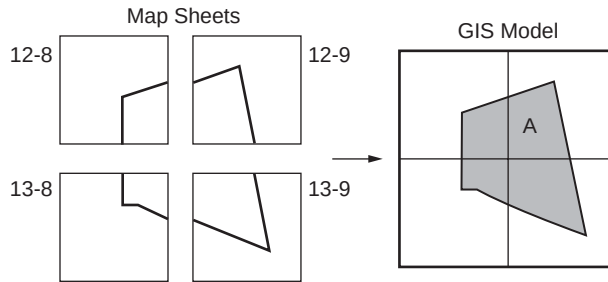


FIGURE 57.4 Map sheets in a GIS. A diagram showing how map sheets stored in separate data files are referenced within a geographic information system in order to model a continuous spatial extent. Polygon A can be accessed as a single geographic feature, regardless of where individual segments of boundary lines are stored.

Attribute Type	Typical Size / Storage Requirements	GIS Feature Example
short integer	2 bytes	land use class
long integer	4 bytes	land value
floating point real	4 bytes	area
double precision real	8 bytes	centroid coordinate
fixed length text	32 bytes	street name
variable length text	80 bytes per line	property description
bulk	30 kbytes for compressed 256x256	scanned image of building
date	6 bytes	land purchase date

FIGURE 57.5 Typical attribute types.

Attributes (Nongraphics)

As already mentioned, the critical distinguishing property of a GIS lies in its ability to relate various types of information within a spatial context. Such a context is provided directly through absolute or relative location or through nongraphic characteristics, also referred to as attributes. Such attributes may act as direct or indirect keys that allow the analysis of otherwise unrelated sets of information.

We should also distinguish between internally defined attributes or keys that typically maintain the link between the geometry component and the nongraphic database system. This key is often the weakest link within a GIS, especially if geometry and attributes are stored and maintained in distinct file or database systems.

Attribute information is typically held in a relational database management system. GIS applications typically allow the retrieval of nongraphic and geographic information in a linked fashion. For example, it is possible to point to a specific location in a graphical map display, to select and identify an individual spatial feature, and to retrieve or update attribute information relating to this feature. Conversely, it is possible to select spatial elements on the basis of attribute matching and to display the selection in a graphical form such as a thematic map. A variety of attribute types are commonly used, as shown in [Fig. 57.5](#).

Vector

The tremendous growth in comprehensive geographic information systems over the past decade reflects the implementation of vector-based GIS software systems that can handle a variety of point, line, and

Vector Element	Parameters	Typical Size / Storage Requirements	GIS Feature Example
node / point	(x,y)	16 bytes (double precision floating point)	centroid of parcel
line	(x ₁ ,y ₁), (x ₂ ,y ₂)	32 bytes	service connection between house and utility main
line string / polyline	(x ₁ ,y ₁), ..., (x _n ,y _n)	20 points require 320 bytes	stream
circular arc / circle	(x ₀ ,y ₀), radius,θ _{start} ,θ _{end}	40 bytes per arc	curved portions of roadway
spline	B ₀ ,B ₁ ,B ₂ ,B ₃ coefficients for each segment	20 segments require 640 bytes	contour line
complex / composite line	combinations of the above	combinations of the above	oddly-shaped parcel boundary
polygon / surface	(x ₁ ,y ₁), ..., (x _n ,y _n)	4-sided polygon requires 64 bytes	boundary of municipal service district
Triangulated Irregular Network (TIN) element	(x ₁ ,y ₁ ,y ₁), (x ₂ ,y ₂ ,z ₂), (x ₃ ,y ₃ ,z ₃)	500 triangles require 3600 bytes	digital terrain model (DTM)

FIGURE 57.6 Vector data components (2-D, 2.5-D applications).

polygon geometries. Based on these structures, as in [Fig. 57.6](#), all geometric information previously depicted on hard copy maps could be modeled, stored, and manipulated. Although in earlier systems it was common to hold a third dimension (the Z coordinate or elevation) only as a nongraphic attribute of the geometric entity, geometric structures in a modern GIS typically hold all three coordinates for each primitive component. The following geometric structures, also shown in [Fig. 57.5](#), are normally available:

1. Point or node elements, each containing X, Y or X, Y, Z coordinates.
2. Line elements, containing or referencing beginning and end nodes, along with a set of intermediate points, sometimes referred to as shape points. In addition to straight-line connections between shape points, other primitive line types may be modeled, including circular arcs defined by three successive points or parametrically (including complete circles) or B-spline connections between all points in a single entity. Such flexibility in line primitives allows more efficient modeling of certain types of spatial features: for example, the outline of a building would be defined by digitizing corner points (vertices) only; a street centerline may be defined using the appropriate parametric geometry, whereas a digitized contour line may be held for cartographic purposes with a B-spline connection.
3. Polygons, surface or area primitives, are closed polyline elements used to represent enclosed areas of the earth's surface, such as property boundaries, lakes, and so on. A GIS typically allows for the formation of polygon boundaries based on one or more line primitives and also maintains special information about "islands" or "holes" within the enclosed area. Such structures allow the appropriate modeling of, for example, a pond on an island on a large inland lake. As this example might indicate, more important than the geometric primitives themselves is the overall data model, including vector-based topology between elements.

Raster

Some of the earliest computer-based GISs were based on grid-cell information. Before the definition of more complex structures and the availability of computing power to support their processing, the simplest way to perform a spatial analysis between multiple layers of information was to subdivide each layer into a grid of small cells, with an associated numeric cell value which denoted a class or set value representing the characteristic of that location. By performing simple Boolean operations between the grids representing each characteristic, useful results could be obtained.

At the same time that raster-based GIS products have been incorporating more vector graphics and database structures, vector-based products have been providing support for raster information. It is now possible to display scanned engineering drawings, maps, digital orthophoto products, and other raster-based imagery, coregistered with vector map information in a common geographic reference system. This can be seen as part of a broader trend to incorporate both vector- and raster-based structures in a single software system where spatial analyses may be performed using the most appropriate technique. For example, cartographic modeling based on raster data allows the simple analysis of such properties as adjacency and proximity, but processing requirements increase according to the spatial precision or resolution required. Future software is likely to allow rule-based conversion of information (from raster to vector, and vector to raster) prior to manipulation, dependent on the specific output requirements.

Topology

The power of a geographic information system lies not only in the data that are held within the system, but also in the data model that provides the fundamental structure or framework for the data. A basic component of a vector-based GIS is a set of topologic structures that allows the appropriate modeling of points, lines, and polygons.

Start- and end-node points of linear elements are often stored as distinct structures, thereby allowing the analysis of adjacent or connected entities through such logical connections, as opposed to simply a graphical operation. For example, two lines representing road segments that cross each other may intersect at an intersection point stored explicitly as an intersection, as opposed to an overpass.

Arc-node topology allows the formation of geometric structures that model elements such as a street centerline that contains straight-line segments, circular, and spiral curves. In addition, aspects such as tangency conditions between connected line elements are also considered a part of the topology. During geometric processing, such as interactive editing, all topology properties would normally be retained or at least restored after processing.

Polygonal geometry stores its own set of topology, including the element of closure between beginning and end points of an enclosing boundary. References to islands or holes within the boundary are also maintained as part of the topology for that feature.

Software tools to create and maintain polygon- and linear-based topology are common required elements of a GIS.

In a polygon coverage — such as a continuous coverage of soils, land use, or similar — it is often of value to store such area classifications with line-polygon topology that allows the analysis of shared boundaries. In such a structure, linear elements that form boundaries between adjacent polygons are referenced by both polygons. Such a structure also allows common geometry to be stored only once, allowing efficient storage and maintenance of related information. In a modern GIS it is possible for single geometric elements to reference not just two areas within a single classification (for example, adjacent areas of soil), but any real-world elements that refer to the same geometry — for example, a property boundary that also forms part of the boundary of municipal jurisdictions (tax districts, school districts) as well as forming a right-of-way boundary.

Another important topologic structure is the composite or complex feature, which allows the grouping of logically related graphical and nongraphic elements. This extends the power of a GIS beyond simple relationships between a single spatial element and one set of attributes to a more sophisticated modeling

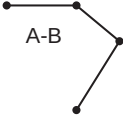
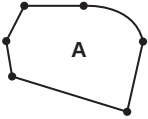
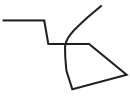
Topologic Type	Description	Properties / Operations	Graphical Representation
Node / point	discrete x,y(z) identifying unique feature	position / proximity	
Line	straight-line (two points), arcs / circles, splines	continuity / connectivity nodes at all intersections and at endpoints	
Polygon / surface	series of connected line segments	areal closure / adjacency	
Spaghetti	sequence of connected points, with no topologic checking	length (line segments may cross/intersect each other)	

FIGURE 57.7 Node-line-polygon topology.

of real-world elements. Through the use of composite features, a logical spatial feature may be defined — such as a school district composed of a school district boundary, a set of school facilities (classrooms, playing fields, administrative offices), school bus routes, and residential catchment areas — that permits more sophisticated modeling of spatial elements.

More sophisticated topologic structures involve those that allow the formation of a single seamless geographic base structure, where graphic or geometric components cross tile or district boundaries that are stored in distinct files or database tables, but referenced. In such systems users may access large or small spatial features.

Topologic structures such as those identified in Fig. 57.7 form one of the most critical aspects of a GIS since they determine how efficiently certain operations or analyses can be performed. (Network analysis depends on connectivity between line elements; polygon analysis requires handling of closed polylines, islands, and so on.)

57.3 Modeling Geographic Information

In this section various methods of modeling spatial information are described. The method chosen has broad implications on the scope and application of GIS.

Layer-Based Approaches

The traditional method of classifying information in a geographic information system derived from the ability to graphically distinguish various layers or levels of data, also affected by the practical limitations of available computing capacity (for example, limits of 256 layers or 32 colors in a palette). Compared to previous means of producing hard copy maps, though, these restrictions did not prevent reasonable modeling of geographic data. Rather, it allowed the storage, maintenance, and manipulation of many more levels of information than previously possible through manual drafting or overlay means.

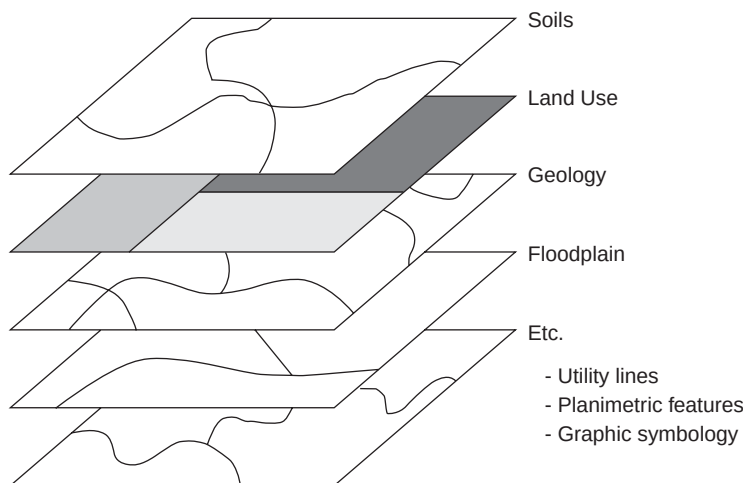


FIGURE 57.8 Layers in a geographic information system. (Source: ESRI. 1992. *ARC/INFO: GIS Today and Tomorrow*. ESRI, Redlands, CA.)

In a layer-based approach, the ability to distinguish different types of information relies on a common definition and usage of layers, the schema or data dictionary. Such a schema, agreed upon by all potential users, defines that graphic primitives (lines, arcs, point symbols, text labels) common to a well-defined set of spatial features are stored in separate and distinct layers. The layer specifications determine how information once collected can be used, since in a graphics system the layer may be the only level at which logical data may be segregated. For example, text labels annotating parcel numbers may be stored in layer 25, and parcel boundaries themselves (the graphical primitives) in layer 24: By separating text and pure graphics, it is then possible to display either graphics, text, or both. In practice, separate layers of information may in fact reside in separate files on disk, physically as well as logically distinct. The layer-based model is closely analogous to logical map overlays or use of color separates as used in map production. Figure 57.8 shows typical graphical layers in a GIS.

Relational Approaches

As a first step in the evolution from a purely graphical system, it was recognized that links to additional nongraphic attributes in a GIS would radically increase the utility and power of such an information system. Although early GISs provided an ability to graphically model and depict cartographic information (much in the same way that paper maps had done previously), such systems became extremely cumbersome and limiting when trying to really apply the power of computers to selective retrieval and analysis of spatial information. This resulted in the development and introduction during the 1980s of a GIS with two distinct components: the graphical database or file system, and an associated nongraphic database system. Early systems used proprietary database structures to store and reference nongraphic information, but the more rapid growth and acceptance of GIS in the past few years has come with the use of standard commercial relational database management systems (RDBMS). Spatial information systems using such an approach are also referred to as *geo-relational* or *all-relational* GIS, the latter term denoting that at least some of the geographic or graphic components are also stored in a relational model.

The simplest relational model used in GIS consists of a graphics-based component that carries with each geometric primitive (point, line, or polygon) a unique internal identifier or tag, which is the means to associate the geometry with additional nongraphic information defining the characteristics of the spatial feature. The geometry identifier is therefore the common key upon which the power of the relational GIS depends. Although the use of such identifiers is often hidden from the casual user, its role is critical: the means that a GIS software system uses to establish and maintain this vital link between graphics and nongraphics determines the practical potential of the system.

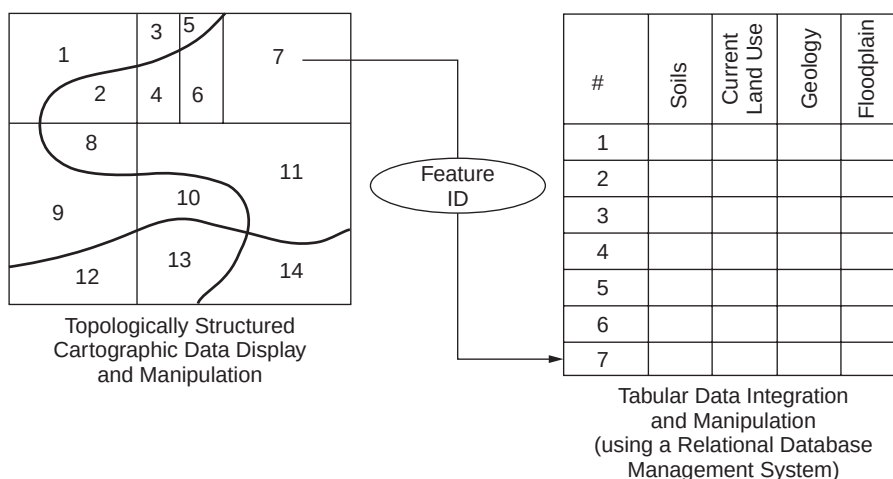


FIGURE 57.9 The geo-relational model. (Source: ESRI. 1992. *ARC/INFO: GIS Today and Tomorrow*. ESRI, Redlands, CA.)

Once linked using the relational model, commonly used and widely understood techniques and tools may be applied in all areas of spatial data management, including data collection, update of nongraphic information, extraction, and reporting. Apart from the key link between geometry (graphics) and primary attribute (nongraphics) tables, it becomes a simple matter to extend the model through additional tables and keys. Common elements in such tables may be used as additional keys through table joins to incorporate the spatial context to a variety of nongraphic information. For example, property information including property addresses can be joined with a table containing address and telephone number in order to provide direct retrieval of phone numbers based on property and location. The relational model can also be used to associate sets of geographic features with common characteristics, such as those properties lying in a specific school district, thereby extending the spatial and logical models as appropriate.

In addition, certain relational terminology can be applied directly in the spatial domain: for example, a polygon overlay can be viewed simply as a “spatial join.”

With the acceptance of standards within the RDBMS industry and with extensions of commercial products to provide transparent access and manipulation of information in a broadly distributed network of computing platforms, the geo-relational model has fit well for those GIS projects seeking to play a key role in a multidepartmental or enterprise-wide information technology environment. In such situations the GIS project manager is often happy to take advantage of these standard commercial products and in turn is able to concentrate on issues of data management specific to the geographic nature of the information.

As already mentioned, it is also possible to model geometry and topology using relational database technology. In such GIS, sometimes referred to as all-relational, such information as start and end nodes, pointers between line entities, and polygon elements may be stored in the form of relational tables. A purely relational approach sometimes adds tremendous computing overhead for simple operations in a GIS (for example, geometric editing of polygonal areas) and may require additional levels of information to be stored in a hybrid fashion to improve performance. Figure 57.9 shows the geo-relational model.

Object-Oriented Approaches

Object-oriented approaches in data modeling were applied early on in CAD products, in such areas as construction and manufacturing. Such a model allowed an architect to apply object-oriented rules to form walls with certain properties and optional or mandatory components (for example, windows and doors). Once a certain type of (standard) window had been modeled, it could be introduced and reused wherever appropriate.

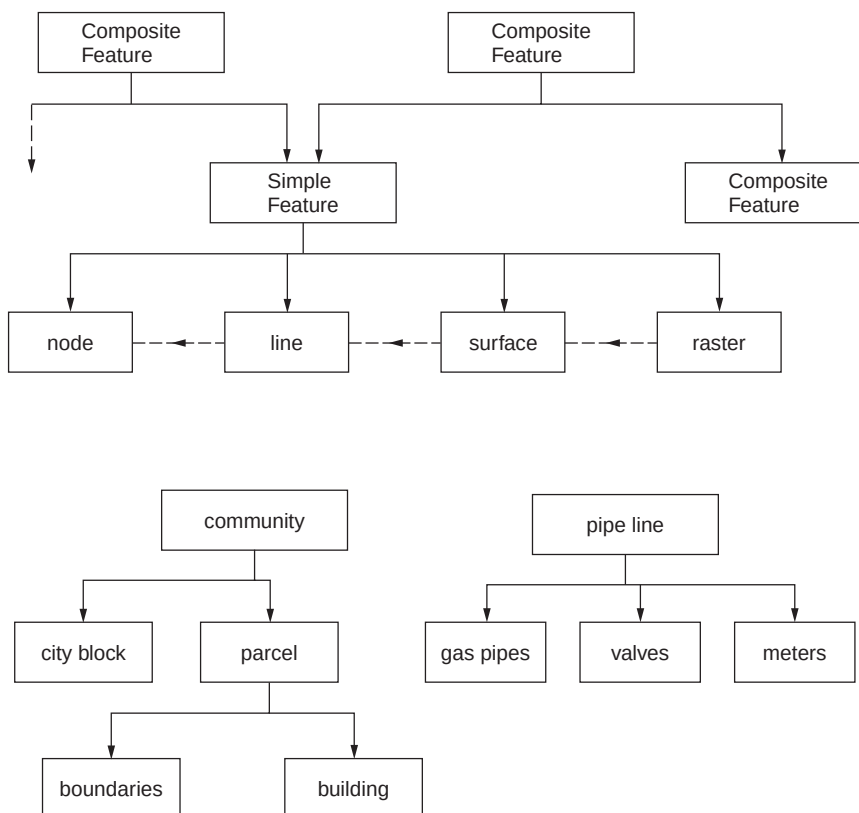


FIGURE 57.10 A GIS data model.

Over the past few years, object-oriented techniques have been applied to GISs in at least two major areas: (1) spatial data modeling, and (2) spatial analysis or programming tools.

In any geographic information system the database and related software attempt to create an abstract model of the real world (as in Fig. 57.10) containing spatially dispersed information. The closer this abstract comes to modeling true characteristics of the physical and human-made world, the more powerful can be the application of computing power to interpret and provide decision support based on the geographic information. We discussed in a previous section the various components of spatial information: An object-oriented approach to spatial data modeling allows all related components to be encapsulated in a single object or feature definition, along with rules that govern how the object may be manipulated or how components may be related or dependent on components of other objects. The terms *feature oriented* or *feature based* are also used to denote software systems in which the modeled entity is intended to closely reflect the real-world element or object. In such an example a land-ownership record may be treated as a single spatial object or feature containing one or more boundary lines with surface (polygon) topology, additional nongraphic characteristics, and logical associations. Relationships based on common boundaries (logical adjacency) or common ownership may be accommodated. The object-oriented paradigm allows for the modeling of sophisticated groupings of geographic and logically related elements, such as a “census tract” composed of a collection of “parcel” objects, or a “gas distribution network” composed of various subelements such as “pipes” and “valves.” Once such elements are defined, the data dictionary contains not just a list of element classes, but the rule base for their use and application. This permits, at least in theory, the population of highly structured geographic data sets in which many real-world relationships and constraints are retained.

Many GIS software products now claim to be object oriented or feature oriented. To a certain degree, such GISs have introduced levels of object-oriented user interfaces or modeling techniques, which may mask a layer-based or fully relational model that the system uses internally. For the time being at least,

there are few GIS products that claim to be fully object oriented in both data modeling and data storage: This situation will probably change as object-oriented database management systems (OODBMS) become more widespread.

Extended Relational Model

Although the object-oriented paradigm carries many advantages — especially in areas of accurately modeling real-world phenomena or features and ability to develop or customize geo-related applications based on reusable blocks of code — the relational model carries with it many positive aspects when considering implementation of GIS projects. These relate to existing (or legacy) information systems that are to be somehow incorporated into or integrated with geographic information, the flexibility and increasing power of commercial RDBMS in handling large amounts of distributed data, and the use of industry-standard products and techniques in accessing and manipulating such information. The use of *structured query language* (SQL) and products such as Oracle and Informix has become widespread in large and small organizations, to the extent that large commercial companies may choose to standardize on the use of a specific RDBMS product. Commercial database products are constantly being extended to address such areas as multimedia (additional and custom data types), the object-component paradigm, as well as specific requirements for spatial data. In this way the growth of OODBMS is being countered by the major RDBMS vendors, many of whom are in the best position to package and offer the best of familiar (relational) technology with additional features addressing these more specialized requirements.

The geographic information system places a considerable burden on data management, from the aspects of both modeling real geographic features and the various types of spatially distributed data. A GIS is often seen as providing the common interface or natural reference system to which attribute, vector, raster, video, sound, and other multimedia data are to be attached. Such an information system was beyond the scope of purely relational database management technology a few years ago, but now several RDBMS provide the facilities to develop such support in an extended relational model.

Security and Information Sharing in a GIS

Early GIS projects were restricted to single users or single departments where data were gathered and processed by at most a small group of individuals. In such implementations little attention was paid to the role of making information secure or of handling multiple user transactions.

However, in many organizations today, the GIS project implementation carries a significant weight as a means of integrating or linking multiple agencies and departments based on (spatial) elements of common interest or value. The general problems of data duplication, inconsistencies, and inaccuracy associated with uncontrolled access to all information within the database system can no longer be ignored. In fact, the introduction of multiuser security to spatially related data elements allows the full benefit of a GIS.

First of all, the layers or classes of information stored in a GIS are typically of interest to more than one user or group of users. However, one user or department is responsible for the creation and maintenance of a single class of data. The goal must be to eliminate duplication both of spatial geometry and of attribute information.

As an example, in a large municipal environment, the private land ownership unit — the “parcel” — may have many fields associated with it. The same parcel is referenced by many departments, but each department views the fields differently. Most importantly, only one department — for example, the department of public works — is empowered with the creation of new property boundaries. This department must work closely with the appraisal district to allow the assignment of a unique property identifier and street address, to allow further information to be applied correctly. [Figure 57.11](#) shows additional examples.

Data inconsistencies can be minimized by the use of controlled procedures and standards for all information handling. However, a database system must also provide controls by allowing tables and basic elements to be accessed in a read-only mode, for read-write, or to be completely restricted. When dealing with geographic data, the example above shows that although one user may require write access to geometric components of a spatial object, other users require write access to specific attribute fields

CLASS / LAYER attributes	Users		
	Public Works / Engineering	Assessor	Planning
PARCEL			
boundary	write	read	read
identifier	read	write	read
street address	write	read	read
land value	-	write	read
improvement value	-	write	read
total value	-	write	read
date of last sale	-	write	read
owner name	-	write	read
owner address	-	write	read
zoning	read	read	write
tax district	read	read	write
STREET			
centerline	write	read	read
right-of-way	write	read	read
name	write	read	read
TAX DISTRICT			
boundary	-	write	read
name	-	write	read
taxes	-	write	read

FIGURE 57.11 Security in a multiuser GIS. This diagram shows read/write access privileges to classes of information for three different groups of users (departments) in a multidepartmental GIS.

for which they are responsible. For example, the assessor is responsible for assigning taxable values to individual properties, whereas the planner may be responsible for the zoning for specific properties. A good GIS provides a unified database management system that permits such combinations for geometric and attribute components. In addition, many GISs that rely on the relational model support and make use of security characteristics of the underlying relational database management system. In this situation individual tables and table records may be locked for read-only access or restricted completely.

In many GISs access to spatial data is managed in conjunction with a checkout procedure, whereby a group of information corresponding to the area of interest (a map sheet for example) is made available to one user only for updating and restricted to read-only access for all others. A transaction-oriented information system often takes this to a more elemental level, in which any database transaction locks out a small set of data immediately prior to the transaction. An AM/FM system often aims for this level of security, in which changes to facilities status and geometric elements are critical. This issue of concurrency — where in theory no two users may access the same data elements for update at the same time — can be achieved only with a sophisticated management of graphic and attribute elements.

57.4 Building and Maintaining a GIS

This section discusses the primary issues in building a geographic information system.

Reference Coordinate Systems

A geographic information system requires the definition and application of a reference coordinate system. The choice of a reference system is typically based on accuracy requirements, geographic scope of the

Type	Extent	Source Data
Latitude/Longitude (geographic coordinates)	global	field geodetic / world maps
Universal Transverse Mercator (UTM)	worldwide zones, each 6 degrees of latitude wide	national / transcontinental
US State Plane Coordinate System	State / portion of State	state / country maps and documents
Project-specific	Projectwide (engineering / construction)	airport / building site

FIGURE 57.12 Reference systems for source data.

proposed GIS, and the predominantly used reference system for source data, as shown by examples in [Fig. 57.12](#).

For example, engineering and property maps within a small municipality are typically referenced to a local grid based on state plane coordinates. An information system that is to be applied for work at national or international levels may require the use of a true geographic reference system (latitude, longitude) to provide true seamlessness and consistency across the entire area of interest. It should be noted that the reference coordinates for the GIS database storage often differ from a local working coordinate system. For example, data entry may derive from large-scale maps referenced to state plane coordinates. In this situation the digitizing and data validation operations take place in state plane, but, once complete, all information is transformed and stored in the project's primary coordinate system. Information may later be extracted and presented in a variety of local coordinate systems, without corrupting the primary data storage.

Consideration of Scale

In a traditional mapping system, consideration of precision of data is closely tied to final map scale. In theory, a GIS requires a much broader view of both data accuracy and precision, because it is open to data at large range of scales and accuracies. Unfortunately, most GISs do not in practice allow the maintenance of data quality and accuracy information. Furthermore, many applications of GISs involve the merging and integration of spatial data of differing resolutions and accuracy. The results of such operations are often used in decision support without regard to their statistical reliability. For example, the results of overlaying parcel boundary information derived from 1" to 50' scale property maps and soil classification boundaries derived from 1" to 1000' orthophotos may be used to determine the taxable value of farm properties.

It is therefore critical that the overall design of a GIS takes into account the accuracy requirements as a function of the application intent.

Data Sources

A variety of data sources are available for input to a GIS, as depicted in [Figs. 57.13](#) and [57.14](#). They can be classified most simply as follows:

- Local maps and related documents
- Existing local information systems
- Commercially available information
- Government sources

GIS implementation is often the driving force behind conversion of existing maps and other documents and records.

Source	Source Scale	Accuracy
Field Survey, total station		.02 m
Field Survey, GPS Carrier Phase Differential		.05 m
Photogrammetry, large scale	1:4,000	0.3 m
Digitized Engineering Map	1:600	0.5 m
Photogrammetry, medium scale	1:16,000	1.5 m
Photogrammetry, small-scale	1:40,000	3 m
Field Survey, GPS Pseudo-range Differential		5 m
Digitized USGS Quad sheet	1:24,000	15 m
Bureau of the Census TIGER files	1:24,000	20 m
Remote Sensing, SPOT panchromatic, rigorous model		20 m
Field Survey, GPS pseudo-range single receiver		50 m
Remote Sensing, Landsat TM panchromatic, rigorous model		60 m
USGS Digital Line Graph	1:2 million	100 m

FIGURE 57.13 Typical data sources. This table indicates typical source materials and methods of spatial data collection, with resultant accuracies.

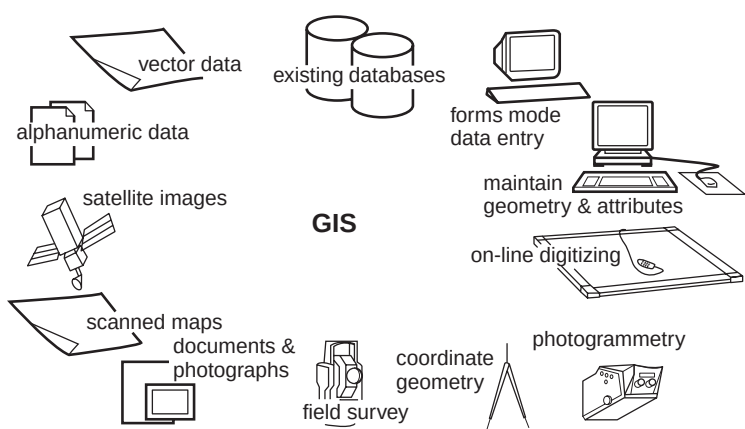


FIGURE 57.14 Data sources.

Data Entry and Processing

The data entry process may be reviewed in terms of:

- Interactive or semiautomated conversion of existing map documents
- Interactive update of attribute information through use of forms
- Batch loading of digital files

Interactive digitizing of existing maps has been superseded in part by a semiautomated process in which documents are scanned, displayed as a registered backdrop to vector data, and converted as necessary. Software that provides line-following and recognition capabilities, augmented by text recognition and a rules-based means to associate printed text labels as descriptors for adjacent graphic elements, may be applied successfully to conversion projects involving a large number of consistent map documents.

In addition, a key component to GIS database implementation is the linking of existing information systems. This is achieved as part of the overall database schema definition by identifying the key that provides the link between the core GIS and the existing data records. For example, the digitizing of parcel maps may include the interactive input of parcel identifier. Once a spatial feature containing the parcel number has been created, existing databases also containing the parcel number can be incorporated and accessed through the GIS. Alternatively, the parcel number may be used as a key to load additional attributes through a batch import process or as a guide to interactive form entry.

Structure/Topology

The data entry process normally includes facilities to check overall consistency and validity of data. The level to which these facilities operate is determined by the source data and the means of entry.

For digitizing operations and import of digital files representing vector geometry, it is critical that the input geometry defines a record that is consistent with both nongraphic descriptors and other graphic elements. For processing of vector line data, operations that determine intersections with adjacent geometry, extend lines to provide polygon closure, or eliminate overshoots are typically applied. Since these operations alter geometric components, they must be used with care and often use a correction tolerance determined by the ultimate accuracy requirements. This operation is often referred to as “building and cleaning,” or “feature assembly.” In addition, the process of “conflation” allows the merging of line data from disparate sources to minimize data storage and permit more consistent assignment of descriptive attributes. For example, street centerlines derived from small-scale TIGER data may be conflated to highway alignment data derived directly from the local engineering department. This process actually “moves” inaccurate TIGER geometry to the more accurate engineering data, but then allows the use of the other TIGER data, such as street names and address ranges. Figure 57.15 identifies database and geometry linkages used to build a GIS.

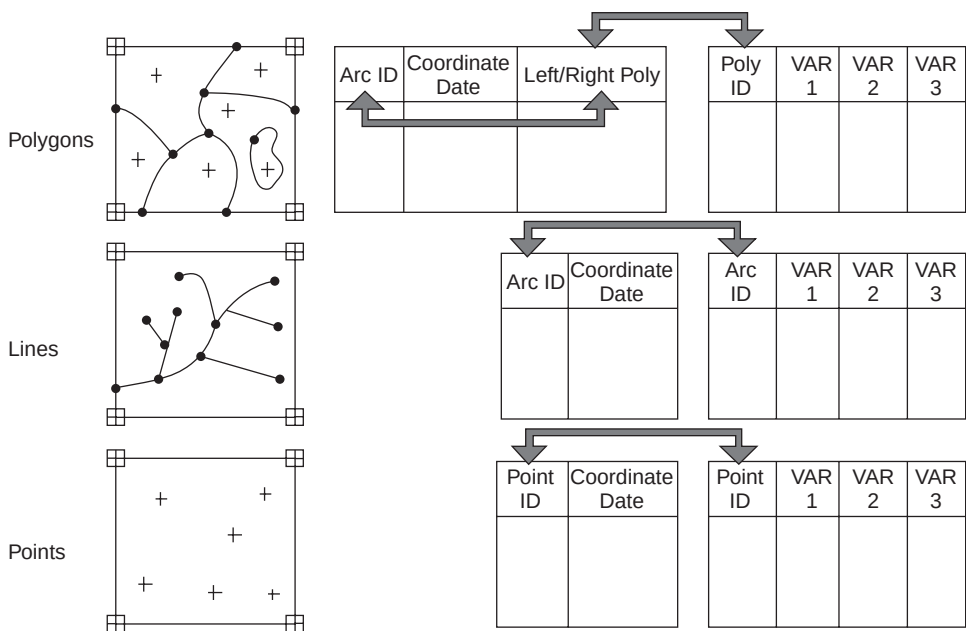


FIGURE 57.15 Database and geometry links. (Source: ESRI. 1992. *ARC/INFO: GIS Today and Tomorrow*. ESRI, Redlands, CA.)

Maintenance Operations

Once the primary spatial database has been created, it is important that maintenance operations may be applied without corrupting either geometric or attribute components. This is the most critical aspect to GIS editing as compared to a graphics-driven CAD system, in which the graphical elements and representation are most critical. For this reason, maintenance operations are most often performed on a primitive spatial object with both graphic and attribute components accessible. Nongraphic updates and edits may take place at a graphics workstation or at a simple alphanumeric screen, where operators update form entries.

57.5 Spatial Analysis

Beyond the implementation of GIS purely for mapping and map production, its key value lies in the ability to perform analysis based on spatial location and relationships between spatial elements. Such spatial analysis is augmented by the incorporation of external database records and application models.

Database Operations

We will now consider three groups of database operations associated with spatial information systems.

The first is simple querying and identification of individual geographic features and all related information. These may be nongraphic queries that return information based on nongraphic characteristics or graphical queries that incorporate a spatial extent to limit the extent of a querying process. Address matching, in which a nongraphic query provides street locations and/or property identification, is a fundamental operation in linking address as a reliable key between geographic and nongraphic information. Conversely, a graphical selection (for example, by pointing to a display) can be used to return a set of nongraphic information or a report. Querying and extraction operations are described later in this chapter.

The second group of operations are based on pure spatial analysis, based on geometric properties of individual elements. These apply to point, line, and polygon data as seen in [Fig. 57.16](#). These include the generation of polygonal buffers around points, lines, or polygons and the overlay of multiple layers of polygonal information to provide a “spatial join” to be used in further processing and analysis. Properties such as proximity, distance, overlapping, and containment are included in this group.

The third group of operations rely on point, line, and polygon topology to provide the basis for analysis. Such properties as adjacency, connectivity, and composed of are included in this group. Linear network analysis provides the basis for shortest path calculations in transportation, distribution, and collection applications. Figure 57.16 identifies some commonly used spatial database operations.

Of course, the typical GIS application may use many database operations, spatial and nonspatial, to provide the desired answer. As an example, the simple question “How many people live within 1/2 mile of the proposed light rail line?” is typically answered by proceeding through the following steps, each one a distinct database operation:

- Identify the light rail line (spatial).
- Create a 1/2-mile corridor/buffer based on the location of the rail line.
- Overlay the rail corridor against census block information.
- Transfer demographic information from census blocks to results of overlay.
- Produce summary statistics for population from the overlay.

Coupling to External Analyses/Applications

As previously discussed, the geographic information provides a powerful, common context for many different applications. In many cases the geographic key provides a much more significant means of interpretation or analysis than a simple numeric or textual key (for example, location versus street

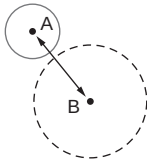
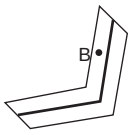
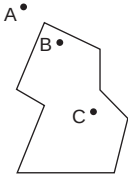
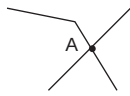
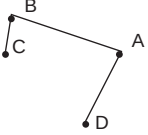
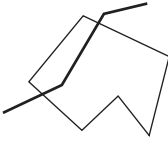
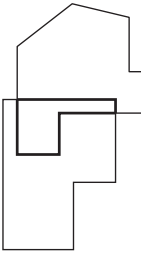
Topologic Type	Node / point	Line	Polygon / surface
Node / point	buffer / proximity 	intersection  corridor / proximity 	containment 
Line		intersection (create new nodes)  connectivity 	intersection 
Polygon / surface			intersection 

FIGURE 57.16 Spatial database operations. Typical operations involving two input geometries (row and column).

address). The structure of a GIS lends itself to linking with additional software and procedures that can be considered “external” applications. Examples are:

- Groundwater pollution models
- Statistical analysis
- Traffic engineering analyses

Data Interchange Standards and Formats

In practice, the exchange of data between geographic information systems has been inefficient and incomplete, except in situations where both systems share a common implementation. Two departments within a single organization, for example, may exchange information in a format recognized only by those departments. Such a practice results in the widespread exchange of information in commercially proprietary formats, such as those of Arc/Info coverage files or Intergraph design files. When geographic data are to be transferred to other agencies or to be imported from other sources, it is common to resort to one of several widely used standards.

Digital Exchange Format (DXF), introduced by Autodesk in conjunction with its AutoCAD software product, is probably the most widely used means to exchange geographic data between CAD systems, desktop mapping systems, and even more sophisticated GISs. It is most effective as a means of encoding graphic elements but has severe limitations in the areas of handling topology, attributes, and more complex data structures. This means that the exchange of any DXF file must be accompanied by additional information describing the layering convention, symbology rules, and coordinate references. Such meta-data is not normally incorporated in the DXF. Because spatial information must be resolved into simple graphic components and separate but related attribute information, both the source and target systems must provide additional utilities to ensure that all data is interpreted correctly.

Other formats, such as DLG (Digital Line Graph) and TIGER (Topologically Integrated Geographic Encoding and Referencing System), have been devised and introduced by federal government agencies as a means to deliver digital geographic data to the public, other agencies, and commercial users. They go beyond standards such as DXF in that they are designed to handle line and polygon topology, common data layer classifications, and real-world coordinate systems (as opposed to a graphics-based standard that often refers to “drawing units”).

More recently, the Spatial Data Transfer Standard (SDTS) has been designed as a means to model, encode, and transfer geographic information. The SDTS goes far beyond previous standards, in that it provides a comprehensive means (1) to define spatial phenomena and spatial objects used to represent phenomena, and (2) to model spatial features in a hierarchical fashion. The exchanged data include not only the raw geographic components, but also the definition of rules that associate the geometric, attribute and topologic elements, and any other metadata that further define the information. First practical implementations of SDTS, involving only a subset of the standard, have been developed for topologically structured vector data. This is referred to as the *vector profile*. Subsequent implementations or profiles will support raster- or grid-based data, such as imagery, digital orthophotos, and terrain models. The SDTS was approved as a Federal Information Processing Standard, Publication 173 (FIPS 173), in July 1992. Federal agencies such as the U.S. Geological Survey and the Bureau of the Census are now required to supply spatial information in an SDTS profile.

The use of structured query language (SQL) and other facilities associated with the relational database management system also provides the means to exchange spatially related data.

57.6 Information Extraction

The GIS has evolved over the past decade from a mainframe-based software system with distributed graphic terminals, through workstation clusters focusing high-performance graphics processors in a departmental fashion, to a more distributed organizational component. The corporate or enterprise-wide information system now typically includes a component dedicated to geographic data handling. Access to and extraction of information from a GIS takes place in various ways, depending on the nature and extent of the data required.

Displays and Reporting

As mentioned in the introduction to this chapter, the most standard product derived from a GIS remains a map. This map may be part of a standard map series, replacing the map product previously derived using manual drafting techniques, but increasingly common is a variety of custom maps and graphical displays that highlight the flexibility of a GIS in presentation of data.

For the typical GIS user the graphical display on a workstation or PC is the typical medium for interaction with the data. Layers or classes of information may be selected from a menu or list, for display at a variety of scales, along with attribute information as textual annotation or as a menu form. Specific layers and attribute information may be switched on or off based on scale of display or other criteria. Multiple windows may be used as a means to guide the user through large amounts of spatial information:

A where_clause may have Boolean expressions of the form:

EXPR	[	is	]	CONDITION
		is not		
null		EXPR		

A CONDITION is:

[OPERATOR]	[EXPR]
in	(EXPR{, EXPR})

Examples of OPERATOR are:

Comparison Operator:
=, !=, >, >=, <, <=, range (between)
Pattern-matching Operators:
like, match, not match
Spatial Operators:
overlap, contain, near, left_of, adjacent to,

Examples of queries in a GIS:

select house where house.address = "5623 South Avalon" and house.zip_code = "43211"
select street where street.address_number range (5600,5700)
select house where house near (select river where river.name = "Colorado")
select house where house overlap (select soil_area where soil_area.type = "clay")

FIGURE 57.17 Query language. The where clause forms part of a query language that is used to retrieve database records satisfying certain conditions specified. The where clause consists of Boolean expressions that can be nested or used in combination, and that evaluate to TRUE or FALSE.

For example, an overview window showing index maps or major features may be used as a constant reference to a larger-scale display in the current window.

In addition to information displays in standard cartographic forms (point symbols, line styles, and area fills and patterns), a thematic display of one class of information may be generated. Such displays rely on nongraphic characteristics such as "property value" or "soil type" or "road classification" to determine the cartographic representation. Thematic maps allow the efficient visualization of information. Also possible are more advanced representations, sometimes referred to as *cartograms*, such as the ability to size symbols based on attribute characteristics (for example, "tree height" or "drain size") or the ability to generate pie charts based on attributes (for example, "registered voters: Democrats" and "registered voters: Republicans").

Increasingly, the output from a GIS is not just a graphical product but a simple report or statistic. Spatial analysis can provide a set of information regarding a subset of spatial information. Thus, a letter to be sent to all property owners within a 1/2-mile buffer of a proposed industrial development may be the final product, based on spatial analysis and reporting.

Spatial Query Languages

In the geo-relational model a critical component in the GIS is the link between nongraphic and graphic elements. In such a model there are normally extensive sets of tools to allow access to, processing of, and

retrieval of attribute information stored in tables. The *structured query language* (SQL) standard provides a common syntax for identifying and extracting specific table entries for further processing. Relational database management systems, such as Oracle and Ingres, are used broadly in conjunction with GIS components.

The critical nature of this link lies in the fact that once a database key is corrupted, the real value of the geographic information is lost and is difficult or impossible to recover. The second difficulty in the geo-relational model lies in the fact that the spatial component requires a distinct management system.

Commercial GIS products have attempted to introduce a geographic query language or spatial query language that incorporates many SQL components but is augmented by specific spatial operators. More recently, the ANSI SQL Committee has proposed spatial extensions to a query language. Commercial RDBMS vendors are also active in this area.

57.7 Applications

In this section we examine further some typical roles of geographic information systems.

Basemap and Infrastructure in Government

In an introductory example we discussed the use of GIS for handling land records and property ownership. In fact, this is but one role for geographic information management in local government. Typically, a set of basemap information — including property boundaries, highways and rights of way, public buildings, and other major transportation and drainage features — form the basis for a multidepartmental information system. During implementation and extension through multiple departments, additional attributes are added to these layers, and additional layers are created. In a sense the GIS allows the modeling in a rigorous, structured fashion of the annotated basemap that previously would have been found in individual agencies of municipal government.

In many situations a GIS is the sole means for providing required services, including meeting local, state, and federal legislative requirements. In Florida, for example, each community is required to submit a master plan that includes land-use zoning, transportation, and residential and commercial zoning plans. In addition, environmental impact regulations require the consideration of multiple levels of information (transportation, residences, zoning, water and wastewater, soils, vegetation, habitats) in producing reports and recommendations: A multilevel GIS is often the only means to meet planning requirements in a timely fashion.

Currently, federal legislation requires all communities to implement a management and control system for storm water runoff. For large cities this requires major work not only in informing property owners and enforcing compliance with regulations, but also in compiling all of the background data required for analysis. This includes terrain-modeling data, as well as complete watershed and drainage network information. However, the timetable for compliance with these regulations requires a computer-based system of information for analysis and prediction.

In general, we can review the capabilities and advantages of a GIS as lying in the areas of:

- *Operation.* Being able to provide response and service requests effectively.
- *Planning.* Being able to anticipate and plan responses, for example, to emergencies.
- *Information sharing.* Allowing multiple agencies to share cost of data collection and maintenance.
- *Decision support.* Providing qualified recommendations to management for review and action.

Facilities Management

For utility companies, such as those dealing with very extensive networks and facilities for the distribution of electricity and gas, a geographic information system is typically an integral part of a program aimed at improving operational efficiency, improvements in planning and engineering work, and higher service

quality. These goals can be met by implementation of a GIS with powerful characteristics, which results in the ability to store and model spatial information at many levels: an upper management level that provides summaries updated on a regular basis; an intermediate level that provides all centralized management information to be used on a daily basis for business control and decision making; and the basic, “on-line” information level used for operational transactions.

A typical implementation is developed around a core facilities database that contains the following:

- Background cartographic, such as highway maps as political boundaries
- Operational facilities, including all information on installations forming the distribution network and data on maintenance, breakdowns, and other incidents
- Network of facilities under construction, showing facilities not yet in service
- Planned network, containing planned facilities

The system provides the following capabilities based on this core database:

- Maintenance of all necessary network diagrams and worksheets
- Maintenance of all alphanumeric (descriptive) information regarding facilities and operations
- Alphanumeric queries on all networks and facilities
- Production of all necessary reports, drawings, plans, and maps

Development Tools (Means to Customize/Build New Applications)

There is always a compromise between implementing a GIS based on standard commercial products, while attempting to maintain the ability to customize a system to accommodate specific and changing requirements over time. There are several levels at which GIS products may offer such capabilities. We may consider these in terms of profiles of the various end-users:

- *The casual user*, with little programming experience
- *The GIS specialist*, with detailed knowledge of spatial data handling and operations
- *The GIS programmer*, with experience in spatial data handling and programming expertise

For the casual user of a GIS standard tools are typically available that allow repetitive or predictable operations to be stored as a macro or command file, which may then be invoked with a single command or menu operation. As commercial GIS products begin to reflect simplistic and intuitive graphical user interfaces — such as those consistent with Windows on a PC platform or X-Windows on a UNIX workstation — facilities such as the ability to “learn” a sequence of menu selections or command-line arguments are those that most simply allow some limited customizing.

The GIS specialist who understands the full capabilities of a specific software product always uses this knowledge to extend its operational use through various combinations of basic functions and operations. For example, a well-known commercial GIS product now offers roughly 800 operations through command-line options and a much smaller number of graphical user interface components such as menus, submenus, and pull-down selection lists. The more advanced macro language permits the extension of simplistic sequences of operations to include flow control, user interaction, and menu components.

The GIS programmer uses knowledge of all programming components — including subroutine libraries supplied with the GIS, user interface libraries, and other programming tools — to develop a high-level application that either can be integrated with standard GIS operations or can completely replace the menus and options made available to the end-users of the information system. In this case the MIS department or an independent software developer may provide a completely customized interface for multiple departments within a large organization. For high levels of integration with other information systems, such programming is typically required. For example, GIS programming tools used in conjunction with high-level tools provided by Oracle relational database management system can provide the optimized functions and operational performance required.

57.8 Summary

This chapter has described the way in which map data are now often being incorporated directly into a computerized information system. A geographic information system typically attempts to link geometric, spatial information with related nongraphic or attribute data from one or many sources, thereby providing a direct spatial context for many databases. Since a large proportion of nongraphic data held by government agencies and certain types of private companies relate to geographic location and context, the introduction of GIS technology has been driven by such organizations in an attempt to make more efficient use of information and to allow more effective information management. GIS provides a more complete model of the real world than can simple (carto)graphic representations or tabular records or reports. In areas of local government and in private utility industries, the GIS approach is synonymous with information sharing, systems integration, and reengineering, in the name of providing more effective service in a competitive environment.

References

- Antenucci, J. et al. 1991. *Geographic Information Systems: A Guide to the Technology*. Van Nostrand Reinhold, New York.
- Aronoff, S. 1989. *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, ON.
- Burrough, P.A. 1986. *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford University Press, New York.
- Calkins, H.W. and Tomlinson, R.F. 1984. *Basic Readings in Geographic Information Systems*. SPAD Systems, Ltd., Williamsville, NY.
- Chance, A., Newell, R.G., and Theriault, D.G. 1990. An object oriented GIS: issues and solutions. *Proceedings EGIS*, Vol. 1. Amsterdam, Netherlands.
- Cowen, D.J. 1988. GIS versus CAD versus DBMS: what are the differences? *Photogrammetric Engineering and Remote Sensing*, Vol. 54.
- Date, G.J. 1987. *An Introduction to Database Systems*. Addison-Wesley, Reading, MA.
- Digital Cartographic Data Standards Task Force. 1988. The proposed standard for digital cartographic data. *The American Cartographer*, Vol. 15.
- Dueker, K.J. 1987. Geographic information systems and computer aided mapping. *Journal of the American Planning Association*. Summer 1987.
- ESRI. 1990. *Understanding GIS: The ARC/Info Way*. ESRI, Redlands, CA.
- ESRI. 1992. *ARC/INFO: GIS Today and Tomorrow*. ESRI, Redlands, CA.
- Fletcher, D. 1987. Modeling GIS transportation networks. *Proceedings URISA 1988*. Los Angeles, CA.
- GIS World. 1993. *GIS International Sourcebook, 1993*. GIS World, Fort Collins, CO.
- Goodchild, M.F. 1988. A spatial analytical perspective on GIS. *International Journal of Geographical Information Systems*, Vol. 1.
- Guptill, S.C. 1988. A process for evaluating GIS. *Proceedings GIS/LIS 1988*. San Antonio, TX.
- Guttman, A. 1984. R-trees: a dynamic index structure for spatial searching. *ACM SIGMOD*.
- Kilborn, K., Rifai, H.S., and Bedient, P.B. 1991. The integration of ground water models with geographic information systems. *Proceedings ACSM-ASPRS*.
- Maguire, D., Goodchild, M.F., and Rhind, D. 1991. *Geographical Information Systems: Principles and Applications*. John Wiley & Sons, New York.
- Montgomery, G., and Schuch, H. 1993. *Data Conversion in GIS*. GIS World, Fort Collins, CO.
- Parker, H.D. 1988. The unique qualities of a geographic information system: a commentary. *Photogrammetric Engineering and Remote Sensing*, Vol. 54.
- Tom, H. 1990. Geographic information systems standards: a federal perspective. *GIS World*, Vol. 3.
- Tomlin, C.D. 1990. *Geographic Information Systems and Cartographic Modeling*. Prentice Hall, Englewood Cliffs, NJ.