Geomatics Engineering and Geographic Information Systems

(Concepts and Applications)

Teresia Cervantes

Janene Tejeda

First Edition, 2012

ISBN 978-81-323-0888-1

Published by:
Academic Studio
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Geodesy

An old geodetic pillar (1855) at Ostend, Belgium

A Munich archive with lithography plates of maps of Bavaria

Geodesy also named **geodetics**, a branch of earth sciences, is the scientific discipline that deals with the measurement and representation of the Earth, including its gravitational field, in a three-dimensional time-varying space. Geodesists also study geodynamical phenomena such as crustal motion, tides, and polar motion. For this they design global and national control networks, using space and terrestrial techniques while relying on datums and coordinate systems.

Definition

Geodesy (from Greek *γεωδαισία – geodaisia*, lit. "division of the Earth") is primarily concerned with positioning within the temporally varying gravity field. Somewhat obsolete nowadays, geodesy in the German speaking world is divided into "Higher Geodesy" ("Erdmessung" or "höhere Geodäsie"), which is concerned with measuring the Earth on the global scale, and "Practical Geodesy" or "Engineering Geodesy" ("Ingenieurgeodäsie"), which is concerned with measuring specific parts or regions of the Earth, and which includes surveying.

The shape of the Earth is to a large extent the result of its rotation, which causes its equatorial bulge, and the competition of geological processes such as the collision of plates and of volcanism, resisted by the Earth's gravity field. This applies to the solid surface, the liquid surface (dynamic sea surface topography) and the Earth's atmosphere. For this reason, the study of the Earth's gravity field is called physical geodesy by some.

History

Geoid and reference ellipsoid

The geoid is essentially the figure of the Earth abstracted from its topographical features. It is an idealized equilibrium surface of sea water, the mean sea level surface in the absence of currents, air pressure variations etc. and continued under the continental masses. The geoid, unlike ellipsoid, is irregular and too complicated to serve as the computational surface on which to solve geometrical problems like point positioning. The geometrical separation between the geoid and the reference ellipsoid is called the geoidal undulation. It varies globally between ±110 m.

A reference ellipsoid, customarily chosen to be the same size (volume) as the geoid, is described by its semi-major axis (equatorial radius) a and flattening f. The quantity $f = (a-b)/a$, where b is the semi-minor axis (polar radius), is a purely geometrical one. The mechanical ellipticity of the Earth (dynamical flattening, symbol J_2) can be determined to high precision by observation of satellite orbit perturbations. Its relationship with the geometrical flattening is indirect. The relationship depends on the internal density distribution, or, in simplest terms, the degree of central concentration of mass.

The 1980 Geodetic Reference System (GRS80) posited a 6,378,137 m semi-major axis and a 1:298.257 flattening. This system was adopted at the XVII General Assembly of the International Union of Geodesy and Geophysics (IUGG). It is essentially the basis for geodetic positioning by the Global Positioning System and is thus also in extremely widespread use outside the geodetic community.

The numerous other systems which have been used by diverse countries for their maps and charts are gradually dropping out of use as more and more countries move to global, geocentric reference systems using the GRS80 reference ellipsoid.

Coordinate systems in space

The locations of points in three-dimensional space are most conveniently described by three cartesian or rectangular coordinates, X,Y and Z. Since the advent of satellite positioning, such coordinate systems are typically geocentric: the Z axis is aligned with the Earth's (conventional or instantaneous) rotation axis.

Prior to satellite geodesy era, the coordinate systems associated with a geodetic datum attempted to be geocentric, but their origins differed from the geocentre by hundreds of metres, due to regional deviations in the direction of the plumbline (vertical). These regional geodetic datums, such as ED50 (European Datum 1950) or NAD83 (North American Datum 1983) have ellipsoids associated with them that are regional 'best fits' to the geoids within their areas of validity, minimising the deflections of the vertical over these areas.

It is only because GPS satellites orbit about the geocentre, that this point becomes naturally the origin of a coordinate system defined by satellite geodetic means, as the satellite positions in space are themselves computed in such a system.

Geocentric coordinate systems used in geodesy can be divided naturally into two classes:

1. Inertial reference systems, where the coordinate axes retain their orientation relative to the fixed stars, or equivalently, to the rotation axes of ideal gyroscopes; the X axis points to the vernal equinox
2. Co-rotating, also ECEF ("Earth Centred, Earth Fixed"), where the axes are attached to the solid body of the Earth. The X axis lies within the Greenwich observatory's meridian plane.

The coordinate transformation between these two systems is described to good approximation by (apparent) sidereal time, which takes into account variations in the Earth's axial rotation (length-of-day variations). A more accurate description also takes polar motion into account, a phenomenon closely monitored by geodesists.

Coordinate systems in the plane

In surveying and mapping, important fields of application of geodesy, two general types of coordinate systems are used in the plane:

1. Plano-polar, in which points in a plane are defined by a distance s from a specified point along a ray having a specified direction α with respect to a base line or axis;
2. Rectangular, points are defined by distances from two perpendicular axes called x and y. It is geodetic practice—contrary to the mathematical convention—to let the x axis point to the North and the y axis to the East.

Rectangular coordinates in the plane can be used intuitively with respect to one's current location, in which case the x axis will point to the local North. More formally, such coordinates can be obtained from three-dimensional coordinates using the artifice of a map projection. It is *not* possible to map the curved surface of the Earth onto a flat map surface without deformation. The compromise most often chosen—called a conformal projection—preserves angles and length ratios, so that small circles are mapped as small circles and small squares as squares.

An example of such a projection is UTM (Universal Transverse Mercator). Within the map plane, we have rectangular coordinates x and y. In this case the North direction used for reference is the *map* North, not the *local* North. The difference between the two is called *meridian convergence*.

It is easy enough to "translate" between polar and rectangular coordinates in the plane: let, as above, direction and distance be α and *s* respectively, then we have

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

The reverse transformation is given by:

$$s = \sqrt{x^2 + y^2}$$
$$\alpha = \arctan(y/x).$$

Heights

In geodesy, point or terrain *heights* are "above sea level", an irregular, physically defined surface. Therefore a height should ideally *not* be referred to as a coordinate. It is more like a physical quantity, and though it can be tempting to treat height as the vertical coordinate z, in addition to the horizontal coordinates x and y, and though this actually is a good approximation of physical reality in small areas, it quickly becomes invalid for regional considerations.

Heights come in the following variants:

1. Orthometric heights
2. Normal heights
3. Geopotential heights

Each has its advantages and disadvantages. Both orthometric and normal heights are heights in metres above sea level, whereas geopotential numbers are measures of potential energy (unit: $m^2\ s^{-2}$) and not metric. Orthometric and normal heights differ in the precise way in which mean sea level is conceptually continued under the continental masses. The reference surface for orthometric heights is the geoid, an equipotential surface approximating mean sea level.

None of these heights is in any way related to **geodetic** or **ellipsoidial** heights, which express the height of a point above the reference ellipsoid. Satellite positioning receivers typically provide ellipsoidal heights, unless they are fitted with special conversion software based on a model of the geoid.

Geodetic data

Because geodetic point coordinates (and heights) are always obtained in a system that has been constructed itself using real observations, geodesists introduce the concept of a *geodetic datum*: a physical realization of a coordinate system used for describing point locations. The realization is the result of *choosing* conventional coordinate values for one or more *datum points*.

In the case of height datums, it suffices to choose *one* datum point: the reference bench mark, typically a tide gauge at the shore. Thus we have vertical datums like the NAP (Normaal Amsterdams Peil), the North American Vertical Datum 1988 (NAVD88), the Kronstadt datum, the Trieste datum, and so on.

In case of plane or spatial coordinates, we typically need several datum points. A regional, ellipsoidal datum like ED50 can be fixed by prescribing the undulation of the geoid and the deflection of the vertical in *one* datum point, in this case the Helmert Tower in Potsdam. However, an overdetermined ensemble of datum points can also be used.

Changing the coordinates of a point set referring to one datum, so to make them refer to another datum, is called a *datum transformation*. In the case of vertical datums, this consists of simply adding a constant shift to all height values. In the case of plane or spatial coordinates, datum transformation takes the form of a similarity or *Helmert transformation*, consisting of a rotation and scaling operation in addition to a simple translation. In the plane, a Helmert transformation has four parameters; in space, seven.

A note on terminology

In the abstract, a coordinate system as used in mathematics and geodesy is, e.g., in ISO terminology, referred to as a *coordinate system*. International geodetic organizations like the IERS (International Earth Rotation and Reference Systems Service) speak of a *reference system*.

When these coordinates are realized by choosing datum points and fixing a geodetic datum, ISO uses the terminology *coordinate reference system*, while IERS speaks of a *reference frame*. A datum transformation again is referred to by ISO as a *coordinate transformation*. (ISO 19111: Spatial referencing by coordinates).

Point positioning

Geodetic Control Mark (example of a deep benchmark)

Point positioning is the determination of the coordinates of a point on land, at sea, or in space with respect to a coordinate system. Point position is solved by computation from measurements linking the known positions of terrestrial or extraterrestrial points with the unknown terrestrial position. This may involve transformations between or among astronomical and terrestrial coordinate systems.

The known points used for point positioning can be triangulation points of a higher order network, or GPS satellites.

Traditionally, a hierarchy of networks has been built to allow point positioning within a country. Highest in the hierarchy were triangulation networks. These were densified into networks of traverses (polygons), into which local mapping surveying measurements, usually with measuring tape, corner prism and the familiar red and white poles, are tied.

Nowadays all but special measurements (e.g., underground or high precision engineering measurements) are performed with GPS. The higher order networks are measured with static GPS, using differential measurement to determine vectors between terrestrial points. These vectors are then adjusted in traditional network fashion. A global polyhedron of permanently operating GPS stations under the auspices of the IERS is used to define a single global, geocentric reference frame which serves as the "zero order" global reference to which national measurements are attached.

For surveying mappings, frequently Real Time Kinematic GPS is employed, tying in the unknown points with known terrestrial points close by in real time.

One purpose of point positioning is the provision of known points for mapping measurements, also known as (horizontal and vertical) control. In every country, thousands of such known points exist and are normally documented by the national mapping agencies. Surveyors involved in real estate and insurance will use these to tie their local measurements to.

Geodetic problems

In geometric geodesy, two standard problems exist:

First geodetic problem

Given a point (in terms of its coordinates) and the direction (azimuth) and distance from that point to a second point, determine (the coordinates of) that second point.

Second (inverse) geodetic problem

Given two points, determine the azimuth and length of the line (straight line, arc or geodesic) that connects them.

In the case of plane geometry (valid for small areas on the Earth's surface) the solutions to both problems reduce to simple trigonometry. On the sphere, the solution is significantly more complex, e.g., in the inverse problem the azimuths will differ between the two end points of the connecting great circle, arc, i.e. the geodesic.

On the ellipsoid of revolution, geodesics may be written in terms of elliptic integrals, which are usually evaluated in terms of a series expansion.

In the general case, the solution is called the geodesic for the surface considered. The differential equations for the geodesic can be solved numerically.

Geodetic observational concepts

Here we define some basic observational concepts, like angles and coordinates, defined in geodesy (and astronomy as well), mostly from the viewpoint of the local observer.

- The *plumbline* or *vertical* is the direction of local gravity, or the line that results by following it. It is slightly curved.

- The *zenith* is the point on the celestial sphere where the direction of the gravity vector in a point, extended upwards, intersects it. More correct is to call it a <direction> rather than a point.
- The *nadir* is the opposite point (or rather, direction), where the direction of gravity extended downward intersects the (invisible) celestial sphere.
- The celestial *horizon* is a plane perpendicular to a point's gravity vector.
- *Azimuth* is the direction angle within the plane of the horizon, typically counted clockwise from the North (in geodesy and astronomy) or South (in France).
- *Elevation* is the angular height of an object above the horizon, Alternatively zenith distance, being equal to 90 degrees minus elevation.
- *Local topocentric coordinates* are azimuth (direction angle within the plane of the horizon) and elevation angle (or zenith angle) and distance.
- The North *celestial pole* is the extension of the Earth's (precessing and nutating) instantaneous spin axis extended Northward to intersect the celestial sphere. (Similarly for the South celestial pole.)
- The *celestial equator* is the intersection of the (instantaneous) Earth equatorial plane with the celestial sphere.
- A *meridian plane* is any plane perpendicular to the celestial equator and containing the celestial poles.
- The *local meridian* is the plane containing the direction to the zenith and the direction to the celestial pole.

Geodetic measurements

The level is used for determining height differences and height reference systems, commonly referred to mean sea level. The traditional spirit level produces these practically most useful heights above sea level directly; the more economical use of GPS instruments for height determination requires precise knowledge of the figure of the geoid, as GPS only gives heights above the GRS80 reference ellipsoid. As geoid knowledge accumulates, one may expect use of GPS heighting to spread.

The theodolite is used to measure horizontal and vertical angles to target points. These angles are referred to the local vertical. The tacheometer additionally determines, electronically or electro-optically, the distance to target, and is highly automated to even robotic in its operations. The method of free station position is widely used.

For local detail surveys, tacheometers are commonly employed although the old-fashioned rectangular technique using angle prism and steel tape is still an inexpensive alternative. Real-time kinematic (RTK) GPS techniques are used as well. Data collected are tagged and recorded digitally for entry into a Geographic Information System (GIS) database.

Geodetic GPS receivers produce directly three-dimensional coordinates in a geocentric coordinate frame. Such a frame is, e.g., WGS84, or the frames that are regularly produced and published by the International Earth Rotation and Reference Systems Service (IERS).

GPS receivers have almost completely replaced terrestrial instruments for large-scale base network surveys. For Planet-wide geodetic surveys, previously impossible, we can still mention Satellite Laser Ranging (SLR) and Lunar Laser Ranging (LLR) and Very Long Baseline Interferometry (VLBI) techniques. All these techniques also serve to monitor Earth rotation irregularities as well as plate tectonic motions.

Gravity is measured using gravimeters. Basically, there are two kinds of gravimeters. *Absolute* gravimeters, which nowadays can also be used in the field, are based directly on measuring the acceleration of free fall (for example, of a reflecting prism in a vacuum tube). They are used for establishing the vertical geospatial control. Most common *relative* gravimeters are spring based. They are used in gravity surveys over large areas for establishing the figure of the geoid over these areas. Most accurate relative gravimeters are *superconducting* gravimeters, and these are sensitive to one thousandth of one billionth of the Earth surface gravity. Twenty-some superconducting gravimeters are used worldwide for studying Earth tides, rotation, interior, and ocean and atmospheric loading, as well as for verifying the Newtonian constant of gravitation.

Units and measures on the ellipsoid

Geographical latitude and longitude are stated in the units degree, minute of arc, and second of arc. They are *angles*, not metric measures, and describe the *direction* of the local normal to the reference ellipsoid of revolution. This is *approximately* the same as the direction of the plumbline, i.e., local gravity, which is also the normal to the geoid surface. For this reason, astronomical position determination – measuring the direction of the plumbline by astronomical means – works fairly well provided an ellipsoidal model of the figure of the Earth is used.

One geographical mile, defined as one minute of arc on the equator, equals 1,855.32571922 m. One nautical mile is one minute of astronomical latitude. The radius of curvature of the ellipsoid varies with latitude, being the longest at the pole and the shortest at the equator as is the nautical mile.

A metre was originally defined as the 40-millionth part of the length of a meridian (the target wasn't quite reached in actual implementation, so that is off by 0.02% in the current definitions). This means that one kilometre is roughly equal to (1/40,000) * 360 * 60 meridional minutes of arc, which equals 0.54 nautical mile, though this is not exact because the two units are defined on different bases (the international nautical mile is defined as exactly 1,852 m, corresponding to a rounding of 1000/0.54 m to four digits).

Temporal change

In geodesy, temporal change can be studied by a variety of techniques. Points on the Earth's surface change their location due to a variety of mechanisms:

- Continental plate motion, plate tectonics
- Episodic motion of tectonic origin, esp. close to fault lines

- Periodic effects due to Earth tides
- Postglacial land uplift due to isostatic adjustment
- Various anthropogenic movements due to, for instance, petroleum or water extraction or reservoir construction.

The science of studying deformations and motions of the Earth's crust and the solid Earth as a whole is called geodynamics. Often, study of the Earth's irregular rotation is also included in its definition.

Techniques for studying geodynamic phenomena on the global scale include:

- satellite positioning by GPS and other such systems,
- Very Long Baseline Interferometry (VLBI)
- satellite and lunar laser ranging
- Regionally and locally, precise levelling,
- precise tacheometers,
- monitoring of gravity change,
- Interferometric synthetic aperture radar (InSAR) using satellite images, etc.

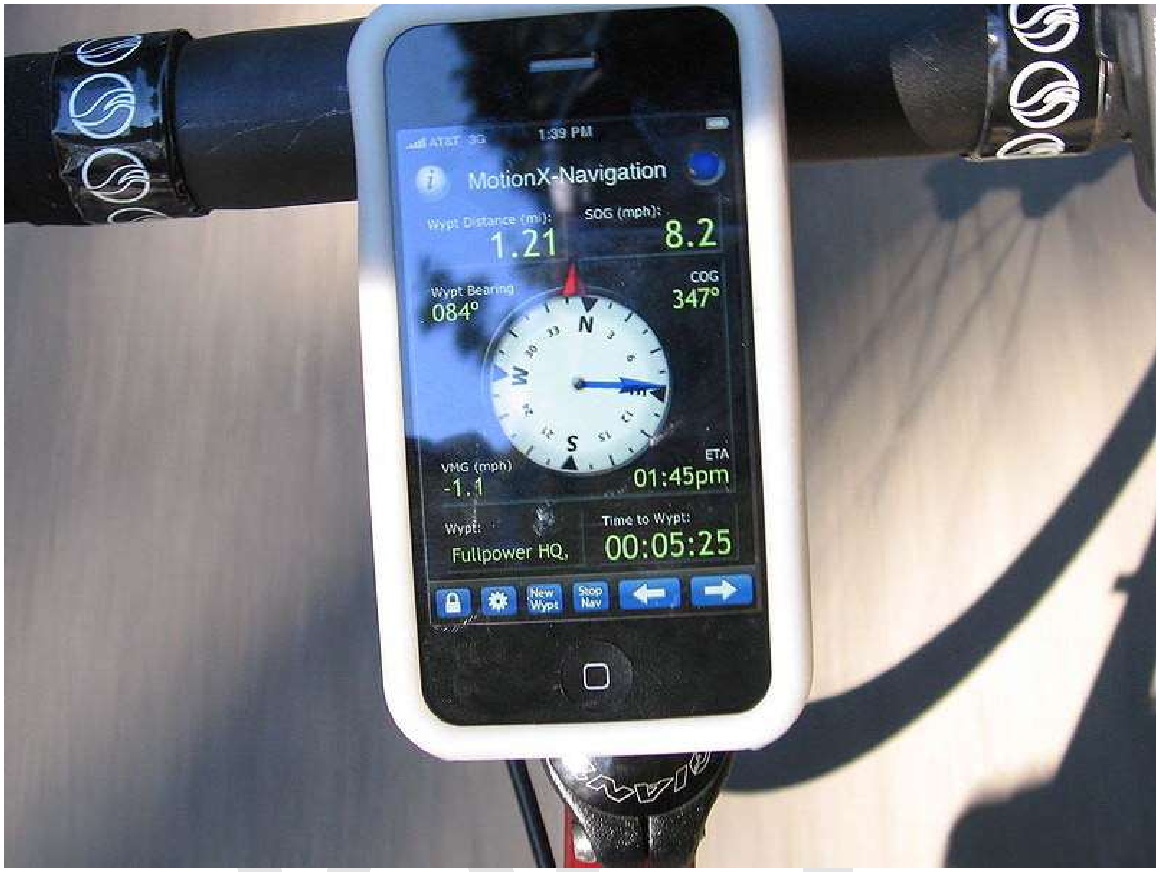
Chapter 2

Global Positioning System

Artist's conception of GPS Block II-F satellite in orbit

Civilian GPS receiver ("GPS navigation device") in a marine application.

Automotive navigation system in a taxicab.



GPS receivers are now integrated in many mobile phones.

The **Global Positioning System** (**GPS**) is a space-based global navigation satellite system (GNSS) that provides reliable location and time information in all weather and at all times and anywhere on or near the Earth when and where there is an unobstructed line of sight to four or more GPS satellites. It is maintained by the United States government and is freely accessible by anyone with a GPS receiver.

The GPS project was started in 1973 to overcome the limitations of previous navigation systems, integrating ideas from several predecessors, including a number of classified engineering design studies from the 1960s. GPS was created and realized by the U.S. Department of Defense (USDOD) and was originally run with 24 satellites. It became fully operational in 1994.

In addition to GPS, other systems are in use or under development. The Russian GLObal NAvigation Satellite System (GLONASS) was in use by the Russian military only until it was made fully available to civilians in 2007. There are also the planned Chinese Compass navigation system and the European Union's Galileo positioning system.

History

The design of GPS is based partly on similar ground-based radio navigation systems, such as LORAN and the Decca Navigator developed in the early 1940s, and used during World War II. In 1956 Friedwardt Winterberg proposed a test of general relativity using accurate atomic clocks placed in orbit in artificial satellites. To achieve accuracy requirements, GPS uses principles of general relativity to correct the satellites' atomic clocks. Additional inspiration for GPS came when the Soviet Union launched the first man-made satellite, Sputnik in 1957. A team of U.S. scientists led by Dr. Richard B. Kershner were monitoring Sputnik's radio transmissions. They discovered that, because of the Doppler effect, the frequency of the signal being transmitted by Sputnik was higher as the satellite approached, and lower as it continued away from them. They realized that because they knew their exact location on the globe, they could pinpoint where the satellite was along its orbit by measuring the Doppler distortion.

The first satellite navigation system, Transit, used by the United States Navy, was first successfully tested in 1960. It used a constellation of five satellites and could provide a navigational fix approximately once per hour. In 1967, the U.S. Navy developed the Timation satellite that proved the ability to place accurate clocks in space, a technology required by GPS. In the 1970s, the ground-based Omega Navigation System, based on phase comparison of signal transmission from pairs of stations, became the first worldwide radio navigation system. Limitations of these systems drove the need for a more universal navigation solution with greater accuracy.

While there were wide needs for accurate navigation in military and civilian sectors, almost none of those were seen as justification for the billions of dollars it would cost in research, development, deployment, and operation for a constellation of navigation satellites. During the Cold War arms race, the nuclear threat to the existence of the United States was the one need that did justify this cost in the view of the United States Congress. This deterrent effect is why GPS was funded. The nuclear triad consisted of the United States Navy's submarine-launched ballistic missiles (SLBMs) along with United States Air Force (USAF) strategic bombers and intercontinental ballistic missiles (ICBMs). Considered vital to the nuclear deterrence posture, accurate determination of the SLBM launch position was a force multiplier.

Precise navigation would enable United States submarines to get an accurate fix of their positions prior to launching their SLBMs. The USAF with two-thirds of the nuclear triad also had requirements for a more accurate and reliable navigation system. The Navy and Air Force were developing their own technologies in parallel to solve what was essentially the same problem. To increase the survivability of ICBMs, there was a proposal to use mobile launch platforms so the need to fix the launch position had similarity to the SLBM situation.

In 1960, the Air Force proposed a radio-navigation system called MOSAIC (Mobile System for Accurate ICBM Control) that was essentially a 3-D LORAN. A follow-on study called Project 57 was worked in 1963 and it was "in this study that the GPS concept was born." That same year the concept was pursued as Project 621B, which had "many of the attributes that you now see in GPS" and promised increased accuracy for Air Force bombers as well as ICBMs. Updates from the Navy Transit system were too slow for the high speeds of Air Force operation. The Navy Research Laboratory continued advancements with their Timation (Time Navigation) satellites, first launched in 1967, and with the third one in 1974 carrying the first atomic clock into orbit.

With these parallel developments in the 1960s, it was realized that a superior system could be developed by synthesizing the best technologies from 621B, Transit, Timation, and SECOR in a multi-service program.

During Labor Day weekend in 1973, a meeting of about 12 military officers at the Pentagon discussed the creation of a *Defense Navigation Satellite System (DNSS)*. It was at this meeting that "the real synthesis that became GPS was created." Later that year, the DNSS program was named *Navstar*. With the individual satellites being associated with the name Navstar (as with the predecessors Transit and Timation), a more fully encompassing name was used to identify the constellation of Navstar satellites, *Navstar-GPS*, which was later shortened simply to GPS.

After Korean Air Lines Flight 007, carrying 269 people, was shot down in 1983 after straying into the USSR's prohibited airspace, in the vicinity of Sakhalin and Moneron Islands, President Ronald Reagan issued a directive making GPS freely available for civilian use, once it was sufficiently developed, as a common good. The first satellite was launched in 1989, and the 24th satellite was launched in 1994.

Initially, the highest quality signal was reserved for military use, and the signal available for civilian use was intentionally degraded ("Selective Availability", SA). This changed with United States President Bill Clinton ordering Selective Availability turned off at midnight May 1, 2000, improving the precision of civilian GPS from 100 meters (about 300 feet) to 20 meters (about 65 feet). The United States military by then had the ability to deny GPS service to potential adversaries on a regional basis.

GPS is owned and operated by the United States Government as a national resource. Department of Defense (USDOD) is the steward of GPS. *Interagency GPS Executive Board (IGEB)* oversaw GPS policy matters from 1996 to 2004. After that the *National Space-Based Positioning, Navigation and Timing Executive Committee* was established by presidential directive in 2004 to advise and coordinate federal departments and agencies on matters concerning the GPS and related systems. The executive committee is chaired jointly by the deputy secretaries of defense and transportation. Its membership includes equivalent-level officials from the departments of state, commerce, and homeland security, the joint chiefs of staff, and NASA. Components of the executive office of the president participate as observers to the executive committee, and the FCC chairman participates as a liaison.

USDOD is required by law to "maintain a Standard Positioning Service (as defined in the federal radio navigation plan and the standard positioning service signal specification) that will be available on a continuous, worldwide basis," and "develop measures to prevent hostile use of GPS and its augmentations without unduly disrupting or degrading civilian uses."

Timeline and modernization

Summary of satellites

Block	Launch Period	Satellite launches				Currently in orbit and healthy
		Suc-cess	**Fail-ure**	**In prep-aration**	**Plan-ned**	
I	1978–1985	10	1	0	0	0
II	1989–1990	9	0	0	0	0
IIA	1990–1997	19	0	0	0	10
IIR	1997–2004	12	1	0	0	12
IIR-M	2005–2009	8	0	0	0	7
IIF	2010–2011	1	0	11	0	1
IIIA	2014–?	0	0	0	12	0
IIIB	Theoretical	0	0	0	8	0
IIIC	Theoretical	0	0	0	16	0
Total		59	2	11	36	30

(Last update: 24 May 2010)

PRN 01 from Block IIR-M is unhealthy
PRN 25 from Block IIA is unhealthy
PRN 32 from Block IIA is unhealthy

- In 1972, the USAF Central Inertial Guidance Test Facility (Holloman AFB), conducted developmental flight tests of two prototype GPS receivers over White Sands Missile Range, using ground-based pseudo-satellites.
- In 1978, the first experimental Block-I GPS satellite was launched.
- In 1983, after Soviet interceptor aircraft shot down the civilian airliner KAL 007 that strayed into prohibited airspace because of navigational errors, killing all

269 people on board, U.S. President Ronald Reagan announced that GPS would be made available for civilian uses once it was completed.

- By 1985, ten more experimental Block-I satellites had been launched to validate the concept.
- On February 14, 1989, the first modern Block-II satellite was launched.
- The Gulf War from 1990 to 1991, was the first conflict where GPS was widely used.
- In 1992, the 2nd Space Wing, which originally managed the system, was de-activated and replaced by the 50th Space Wing.
- By December 1993, GPS achieved initial operational capability (IOC), indicating a full constellation (24 satellites) was available and providing the Standard Positioning Service (SPS).
- Full Operational Capability (FOC) was declared by Air Force Space Command (AFSPC) in April 1995, signifying full availability of the military's secure Precise Positioning Service (PPS).
- In 1996, recognizing the importance of GPS to civilian users as well as military users, U.S. President Bill Clinton issued a policy directive declaring GPS to be a dual-use system and establishing an Interagency GPS Executive Board to manage it as a national asset.
- In 1998, United States Vice President Al Gore announced plans to upgrade GPS with two new civilian signals for enhanced user accuracy and reliability, particularly with respect to aviation safety and in 2000 the United States Congress authorized the effort, referring to it as *GPS III*.
- In 1998, GPS technology was inducted into the Space Foundation Space Technology Hall of Fame.
- On May 2, 2000 "Selective Availability" was discontinued as a result of the 1996 executive order, allowing users to receive a non-degraded signal globally.
- In 2004, the United States Government signed an agreement with the European Community establishing cooperation related to GPS and Europe's planned Galileo system.
- In 2004, United States President George W. Bush updated the national policy and replaced the executive board with the National Executive Committee for Space-Based Positioning, Navigation, and Timing.
- November 2004, QUALCOMM announced successful tests of assisted GPS for mobile phones.
- In 2005, the first modernized GPS satellite was launched and began transmitting a second civilian signal (L2C) for enhanced user performance.
- On September 14, 2007, the aging mainframe-based Ground Segment Control System was transferred to the new Architecture Evolution Plan.
- On May 19, 2009, the United States Government Accountability Office issued a report warning that some GPS satellites could fail as soon as 2010.
- On May 21, 2009, the Air Force Space Command allayed fears of GPS failure saying "There's only a small risk we will not continue to exceed our performance standard."

- On January 11, 2010, an update of ground control systems caused a software incompatibility with 8000 to 10000 military receivers manufactured by a division of Trimble Navigation Limited of Sunnyvale, Calif.
- The most recent launch was on May 28, 2010. The oldest GPS satellite still in operation was launched on November 26, 1990, and became operational on December 10, 1990.

Awards

On February 10, 1993, the National Aeronautic Association selected the GPS Team as winners of the 1992 Robert J. Collier Trophy, the nation's most prestigious aviation award. This team combines researchers from the Naval Research Laboratory, the USAF, the Aerospace Corporation, Rockwell International Corporation, and IBM Federal Systems Company. The citation honors them "for the most significant development for safe and efficient navigation and surveillance of air and spacecraft since the introduction of radio navigation 50 years ago."

Two GPS developers received the National Academy of Engineering Charles Stark Draper Prize for 2003:

- Ivan Getting, emeritus president of The Aerospace Corporation and an engineer at the Massachusetts Institute of Technology, established the basis for GPS, improving on the World War II land-based radio system called LORAN (*Lo*ng-range *R*adio *A*id to *N*avigation).
- Bradford Parkinson, professor of aeronautics and astronautics at Stanford University, conceived the present satellite-based system in the early 1960s and developed it in conjunction with the U.S. Air Force. Parkinson served twenty-one years in the Air Force, from 1957 to 1978, and retired with the rank of colonel.

GPS developer Roger L. Easton received the National Medal of Technology on February 13, 2006.

Francis X. Kane (Col. USAF, ret.) was inducted into the U.S. Air Force Space and Missile Pioneers Hall of Fame at Lackland A.F.B., San Antonio, Texas, March 2, 2010 for his role in space technology development and the engineering design concept of GPS conducted as part of Project 621B.

Basic concept of GPS

A GPS receiver calculates its position by precisely timing the signals sent by GPS satellites high above the Earth. Each satellite continually transmits messages that include

- the time the message was transmitted
- precise orbital information (the ephemeris)
- the general system health and rough orbits of all GPS satellites (the almanac).

The receiver uses the messages it receives to determine the transit time of each message and computes the distance to each satellite. These distances along with the satellites' locations are used with the possible aid of trilateration, depending on which algorithm is used, to compute the position of the receiver. This position is then displayed, perhaps with a moving map display or latitude and longitude; elevation information may be included. Many GPS units show derived information such as direction and speed, calculated from position changes.

Three satellites might seem enough to solve for position since space has three dimensions and a position near the Earth's surface can be assumed. However, even a very small clock error multiplied by the very large speed of light — the speed at which satellite signals propagate — results in a large positional error. Therefore receivers use four or more satellites to solve for the receiver's location and time. The very accurately computed time is effectively hidden by most GPS applications, which use only the location. A few specialized GPS applications do however use the time; these include time transfer, traffic signal timing, and synchronization of cell phone base stations.

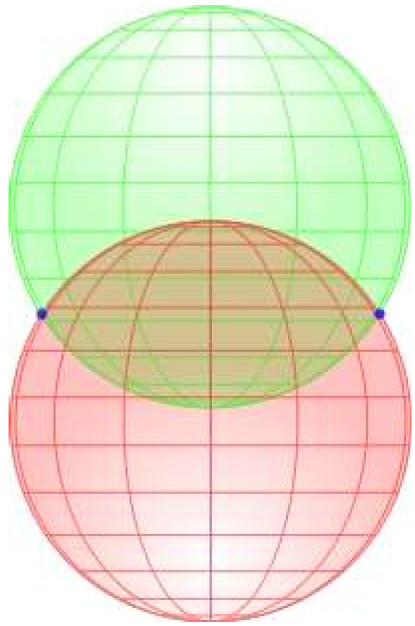
Although four satellites are required for normal operation, fewer apply in special cases. If one variable is already known, a receiver can determine its position using only three satellites. For example, a ship or aircraft may have known elevation. Some GPS receivers may use additional clues or assumptions (such as reusing the last known altitude, dead reckoning, inertial navigation, or including information from the vehicle computer) to give a less accurate (degraded) position when fewer than four satellites are visible.

Position calculation introduction

To provide an introductory description of how a GPS receiver works, error effects are deferred to a later section. Using messages received from a minimum of four visible satellites, a GPS receiver is able to determine the times sent and then the satellite positions corresponding to these times sent. The x, y, and z components of position, and the time sent, are designated as $[x_i, y_i, z_i, t_i]$where the subscript *i* is the satellite number and has the value 1, 2, 3, or 4. Knowing the indicated, or uncorrected, time the message was received $t_{\text{r, uncorr}}$, the GPS receiver can compute the uncorrected transit time of the message as $(t_{\text{r, uncorr}} - t_i)$. Assuming the message traveled at the speed of light, *c*, the uncorrected distance traveled or pseudorange, $p_i$can be computed as $(t_{\text{r, uncorr}} - t_i)c$.

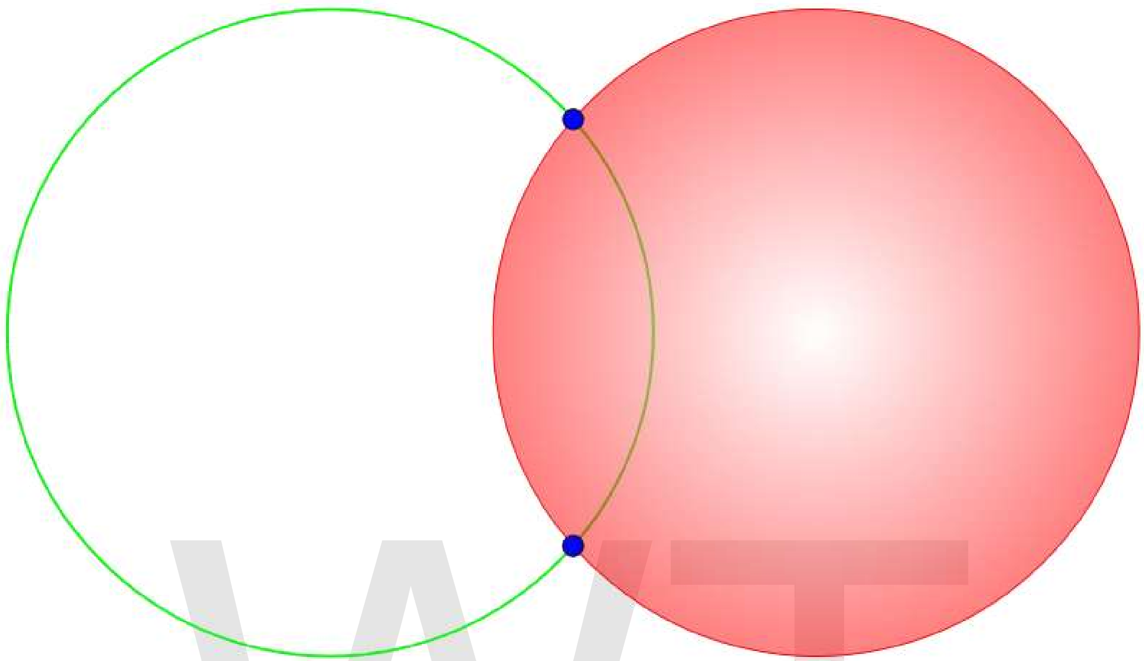
A satellite's position and pseudorange define a sphere, centered on the satellite with radius equal to the pseudorange. The position of the receiver is somewhere on the surface of this sphere. Thus with four satellites, the indicated position of the GPS receiver is at or near the intersection of the surfaces of four spheres. In the ideal case of no errors, the GPS receiver would be at a precise intersection of the four surfaces.

If the surfaces of two spheres intersect at more than one point, they intersect in a circle. A figure, *Two Sphere Surfaces Intersecting in a Circle*, is shown below. Two points where the surfaces of the spheres intersect are clearly shown in the figure. The distance between these two points is the diameter of the circle of intersection.



Two sphere surfaces intersecting in a circle

The intersection of a third spherical surface with the first two will be its intersection with that circle; in most cases of practical interest, this means they intersect at two points. Another figure, *Surface of Sphere Intersecting a Circle (not a solid disk) at Two Points*, illustrates the intersection. The two intersections are marked with dots.



Surface of sphere Intersecting a circle (not a solid disk) at two points

For automobiles and other near-earth vehicles, the correct position of the GPS receiver is the intersection closest to the Earth's surface. For space vehicles, the intersection farthest from Earth may be the correct one.

The correct position for the GPS receiver is also the intersection closest to the surface of the sphere corresponding to the fourth satellite.

Correcting a GPS receiver's clock

One of the most significant error sources is the GPS receiver's clock. Because of the very large value of the speed of light, c, the estimated distances from the GPS receiver to the satellites, the pseudoranges, are very sensitive to errors in the GPS receiver clock; for example an error of one microsecond (0.000 001 second) corresponds to an error of 300 metres (980 ft). This suggests that an extremely accurate and expensive clock is required for the GPS receiver to work. Because manufacturers prefer to build inexpensive GPS receivers for mass markets, the solution for this dilemma is based on the way sphere surfaces intersect in the GPS problem.

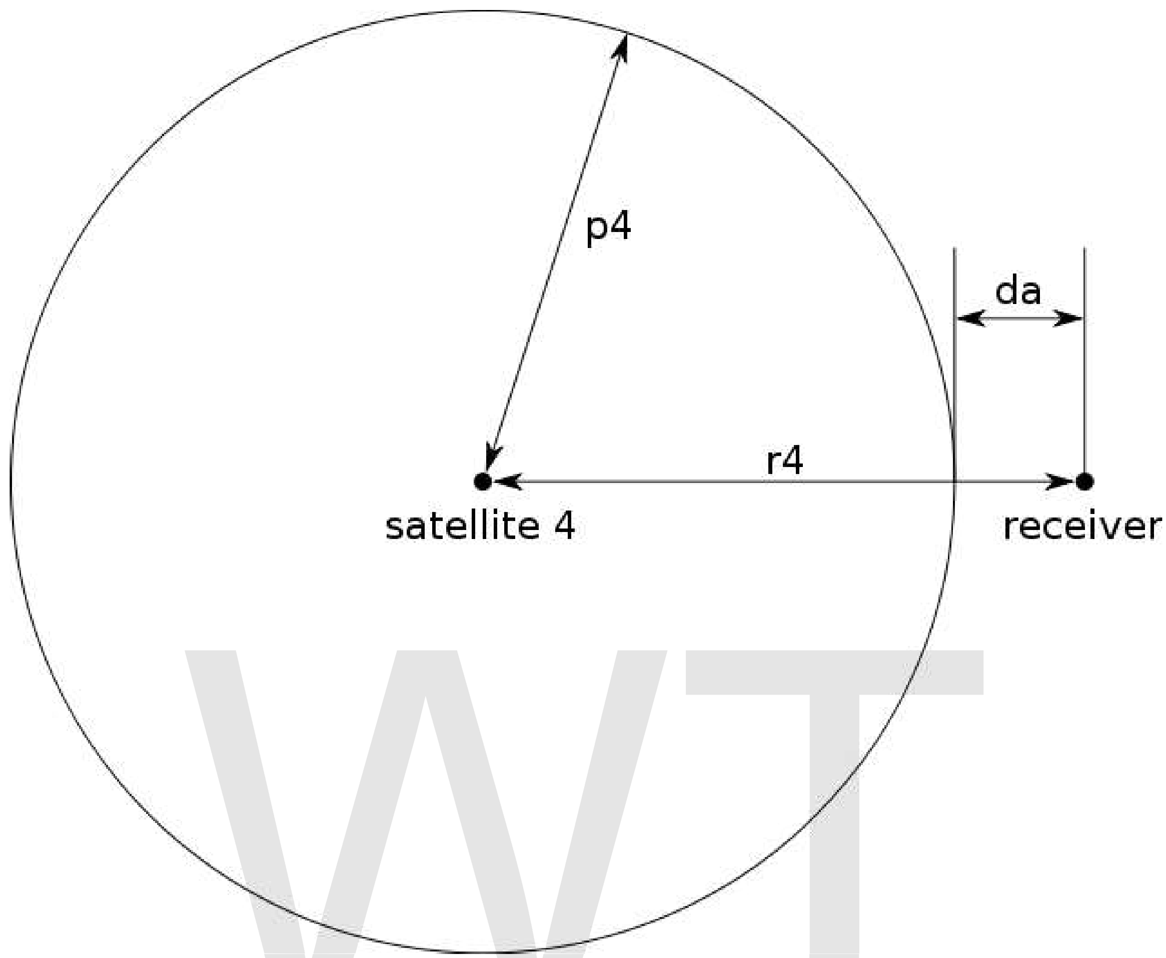


Diagram depicting satellite 4, sphere, p4, r4, and da

It is likely that the surfaces of the three spheres intersect, because the circle of intersection of the first two spheres is normally quite large, and thus the third sphere surface is likely to intersect this large circle. It is very unlikely that the surface of the sphere corresponding to the fourth satellite will intersect either of the two points of intersection of the first three, because any clock error could cause it to miss intersecting a point. However, the distance from the valid estimate of GPS receiver position to the surface of the sphere corresponding to the fourth satellite can be used to compute a clock correction. Let $r_4$denote the distance from the valid estimate of GPS receiver position to the fourth satellite and let $p_4$denote the pseudorange of the fourth satellite. Let $da = r_4 - p_4$. $da$is the distance from the computed GPS receiver position to the surface of the sphere corresponding to the fourth satellite. Thus the quotient, $b_r = -da/c$, provides an estimate of

(time indicated by the receiver's on-board clock) - (correct time),

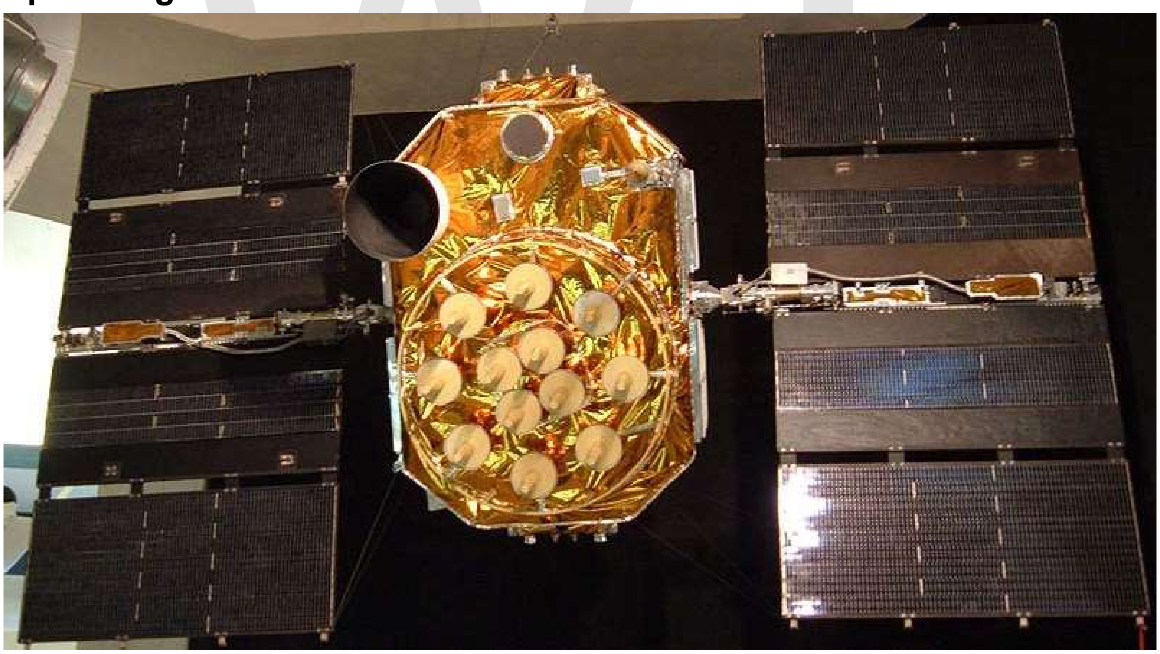
and the GPS receiver clock can be advanced if $b_r$is positive or delayed if $b_r$is negative. However, it should be kept in mind that a less simple function of damay be needed to estimate the time error in an iterative algorithm as discussed in the Navigation equations section.

Structure

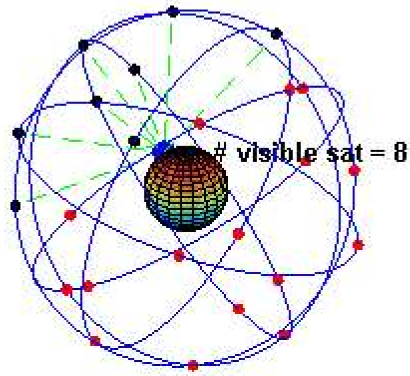
The current GPS consists of three major segments. These are the space segment (SS), a control segment (CS), and a user segment (U.S.). The U.S. Air Force develops, maintains, and operates the space and control segments. GPS satellites broadcast signals from space, and each GPS receiver uses these signals to calculate its three-dimensional location (latitude, longitude, and altitude) and the current time.

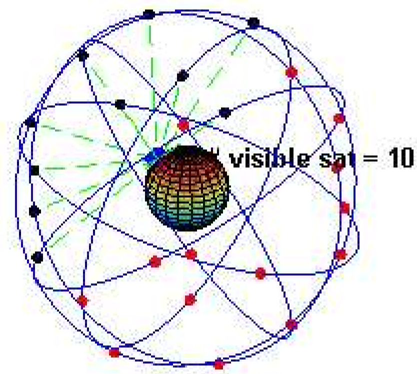
The space segment is composed of 24 to 32 satellites in medium Earth orbit and also includes the payload adapters to the boosters required to launch them into orbit. The control segment is composed of a master control station, an alternate master control station, and a host of dedicated and shared ground antennas and monitor stations. The user segment is composed of hundreds of thousands of U.S. and allied military users of the secure GPS Precise Positioning Service, and tens of millions of civil, commercial, and scientific users of the Standard Positioning Service.

Space segment



Unlaunched GPS satellite on display at the San Diego Air & Space Museum





A example of the GPS constellation in motion with the Earth rotating. Notice how the number of *satellites in view* from a given point on the Earth's surface, in this example at 45°N, changes with time.

The space segment (SS) is composed of the orbiting GPS satellites, or Space Vehicles (SV) in GPS parlance. The GPS design originally called for 24 SVs, eight each in three circular orbital planes, but this was modified to six planes with four satellites each. The orbital planes are centered on the Earth, not rotating with respect to the distant stars. The six planes have approximately 55° inclination (tilt relative to Earth's equator) and are separated by 60° right ascension of the ascending node (angle along the equator from a reference point to the orbit's intersection). The orbits are arranged so that at least six satellites are always within line of sight from almost everywhere on Earth's surface. The result of this objective is that the four satellites are not evenly spaced (90 degrees) apart within each orbit. In general terms, the angular difference between satellites in each orbit is 30, 105, 120, and 105 degrees apart which, of course, sum to 360 degrees.

Orbiting at an altitude of approximately 20,200 kilometers (about 12,550 miles or 10,900 nautical miles; orbital radius of approximately 26,600 km (about 16,500 mi or 14,400 NM)), each SV makes two complete orbits each sidereal day, repeating the same ground track each day. This was very helpful during development because even with only four satellites, correct alignment means all four are visible from one spot for a few hours each day. For military operations, the ground track repeat can be used to ensure good coverage in combat zones.

As of March 2008, there are 31 actively broadcasting satellites in the GPS constellation, and two older, retired from active service satellites kept in the constellation as orbital spares. The additional satellites improve the precision of GPS receiver calculations by providing redundant measurements. With the increased number of satellites, the constellation was changed to a nonuniform arrangement. Such an arrangement was shown to improve reliability and availability of the system, relative to a uniform system, when multiple satellites fail. About eight satellites are visible from any point on the ground at any one time.

Control segment

Ground monitor station used from 1984 to 2007, on display at the Air Force Space & Missile Museum

The control segment is composed of

1. a master control station (MCS),
2. an alternate master control station,
3. four dedicated ground antennas and
4. six dedicated monitor stations

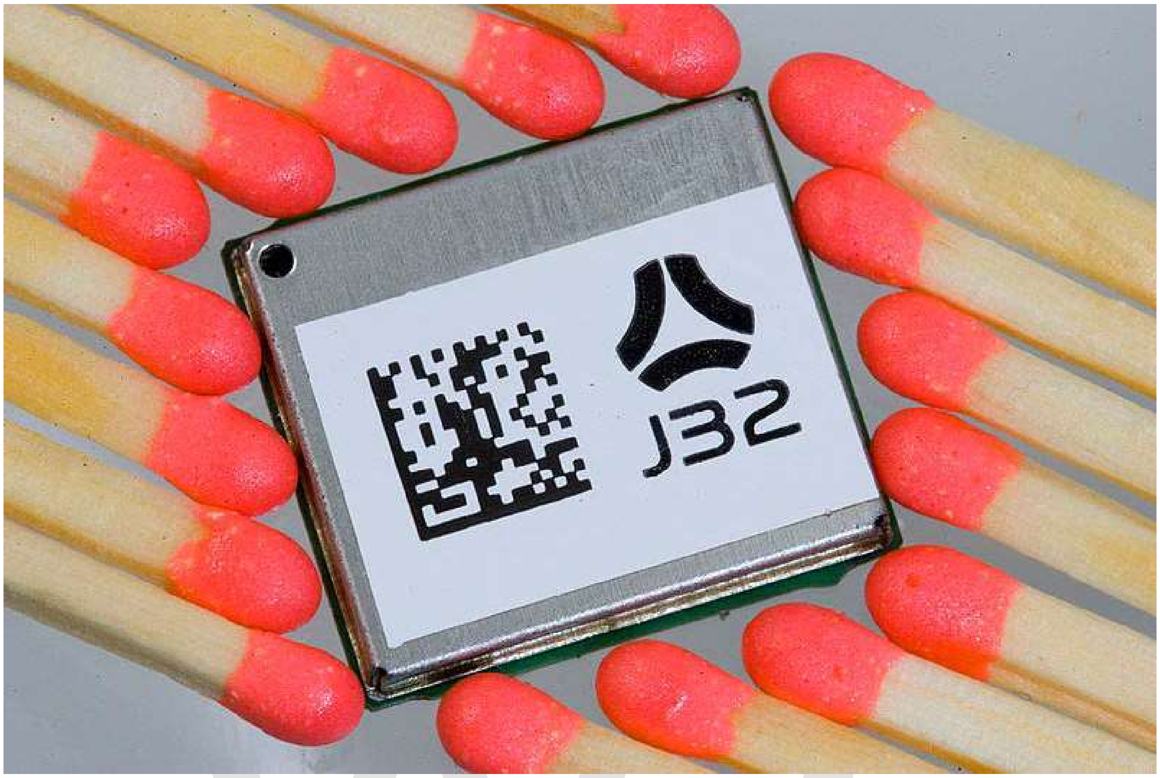
The MCS can also access U.S. Air Force Satellite Control Network (AFSCN) ground antennas (for additional command and control capability) and NGA (National Geospatial-Intelligence Agency) monitor stations. The flight paths of the satellites are tracked by dedicated U.S. Air Force monitoring stations in Hawaii, Kwajalein, Ascension Island, Diego Garcia, Colorado Springs, Colorado and Cape Canaveral, along with shared NGA monitor stations operated in England, Argentina, Ecuador, Bahrain, Australia and Washington DC. The tracking information is sent to the Air Force Space Command's MCS at Schriever Air Force Base 25 km (16 miles) ESE of Colorado Springs, which is operated by the 2nd Space Operations Squadron (2 SOPS) of the U.S. Air Force. Then 2 SOPS contacts each GPS satellite regularly with a navigational update using dedicated or shared (AFSCN) ground antennas (GPS dedicated ground antennas are located at Kwajalein, Ascension Island, Diego Garcia, and Cape Canaveral). These updates synchronize the atomic clocks on board the satellites to within a few nanoseconds of each other, and adjust the ephemeris of each satellite's internal orbital model. The updates are created by a Kalman filter that uses inputs from the ground monitoring stations, space weather information, and various other inputs.

Satellite maneuvers are not precise by GPS standards. So to change the orbit of a satellite, the satellite must be marked *unhealthy*, so receivers will not use it in their calculation. Then the maneuver can be carried out, and the resulting orbit tracked from the ground. Then the new ephemeris is uploaded and the satellite marked healthy again.

User segment

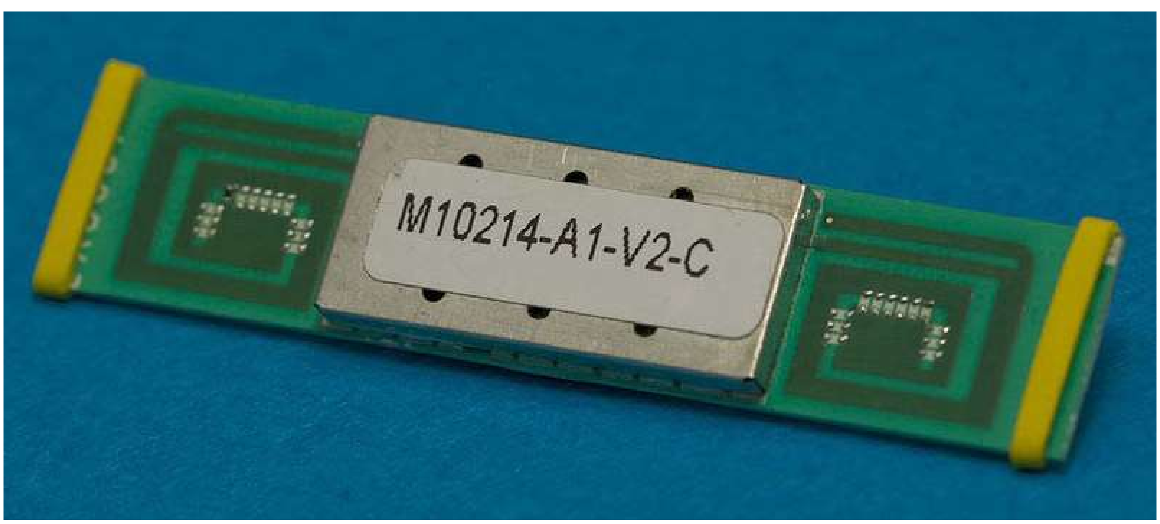
GPS receivers come in a variety of formats, from devices integrated into cars, phones, and watches, to dedicated devices such as those shown here from manufacturers Trimble, Garmin and Leica (left to right).

The user segment is composed of hundreds of thousands of U.S. and allied military users of the secure GPS Precise Positioning Service, and tens of millions of civil, commercial and scientific users of the Standard Positioning Service. In general, GPS receivers are composed of an antenna, tuned to the frequencies transmitted by the satellites, receiver-processors, and a highly stable clock (often a crystal oscillator). They may also include a display for providing location and speed information to the user. A receiver is often described by its number of channels: this signifies how many satellites it can monitor simultaneously. Originally limited to four or five, this has progressively increased over the years so that, as of 2007, receivers typically have between 12 and 20 channels.



A typical OEM GPS receiver module measuring 15×17 mm.

GPS receivers may include an input for differential corrections, using the RTCM SC-104 format. This is typically in the form of an RS-232 port at 4,800 bit/s speed. Data is actually sent at a much lower rate, which limits the accuracy of the signal sent using RTCM. Receivers with internal DGPS receivers can outperform those using external RTCM data. As of 2006, even low-cost units commonly include Wide Area Augmentation System (WAAS) receivers.



A typical GPS receiver with integrated antenna.

Many GPS receivers can relay position data to a PC or other device using the NMEA 0183 protocol. Although this protocol is officially defined by the National Marine Electronics Association (NMEA), references to this protocol have been compiled from public records, allowing open source tools like gpsd to read the protocol without violating intellectual property laws. Other proprietary protocols exist as well, such as the SiRF and MTK protocols. Receivers can interface with other devices using methods including a serial connection, USB, or Bluetooth.

Applications

While originally a military project, GPS is considered a *dual-use* technology, meaning it has significant military and civilian applications.

GPS has become a widely deployed and useful tool for commerce, scientific uses, tracking, and surveillance. GPS's accurate time facilitates everyday activities such as banking, mobile phone operations, and even the control of power grids by allowing well synchronized hand-off switching.

Civilian

This antenna is mounted on the roof of a hut containing a scientific experiment needing precise timing.

Many civilian applications use one or more of GPS's three basic components: absolute location, relative movement, and time transfer.

- Cellular telephony: Clock synchronization enables time transfer, which is critical for synchronizing its spreading codes with other base stations to facilitate inter-cell handoff and support hybrid GPS/cellular position detection for mobile emergency calls and other applications. The first handsets with integrated GPS launched in the late 1990s. The U.S. Federal Communications Commission (FCC) mandated the feature in either the handset or in the towers (for use in triangulation) in 2002 so emergency services could locate 911 callers. Third-party software developers later gained access to GPS APIs from Nextel upon launch, followed by Sprint in 2006, and Verizon soon thereafter.
- Disaster relief/emergency services: Depend upon GPS for location and timing capabilities.
- Geofencing: Vehicle tracking systems, person tracking systems, and pet tracking systems use GPS to locate a vehicle, person, or pet. These devices are attached to the vehicle, person, or the pet collar. The application provides continuous tracking and mobile or Internet updates should the target leave a designated area.
- Geotagging: Applying location coordinates to digital objects such as photographs and other documents for purposes such as creating map overlays.
- GPS Aircraft Tracking
- GPS tours: Location determines what content to display; for instance, information about an approaching point of interest.
- Map-making: Both civilian and military cartographers use GPS extensively.
- Navigation: Navigators value digitally precise velocity and orientation measurements.
- Phasor measurement units: GPS enables highly accurate timestamping of power system measurements, making it possible to compute phasors.
- Recreation: For example, geocaching, geodashing, GPS drawing and waymarking.
- Surveying: Surveyors use absolute locations to make maps and determine property boundaries.
- Tectonics: GPS enables direct fault motion measurement in earthquakes.

Restrictions on civilian use

The U.S. Government controls the export of some civilian receivers. All GPS receivers capable of functioning above 18 kilometers (11 mi) altitude and 515 metres per second (1,001 kn) are classified as munitions (weapons) for which U.S. State Department export licenses are required. These limits attempt to prevent use of a receiver in a ballistic missile. They would not prevent use in a cruise missile because their altitudes and speeds are similar to those of ordinary aircraft.

This rule applies even to otherwise purely civilian units that only receive the L1 frequency and the C/A (Clear/Acquisition) code and cannot correct for Selective Availability (SA), etc.

Disabling operation above these limits exempts the receiver from classification as a munition. Vendor interpretations differ. The rule targets operation given the combination of altitude and speed, while some receivers stop operating even when stationary. This has caused problems with some amateur radio balloon launches that regularly reach 30 kilometers (19 mi).

Military

Attaching a GPS guidance kit to a 'dumb' bomb, March 2003

As of 2009, military applications of GPS include:

- Navigation: GPS allows soldiers to find objectives, even in the dark or in unfamiliar territory, and to coordinate troop and supply movement. In the United States armed forces, commanders use the *Commanders Digital Assistant* and lower ranks use the *Soldier Digital Assistant*.
- Target tracking: Various military weapons systems use GPS to track potential ground and air targets before flagging them as hostile. These weapon systems pass target coordinates to precision-guided munitions to allow them to engage targets accurately. Military aircraft, particularly in air-to-ground roles, use GPS to find targets (for example, gun camera video from AH-1 Cobras in Iraq show GPS co-ordinates that can be viewed with special software.)
- Missile and projectile guidance: GPS allows accurate targeting of various military weapons including ICBMs, cruise missiles and precision-guided munitions. Artillery projectiles. Embedded GPS receivers able to withstand accelerations of 12,000 g or about 118 km/s^2 have been developed for use in 155 millimeters (6.1 in) howitzers.
- Search and Rescue: Downed pilots can be located faster if their position is known.

- Reconnaissance: Patrol movement can be managed more closely.
- GPS satellites carry a set of nuclear detonation detectors consisting of an optical sensor (Y-sensor), an X-ray sensor, a dosimeter, and an electromagnetic pulse (EMP) sensor (W-sensor), that form a major portion of the United States Nuclear Detonation Detection System.

Communication

The navigational signals transmitted by GPS satellites encode a variety of information including satellite positions, the state of the internal clocks, and the health of the network. These signals are transmitted on two separate carrier frequencies that are common to all satellites in the network. Two different encodings are used, a public encoding that enables lower resolution navigation, and an encrypted encoding used by the U.S. military.

Message format

GPS message format

Subframes	**Description**
1	Satellite clock, GPS time relationship
2–3	Ephemeris (precise satellite orbit)
4–5	Almanac component (satellite network synopsis, error correction)

Each GPS satellite continuously broadcasts a *navigation message* at a rate of 50 bits per second. Each complete message is composed of 30-second frames, distinct groupings of 1,500 bits of information. Each frame is further subdivided into 5 subframes of length 6 seconds and with 300 bits each. Each subframe contains 10 words of 30 bits with length 0.6 seconds each. Each 30 second frame begins precisely on the minute or half minute as indicated by the atomic clock on each satellite.

The first part of the message encodes the week number and the time within the week, as well as the data about the health of the satellite. The second part of the message, the *ephemeris*, provides the precise orbit for the satellite. The last part of the message, the *almanac*, contains coarse orbit and status information for all satellites in the network as well as data related to error correction.

All satellites broadcast at the same frequencies. Signals are encoded using code division multiple access (CDMA) allowing messages from individual satellites to be distinguished from each other based on unique encodings for each satellite (that the receiver must be aware of). Two distinct types of CDMA encodings are used: the coarse/acquisition (C/A) code, which is accessible by the general public, and the precise (P) code, that is encrypted so that only the U.S. military can access it.

The ephemeris is updated every 2 hours and is generally valid for 4 hours, with provisions for updates every 6 hours or longer in non-nominal conditions. The almanac is updated typically every 24 hours. Additionally data for a few weeks following is uploaded in case of transmission updates that delay data upload.

Satellite frequencies

GPS frequency overview

Band	Frequency	Description
L1	1575.42 MHz	Coarse-acquisition (C/A) and encrypted precision P(Y) codes, plus the L1 civilian (L1C) and military (M) codes on future Block III satellites.
L2	1227.60 MHz	P(Y) code, plus the L2C and military codes on the Block IIR-M and newer satellites.
L3	1381.05 MHz	Used for nuclear detonation (NUDET) detection.
L4	1379.913 MHz	Being studied for additional ionospheric correction.
L5	1176.45 MHz	Proposed for use as a civilian safety-of-life (SoL) signal.

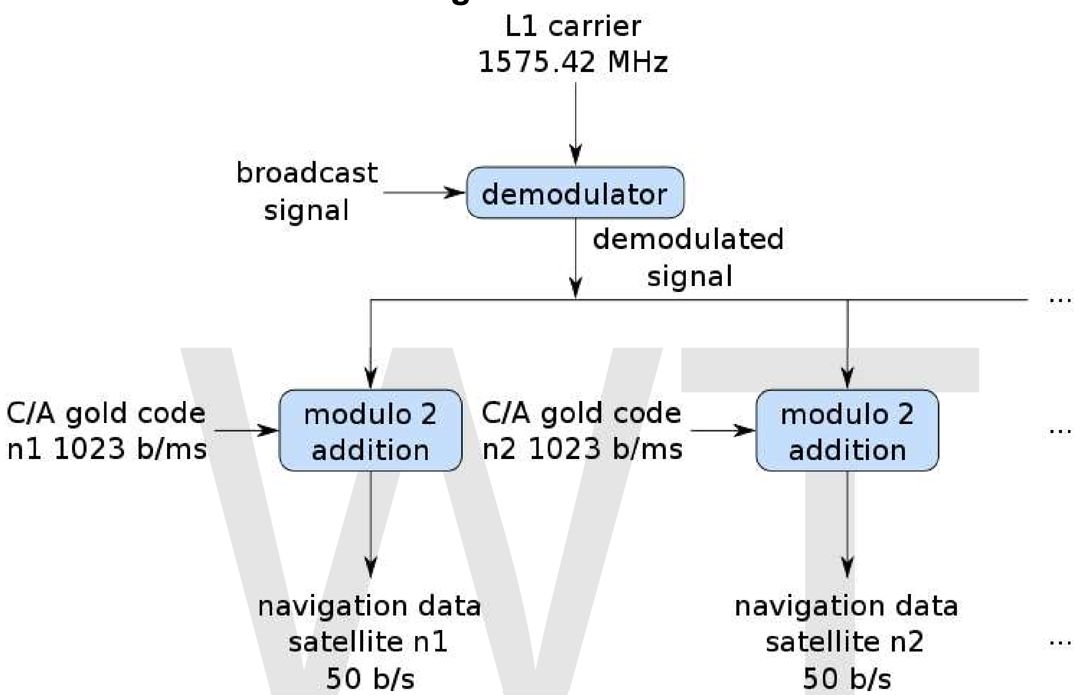
All satellites broadcast at the same two frequencies, 1.57542 GHz (L1 signal) and 1.2276 GHz (L2 signal). The satellite network uses a CDMA spread-spectrum technique where the low-bitrate message data is encoded with a high-rate pseudo-random (PRN) sequence that is different for each satellite. The receiver must be aware of the PRN codes for each satellite to reconstruct the actual message data. The C/A code, for civilian use, transmits data at 1.023 million chips per second, whereas the P code, for U.S. military use, transmits at 10.23 million chips per second. The L1 carrier is modulated by both the C/A and P codes, while the L2 carrier is only modulated by the P code. The P code can be encrypted as a so-called P(Y) code that is only available to military equipment with a proper decryption key. Both the C/A and P(Y) codes impart the precise time-of-day to the user.

The L3 signal at a frequency of 1.38105 GHz is used by the United States Nuclear Detonation (NUDET) Detection System (USNDS) to detect, locate, and report nuclear detonations (NUDETs) in the Earth's atmosphere and near space. One usage is the enforcement of nuclear test ban treaties.

The L4 band at 1.379913 GHz is being studied for additional ionospheric correction.

The L5 frequency band at 1.17645 GHZ was added in the process of GPS modernization. This frequency falls into an internationally protected range for aeronautical navigation, promising little or no interference under all circumstances. The first Block IIF satellite that would provide this signal is set to be launched in 2009. The L5 consists of two carrier components that are in phase quadrature with each other. Each carrier component is bi-phase shift key (BPSK) modulated by a separate bit train.

Demodulation and decoding



Demodulating and Decoding GPS Satellite Signals using the Coarse/Acquisition Gold code.

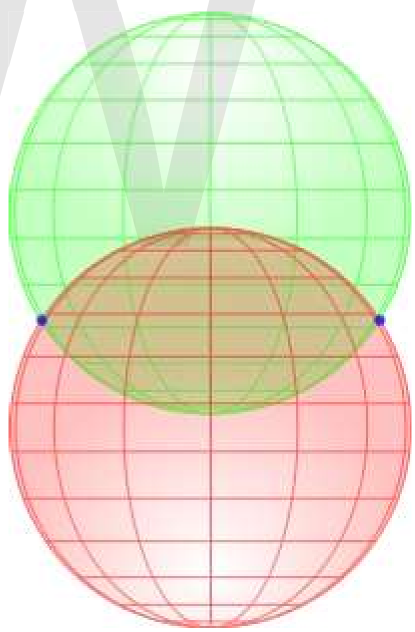
Because all of the satellite signals are modulated onto the same L1 carrier frequency, the signals must be separated after demodulation. This is done by assigning each satellite a unique binary sequence known as a Gold code. The signals are decoded after demodulation using addition of the Gold codes corresponding to the satellites monitored by the receiver.

If the almanac information has previously been acquired, the receiver picks the satellites to listen for by their PRNs, unique numbers in the range 1 through 32. If the almanac information is not in memory, the receiver enters a search mode until a lock is obtained on one of the satellites. To obtain a lock, it is necessary that there be an unobstructed line of sight from the receiver to the satellite. The receiver can then acquire the almanac and determine the satellites it should listen for. As it detects each satellite's signal, it identifies it by its distinct C/A code pattern. There can be a delay of up to 30 seconds before the first estimate of position because of the need to read the ephemeris data.

Processing of the navigation message enables the determination of the time of transmission and the satellite position at this time.

Navigation equations

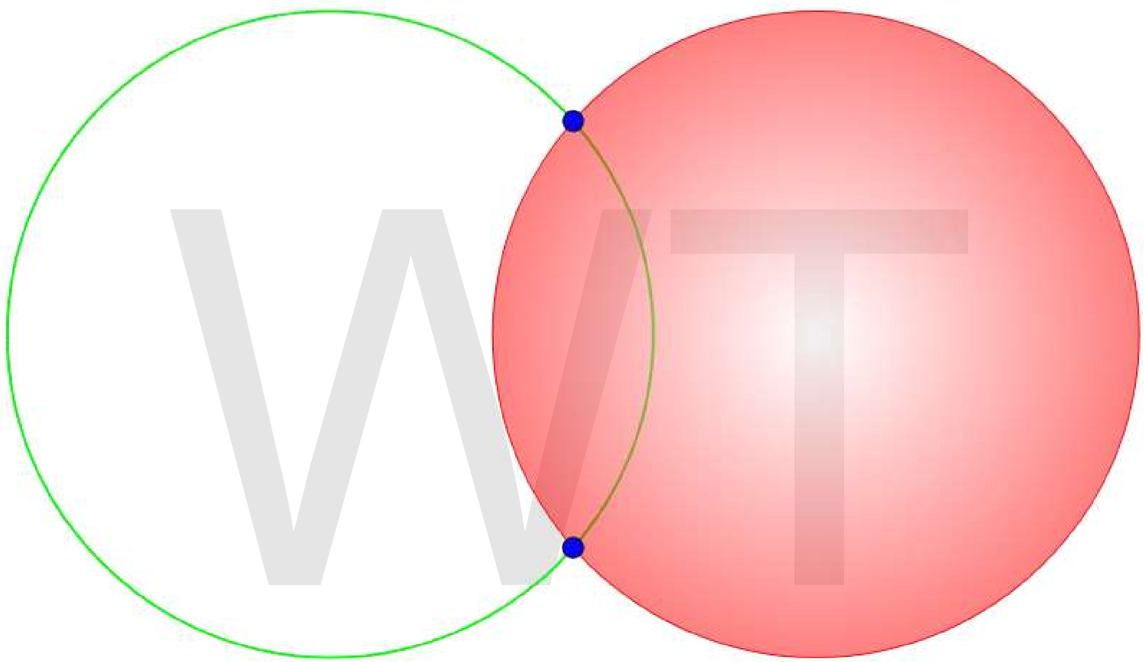
The receiver uses messages received from four satellites to determine the satellite positions and time sent. The *x, y,* and *z* components of position and the time sent are designated as $[x_i, y_i, z_i, t_i]$ where the subscript *i* denotes the satellite and has the value 1, 2, 3, or 4. Knowing when the message was received t_{r}, the receiver computes the message's transit time as $(t_{\text{r}} - t_i)$. Assuming the message traveled at the speed of light (c) the distance traveled is $(t_{\text{r}} - t_i)\,c$. Knowing the distance from receiver to satellite and the satellite's position implies that the receiver is on the surface of a sphere centered at the satellite's position. Thus the receiver is at or near the intersection of the surfaces of four spheres. In the ideal case of no errors, the receiver is at the intersection of the surfaces of four spheres. Excluding the unrealistic case (for GPS purposes) of two coincident spheres, the surfaces of two intersecting spheres is either a point (if they merely touch) or a circle as depicted in the illustration below. Two of the points where the surfaces of the spheres intersect are clearly marked on the figure. The distance between these two points is the diameter of the circle of intersection.



Two sphere surfaces intersecting in a circle

This can be seen more clearly by considering a side view of the intersecting spheres. This view would match the figure because of the symmetry of the spheres. A view from any horizontal direction would look exactly the same. Therefore the diameter as seen from all directions is the same and thus the surfaces actually do intersect in a circle.

Having found that two sphere surfaces intersect in a circle, we now consider how the intersection of the first two sphere surfaces, the circle, intersect with the third sphere. A circle and sphere surface intersect at zero, one or two points. For the GPS problem we are concerned with the case of two points of intersection. Another figure, Surface of Sphere Intersecting a Circle (not a solid disk) at Two Points, is shown below to aid in visualizing this intersection. Trilateration algebraically confirms this geometric observation. The ambiguity of two points of intersection of three sphere surfaces can be resolved by noting the point that is closest to the fourth sphere surface.



Surface of a sphere intersecting a circle (i.e., the edge of a disk) at two points

Having provided a discussion of how sphere surfaces intersect, we now formulate the equations for the case when errors are present.

Let b denote the clock error or bias, the amount that the receiver's clock is off. The receiver has four unknowns, the three components of GPS receiver position and the clock bias $[x, y, z, b]$. The equation of the sphere surfaces are given by:

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = \left([t_r + b - t_i]c\right)^2,$$

$$i = 1, 2, 3, 4$$

Another useful form of these equations is in terms of *pseudoranges*, which are the approximate ranges based on the receiver clock's uncorrected time so that $p_i = (t_\text{r} - t_i)\, c$. Then the equations becomes:

$$p_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - bc,\ \ i = 1, 2, 3, 4.$$

Methods of solution of navigation equations

The navigation equations can be solved by an algebraic method, called the Bancroft Method or by numerical methods involving trilateration or multidimensional root finding.

Bancroft's method

Bancroft's method is perhaps the most important method of solving the navigation equations because it involves an algebraic as opposed to numerical method. The method requires at least four satellites but more can be used.

Trilateration

The receiver can use trilateration and one dimensional numerical root finding. Trilateration is used to determine the intersection of the surfaces of three spheres. In the usual case of two intersections, the point nearest the surface of the sphere corresponding to the fourth satellite is chosen. The Earth's surface can also sometimes be used instead, especially by civilian GPS receivers, because it is illegal in the United States to track vehicles more than 60,000 feet (18,000 m) in altitude. Let da denote the signed magnitude of the vector from the receiver position to the fourth satellite (i.e. da = r4 - p4) as defined in the section, Correcting a GPS receiver's clock. da is a function of the correction because the correction changes the satellite transmission times and thus the pseudoranges. The notation, da(correction) denotes this function. The problem is to determine the correction such that

$$da\,(correction) = 0.$$

This is the familiar problem of finding the zeroes of a one dimensional non-linear function of a scalar variable. Iterative numerical methods, such as those found in the chapter on root finding in *Numerical Recipes* can solve this type of problem. One advantage of this method is that it involves one dimensional as opposed to multidimensional numerical root finding.

Multidimensional Newton-Raphson calculations

- Alternatively, multidimensional root finding method such as Newton-Raphson method can be used. The approach is to linearize around an approximate solution, say $\left[x_\text{r}^{(k)}, y_\text{r}^{(k)}, z_\text{r}^{(k)}, b_\text{r}^{(k)}\right]$ from iteration *k,* then solve four linear equations derived from the quadratic equations above to obtain $\left[x_r^{(k+1)}, y_r^{(k+1)}, z_r^{(k+1)}, b_r^{(k+1)}\right]$. The Newton-Raphson method is more rapidly convergent than other methods of numerical root finding. A disadvantage of this multidimensional root finding method as compared to single dimensional root findiing is that, "There are no good general methods for solving systems of more than one nonlinear equations."

- When more than four satellites are available, the calculation can use the four best or more than four, considering number of channels, processing capability, and geometric dilution of precision (GDOP). Using more than four is an over-determined system of equations with no unique solution, which must be solved by least-squares or a similar technique. If all visible satellites are used, the results are as good as or better than using the four best. Errors can be estimated through the residuals. With each combination of four or more satellites, a GDOP factor can be calculated, based on the relative sky directions of the satellites used. As more satellites are picked up, pseudoranges from various 4-way combinations can be processed to add more estimates to the location and clock offset. The receiver then takes the weighted average of these positions and clock offsets. After the final location and time are calculated, the location is expressed in a specific coordinate system such as latitude and longitude, using the WGS 84 geodetic datum or a country-specific system.
- Finally, results from other positioning systems such as GLONASS or the upcoming Galileo can be incorporated or used to check the result. (By design, these systems use the same frequency bands, so much of the receiver circuitry can be shared, though the decoding is different.)

Error sources and analysis

Error analysis for the Global Positioning System is an important aspect for determining what errors and their magnitude are to be expected. GPS errors are affected by geometric dilution of precision and depend on signal arrival time errors, numerical errors, atmospherics effects, ephemeris erros, multipath errors and other effects.

Accuracy enhancement and surveying

Augmentation

Integrating external information into the calculation process can materially improve accuracy. Such augmentation systems are generally named or described based on how the information arrives. Some systems transmit additional error information (such as clock drift, ephemera, or ionospheric delay), others characterize prior errors, while a third group provides additional navigational or vehicle information.

Examples of augmentation systems include the Wide Area Augmentation System (WAAS), European Geostationary Navigation Overlay Service (EGNOS), Differential GPS, Inertial Navigation Systems (INS) and Assisted GPS.

Precise monitoring

Accuracy can be improved through precise monitoring and measurement of existing GPS signals in additional or alternate ways.

The largest remaining error is usually the unpredictable delay through the ionosphere. The spacecraft broadcast ionospheric model parameters, but errors remain. This is one reason GPS spacecraft transmit on at least two frequencies, L1 and L2. Ionospheric delay is a well-defined function of frequency and the total electron content (TEC) along the path, so measuring the arrival time difference between the frequencies determines TEC and thus the precise ionospheric delay at each frequency.

Military receivers can decode the P(Y)-code transmitted on both L1 and L2. Without decryption keys, it is still possible to use a *codeless* technique to compare the P(Y) codes on L1 and L2 to gain much of the same error information. However, this technique is slow, so it is currently available only on specialized surveying equipment. In the future, additional civilian codes are expected to be transmitted on the L2 and L5 frequencies. Then all users will be able to perform dual-frequency measurements and directly compute ionospheric delay errors.

A second form of precise monitoring is called *Carrier-Phase Enhancement* (CPGPS). This corrects the error that arises because the pulse transition of the PRN is not instantaneous, and thus the correlation (satellite-receiver sequence matching) operation is imperfect. CPGPS uses the L1 carrier wave, which has a period of $\frac{1\ \text{sec}}{1575.42 * 10^6} = 0.63475\ \text{nanoseconds} \approx 1\ \text{nanosecond}$, which is about one-thousandth of the C/A Gold code bit period of $\frac{1\ \text{sec}}{1023 * 10^3} = 977.5\ \text{nanosecond} \approx 1000\ \text{nanosecond}$, to act as an additional clock signal and resolve the uncertainty. The phase difference error in the normal GPS amounts to 2–3 metres (6.6–9.8 ft) of ambiguity. CPGPS working to within 1% of perfect transition reduces this error to 3 centimeters (1.2 in) of ambiguity. By eliminating this error source, CPGPS coupled with DGPS normally realizes between 20–30 centimetres (7.9–12 in) of absolute accuracy.

Relative Kinematic Positioning (RKP) is a third alternative for a precise GPS-based positioning system. In this approach, determination of range signal can be resolved to a precision of less than 10 centimeters (3.9 in). This is done by resolving the number of cycles that the signal is transmitted and received by the receiver by using a combination of differential GPS (DGPS) correction data, transmitting GPS signal phase information and ambiguity resolution techniques via statistical tests—possibly with processing in real-time (real-time kinematic positioning, RTK).

Timekeeping

Timekeeping and leap seconds

While most clocks are synchronized to Coordinated Universal Time (UTC), the atomic clocks on the satellites are set to *GPS time*. The difference is that GPS time is not corrected to match the rotation of the Earth, so it does not contain leap seconds or other corrections that are periodically added to UTC. GPS time was set to match Coordinated Universal Time (UTC) in 1980, but has since diverged. The lack of corrections means that GPS time remains at a constant offset with International Atomic Time (TAI) (TAI - GPS = 19 seconds). Periodic corrections are performed on the on-board clocks to correct relativistic effects and keep them synchronized with ground clocks.

The GPS navigation message includes the difference between GPS time and UTC, which as of 2011 is 15 seconds because of the leap second added to UTC December 31, 2008. Receivers subtract this offset from GPS time to calculate UTC and specific timezone values. New GPS units may not show the correct UTC time until after receiving the UTC offset message. The GPS-UTC offset field can accommodate 255 leap seconds (eight bits) that, given the current period of the Earth's rotation (with one leap second introduced approximately every 18 months), should be sufficient to last until approximately the year 2300.

Timekeeping accuracy

GPS time is accurate to about 14ns.

Timekeeping format

As opposed to the year, month, and day format of the Gregorian calendar, the GPS date is expressed as a week number and a seconds-into-week number. The week number is transmitted as a ten-bit field in the C/A and P(Y) navigation messages, and so it becomes zero again every 1,024 weeks (19.6 years). GPS week zero started at 00:00:00 UTC (00:00:19 TAI) on January 6, 1980, and the week number became zero again for the first time at 23:59:47 UTC on August 21, 1999 (00:00:19 TAI on August 22, 1999). To determine the current Gregorian date, a GPS receiver must be provided with the approximate date (to within 3,584 days) to correctly translate the GPS date signal. To address this concern the modernized GPS navigation message uses a 13-bit field that only repeats every 8,192 weeks (157 years), thus lasting until the year 2137 (157 years after GPS week zero).

Carrier phase tracking (surveying)

Another method that is used in surveying applications is carrier phase tracking. The period of the carrier frequency times the speed of light gives the wavelength, which is about 0.19 meters for the L1 carrier. Accuracy within 1% of wavelength in detecting the leading edge, reduces this component of pseudorange error to as little as 2 millimeters. This compares to 3 meters for the C/A code and 0.3 meters for the P code.

However, 2 millimeter accuracy requires measuring the total phase—the number of waves times the wavelength plus the fractional wavelength, which requires specially equipped receivers. This method has many surveying applications.

Triple differencing followed by numerical root finding, and a mathematical technique called least squares can estimate the position of one receiver given the position of another. First, compute the difference between satellites, then between receivers, and finally between epochs. Other orders of taking differences are equally valid. Detailed discussion of the errors is omitted.

The satellite carrier total phase can be measured with ambiguity as to the number of cycles. Let $\phi(r_i, s_j, t_k)$ denote the phase of the carrier of satellite *j* measured by receiver *i* at time t_k. This notation shows the meaning of the subscripts *i, j,* and *k.* The receiver (*r*), satellite (*s*), and time (*t*) come in alphabetical order as arguments of ϕ and to balance readability and conciseness, let $\phi_{i,j,k} = \phi(r_i, s_j, t_k)$ be a concise abbreviation. Also we define three functions, :$\Delta^r, \Delta^s, \Delta^t$, which return differences between receivers, satellites, and time points, respectively. Each function has variables with three subscripts as its arguments. These three functions are defined below. If $\alpha_{i,j,k}$ is a function of the three integer arguments, *i, j,* and *k* then it is a valid argument for the functions, :$\Delta^r, \Delta^s, \Delta^t$, with the values defined as

$$\Delta^r(\alpha_{i,j,k}) = \alpha_{i+1,j,k} - \alpha_{i,j,k},$$
$$\Delta^s(\alpha_{i,j,k}) = \alpha_{i,j+1,k} - \alpha_{i,j,k}, \text{ and}$$
$$\Delta^t(\alpha_{i,j,k}) = \alpha_{i,j,k+1} - \alpha_{i,j,k}.$$

Also if $\alpha_{i,j,k}\ and\ \beta_{l,m,n}$ are valid arguments for the three functions and *a* and *b* are constants then $(a\ \alpha_{i,j,k} + b\ \beta_{l,m,n})$ is a valid argument with values defined as

$$\Delta^r(a\ \alpha_{i,j,k} + b\ \beta_{l,m,n}) = a\ \Delta^r(\alpha_{i,j,k}) + b\ \Delta^r(\beta_{l,m,n}),$$
$$\Delta^s(a\ \alpha_{i,j,k} + b\ \beta_{l,m,n}) = a\ \Delta^s(\alpha_{i,j,k}) + b\ \Delta^s(\beta_{l,m,n}), \text{ and}$$
$$\Delta^t(a\ \alpha_{i,j,k} + b\ \beta_{l,m,n}) = a\ \Delta^t(\alpha_{i,j,k}) + b\ \Delta^t(\beta_{l,m,n}).$$

Receiver clock errors can be approximately eliminated by differencing the phases measured from satellite 1 with that from satellite 2 at the same epoch. This difference is designated as $\Delta^s(\phi_{1,1,1}) = \phi_{1,2,1} - \phi_{1,1,1}$

Double differencing computes the difference of receiver 1's satellite difference from that of receiver 2. This approximately eliminates satellite clock errors. This double difference is:

$$\Delta^r(\Delta^s(\phi_{1,1,1})) = \Delta^r(\phi_{1,2,1} - \phi_{1,1,1}) = \Delta^r(\phi_{1,2,1}) - \Delta^r(\phi_{1,1,1}) = (\phi_{2,2,1} - \phi_{1,2,1}) - (\phi_{2,1,1} - \phi_{1,1,1})$$

Triple differencing subtracts the receiver difference from time 1 from that of time 2. This eliminates the ambiguity associated with the integral number of wave lengths in carrier phase provided this ambiguity does not change with time. Thus the triple difference result eliminates practically all clock bias errors and the integer ambiguity. Atmospheric delay and satellite ephemeris errors have been significantly reduced. This triple difference is:

$$\Delta^t(\Delta^r(\Delta^s(\phi_{1,1,1})))$$

Triple difference results can be used to estimate unknown variables. For example if the position of receiver 1 is known but the position of receiver 2 unknown, it may be possible to estimate the position of receiver 2 using numerical root finding and least squares. Triple difference results for three independent time pairs quite possibly will be sufficient to solve for receiver 2's three position components. This may require the use of a numerical procedure. An approximation of receiver 2's position is required to use such a numerical method. This initial value can probably be provided from the navigation message and the intersection of sphere surfaces. Such a reasonable estimate can be key to successful multidimensional root finding. Iterating from three time pairs and a fairly good initial value produces one observed triple difference result for receiver 2's position. Processing additional time pairs can improve accuracy, overdetermining the answer with multiple solutions. Least squares can estimate an overdetermined system. Least squares determines the position of receiver 2 which best fits the observed triple difference results for receiver 2 positions under the criterion of minimizing the sum of the squares.

Other systems

Other satellite navigation systems in use or various states of development include:

- Galileo – a global system being developed by the European Union and other partner countries, planned to be operational by 2014
- Beidou – People's Republic of China's regional system, covering Asia and the West Pacific
- COMPASS – People's Republic of China's global system, planned to be operational by 2020
- GLONASS – Russia's global navigation system
- IRNSS – India's regional navigation system, planned to be operational by 2012, covering India and Northern Indian Ocean
- QZSS – Japanese regional system covering Asia and Oceania

Chapter 3

Surveying

US Navy Surveyor at work with a leveling instrument

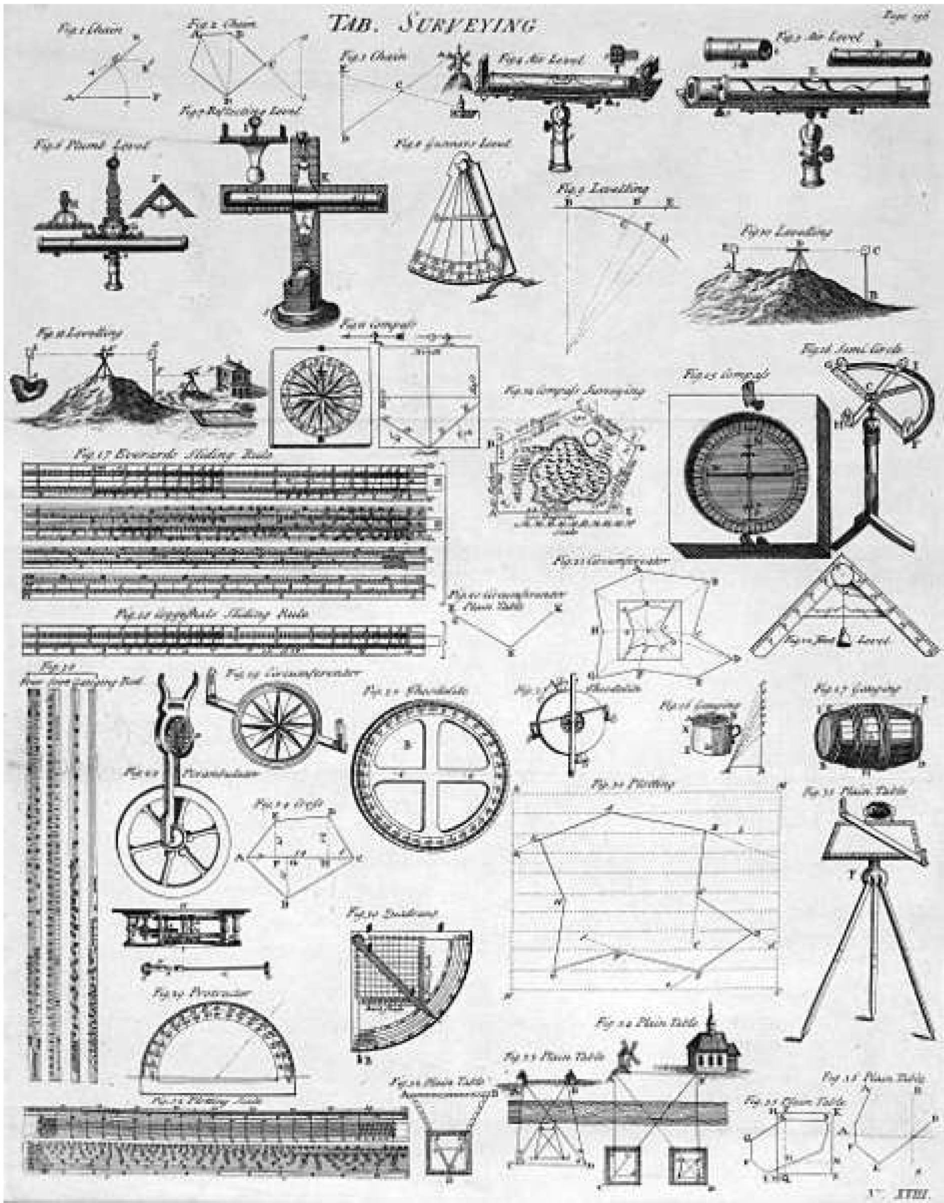


Table of Surveying, 1728 *Cyclopaedia*

Surveying or land surveying is the technique and science of accurately determining the terrestrial or three-dimensional position of points and the distances and angles between them. These points are usually on the surface of the Earth, and they are often used to establish land maps and boundaries for ownership or governmental purposes.

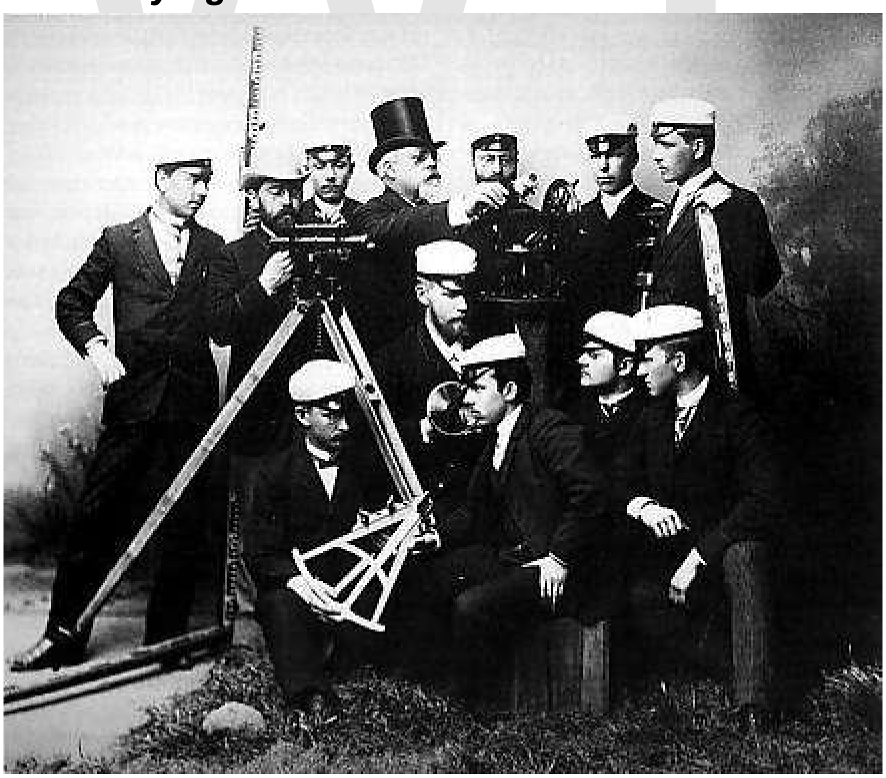
To accomplish their objective, **surveyors** use elements of geometry, engineering, trigonometry, mathematics, physics, and law.

An alternative definition, per the American Congress on Surveying and Mapping (ACSM), is the science and art of making all essential measurements to determine the relative position of points and/or physical and cultural details above, on, or beneath the surface of the Earth, and to depict them in a usable form, or to establish the position of points and/or details.

Furthermore, as alluded to above, a particular type of surveying known as "land surveying" (also per ACSM) is the detailed study or inspection, as by gathering information through observations, measurements in the field, questionnaires, or research of legal instruments, and data analysis in the support of planning, designing, and establishing of property boundaries. It involves the re-establishment of cadastral surveys and land boundaries based on documents of record and historical evidence, as well as certifying surveys (as required by statute or local ordinance) of subdivision plats/maps, registered land surveys, judicial surveys, and space delineation. Land surveying can include associated services such as mapping and related data accumulation, construction layout surveys, precision measurements of length, angle, elevation, area, and volume, as well as horizontal and vertical control surveys, and the analysis and utilization of land survey data.

Surveying has been an essential element in the development of the human environment since the beginning of recorded history (about 5,000 years ago). It is required in the planning and execution of nearly every form of construction. Its most familiar modern uses are in the fields of transport, building and construction, communications, mapping, and the definition of legal boundaries for land ownership.

History of surveying



Surveying students with professor at the Helsinki University of Technology in the early 20th century.

Surveying techniques have existed throughout much of recorded history. In ancient Egypt, when the Nile River overflowed its banks and washed out farm boundaries, boundaries were re-established by a rope stretcher, or surveyor, through the application of simple geometry. The nearly perfect squareness and north-south orientation of the Great Pyramid of Giza, built c. 2700 BC, affirm the Egyptians' command of surveying.

- The Egyptian land register (3000 BC).
- A recent reassessment of Stonehenge (c. 2500 BC) indicates that the monument was set out by prehistoric surveyors using peg and rope geometry.
- The Groma surveying instrument originated in Mesopotamia (early 1st millennium BC).
- Under the Romans, land surveyors were established as a profession, and they established the basic measurements under which the Roman Empire was divided, such as a tax register of conquered lands (300 AD).
- The rise of the Caliphate led to extensive surveying throughout the Arab Empire. Arabic surveyors invented a variety of specialized instruments for surveying, including:
    - Instruments for accurate leveling: A wooden board with a plumb line and two hooks, an equilateral triangle with a plumb line and two hooks, and a reed level.
    - A rotating alhidade, used for accurate alignment.
    - A surveying astrolabe, used for alignment, measuring angles, triangulation, finding the width of a river, and the distance between two points separated by an impassable obstruction.
- In England, The Domesday Book by William the Conqueror (1086)
    - covered all England
    - contained names of the land owners, area, land quality, and specific information of the area's content and inhabitants.
    - did not include maps showing exact locations

In the 18th century in Europe triangulation was used to build a hierarchy of networks to allow point positioning within a country. Highest in the hierarchy were triangulation networks. These were densified into networks of traverses (polygons), into which local mapping surveying measurements, usually with measuring tape, corner prism and the familiar red and white poles, are tied. For example, in the late 1780s, a team from the Ordnance Survey of Great Britain, originally under General William Roy began the Principal Triangulation of Britain using the specially built Ramsden theodolite. Large scale surveys are known as geodetic surveys.

- Continental Europe's Cadastre was created in 1808
    - founded by Napoleon I (Bonaparte)
    - contained numbers of the parcels of land (or just land), land usage, names etc., and value of the land
    - 100 million parcels of land, triangle survey, measurable survey, map scale: 1:2500 and 1:1250

- spread fast around Europe, but faced problems especially in Mediterranean countries, Balkan, and Eastern Europe due to cadastre upkeep costs and troubles.

A cadastre loses its value if register and maps are not constantly updated. Because of the fundamental value of land and real estate to the local and global economy, land surveying was one of the first professions to require Professional Licensure. In many jurisdictions, the land surveyors license was the first Professional Licensure issued by the state, province, or federal government.

Surveying techniques

A standard Brunton Geo compass, still used commonly today by geologists and surveyors for field-based measurements.

Example of modern hardware for surveying (Field-Map technology): GPS, laser rangefinder and field computer allows surveying as well as cartography (creation of map in real-time) and field data collection.

Historically, distances were measured using a variety of means, such as with chains having links of a known length, for instance a Gunter's chain, or measuring tapes made of steel or invar. To measure horizontal distances, these chains or tapes were pulled taut according to temperature, to reduce sagging and slack. Additionally, attempts to hold the measuring instrument level would be made. In instances of measuring up a slope, the surveyor might have to "break" (break chain) the measurement- use an increment less than the total length of the chain.

Historically, horizontal angles were measured using a compass, which would provide a magnetic bearing, from which deflections could be measured. This type of instrument was later improved, with more carefully scribed discs providing better angular resolution, as well as through mounting telescopes with reticles for more-precise sighting atop the disc. Additionally, levels and calibrated circles allowing measurement of vertical angles were added, along with verniers for measurement to a fraction of a degree—such as with a turn-of-the-century transit.

The simplest method for measuring height is with an altimeter — basically a barometer — using air pressure as an indication of height. But surveying requires greater precision. A variety of means, such as precise levels (also known as differential leveling), have been developed to do this. With precise leveling, a series of measurements between two points are taken using an instrument and a measuring rod. Differentials in height between the measurements are added and subtracted in a series to derive the net difference in elevation between the two endpoints of the series. With the advent of the Global Positioning System (GPS), elevation can also be derived with sophisticated satellite receivers, but usually with somewhat less accuracy than with traditional precise leveling. However, the accuracies may be similar if the traditional leveling would have to be run over a long distance.

Triangulation is another method of horizontal location made almost obsolete by GPS. With the triangulation method, distances, elevations and directions between objects at great distance from one another can be determined. Since the early days of surveying, this was the primary method of determining accurate positions of objects for topographic maps of large areas. A surveyor first needs to know the horizontal distance between two of the objects. Then the height, distances and angular position of other objects can be derived, as long as they are visible from one of the original objects. High-accuracy transits or theodolites were used for this work, and angles between objects were measured repeatedly for increased accuracy.

Surveying equipment

A German engineer surveying during the First World War, 1918

As late as the 1990s, the basic tools used in planar surveying were a tape measure for determining shorter distances, a level to determine height or elevation differences, and a theodolite, set on a tripod, to measure angles (horizontal and vertical), combined with the process of triangulation. Starting from a position with known location and elevation, the distance and angles to the unknown point are measured.

A more modern instrument is a total station, which is a theodolite with an electronic distance measurement device (EDM). A total station can also be used for leveling when set to the horizontal plane. Since their introduction, total stations have made the technological shift from being optical-mechanical devices to being fully electronic with an onship computer and software as well as humans.

Modern top-of-the-line total stations no longer require a reflector or prism (used to return the light pulses used for distancing) to return distance measurements, are fully robotic, and can even e-mail point data to the office computer and connect to satellite positioning systems, such as a Global Positioning System (GPS). Though real-time kinematic GPS systems have increased the speed of surveying, they are still horizontally accurate to only about 20 mm and vertically accurate to about 30–40 mm.

Total stations are still used widely, along with other types of surveying instruments. However, GPS systems do not work well in areas with dense tree cover or constructions. One-person robotic-guided total stations allow surveyors to gather precise measurements without extra workers to look through and turn the telescope or record data. A faster but expensive way to measure large areas (not details, and no obstacles) is with a helicopter, equipped with a laser scanner, combined with a GPS to determine the position and elevation of the helicopter. To increase precision, beacons are placed on the ground (about 20 km (12 mi) apart). This method reaches precisions between 5–40 cm (depending on flight height).

Types of surveys and applicability

- *ALTA/ACSM Survey*: a surveying standard jointly proposed by the American Land Title Association and the American Congress on Surveying and Mapping that incorporates elements of the boundary survey, mortgage survey, and topographic survey.
- *Archaeological survey*: used to accurately assess the relationship of archaeological sites in a landscape or to accurately record finds on an archaeological site.
- *As-built survey*: a survey carried out during or immediately after a construction project for record, completion evaluation and payment purposes. An as-built survey also known as a 'works as executed survey' documents the location of the recently constructed elements that are subject to completion evaluation. As built surveys are typically presented in red or redline and overlayed over existing design plans for direct comparison with design information.
- *Bathymetric survey:* a survey carried out to map the topography and features of the bed of an ocean, lake, river or other body of water.
- *Boundary survey*: a survey that establishes boundaries of a parcel using its legal description, which typically involves the setting or restoration of monuments or markers at the corners or along the lines of the parcel, often in the form of iron rods, pipes, or concrete monuments in the ground, or nails set in concrete or asphalt.
- *Deformation survey:* a survey to determine if a structure or object is changing shape or moving. The three-dimensional positions of specific points on an object are determined, a period of time is allowed to pass, these positions are then re-measured and calculated, and a comparison between the two sets of positions is made.
- *Engineering surveys*: those surveys associated with the engineering design (topographic, layout and as-built) often requiring geodetic computations beyond normal civil engineering practise.
- *Foundation survey*: a survey done to collect the positional data on a foundation that has been poured and is cured. This is done to ensure that the foundation was constructed in the location, and at the elevation, authorized in the *plot plan*, *site plan*, or *subdivision plan*.
- *Geological survey*: generic term for a survey conducted for the purpose of recording the geologically significant features of the area under investigation. .

- *Hydrographic survey*: a survey conducted with the purpose of mapping the coastline and seabed for navigation, engineering, or resource management purposes.
- *Measured survey* : a building survey to produce plans of the building. such a survey may be conducted before renovation works, for commercial purpose, or at end of the construction process "as built survey"
- *Mortgage survey or physical survey*: a simple survey that delineates land boundaries and building locations. In many places a mortgage survey is required by lending institutions as a precondition for a mortgage loan.
- *Soil survey*, or soil mapping, is the process of determining the soil types or other properties of the soil cover over a landscape, and mapping them for others to understand and use.
- *Structural survey*: a detailed inspection to report upon the physical condition and structural stability of a building or other structure and to highlight any work needed to maintain it in good repair.
- *Tape survey*: this type of survey is the most basic and inexpensive type of land survey. Popular in the middle part of the 20th century, tape surveys while being accurate for distance lack substantially in their accuracy of measuring angle and bearing. Standards that are practiced by professional land surveyors.
- *Topographic survey*: a survey that measures the elevation of points on a particular piece of land, and presents them as contour lines on a plot.

Surveying as a career

The basic principles of surveying have changed little over the ages, but the tools used by surveyors have evolved tremendously. Engineering, especially civil engineering, depends heavily on surveyors.

Whenever there are roads, railways, reservoir, dams, retaining walls, bridges or residential areas to be built, surveyors are involved. They establish the boundaries of legal descriptions and the boundaries of various lines of political divisions. They also provide advice and data for *geographical information systems* (GIS), computer databases that contain data on land features and boundaries.

Surveyors must have a thorough knowledge of algebra, basic calculus, geometry, and trigonometry. They must also know the laws that deal with surveys, property, and contracts.

In addition, they must be able to use delicate instruments with accuracy and precision. In the United States, surveyors and civil engineers use units of feet wherein a survey foot is broken down into 10ths and 100ths. Many deed descriptions requiring distance calls are often expressed using these units (125.25 ft). On the subject of accuracy, surveyors are often held to a standard of one one-hundredth of a foot; about 1/8th inch. Calculation and mapping tolerances are much smaller wherein achieving near-perfect closures are desired. Though tolerances such as this will vary from project to project, in the field and day to day usage beyond a 100th of a foot is often impractical.

In most of the United States, surveying is recognized as a distinct profession apart from engineering. Licensing requirements vary by state, but they generally have components of education, experience and examinations. In the past, experience gained through an apprenticeship, together with passing a series of state-administered examinations, was required to attain licensure. Now, most states insist upon basic qualification of a degree in surveying, plus experience and examination requirements.

The licensing process typically follows two phases. First, upon graduation, the candidate may be eligible to take the Fundamentals of Land Surveying exam, to be certified upon passing and meeting all other requirements as a surveyor in training (SIT). Upon being certified as an SIT, the candidate then needs to gain additional experience to become eligible for the second phase. That typically consists of the Principles and Practice of Land Surveying exam along with a state-specific examination.

An all-female surveying crew in Idaho, 1918

Licensed surveyors usually denote themselves with the letters P.S. (professional surveyor), L.S. (land surveyor), P.L.S. (professional land surveyor), R.L.S. (registered land surveyor), R.P.L.S. (Registered Professional Land Surveyor), or P.S.M. (professional surveyor and mapper) following their names, depending upon the dictates of their particular state of registration.

In Canada, land Surveyors are registered to work in their respective province. The designation for a land surveyor breaks down by province, but follows the rule whereby the first letter indicates the province, followed by L.S. There is also a designation as a C.L.S. or Canada lands surveyor, who has the authority to work on Canada Lands, which include Indian Reserves, National Parks, the three territories and offshore lands.

In many Commonwealth countries, the term Chartered Land Surveyor is used for someone holding a professional license to conduct surveys.

A licensed land surveyor is typically required to sign and seal all plans, the format of which is dictated by their state jurisdiction, which shows their name and registration number. In many states, when setting boundary corners land surveyors are also required to place survey monuments bearing their registration numbers, typically in the form of capped iron rods, concrete monuments, or nails with washers.

Building surveying

Building surveying emerged in the 1970s as a profession in the United Kingdom by a group of technically minded General Practice Surveyors. Building surveying is a recognised profession in Britain, Australia and Hong Kong. In Australia in particular, due to risk mitigation and limitation factors, the employment of surveyors at all levels of the construction industry is widespread. There are still many countries where it is not widely recognized as a profession.

Services that building surveyors undertake are broad but can include:

- Construction design and building works
- Project management and monitoring
- Property Legislation advice
- Insurance assessment and claims assistance
- Defect investigation and maintenance advice
- Building surveys and measured surveys
- Handling planning applications
- Building inspection to ensure compliance with building regulations
- Pre-acquisition surveys
- Negotiating dilapidations claims

Building surveyors also advise on many aspects of construction including:

- design
- maintenance
- repair
- refurbishment
- restoration and preservation of buildings and monuments

Clients of a building surveyor can be the government agencies, businesses and individuals. Surveyors work closely with architects, planners, homeowners and tenants groups. Building surveyors may also be called to act as an expert witnesses. It is usual for building surveyors to earn a college degree before undertaking structured training to become a member of a professional organisation.

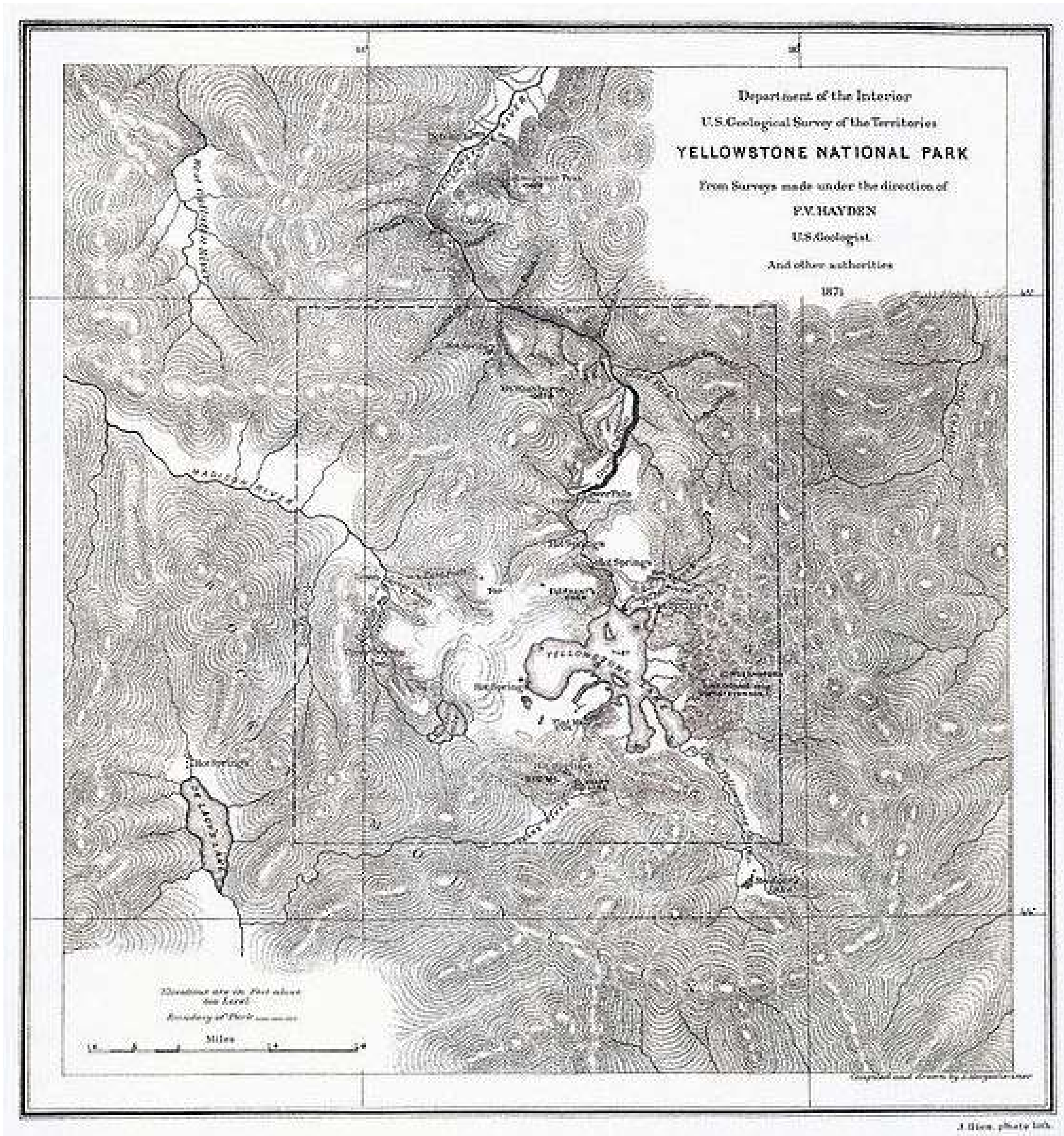
With the enlargement of the European community, the profession of the building surveyor is becoming more widely known in other European states, particularly France, where many English-speaking people buy second homes.

Lidar Surveying - Three-dimensional laser scanning provides high definition surveying for architectural, as-built, and engineering surveys. Recent technological advances make it the most cost-effective and time-sensitive solution for providing the highest level of detail available for interior and exterior building work.

Land surveyor

The job of the land surveyor is to retrace legal description(s) from the deed belonging to the subject property by locating actual reference monumentation and verifying its correct position. Over time, development, vandalism and acts of nature often wreak havoc on monuments, so the land surveyor is often forced to consider other evidence such as fence locations, woodlines, monuments on neighboring property, parole evidence and other evidence.

Reference monumentation refers to actual physical points on the ground that define location of boundary lines that divide neighboring parcels as well as their respective corners. Also called survey control, they are most often 1/2" or 5/8" iron rebar rods or pipes placed at 18" minimum depth. These rods and/or pipes usually have an affixed plastic cap over the top bearing the responsible surveyors' name and license number In addition to rods/pipes, 4x4" concrete posts are often used at corners of large parcels or anywhere that would require more stability (ex. beach sand). They are placed at a depth of 3 feet. In places where there is asphalt or concrete, it is common to place nails or aluminum alloy caps to re-establish boundary corners. Marks should be durable, stable, and as "permanent" as possible. The aim is to provide sufficient marks so some marks will remain for future re-establishment of boundaries. The material and marking used on monuments placed to mark boundary corners are often subject to state laws.



F.V. Hayden's map of Yellowstone National Park, 1871. His surveys were a significant factor toward establishing the park in 1872.

Cadastral land surveyors are licensed by state governments. In the United States, cadastral surveys are typically conducted by the federal government, specifically through the Cadastral Surveys branch of the Bureau of Land Management (BLM), formerly the General Land Office (GLO). This includes consultation and boundary determination expertise for USFS, Park Service, Corps of Engineers, BIA, Fish and Wildlife Service, Bureau of Reclamation, etc. In states that have been subdivided as per the Public Land Survey System (PLSS), the BLM Cadastral Surveys are carried out in accordance with that system. This information is required to define ownership and rights in real property (such as land, water, mineral, easements, rights-of-way), to resolve boundary disputes between neighbours, and for any subdivision of land, building development, road boundary realignment, etc.

The aim of cadastral surveys is normally to re-establish and mark the corners of original land boundaries. The first stage is to research relevant records such as land titles (deeds), easements, survey monumentation (marks on the ground) and any public or private records that provide relevant data.

In order to properly establish accurate position of survey markers, it is then necessary for measurements to be taken. This is achieved by placing a [total station] over the points and recording distances taken with the [EDM].

The data is analysed and comparisons made with existing records to determine evidence that can be used to establish boundary positions. The bearing and distance of lines between the boundary corners and total station positions are calculated and used to set out and mark the corners in the field. Checks are made by measuring directly between pegs places using a flexible tape. Subdivision of land generally requires that the external boundary is re-established and marked using pegs, and the new internal boundaries are then marked.

The art of surveying

Many properties have considerable problems with regards to improper bounding, miscalculations in past surveys, titles, easements, and wildlife crossings. Also many properties are created from multiple divisions of a larger piece over the course of years, and with every additional division the risk of miscalculation increases. The result can be abutting properties not coinciding with adjacent parcels, resulting in hiatuses (gaps) and overlaps. The art plays a role when a surveyor must solve a puzzle using pieces that do not exactly fit together. In these cases, the solution is based upon the surveyor's research and interpretation, along with established procedures for resolving discrepancies.

Senior Evidence - Priority #1

A land surveyor is an investigator of evidence. The land surveyor creates evidence on and under the ground to reference/preserve/perpetuate existing evidence. It is not the position of a land surveyor to make legal determinations; instead, the surveyor provides evidence that can be ruled on by a judge in a court of law (regarding legal decisions as to boundary lines). Evidence found and set by the land surveyor can be filed of record to be used for decisions as to land boundaries. In most States, it is the first one to the court house with that evidence that wins a boundary dispute unless other evidence is found with senior importance to the contrary. In most circumstances, the survey marker is of the highest priority as evidence of the boundary, unless evidence exists to show it was moved.

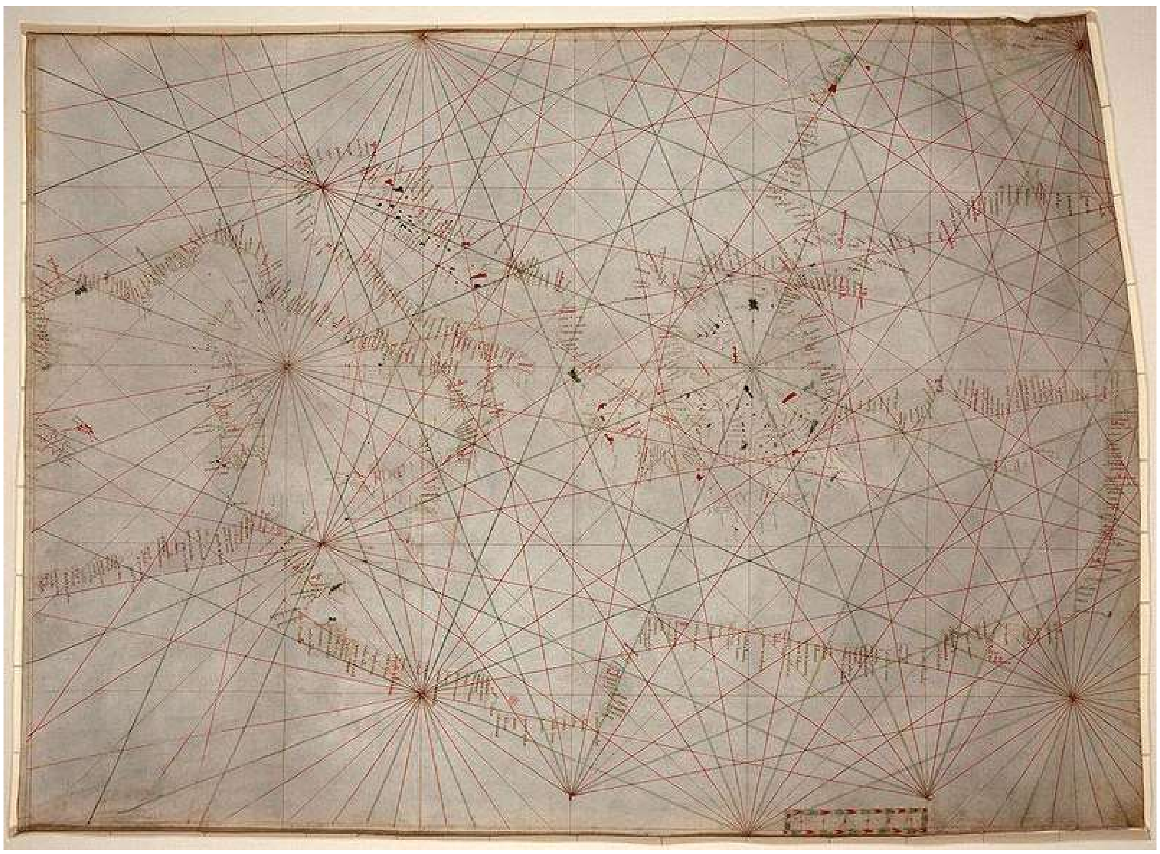
References to nearby survey markers are important when determining the "preponderance of evidence" for use by a judge who can set the legal boundary of land. A surveyor's opinion is valuable as evidence for legal boundary decisions, by those in authority, to legally settle a boundary dispute. In most cases, it is wiser for both parties to obtain the evidence and settle the dispute with the help of a land surveyor, than to press a suit in court.

It has become more common for title companies to try forcing a surveyor to change the measured distances and bearings to match existing documentation. If the title company is invested in the closing and promoting a faster "close" to avoid the cost of record corrections, this is a conflict of interest. Title company employees may not understand the importance of a bearing base and measured boundary, based on points found and missing points set, and based on the best fit provided by the surveyor. This has become more and more of a problem with the lack of common knowledge of the importance of land surveying evidence. The survey boundary based on survey field evidence, especially measured boundary markers, should overrule previous written documentation that does not include the description of the survey markers found by the land surveyor.

Many do not understand the true meaning of a "metes and bounds" boundary description. The "bounds" or physical location and relationship of the survey markers has priority over the "metes" or measurements in the recorded description of a boundary. For example, an old measurement of 420 yards at a bearing of 120 degrees does not take priority over the actual positions of the survey markers on both ends, unless a marker is missing and needs to be re-set using that information. Other evidence that will verify the position of the missing marker, based on the senior evidence nearby as first priority, is preferred in such a case.

Chapter 4

Cartography



A nautical chart of the Mediterranean Sea from the second quarter of the 14th century. It is the oldest original cartographic artifact in the Library of Congress.

Cartography (from Greek *chartis* = map and *graphein* = write) is the study and practice of making maps. Combining science, aesthetics, and technique, cartography builds on the premise that reality can be modeled in ways that communicate spatial information effectively.

The fundamental problems of traditional cartography are to:

- Set the map's agenda and select traits of the object to be mapped. This is the concern of map editing. Traits may be physical, such as roads or land masses, or may be abstract, such as toponyms or political boundaries.
- Represent the terrain of the mapped object on flat media. This is the concern of map projections.
- Eliminate characteristics of the mapped object that are not relevant to the map's purpose. This is the concern of generalization.
- Reduce the complexity of the characteristics that will be mapped. This is also the concern of generalization.
- Orchestrate the elements of the map to best convey its message to its audience. This is the concern of map design.

Modern cartography is closely integrated with geographic information science (GIScience) and constitutes many theoretical and practical foundations of geographic information systems.

History

Copy (1475) of St. Isidore's TO map of the world.

The earliest known map is a matter of some debate, both because the definition of "map" is not sharp and because some artifacts speculated to be maps might actually be something else. A wall painting, which may depict the ancient Anatolian city of Çatalhöyük (previously known as Catal Huyuk or Çatal Hüyük), has been dated to the late 7th millennium BCE. Other known maps of the ancient world include the Minoan “House of the Admiral” wall painting from c. 1600 BCE, showing a seaside community in an oblique perspective and an engraved map of the holy Babylonian city of Nippur, from the Kassite period (14th – 12th centuries BCE). The oldest surviving world maps are the Babylonian world maps from the 9th century BCE. One shows Babylon on the Euphrates, surrounded by a circular landmass showing Assyria, Urartu and several cities, in turn surrounded by a "bitter river" (Oceanus), with seven islands arranged around it. Another depicts Babylon as being further north from the center of the world.

The ancient Greeks and Romans created maps, beginning at latest with Anaximander in the 6th century BC. In the 2nd century AD, Ptolemy produced his treatise on cartography, Geographia. This contained Ptolemy's world map - the world then known to Western society *(Ecumene)*. As early as the 8th century, Arab scholars were translating the works of the Greek geographers into Arabic.

In ancient China, geographical literature spans back to the 5th century BC. The oldest extant Chinese maps come from the State of Qin, dated back to the 4th century BC, during the Warring States Period. In the book of the *Xin Yi Xiang Fa Yao*, published in 1092 by the Chinese scientist Su Song, a star map on the equidistant cylindrical projection. Although this method of charting seems to have existed in China even prior to this publication and scientist, the greatest significance of the star maps by Su Song is that they represent the oldest existent star maps in printed form.

Early forms of cartography of India included the locations of the Pole star and other constellations of use. These charts may have been in use by the beginning of the Common Era for purposes of navigation.

Mappa mundi is the general term used to describe Medieval European maps of the world. Approximately 1,100 mappae mundi are known to have survived from the Middle Ages. Of these, some 900 are found illustrating manuscripts and the remainder exist as stand-alone documents.

The *Tabula Rogeriana*, drawn by Muhammad al-Idrisi for Roger II of Sicily in 1154

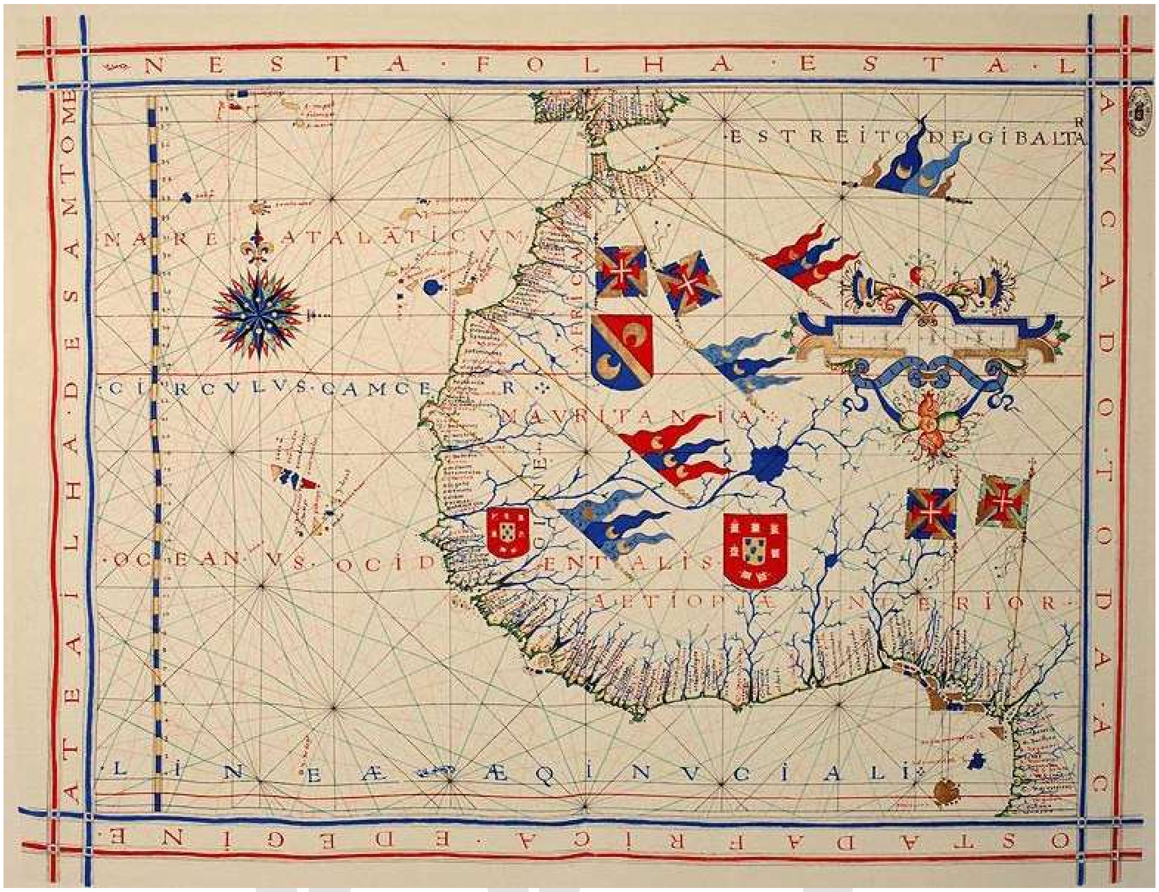
The Arab geographer Muhammad al-Idrisi produced his medieval atlas *Tabula Rogeriana* in 1154. He incorporated the knowledge of Africa, the Indian Ocean and the Far East, gathered by Arab merchants and explorers with the information inherited from the classical geographers to create the most accurate map of the world up until his time. It remained the most accurate world map for the next three centuries.

In the Age of Exploration, from the 15th century to the 17th century, European cartographers both copied earlier maps (some of which had been passed down for centuries) and drew their own based on explorers' observations and new surveying techniques. The invention of the magnetic compass, telescope and sextant enabled increasing accuracy. In 1492, Martin Behaim, a German cartographer, made the oldest extant globe of the Earth.

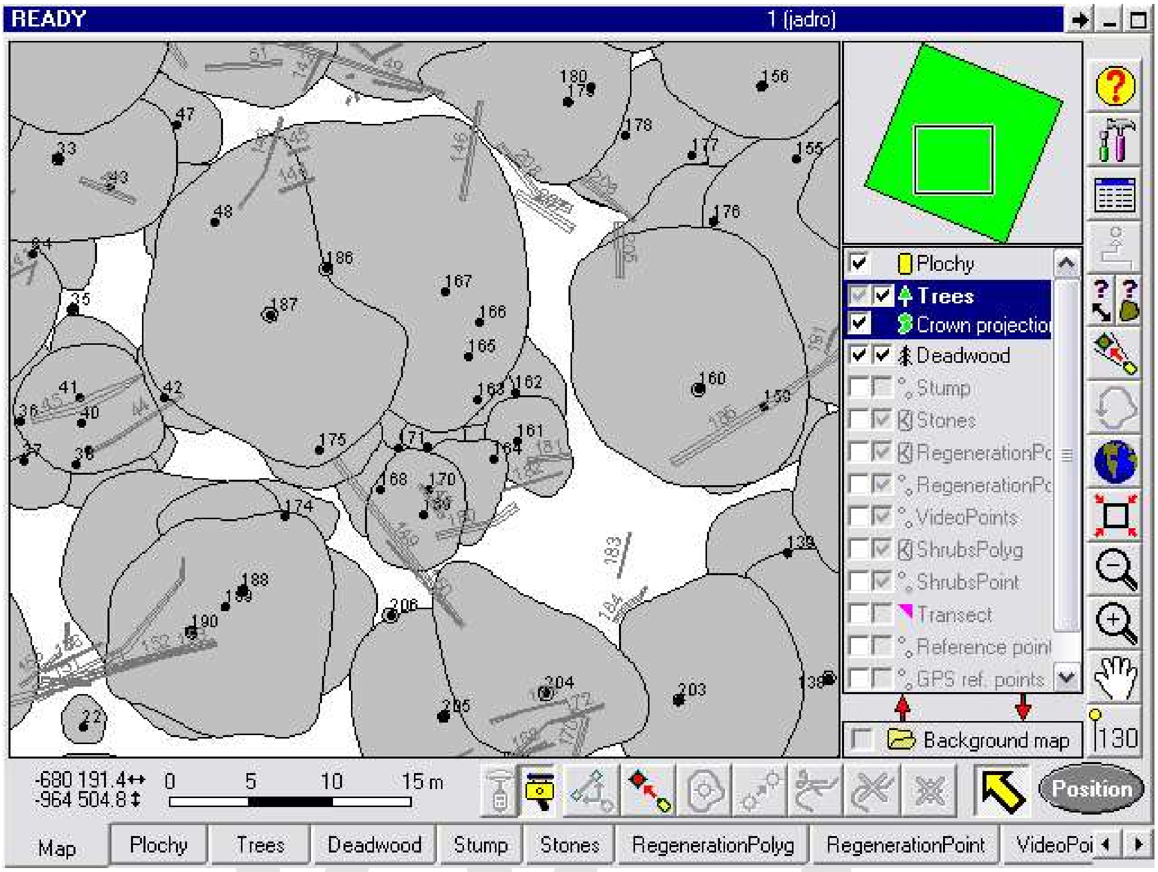
Johannes Werner refined and promoted the *Werner map projection*. In 1507, Martin Waldseemüller produced a globular world map and a large 12-panel world wall map (*Universalis Cosmographia*) bearing the first use of the name "America". Portuguese cartographer Diego Ribero was the author of the first known planisphere with a graduated Equator (1527). Italian cartographer Battista Agnese produced at least 71 manuscript atlases of sea charts.

Due to the sheer physical difficulties inherent in cartography, map-makers frequently lifted material from earlier works without giving credit to the original cartographer. For example, one of the most famous early maps of North America is unofficially known as the "Beaver Map", published in 1715 by Herman Moll. This map is an exact reproduction of a 1698 work by Nicolas de Fer. De Fer in turn had copied images that were first printed in books by Louis Hennepin, published in 1697, and François Du Creux, in 1664. By the 18th century, map-makers started to give credit to the original engraver by printing the phrase "After [the original cartographer]" on the work.

Technological changes



A pre-Mercator nautical chart of 1571, from Portuguese cartographer Fernão Vaz Dourado (c. 1520–c. 1580). It belongs to the so-called *plane chart* model, where observed latitudes and magnetic directions are plotted directly into the plane, with a constant scale, as if the Earth were plane (Portuguese National Archives of Torre do Tombo, Lisbon).



Mapping can be done with GPS and laser rangefinder directly in the field (for example by Field-Map technology). Real-time map construction improve productivity and quality of mapping. Image is showing mapping of forest structure (position of trees, dead wood and canopy).

In cartography, technology has continually changed in order to meet the demands of new generations of mapmakers and map users. The first maps were manually constructed with brushes and parchment; therefore, varied in quality and were limited in distribution. The advent of magnetic devices, such as the compass and much later, magnetic storage devices, allowed for the creation of far more accurate maps and the ability to store and manipulate them digitally.

Advances in mechanical devices such as the printing press, quadrant and vernier, allowed for the mass production of maps and the ability to make accurate reproductions from more accurate data. Optical technology, such as the telescope, sextant and other devices that use telescopes, allowed for accurate surveying of land and the ability of mapmakers and navigators to find their latitude by measuring angles to the North Star at night or the sun at noon.

Advances in photochemical technology, such as the lithographic and photochemical processes, have allowed for the creation of maps that have fine details, do not distort in shape and resist moisture and wear. This also eliminated the need for engraving, which further shortened the time it takes to make and reproduce maps.

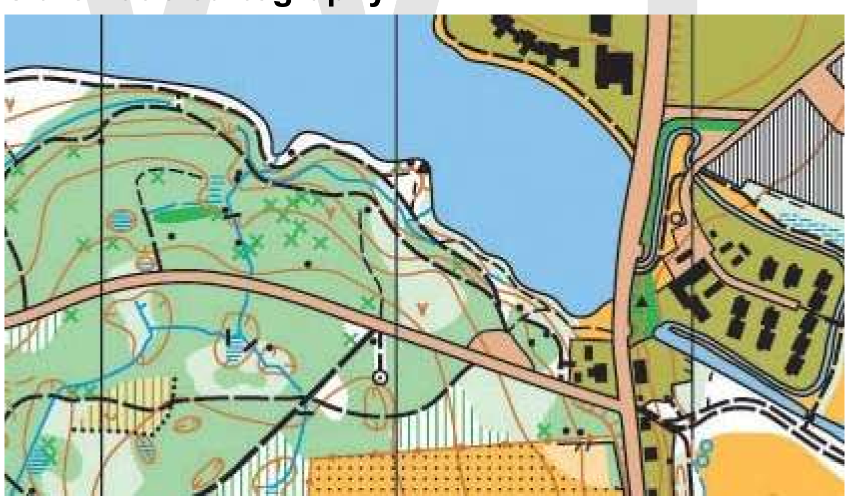
Advances in electronic technology in the 20th century ushered in another revolution in cartography. Ready availability of computers and peripherals such as monitors, plotters, printers, scanners (remote and document) and analytic stereo plotters, along with computer programs for visualization, image processing, spatial analysis, and database management, have democratized and greatly expanded the making of maps. The ability to superimpose spatially located variables onto existing maps created new uses for maps and new industries to explore and exploit these potentials.

These days most commercial-quality maps are made using software that falls into one of three main types: CAD, GIS and specialized illustration software. Spatial information can be stored in a database, from which it can be extracted on demand. These tools lead to increasingly dynamic, interactive maps that can be manipulated digitally.

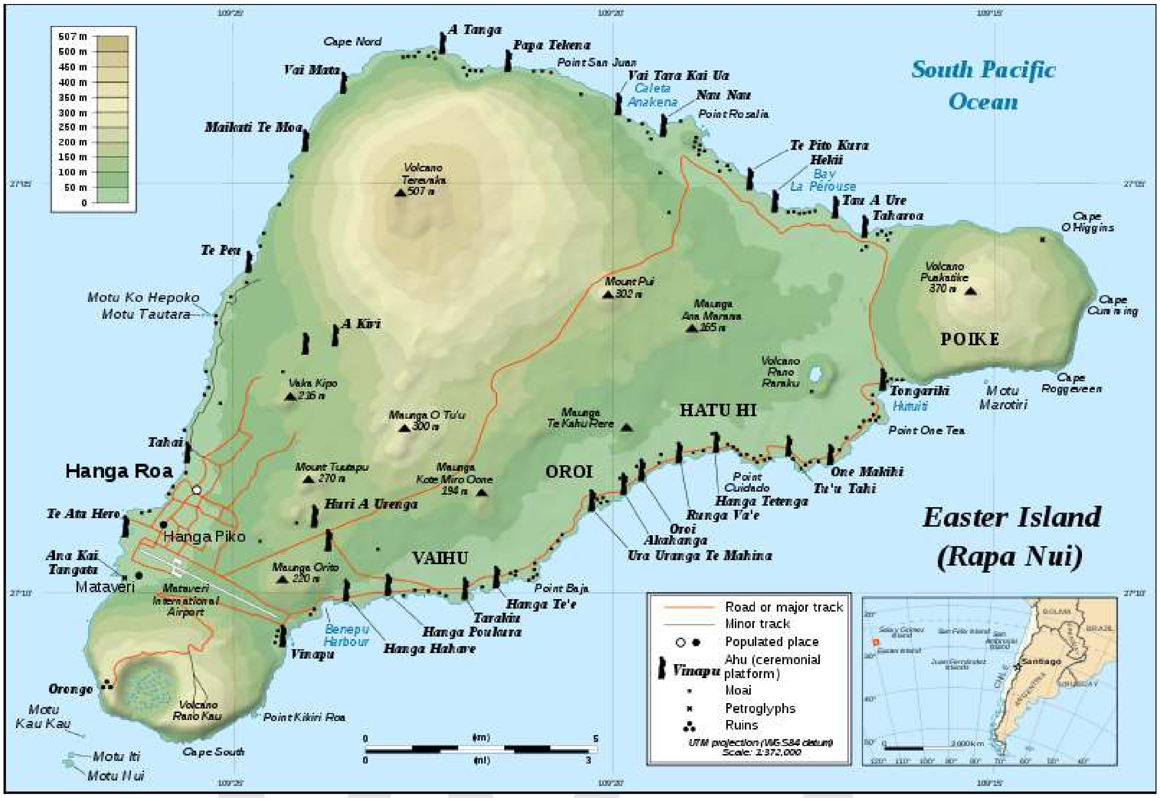
With the field rugged computers, GPS and laser rangefinders, it is possible to perform mapping directly in the terrain. The construction of the map in real time improve productivity and quality of the result. **Real time mapping** is done for example with Field-map technology.

Map types

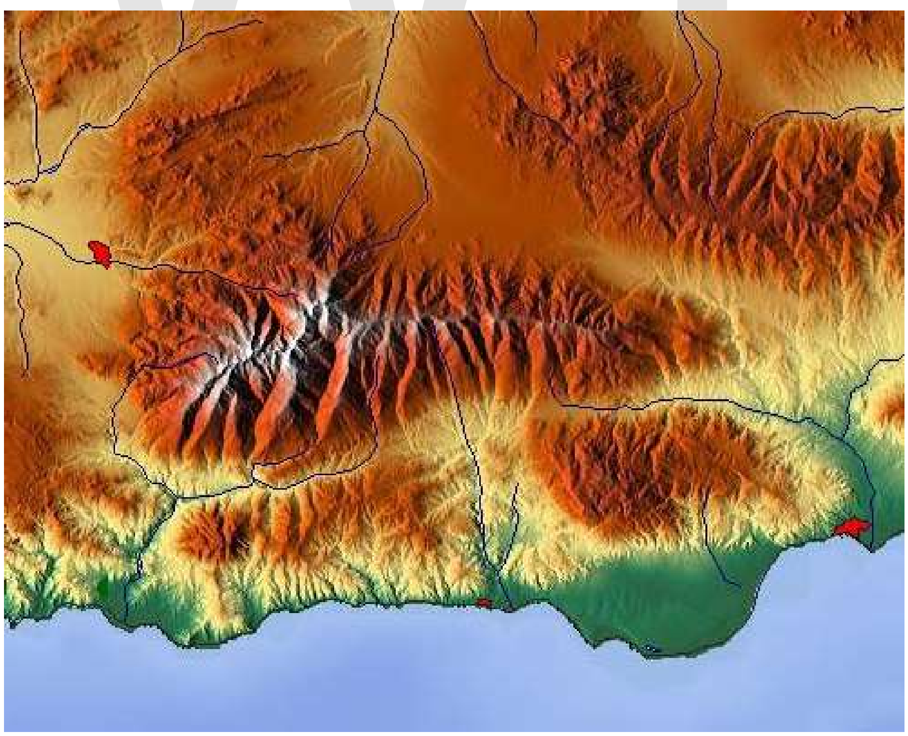
General vs thematic cartography



Small section of an orienteering map.



Topographic map of Easter Island



Relief map Sierra Nevada

In understanding basic maps, the field of cartography can be divided into two general categories: general cartography and thematic cartography. General cartography involves those maps that are constructed for a general audience and thus contain a variety of features. General maps exhibit many reference and location systems and often are produced in a series. For example, the 1:24,000 scale topographic maps of the United States Geological Survey (USGS) are a standard as compared to the 1:50,000 scale Canadian maps. The government of the UK produces the classic 1:50,000 (replacing the older 1 inch to 1 mile) "Ordnance Survey" maps of the entire UK and with a range of correlated larger- and smaller-scale maps of great detail.

Thematic cartography involves maps of specific geographic themes, oriented toward specific audiences. A couple of examples might be a dot map showing corn production in Indiana or a shaded area map of Ohio counties, divided into numerical choropleth classes. As the volume of geographic data has exploded over the last century, thematic cartography has become increasingly useful and necessary to interpret spatial, cultural and social data.

An orienteering map combines both general and thematic cartography, designed for a very specific user community. The most prominent thematic element is shading, that indicates degrees of difficulty of travel due to vegetation. The vegetation itself is not identified, merely classified by the difficulty ("fight") that it presents.

Topographic vs topological

A topographic map is primarily concerned with the topographic description of a place, including (especially in the 20th century) the use of contour lines showing elevation. Terrain or relief can be shown in a variety of ways.

A topological map is a very general type of map, the kind you might sketch on a napkin. It often disregards scale and detail in the interest of clarity of communicating specific route or relational information. Beck's London Underground map is an iconic example. Though the most widely used map of "The Tube," it preserves little of reality: It varies scale constantly and abruptly, it straightens curved tracks, and it contorts directions haphazardly. The only topography on it is the River Thames, letting the reader know whether a station is north or south of the river. That and the topology of station order and interchanges between train lines are all that is left of the geographic space. Yet those are all a typical passenger wishes to know, so the map fulfills its purpose.

Map design

Illustrated map

Map purpose and informations' selection

Arthur H. Robinson, an American cartographer influential in thematic cartography, stated that a map not properly designed "will be a cartographic failure." He also claimed, when considering all aspects of cartography, that "map design is perhaps the most complex." Robinson codified the mapmaker's understanding that a map must be designed foremost with consideration to the audience and its needs.

From the very beginning of mapmaking, maps "have been made for some particular purpose or set of purposes". The intent of the map should be illustrated in a manner in which the percipient acknowledges its purpose in a timely fashion. The term *percipient* refers to the person receiving information and was coined by Robinson. The principle of figure-ground refers to this notion of engaging the user by presenting a clear presentation, leaving no confusion concerning the purpose of the map. This will enhance the user's experience and keep his attention. If the user is unable to identify what is being demonstrated in a reasonable fashion, the map may be regarded as useless.

Making a meaningful map is the ultimate goal. Alan MacEachren explains that a well designed map "is convincing because it implies authenticity" (1994, pp. 9). An interesting map will no doubt engage a reader. Information richness or a map that is multivariate shows relationships within the map. Showing several variables allows comparison, which adds to the meaningfulness of the map. This also generates hypothesis and stimulates ideas and perhaps further research. In order to convey the message of the map, the creator must design it in a manner which will aid the reader in the overall understanding of its purpose. The title of a map may provide the "needed link" necessary for communicating that message, but the overall design of the map fosters the manner in which the reader interprets it (Monmonier, 1993, pp. 93).

In the 21st century it is possible to find a map of virtually anything from the inner workings of the human body to the virtual worlds of cyberspace. Therefore there are now a huge variety of different styles and types of map - for example, one area which has evolved a specific and recognisable variation are those used by public transport organisations to guide passengers, namely urban rail and metro maps, many of which are loosely based on 45 degree angles as originally perfected by Harry Beck and George Dow.

Naming conventions

Most maps use text to label places and for such things as a map title, legend and other information. Maps are often made in specific languages, though names of places often differ between languages. So a map made in English may use the name *Germany* for that country, while a German map would use *Deutschland* and a French map *Allemagne*. A word that describes a place, using a non-native terminology or language is referred to as an exonym.

In some cases the proper name is not clear. For example, the nation of Burma officially changed its name to Myanmar, but many nations do not recognize the ruling junta and continue to use *Burma*. Sometimes an official name change is resisted in other languages and the older name may remain in common use. Examples include the use of *Saigon* for Ho Chi Minh City, *Bangkok* for Krung Thep and *Ivory Coast* for Côte d'Ivoire.

Difficulties arise, when transliteration or transcription between writing systems is required. National names tend to have well established names in other languages and writing systems, such as *Russia* for Росси́я, but for many placenames a system of transliteration or transcription is required. In transliteration, the symbols of one language are represented by symbols in another. For example, the Cyrillic letter *Р* is traditionally written as *R* in the Latin alphabet. Systems exist for transliteration of Arabic, but the results may vary. For example, the Yemeni city of Mocha is written variously in English as Mocha, Al Mukha, al-Mukhā, Mocca and Moka. Transliteration systems are based on relating written symbols to one another, while transcription is the attempt to spell in one language the phonetic sounds of another. Chinese writing is transformed into the Latin alphabet through the Pinyin phonetic transcription systems. Other systems were used in the past, such as Wade-Giles, resulting in the city being spelled *Beijing* on newer English maps and *Peking* on older ones.

Further difficulties arise when countries, especially former colonies, do not have a strong national geographic naming standard. In such cases, cartographers may have to choose between various phonetic spellings of local names versus older imposed, sometimes resented, colonial names. Some countries have multiple official languages, resulting in multiple official placenames. For example, the capital of Belgium is both *Brussel* and *Bruxelles*. In Canada, English and French are official languages and places have names in both languages. British Columbia is also officially named *la Colombie-Britannique*. English maps rarely show the French names outside of Quebec, which itself is spelled *Québec* in French.

The study of placenames is called toponymy, while that of the origin and historical usage of placenames as words is etymology.

In order to improve legibility or to aid the illiterate, some maps have been produced using pictograms to represent places. The iconic example of this practice is Lance Wyman's early plans for the Mexico City Metro, on which stations were shown simply as stylized logos. Wyman also prototyped such a map for the Washington Metro, though ultimately the idea was rejected. Other cities experimenting with such maps are Fukuoka, Guadalajara and Monterrey.

Map symbology

A map of the southwest coast of Ireland created in the early 18th century. Notice the north arrow at the bottom of the map. Also, colors are used in the map to distinguish different geographical areas.

The quality of a map's design affects its reader's ability to extract information and to learn from the map. Cartographic symbology has been developed in an effort to portray the world accurately and effectively convey information to the map reader. A legend explains the pictorial language of the map, known as its symbology. The title indicates the region the map portrays; the map image portrays the region and so on. Although every map element serves some purpose, convention only dictates inclusion of some elements, while others are considered optional. A menu of map elements includes the neatline (border), compass rose or north arrow, overview map, bar scale, projection and information about the map sources, accuracy and publication.

When examining a landscape, scale can be intuited from trees, houses and cars. Not so with a map. Even such a simple thing as a north arrow is crucial. It may seem obvious that the top of a map should point north, but this might not be the case.

Color, likewise, is equally important. How the cartographer displays the data in different hues can greatly affect the understanding or feel of the map. Different intensities of hue portray different objectives the cartographer is attempting to get across to the audience. Today, personal computers can display up to 16 million distinct colors at a time (Jeer, 1997). This fact allows for a multitude of color options for even for the most demanding maps. Moreover, computers can easily hatch patterns in colors to give even more options. This is very beneficial, when symbolizing data in categories such as quintile and equal interval classifications.

Quantitative symbols give a visual measure of the relative size/importance/number that a symbol represents and to symbolize this data on a map, there are two major classes of symbols used for portraying quantitative properties. Proportional symbols change their visual weight according to a quantitative property. These are appropriate for extensive statistics. Choropleth maps portray data collection areas, such as counties or census tracts, with color. Using color this way, the darkness and intensity (or value) of the color is evaluated by the eye as a measure of intensity or concentration (Harvard Graduate School of Design, 2005).

Map generalization

A good map has to provide a compromise between portraying the items of interest (or themes) in the *right place* for the map scale used, against the need to annotate that item with text or a symbol, which takes up space on the map medium and very likely will cause some other item of interest to be displaced. The cartographer is thus constantly making judgements about what to include, what to leave out and what to show in a *slightly* incorrect place - because of the demands of the annotation. This issue assumes more importance as the scale of the map gets smaller (i.e. the map shows a larger area), because relatively, the annotation on the map takes up more space *on the ground*. A good example from the late 1980s was the Ordnance Survey's first digital maps, where the *absolute* positions of major roads shown at scales of 1:1250 and 1:2500 were sometimes a scale distance of hundreds of metres away from ground truth, when shown on digital maps at scales of 1:250000 and 1:625000, because of the overriding need to annotate the features.

Cartographic errors

Some maps contain deliberate errors or distortions, either as propaganda or as a "watermark" helping the copyright owner identify infringement if the error appears in competitors' maps. The latter often come in the form of nonexistent, misnamed, or misspelled "trap streets". Other names and forms for this are paper townsites, fictitious entries, and copyright easter eggs.

Another motive for deliberate errors is simply cartographic graffiti or prank: a mapmaker wishing to leave his or her mark on the work. Mount Richard, for example, was a fictitious peak on the Rocky Mountains' continental divide that appeared on a Boulder County, Colorado map in the early 1970s. It is believed to be the work of drafts man Richard Ciacci. The fiction was not discovered until two years later.

Chapter 5

Geographic Information System

A **geographic information system (GIS)**, **geographical information system**, or **geospatial information system** is a system that captures, stores, analyzes, manages and presents data with reference to geographic location data. In the simplest terms, GIS is the merging of cartography, statistical analysis and database technology. GIS may be used in archaeology, geography, cartography, remote sensing, land surveying, public utility management, natural resource management, precision agriculture, photogrammetry, urban planning, emergency management, GIS in Environmental Contamination, landscape architecture, navigation, aerial video and localized search engines.

A GIS can be thought of as a system - it digitally creates and "manipulates" spatial areas that may be jurisdictional, purpose or application-oriented for which a specific GIS is developed. Hence, a GIS developed for an application, jurisdiction, enterprise or purpose may not be necessarily interoperable or compatible with a GIS that has been developed for some other application, jurisdiction, enterprise, or purpose. What goes beyond a GIS is a spatial data infrastructure (SDI), a concept that has no such restrictive boundaries.

Therefore, in a general sense, the term describes any information system that integrates, stores, edits, analyzes, shares and displays geographic information for informing decision making. GIS applications are tools that allow users to create interactive queries (user-created searches), analyze spatial information, edit data, maps, and present the results of all these operations. Geographic information science is the science underlying the geographic concepts, applications and systems.

Applications

GIS technology can be used for:

- earth surface-based scientific investigations;
- resource management
- reference and projections of a geospatial nature, both man-made and natural;
- asset management and location planning
- archaeology;
- environmental impact-assessment;
- infrastructure assessment and development;
- urban planning;

- cartography, for a thematic and/or time-based purpose;
- criminology;
- geospatial intelligence;
- GIS data development;
- geographic history;
- marketing;
- logistics;
- population and demographic studies;
- prospectivity mapping;
- statistical analysis;
- GIS in Environmental Contamination;
- military planning; and
- other purposes.

Examples of use are:

- GIS may allow emergency planners to easily calculate emergency response times and the movement of response resources (for logistics) in the case of a natural disaster;
- GIS might be used to find wetlands that need protection strategies regarding pollution; and
- GIS can be used by a company to site a new business location to take advantage of GIS data identified trends to respond to a previously under-served market. Most city and transportation systems planning offices have GIS sections.

History of development

In 1854, John Snow depicted a cholera outbreak in London using points to represent the locations of some individual cases, possibly the earliest use of the geographic method. His study of the distribution of cholera led to the source of the disease, a contaminated water pump (the Broad Street Pump, whose handle he had disconnected, thus terminating the outbreak) within the heart of the cholera outbreak.

E. W. Gilbert's version (1958) of John Snow's 1855 map of the Soho cholera outbreak showing the clusters of cholera cases in the London epidemic of 1854

While the basic elements of topography and theme existed previously in cartography, the John Snow map was unique, using cartographic methods not only to depict but also to analyze clusters of geographically-dependent phenomena for the first time.

The early 20th century saw the development of photozincography, which allowed maps to be split into layers, for example one layer for vegetation and another for water. This was particularly used for printing contours - drawing these was a labour intensive task but having them on a separate layer meant they could be worked on without the other layers to confuse the draughtsman. This work was originally drawn on glass plates but later, plastic film was introduced, being lighter, using less storage space and being less brittle were several of its advantages over glass plates. When all the layers were finished, they were combined into one image using a large process camera. Once colour printing came in, the layers idea was also used for creating separate printing plates for each colour. While the use of layers much later became one of the main typical features of a contemporary GIS, the photographic process just described is not considered to be a GIS in itself - as the maps were just images with no database to link them to.

Computer hardware development spurred by nuclear weapon research led to general-purpose computer 'mapping' applications by the early 1960s.

The year 1960 saw the development of the world's first true operational GIS in Ottawa, Ontario, Canada by the federal Department of Forestry and Rural Development. Developed by Dr. Roger Tomlinson, it was called the Canada Geographic Information System (CGIS) and was used to store, analyze, and manipulate data collected for the Canada Land Inventory (CLI) – an effort to determine the land capability for rural Canada by mapping information about soils, agriculture, recreation, wildlife, waterfowl, forestry and land use at a scale of 1:50,000. A rating classification factor was also added to permit analysis.

CGIS was an improvement over 'computer mapping' applications as it provided capabilities for overlay, measurement and digitizing/scanning. It supported a national coordinate system that spanned the continent, coded lines as arcs having a true embedded topology and it stored the attribute and locational information in separate files. As a result of this, Tomlinson has become known as the 'father of GIS', particularly for his use of overlays in promoting the spatial analysis of convergent geographic data.

CGIS lasted into the 1990s and built a large digital land resource database in Canada. It was developed as a mainframe-based system in support of federal and provincial resource planning and management. Its strength was continent-wide analysis of complex datasets. The CGIS was never available in a commercial form.

In 1964, Howard T. Fisher formed the Laboratory for Computer Graphics and Spatial Analysis at the Harvard Graduate School of Design (LCGSA 1965-1991), where a number of important theoretical concepts in spatial data handling were developed, and which by the 1970s had distributed seminal software code and systems, such as 'SYMAP', 'GRID' and 'ODYSSEY' - that served as sources for subsequent commercial development — to universities, research centers and corporations worldwide.

By the early 1980s, M&S Computing (later Intergraph), Environmental Systems Research Institute (ESRI), CARIS (Computer Aided Resource Information System) and ERDAS emerged as commercial vendors of GIS software, successfully incorporating many of the CGIS features, combining the first generation approach to separation of spatial and attribute information with a second generation approach to organizing attribute data into database structures. In parallel, the development of two public domain systems began in the late 1970s and early 1980s.

The Map Overlay and Statistical System (MOSS) project started in 1977 in Fort Collins, Colorado under the auspices of the Western Energy and Land Use Team (WELUT) and the US Fish and Wildlife Service. GRASS GIS was introduced in 1982 by the US Army Corps of Engineering Research Laboratory (USA-CERL) in Champaign, Illinois, a branch of the US Army Corps of Engineers to meet the need of the US military for software for land management and environmental planning.

In the later 1980s and 1990s, industry growth was spurred on by the growing use of GIS on Unix workstations and the personal computer. By the end of the 20th century, the rapid growth in various systems had been consolidated and standardized on relatively few platforms and users were beginning to explore the concept of viewing GIS data over the Internet, requiring data format and transfer standards. More recently, a growing number of free, open-source GIS packages run on a range of operating systems and can be customized to perform specific tasks. Increasingly geospatial data and mapping applications are being made available via the world wide web.

Several authoritative books on the history of GIS have been published.

GIS techniques and technology

Modern GIS technologies use digital information, for which various digitized data creation methods are used. The most common method of data creation is digitization, where a hard copy map or survey plan is transferred into a digital medium through the use of a computer-aided design (CAD) program, and geo-referencing capabilities. With the wide availability of ortho-rectified imagery (both from satellite and aerial sources), heads-up digitizing is becoming the main avenue through which geographic data is extracted. Heads-up digitizing involves the tracing of geographic data directly on top of the aerial imagery instead of by the traditional method of tracing the geographic form on a separate digitizing tablet (heads-down digitizing).

Relating information from different sources

GIS uses spatio-temporal (space-time) location as the key index variable for all other information. Just as a relational database containing text or numbers can relate many different tables using common key index variables, GIS can relate otherwise unrelated information by using location as the key index variable. The key is the location and/or extent in space-time.

Any variable that can be located spatially, and increasingly also temporally, can be referenced using a GIS. Locations or extents in Earth space-time may be recorded as dates/times of occurrence, and x, y, and z coordinates representing, longitude, latitude, and elevation, respectively. These GIS coordinates may represent other quantified systems of temporo-spatial reference (for example, film frame number, stream gage station, highway mile marker, surveyor benchmark, building address, street intersection, entrance gate, water depth sounding, POS or CAD drawing origin/units). Units applied to recorded temporal-spatial data can vary widely (even when using exactly the same data), but all Earth-based spatial-temporal location and extent references should, ideally, be relatable to one another and ultimately to a "real" physical location or extent in space-time.

Related by accurate spatial information, an incredible variety of real-world and projected past or future data can be analyzed, interpreted and represented to facilitate education and decision making. This key characteristic of GIS has begun to open new avenues of scientific inquiry into behaviors and patterns of previously considered unrelated real-world information.

GIS Uncertainties

GIS accuracy depends upon source data, and how it is encoded to be data referenced. Land Surveyors have been able to provide a high level of positional accuracy utilizing the GPS derived positions. [Retrieved from Federal Geographic Data Committee] the high-resolution digital terrain and aerial imagery, [Retrieved NJGIN] the powerful computers, Web technology, are changing the quality, utility, and expectations of GIS to serve society on a grand scale, but nevertheless there are other source data that has an impact on the overall GIS accuracy like: paper maps that are not found to be very suitable to achieve the desired accuracy since the aging of maps affects their dimensional stability.

In developing a Digital Topographic Data Base for a GIS, topographical maps are the main source of data. Aerial photography and satellite images are extra sources for collecting data and identifying attributes which can be mapped in layers over a location facsimile of scale. The scale of a map and geographical rendering area representation type are very important aspects since the information content depends mainly on the scale set and resulting locatability of the map's representations. In order to digitize a map, the map has to be checked within theoretical dimensions, then scanned into a raster format, and resulting raster data has to be given a theoretical dimension by a rubber sheeting/warping technology process.

Uncertainty is a significant problem in designing a GIS because spatial data tend to be used for purposes for which they were never intended. Some maps were made many decades ago, where at that time the computer industry was not even in its perspective establishments. This has led to historical reference maps without common norms. Map accuracy is a relative issue of minor importance in cartography. All maps are established for communication ends. Maps use a historically constrained technology of pen and paper to communicate a view of the world to their users. Cartographers feel little need to communicate information based on accuracy, for when the same map is digitized and input into a GIS, the mode of use often changes. The new uses extend well beyond a determined domain for which the original map was intended and designed.

A quantitative analysis of maps brings accuracy issues into focus. The electronic and other equipment used to make measurements for GIS is far more precise than the machines of conventional map analysis. [Retrieved USGS]. The truth is that all geographical data are inherently inaccurate, and these inaccuracies will propagate through GIS operations in ways that are difficult to predict, yet have goals of conveyance in mind for original design. Accuracy Standards for 1:24000 Scales Map: 1:24,000 ± 40.00 feet

This means that when we see a point or attribute on a map, its "probable" location is within a +/- 40 foot area of its rendered reference, according to area representations and scale.

A GIS can also convert existing digital information, which may not yet be in map form, into forms it can recognize, employ for its data analysis processes, and use in forming mapping output. For example, digital satellite images generated through remote sensing can be analyzed to produce a map-like layer of digital information about vegetative covers on land locations. Another fairly recently developed resource for naming GIS location objects is the Getty Thesaurus of Geographic Names (GTGN), which is a structured vocabulary containing about 1,000,000 names and other information about places.

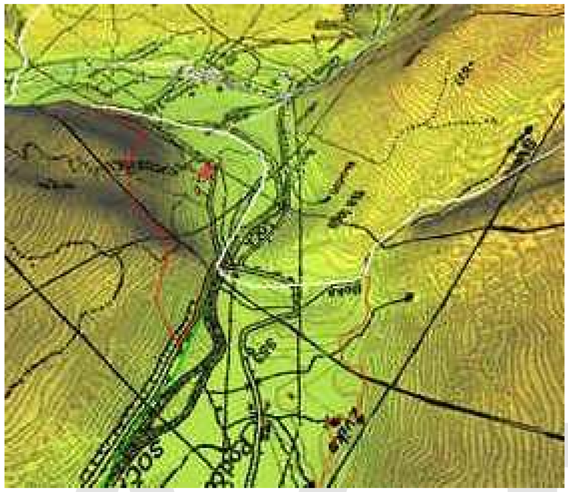
Likewise, researched census or hydrological tabular data can be displayed in map-like form, serving as layers of thematic information for forming a GIS map.

Data representation

GIS data represents real objects (such as roads, land use, elevation, trees, waterways, etc.) with digital data determining the mix. Real objects can be divided into two abstractions: discrete objects (e.g., a house) and continuous fields (such as rainfall amount, or elevations). Traditionally, there are two broad methods used to store data in a GIS for both kinds of abstractions mapping references: raster images and vector. Points, lines, and polygons are the stuff of mapped location attribute references. A new hybrid method of storing data is that of identifying point clouds, which combine three-dimensional points with RGB information at each point, returning a "3D color image". GIS Thematic maps then are becoming more and more realistically visually descriptive of what they set out to show or determine.

Raster

A raster data type is, in essence, any type of digital image represented by reducible and enlargeable grids. Anyone who is familiar with digital photography will recognize the Raster graphics pixel as the smallest individual grid unit building block of an image, usually not readily identified as an artifact shape until an image is produced on a very large scale. A combination of the pixels making up an image color formation scheme will compose details of an image, as is distinct from the commonly used points, lines, and polygon area location symbols of scalable vector graphics as the basis of the vector model of area attribute rendering. While a digital image is concerned with its output blending together its grid based details as an identifiable representation of reality, in a photograph or art image transferred into a computer, the raster data type will reflect a digitized abstraction of reality dealt with by grid populating tones or objects, quantities, cojoined or open boundaries, and map relief schemas. Aerial photos are one commonly used form of raster data, with one primary purpose in mind: to display a detailed image on a map area, or for the purposes of rendering its identifiable objects by digitization. Additional raster data sets used by a GIS will contain information regarding elevation, a digital elevation model, or reflectance of a particular wavelength of light, Landsat, or other electromagnetic spectrum indicators.



Digital elevation model, map (image), and vector data

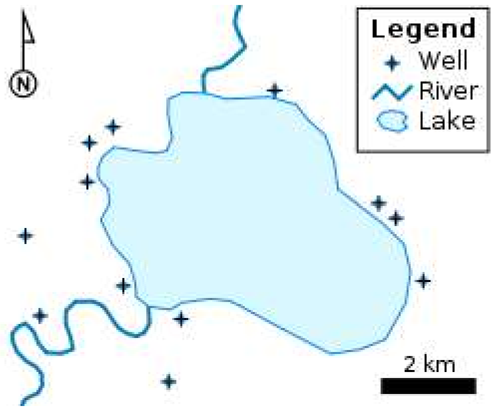
Raster data type consists of rows and columns of cells, with each cell storing a single value. Raster data can be images (raster images) with each pixel (or cell) containing a color value. Additional values recorded for each cell may be a discrete value, such as land use, a continuous value, such as temperature, or a null value if no data is available. While a raster cell stores a single value, it can be extended by using raster bands to represent RGB (red, green, blue) colors, colormaps (a mapping between a thematic code and RGB value), or an extended attribute table with one row for each unique cell value. The resolution of the raster data set is its cell width in ground units.

Raster data is stored in various formats; from a standard file-based structure of TIF, JPEG, etc. to binary large object (BLOB) data stored directly in a relational database management system (RDBMS) similar to other vector-based feature classes. Database storage, when properly indexed, typically allows for quicker retrieval of the raster data but can require storage of millions of significantly sized records.

Vector

In a GIS, geographical features are often expressed as vectors, by considering those features as geometrical shapes. Different geographical features are expressed by different types of geometry:

- Points



A simple vector map, using each of the vector elements: points for wells, lines for rivers, and a polygon for the lake.

Zero-dimensional points are used for geographical features that can best be expressed by a single point reference — in other words, by simple location. Examples include wells, peaks, features of interest, and trailheads. Points convey the least amount of information of these file types. Points can also be used to represent areas when displayed at a small scale. For example, cities on a map of the world might be represented by points rather than polygons. No measurements are possible with point features.

- Lines or polylines

One-dimensional lines or polylines are used for linear features such as rivers, roads, railroads, trails, and topographic lines. Again, as with point features, linear features displayed at a small scale will be represented as linear features rather than as a polygon. Line features can measure distance.

- Polygons

Two-dimensional polygons are used for geographical features that cover a particular area of the earth's surface. Such features may include lakes, park boundaries, buildings, city boundaries, or land uses. Polygons convey the most amount of information of the file types. Polygon features can measure perimeter and area.

Each of these geometries are linked to a row in a database that describes their attributes. For example, a database that describes lakes may contain a lake's depth, water quality, pollution level. This information can be used to make a map to describe a particular attribute of the dataset. For example, lakes could be coloured depending on level of pollution. Different geometries can also be compared. For example, the GIS could be used to identify all wells (point geometry) that are within one kilometre of a lake (polygon geometry) that has a high level of pollution.

Vector features can be made to respect spatial integrity through the application of topology rules such as 'polygons must not overlap'. Vector data can also be used to represent continuously varying phenomena. Contour lines and triangulated irregular networks (TIN) are used to represent elevation or other continuously changing values. TINs record values at point locations, which are connected by lines to form an irregular mesh of triangles. The face of the triangles represent the terrain surface.

Advantages and disadvantages

There are some important advantages and disadvantages to using a raster or vector data model to represent reality:

- Raster datasets record a value for all points in the area covered which may require more storage space than representing data in a vector format that can store data only where needed.
- Raster data allows easy implementation of overlay operations, which are more difficult with vector data.
- Vector data can be displayed as vector graphics used on traditional maps, whereas raster data will appear as an image that may have a blocky appearance for object boundaries. (depending on the resolution of the raster file)
- Vector data can be easier to register, scale, and re-project, which can simplify combining vector layers from different sources.
- Vector data is more compatible with relational database environments, where they can be part of a relational table as a normal column and processed using a multitude of operators.
- Vector file sizes are usually smaller than raster data, which can be 10 to 100 times larger than vector data (depending on resolution).
- Vector data is simpler to update and maintain, whereas a raster image will have to be completely reproduced. (Example: a new road is added).
- Vector data allows much more analysis capability, especially for "networks" such as roads, power, rail, telecommunications, etc. (Examples: Best route, largest port, airfields connected to two-lane highways). Raster data will not have all the characteristics of the features it displays.

Non-spatial data

Additional non-spatial data can also be stored along with the spatial data represented by the coordinates of a vector geometry or the position of a raster cell. In vector data, the additional data contains attributes of the feature. For example, a forest inventory polygon may also have an identifier value and information about tree species. In raster data the cell value can store attribute information, but it can also be used as an identifier that can relate to records in another table.

Software is currently being developed to support spatial and non-spatial decision-making, with the solutions to spatial problems being integrated with solutions to non-spatial problems. The end result with these Flexible Spatial Decision-Making Support Systems

(FSDSS) is expected to be that non-experts will be able to use GIS, along with spatial criteria, and simply integrate their non-spatial criteria to view solutions to multi-criteria problems. This system is intended to assist decision-making.

Data capture

Example of hardware for mapping (GPS and laser rangefinder) and data collection (rugged computer). Field GIS are current trend, accurate mapping and data analysis are done directly in the field. Presented hardware (Field-Map technology) is used mainly for forest inventories, monitoring and mapping.

Data capture—entering information into the system—consumes much of the time of GIS practitioners. There are a variety of methods used to enter data into a GIS where it is stored in a digital format.

Existing data printed on paper or PET film maps can be digitized or scanned to produce digital data. A digitizer produces vector data as an operator traces points, lines, and polygon boundaries from a map. Scanning a map results in raster data that could be further processed to produce vector data.

Survey data can be directly entered into a GIS from digital data collection systems on survey instruments using a technique called Coordinate Geometry (COGO). Positions from a Global Navigation Satellite System (GNSS) like Global Positioning System (GPS), another survey tool, can also be directly entered into a GIS. Current trend is data collection and field mapping carried out directly with field computers (position from GPS and/or laser rangefinder). New technologies allow to create maps as well as analysis directly in the field, projects are more efficient and mapping is more accurate.

Remotely sensed data also plays an important role in data collection and consist of sensors attached to a platform. Sensors include cameras, digital scanners and LIDAR, while platforms usually consist of aircraft and satellites.

The majority of digital data currently comes from photo interpretation of aerial photographs. Soft copy workstations are used to digitize features directly from stereo pairs of digital photographs. These systems allow data to be captured in two and three dimensions, with elevations measured directly from a stereo pair using principles of photogrammetry. Currently, analog aerial photos are scanned before being entered into a soft copy system, but as high quality digital cameras become cheaper this step will be skipped.

Satellite remote sensing provides another important source of spatial data. Here satellites use different sensor packages to passively measure the reflectance from parts of the electromagnetic spectrum or radio waves that were sent out from an active sensor such as radar. Remote sensing collects raster data that can be further processed using different bands to identify objects and classes of interest, such as land cover.

When data is captured, the user should consider if the data should be captured with either a relative accuracy or absolute accuracy, since this could not only influence how information will be interpreted but also the cost of data capture.

In addition to collecting and entering spatial data, attribute data is also entered into a GIS. For vector data, this includes additional information about the objects represented in the system.

After entering data into a GIS, the data usually requires editing, to remove errors, or further processing. For vector data it must be made "topologically correct" before it can be used for some advanced analysis. For example, in a road network, lines must connect with nodes at an intersection. Errors such as undershoots and overshoots must also be removed. For scanned maps, blemishes on the source map may need to be removed from the resulting raster. For example, a fleck of dirt might connect two lines that should not be connected.

Raster-to-vector translation

Data restructuring can be performed by a GIS to convert data into different formats. For example, a GIS may be used to convert a satellite image map to a vector structure by generating lines around all cells with the same classification, while determining the cell spatial relationships, such as adjacency or inclusion.

More advanced data processing can occur with image processing, a technique developed in the late 1960s by NASA and the private sector to provide contrast enhancement, false colour rendering and a variety of other techniques including use of two dimensional Fourier transforms.

Since digital data is collected and stored in various ways, the two data sources may not be entirely compatible. So a GIS must be able to convert geographic data from one structure to another.

Projections, coordinate systems and registration

A property ownership map and a soils map might show data at different scales. Map information in a GIS must be manipulated so that it registers, or fits, with information gathered from other maps. Before the digital data can be analyzed, they may have to undergo other manipulations—projection and coordinate conversions, for example—that integrate them into a GIS.

The earth can be represented by various models, each of which may provide a different set of coordinates (e.g., latitude, longitude, elevation) for any given point on the Earth's surface. The simplest model is to assume the earth is a perfect sphere. As more measurements of the earth have accumulated, the models of the earth have become more sophisticated and more accurate. In fact, there are models that apply to different areas of the earth to provide increased accuracy (e.g., North American Datum, 1927 - NAD27 - works well in North America, but not in Europe).

Projection is a fundamental component of map making. A projection is a mathematical means of transferring information from a model of the Earth, which represents a three-dimensional curved surface, to a two-dimensional medium—paper or a computer screen. Different projections are used for different types of maps because each projection particularly suits specific uses. For example, a projection that accurately represents the shapes of the continents will distort their relative sizes.

Since much of the information in a GIS comes from existing maps, a GIS uses the processing power of the computer to transform digital information, gathered from sources with different projections and/or different coordinate systems, to a common projection and coordinate system. For images, this process is called rectification.

Spatial analysis with GIS

Given the vast range of spatial analysis techniques that have been developed over the past half century, any summary or review can only cover the subject to a limited depth. This is a rapidly changing field, and GIS packages are increasingly including analytical tools as standard built-in facilities or as optional toolsets, add-ins or 'analysts'. In many instances such facilities are provided by the original software suppliers (commercial vendors or collaborative non commercial development teams), whilst in other cases facilities have been developed and are provided by third parties. Furthermore, many products offer software development kits (SDKs), programming languages and language support, scripting facilities and/or special interfaces for developing one's own analytical tools or variants. The website Geospatial Analysis and associated book/ebook attempt to provide a reasonably comprehensive guide to the subject. The impact of these myriad paths to perform spatial analysis create a new dimension to business intelligence termed "spatial intelligence" which, when delivered via intranet, democratizes access to operational sorts not usually privy to this type of information.

Slope and Aspect

Slope, aspect and surface curvature in terrain analysis are all derived from neighbourhood operations using elevation values of a cell's adjacent neighbours. Authors such as Skidmore, Jones and Zhou and Liu have compared techniques for calculating slope and aspect. Slope is a function of resolution, and the spatial resolution used to calculate slope and aspect should always be specified

The elevation at a point will have perpendicular tangents (slope) passing through the point, in an east-west and north-south direction. These two tangents give two components, $\partial z/\partial x$ and $\partial z/\partial y$, which then be used to determine the overall direction of slope, and the aspect of the slope. The gradient is defined as a vector quantity with components equal to the partial derivatives of the surface in the x and y directions.

The calculation of the overall 3x3 grid slope and aspect for methods that determine east-west and north-south component use the following formulas respectively:

$$\tan S = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}$$

$$\tan A = \left(\frac{\left(\frac{-\partial z}{\partial y}\right)}{\left(\frac{\partial z}{\partial x}\right)}\right)$$

Zhou and Liu describe another algorithm for calculating aspect, as follows:

$$A = 270^\circ + \arctan\left(\frac{\left(\frac{\partial z}{\partial x}\right)}{\left(\frac{\partial z}{\partial y}\right)}\right) - 90^\circ\left(\frac{\left(\frac{\partial z}{\partial y}\right)}{\left|\frac{\partial z}{\partial y}\right|}\right)$$

Data modeling

It is difficult to relate wetlands maps to rainfall amounts recorded at different points such as airports, television stations, and high schools. A GIS, however, can be used to depict two- and three-dimensional characteristics of the Earth's surface, subsurface, and atmosphere from information points. For example, a GIS can quickly generate a map with isopleth or contour lines that indicate differing amounts of rainfall.

Such a map can be thought of as a rainfall contour map. Many sophisticated methods can estimate the characteristics of surfaces from a limited number of point measurements. A two-dimensional contour map created from the surface modeling of rainfall point measurements may be overlaid and analyzed with any other map in a GIS covering the same area.

Additionally, from a series of three-dimensional points, or digital elevation model, isopleth lines representing elevation contours can be generated, along with slope analysis, shaded relief, and other elevation products. Watersheds can be easily defined for any given reach, by computing all of the areas contiguous and uphill from any given point of interest. Similarly, an expected thalweg of where surface water would want to travel in intermittent and permanent streams can be computed from elevation data in the GIS.

Topological modeling

A GIS can recognize and analyze the spatial relationships that exist within digitally stored spatial data. These topological relationships allow complex spatial modelling and analysis to be performed. Topological relationships between geometric entities traditionally include adjacency (what adjoins what), containment (what encloses what), and proximity (how close something is to something else).

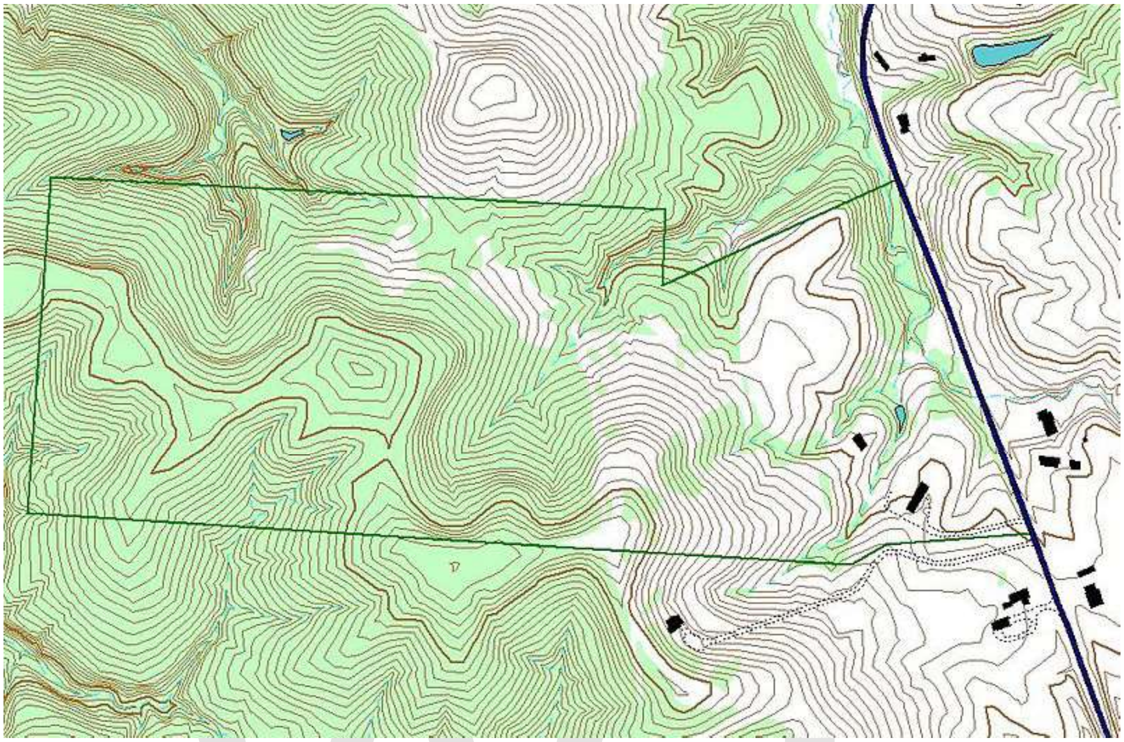
Networks

Geometric networks are linear networks of objects that can be used to represent interconnected features, and to perform special spatial analysis on them. A geometric network is composed of edges, which are connected at junction points, similar to graphs in mathematics and computer science. Just like graphs, networks can have weight and flow assigned to its edges, which can be used to represent various interconnected features more accurately. Geometric networks are often used to model road networks and public utility networks, such as electric, gas, and water networks. Network modeling is also commonly employed in transportation planning, hydrology modeling, and infrastructure modeling.

Hydrological Modeling

GIS hydrological models can provide a spatial element that other hydrological models lack, with the analysis of variables such as slope, aspect and watershed or catchment area. Terrain analysis is fundamental to hydrology, since water always flows down a slope. As basic terrain analysis of a DEM involves calculation of slope and aspect, DEMs are very useful for hydrological analysis. Slope and aspect can then be used to determine direction of surface runoff, and hence flow accumulation for the formation of streams, rivers and lakes. Areas of divergent flow can also give a clear indication of the boundaries of a catchment. Once a flow direction and accumulation matrix has been created, queries can be performed that show contributing or dispersal areas at a certain point. More detail can be added to the model, such as terrain roughness, vegetation types and soil types, which can influence infiltration and evapotranspiration rates, and hence influencing surface flow. These extra layers of detail ensures a more accurate model. Also, check out GIS in Water Contamination and GIS in Environmental Contamination.

Cartographic modeling



An example of use of layers in a GIS application. In this example, the forest cover layer (light green) is at the bottom, with the topographic layer over it. Next up is the stream layer, then the boundary layer, then the road layer. The order is very important in order to properly display the final result. Note that the pond layer was located just below the stream layer, so that a stream line can be seen overlying one of the ponds.

The term "cartographic modeling" was (probably) coined by Dana Tomlin in his PhD dissertation and later in his book which has the term in the title. Cartographic modeling refers to a process where several thematic layers of the same area are produced, processed, and analyzed. Tomlin used raster layers, but the overlay method (see below) can be used more generally. Operations on map layers can be combined into algorithms, and eventually into simulation or optimization models.

Map overlay

The combination of several spatial datasets (points, lines or polygons) creates a new output vector dataset, visually similar to stacking several maps of the same region. These overlays are similar to mathematical Venn diagram overlays. A union overlay combines the geographic features and attribute tables of both inputs into a single new output. An intersect overlay defines the area where both inputs overlap and retains a set of attribute fields for each. A symmetric difference overlay defines an output area that includes the total area of both inputs except for the overlapping area.

Data extraction is a GIS process similar to vector overlay, though it can be used in either vector or raster data analysis. Rather than combining the properties and features of both datasets, data extraction involves using a "clip" or "mask" to extract the features of one data set that fall within the spatial extent of another dataset.

In raster data analysis, the overlay of datasets is accomplished through a process known as "local operation on multiple rasters" or "map algebra," through a function that combines the values of each raster's matrix. This function may weigh some inputs more than others through use of an "index model" that reflects the influence of various factors upon a geographic phenomenon.

Automated cartography

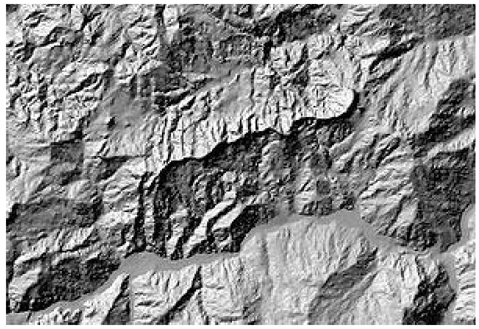
Digital cartography and GIS both encode spatial relationships in structured formal representations. GIS is used in digital cartography modeling as a (semi)automated process of making maps, so called Automated Cartography. In practice, it can be a subset of a GIS, within which it is equivalent to the stage of visualization, since in most cases not all of the GIS functionality is used. Cartographic products can be either in a digital or in a hardcopy format. Powerful analysis techniques with different data representation can produce high-quality maps within a short time period. The main problem in Automated Cartography is to use a single set of data to produce multiple products at a variety of scales, a technique known as cartographic generalization.

Geostatistics

Geostatistics is a point-pattern analysis that produces field predictions from data points. It is a way of looking at the statistical properties of those special data. It is different from general applications of statistics because it employs the use of graph theory and matrix algebra to reduce the number of parameters in the data. Only the second-order properties of the GIS data are analyzed.

When phenomena are measured, the observation methods dictate the accuracy of any subsequent analysis. Due to the nature of the data (e.g. traffic patterns in an urban environment; weather patterns over the Pacific Ocean), a constant or dynamic degree of precision is always lost in the measurement. This loss of precision is determined from the scale and distribution of the data collection.

To determine the statistical relevance of the analysis, an average is determined so that points (gradients) outside of any immediate measurement can be included to determine their predicted behavior. This is due to the limitations of the applied statistic and data collection methods, and interpolation is required to predict the behavior of particles, points, and locations that are not directly measurable.



Hillshade model derived from a Digital Elevation Model (DEM) of the Valestra area in the northern Apennines (Italy)

Interpolation is the process by which a surface is created, usually a raster dataset, through the input of data collected at a number of sample points. There are several forms of interpolation, each which treats the data differently, depending on the properties of the data set. In comparing interpolation methods, the first consideration should be whether or not the source data will change (exact or approximate). Next is whether the method is subjective, a human interpretation, or objective. Then there is the nature of transitions between points: are they abrupt or gradual. Finally, there is whether a method is global (it uses the entire data set to form the model), or local where an algorithm is repeated for a small section of terrain.

Interpolation is a justified measurement because of a spatial autocorrelation principle that recognizes that data collected at any position will have a great similarity to, or influence of those locations within its immediate vicinity.

Digital elevation models (DEM), triangulated irregular networks (TIN), edge finding algorithms, Thiessen polygons, Fourier analysis, (weighted) moving averages, inverse distance weighting, kriging, spline, and trend surface analysis are all mathematical methods to produce interpolative data.

Address geocoding

Geocoding is interpolating spatial locations (X,Y coordinates) from street addresses or any other spatially referenced data such as ZIP Codes, parcel lots and address locations. A reference theme is required to geocode individual addresses, such as a road centerline file with address ranges. The individual address locations have historically been interpolated, or estimated, by examining address ranges along a road segment. These are usually provided in the form of a table or database. The GIS will then place a dot approximately where that address belongs along the segment of centerline. For example, an address point of 500 will be at the midpoint of a line segment that starts with address 1 and ends with address 1000. Geocoding can also be applied against actual parcel data, typically from municipal tax maps. In this case, the result of the geocoding will be an actually positioned space as opposed to an interpolated point. This approach is being increasingly used to provide more precise location information.

There are several potentially dangerous caveats that are often overlooked when using interpolation.

Various algorithms are used to help with address matching when the spellings of addresses differ. Address information that a particular entity or organization has data on, such as the post office, may not entirely match the reference theme. There could be variations in street name spelling, community name, etc. Consequently, the user generally has the ability to make matching criteria more stringent, or to relax those parameters so that more addresses will be mapped. Care must be taken to review the results so as not to map addresses incorrectly due to overzealous matching parameters.

Reverse geocoding

Reverse geocoding is the process of returning an estimated street address number as it relates to a given coordinate. For example, a user can click on a road centerline theme (thus providing a coordinate) and have information returned that reflects the estimated house number. This house number is interpolated from a range assigned to that road segment. If the user clicks at the midpoint of a segment that starts with address 1 and ends with 100, the returned value will be somewhere near 50. Note that reverse geocoding does not return actual addresses, only estimates of what should be there based on the predetermined range.

Data output and cartography

Cartography is the design and production of maps, or visual representations of spatial data. The vast majority of modern cartography is done with the help of computers, usually using a GIS but production quality cartography is also achieved by importing layers into a design program to refine it. Most GIS software gives the user substantial control over the appearance of the data.

Cartographic work serves two major functions:

First, it produces graphics on the screen or on paper that convey the results of analysis to the people who make decisions about resources. Wall maps and other graphics can be generated, allowing the viewer to visualize and thereby understand the results of analyses or simulations of potential events. Web Map Servers facilitate distribution of generated maps through web browsers using various implementations of web-based application programming interfaces (AJAX, Java, Flash, etc.).

Second, other database information can be generated for further analysis or use. An example would be a list of all addresses within one mile (1.6 km) of a toxic spill.

Graphic display techniques

Traditional maps are abstractions of the real world, a sampling of important elements portrayed on a sheet of paper with symbols to represent physical objects. People who use maps must interpret these symbols. Topographic maps show the shape of land surface with contour lines or with shaded relief.

Today, graphic display techniques such as shading based on altitude in a GIS can make relationships among map elements visible, heightening one's ability to extract and analyze information. For example, two types of data were combined in a GIS to produce a perspective view of a portion of San Mateo County, California.

- The digital elevation model, consisting of surface elevations recorded on a 30-meter horizontal grid, shows high elevations as white and low elevation as black.
- The accompanying Landsat Thematic Mapper image shows a false-color infrared image looking down at the same area in 30-meter pixels, or picture elements, for the same coordinate points, pixel by pixel, as the elevation information.

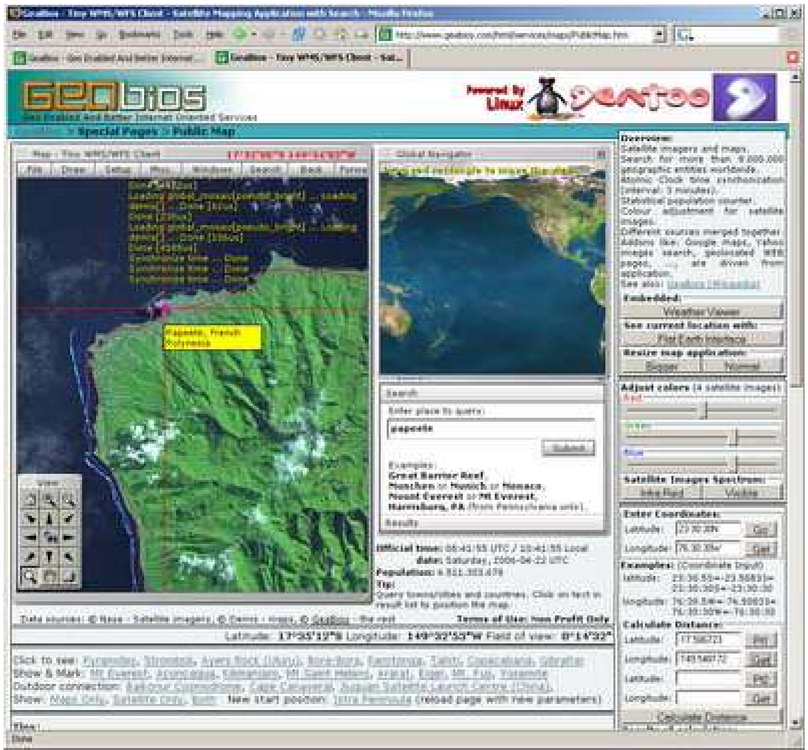
A GIS was used to register and combine the two images to render the three-dimensional perspective view looking down the San Andreas Fault, using the Thematic Mapper image pixels, but shaded using the elevation of the landforms. The GIS display depends on the viewing point of the observer and time of day of the display, to properly render the shadows created by the sun's rays at that latitude, longitude, and time of day.

An archeochrome is a new way of displaying spatial data. It is a thematic on a 3D map that is applied to a specific building or a part of a building. It is suited to the visual display of heat loss data.

Spatial ETL

Spatial ETL tools provide the data processing functionality of traditional Extract, Transform, Load (ETL) software, but with a primary focus on the ability to manage spatial data. They provide GIS users with the ability to translate data between different standards and proprietary formats, whilst geometrically transforming the data en-route.

GIS developments



GeaBios - tiny WMS/WFS client (Flash/DHTML)

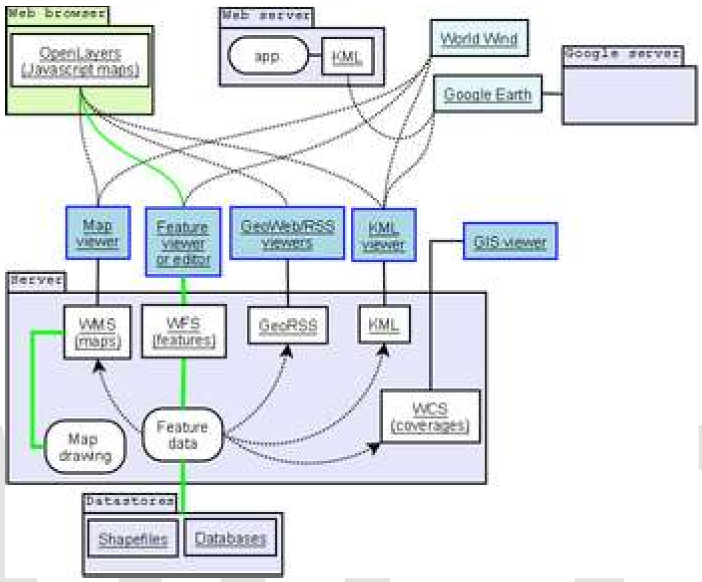
Many disciplines can benefit from GIS technology. An active GIS market has resulted in lower costs and continual improvements in the hardware and software components of GIS. These developments will, in turn, result in a much wider use of the technology throughout science, government, business, and industry, with applications including real estate, public health, crime mapping, national defense, sustainable development, natural resources, landscape architecture, archaeology, regional and community planning, transportation and logistics. GIS is also diverging into location-based services (LBS). LBS allows GPS enabled mobile devices to display their location in relation to fixed assets (nearest restaurant, gas station, fire hydrant), mobile assets (friends, children, police car) or to relay their position back to a central server for display or other processing. These services continue to develop with the increased integration of GPS functionality with increasingly powerful mobile electronics (cell phones, PDAs, laptops).

OGC standards

The Open Geospatial Consortium (OGC) is an international industry consortium of 384 companies, government agencies, universities and individuals participating in a consensus process to develop publicly available geoprocessing specifications. Open interfaces and protocols defined by OpenGIS Specifications support interoperable solutions that "geo-enable" the Web, wireless and location-based services, and mainstream IT, and empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications. Open

Geospatial Consortium (OGC) protocols include Web Map Service (WMS) and Web Feature Service (WFS).

GIS products are broken down by the OGC into two categories, based on how completely and accurately the software follows the OGC specifications.



OGC standards help GIS tools communicate.

Compliant Products are software products that comply to OGC's OpenGIS Specifications. When a product has been tested and certified as compliant through the OGC Testing Program, the product is automatically registered as "compliant" on this site.

Implementing Products are software products that implement OpenGIS Specifications but have not yet passed a compliance test. Compliance tests are not available for all specifications. Developers can register their products as implementing draft or approved specifications, though OGC reserves the right to review and verify each entry.

Web mapping

In recent years there has been an explosion of mapping applications on the web such as Google Maps and Bing Maps. These websites give the public access to huge amounts of geographic data.

Some of them, like Google Maps and OpenLayers, expose an API that enable users to create custom applications. These toolkits commonly offer street maps, aerial/satellite imagery, geocoding, searches, and routing functionality.

Other applications for publishing geographic information on the web include GeoBase (Telogis GIS software), Smallworld's SIAS or GSS, MapInfo's MapXtreme or PlanAcess or Stratus Connect, Cadcorp's GeognoSIS, Intergraph's GeoMedia WebMap (TM), ESRI's ArcIMS, ArcGIS Server, Autodesk's Mapguide, SeaTrails' AtlasAlive, ObjectFX's Web Mapping Tools, ERDAS APOLLO Suite, Google Earth, Google Fusion Tables, and the open source MapServer or GeoServer.

In recent years web mapping services have begun to adopt features more common in GIS. Services such as Google Maps and Bing Maps allow users to access and annotate maps and share the maps with others.

Global change, climate history program and prediction of its impact

Maps have traditionally been used to explore the Earth and to exploit its resources. GIS technology, as an expansion of cartographic science, has enhanced the efficiency and analytic power of traditional mapping. Now, as the scientific community recognizes the environmental consequences of anthropogenic activities influencing climate change, GIS technology is becoming an essential tool to understand the impacts of this change over time. GIS enables the combination of various sources of data with existing maps and up-to-date information from earth observation satellites along with the outputs of climate change models. This can help in understanding the effects of climate change on complex natural systems. One of the classic examples of this is the study of Arctic Ice Melting.

The outputs from a GIS in the form of maps combined with satellite imagery allow researchers to view their subjects in ways that literally never have been seen before. The images are also invaluable for conveying the effects of climate change to non-scientists.

Adding the dimension of time

The condition of the Earth's surface, atmosphere, and subsurface can be examined by feeding satellite data into a GIS. GIS technology gives researchers the ability to examine the variations in Earth processes over days, months, and years.

As an example, the changes in vegetation vigor through a growing season can be animated to determine when drought was most extensive in a particular region. The resulting graphic, known as a normalized vegetation index, represents a rough measure of plant health. Working with two variables over time would then allow researchers to detect regional differences in the lag between a decline in rainfall and its effect on vegetation.

GIS technology and the availability of digital data on regional and global scales enable such analyses. The satellite sensor output used to generate a vegetation graphic is produced for example by the Advanced Very High Resolution Radiometer (AVHRR). This sensor system detects the amounts of energy reflected from the Earth's surface across various bands of the spectrum for surface areas of about 1 square kilometer. The satellite sensor produces images of a particular location on the Earth twice a day. AVHRR and more recently the Moderate-Resolution Imaging Spectroradiometer (MODIS) are only two of many sensor systems used for Earth surface analysis. More sensors will follow, generating ever greater amounts of data.

GIS and related technology will help greatly in the management and analysis of these large volumes of data, allowing for better understanding of terrestrial processes and better management of human activities to maintain world economic vitality and environmental quality.

In addition to the integration of time in environmental studies, GIS is also being explored for its ability to track and model the progress of humans throughout their daily routines. A concrete example of progress in this area is the recent release of time-specific population data by the US Census. In this data set, the populations of cities are shown for daytime and evening hours highlighting the pattern of concentration and dispersion generated by North American commuting patterns. The manipulation and generation of data required to produce this data would not have been possible without GIS.

Using models to project the data held by a GIS forward in time have enabled planners to test policy decisions. These systems are known as Spatial Decision Support Systems.

Semantics

Tools and technologies emerging from the W3C's Semantic Web Activity are proving useful for data integration problems in information systems. Correspondingly, such technologies have been proposed as a means to facilitate interoperability and data reuse among GIS applications and also to enable new analysis mechanisms.

Ontologies are a key component of this semantic approach as they allow a formal, machine-readable specification of the concepts and relationships in a given domain. This in turn allows a GIS to focus on the intended meaning of data rather than its syntax or structure. For example, reasoning that a land cover type classified as *deciduous needleleaf trees* in one dataset is a specialization or subset of land cover type *forest* in another more roughly classified dataset can help a GIS automatically merge the two datasets under the more general land cover classification. Tentative ontologies have been developed in areas related to GIS applications, for example the hydrology ontology developed by the Ordnance Survey in the United Kingdom and the SWEET ontologies developed by NASA's Jet Propulsion Laboratory. Also, simpler ontologies and semantic metadata standards are being proposed by the W3C Geo Incubator Group to represent geospatial data on the web.

Recent research results in this area can be seen in the International Conference on Geospatial Semantics and the Terra Cognita -- Directions to the Geospatial Semantic Web workshop at the International Semantic Web Conference.

Chapter 6

Geostatistics

Geostatistics is a branch of statistics focusing on spatial or spatiotemporal datasets. Developed originally to predict probability distributions of ore grades for mining operations, it is currently applied in diverse disciplines including petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geometallurgy, geography, forestry, environmental control, landscape ecology, soil science, and agriculture (esp. in precision farming). Geostatistics is applied in varied branches of geography, particularly those involving the spread of diseases (epidemiology), the practice of commerce and military planning (logistics), and the development of efficient spatial networks. Geostatistical algorithms are incorporated in many places, including geographic information systems (GIS) and the R statistical environment.

Background

Geostatistics is intimately related to interpolation methods, but extends far beyond simple interpolation problems. It consists of a collection of numerical and mathematical techniques dealing with the characterization of spatial phenomena. Geostatistical techniques rely on statistical model that is based on random function (or random variable) theory to model the uncertainty associated with spatial estimation.

A number of simpler interpolation methods/algorithms, such as inverse distance weighting, linear interpolation and nearest-neighbor interpolation, were already well known before geostatistics. Geostatistics goes beyond the interpolation problem by considering the studied phenomenon at unknown locations as a random variable.

Let $Z(\mathbf{x})$ be the value of the variable of interest at certain location. This value is unknown (e.g. temperature, rainfall, piezometric level, geological facies, etc). Although there exists a value at location $\mathbf{x}$ that could be measured, geostatistics consider this value as random since it was not measured, or has not been measured yet. However, the randomness of $Z(\mathbf{x})$ is not complete, but defined by a cumulative distribution function (cdf) that depends on certain information that is known about the value $Z(\mathbf{x})$:

$$F(z, \mathbf{x}) = \text{Prob}\{Z(\mathbf{x}) \leqslant z \mid \text{information}\}.$$

Typically, if the value of Z is known at locations close to $\mathbf{x}$ (or in the neighborhood of $\mathbf{x}$) one can constrain the pdf of $Z(\mathbf{x})$ to this neighborhood. If a high spatial continuity is assumed, $Z(\mathbf{x})$ can only have values similar to the ones found in the neighborhood. Conversely, in the absence of spatial continuity $Z(\mathbf{x})$ can take any value. The spatial continuity of the random variables is described by a model of spatial continuity that can be either a parametric function in the case of variogram-based geostatistics, or have a non-parametric form when using other methods such as multiple-point simulation or pseudo-genetic techniques.

By applying a single spatial model on an entire domain, one makes the assumption that Z is a stationary process. It means that the same statistical properties are applicable on the entire domain. Several geostatistical methods provide ways of relaxing this stationarity assumption.

In this framework, one can distinguish two modeling goals:

- 1) Estimating the most probable value for $Z(\mathbf{x})$, which is equivalent to finding the mode of the pdf $f(z,\mathbf{x})$. This is usually denoted as an estimation problem.

- 2) Working with the entire probability density function $f(z,\mathbf{x})$ by actually considering each possible outcome of it at each location. This is generally done by creating several alternative maps of Z, called realizations. Consider a domain discretized in N grid nodes (or pixels). Each realization is a sample of the complete N-dimensional joint distribution function

  $$F(\mathbf{z}, \mathbf{x}) = \text{Prob}\{Z(\mathbf{x}_1) \leqslant z_1, Z(\mathbf{x}_2) \leqslant z_2, ..., Z(\mathbf{x}_N) \leqslant z_N\}.$$

  In this approach, the presence of multiple solutions to the interpolation problem is acknowledged. Each realization is considered as a possible scenario of what the real variable could be. All associated workflows are then considering ensemble of realizations, and consequently ensemble of predictions that allow for probabilistic forecasting.

A number of methods exist for both geostatistical estimation and multiple realizations approaches. Several reference books provide a comprehensive overview of the discipline.

Methods

Exploratory data analysis

Estimation

Kriging
Indicator kriging

Simulation

Aggregation
Dissagregation
Turning bands
Spectral simulation
SGS
Transition probabilities
Markov chain geostatistics
Support vector machine
Boolean simulation
Genetic models
Pseudo-genetic models
Cellular automata
Multiple-Point Geostatistics (MPS)

Definitions and tools

- Regionalized variable theory
- Covariance function
- Semi-variance
- Variogram
- Kriging
- Range (geostatistics)
- Sill (geostatistics)
- Nugget effect
- Training image

Main scientific journals related to geostatistics

- Water Resources Research
- Advances in Water Resources
- Ground Water
- Mathematical Geosciences
- Computers & Geosciences
- Computational Geosciences

- J. Soil Science Society of America
- Environmetrics
- Remote Sensing of the Environment
- Stochastic Environmental Research and Risk Assessment

Related software

- gslib is a set of Fortran 77 routines (open source) implementing most of the classical geostatistics estimation and simulation algorithms
- sgems is a cross-platform (Windows, Unix), open-source software that implements most of the classical geostatistics algorithms (kriging, Gaussian and indicator simulation, etc.) as well as new developments (multiple-points geostatistics). It also provides an interactive 3D visualization and offers the scripting capabilities of Python.
- mgstat is a free MATLAB toolbox that allows calling sgems from MATLAB and transparent import/export of objects.
- gstat is an open source computer code for multivariable geostatistical modelling, prediction and simulation. The gstat functionality is also available as an S extension, either as R package or S-Plus library.
- besides gstat, R has at least six other packages dedicated to geostatistics and other areas in spatial statistics.

Chapter 7

Geocoding & Reverse Geocoding

Geocoding

Geocoding is the process of finding associated geographic coordinates (often expressed as latitude and longitude) from other geographic data, such as street addresses, or zip codes (postal codes). With geographic coordinates the features can be mapped and entered into Geographic Information Systems, or the coordinates can be embedded into media such as digital photographs via geotagging.

Reverse geocoding is the opposite: finding an associated textual location such as a street address, from geographic coordinates.

A **geocoder** is a piece of software or a (web) service that helps in this process.

Address interpolation

A simple method of geocoding is address interpolation. This method makes use of data from a street geographic information system where the street network is already mapped within the geographic coordinate space. Each street segment is attributed with address ranges (e.g. house numbers from one segment to the next). Geocoding takes an address, matches it to a street and specific segment (such as a block, in towns that use the "block" convention). Geocoding then interpolates the position of the address, within the range along the segment.

Example

Take for example: *742 Evergreen Terrace*

Let's say that this segment (for instance, a block) of Evergreen Terrace runs from 700 to 799. Even-numbered addresses would fall on one side (e.g. west side) of Evergreen Terrace, with odd-numbered addresses on the other side (e.g. east side). 742 Evergreen Terrace would (probably) be located slightly less than halfway up the block, on the west side of the street. A point would be mapped at that location along the street, perhaps offset some distance to the west of the street centerline.

Complicating factors

However, this process is not always as straightforward as in this example.

Difficulties arise when

- distinguishing between ambiguous addresses such as 742 Evergreen Terrace and 742 W Evergreen Terrace.
- attempting to geocode new addresses for a street that is not yet added to the geographic information system database.

While there might be 742 Evergreen Terrace in Springfield, there might also be a 742 Evergreen Terrace in Shelbyville. Asking for the city name (and state, province, country, etc. as needed) can solve this problem. Some situations require use of postal codes or district name for disambiguation. For example, there are multiple 100 Washington Streets in Boston, Massachusetts because several cities have been annexed without changing street names.

Finally, several caveats on using interpolation:

- The typical attribution of a street segment assumes that all "even" numbered parcels are on one side of the segment, and all "odd" numbered parcels are on the other. This is often not true in real life.
- Interpolation assumes that the given parcels are evenly distributed along the length of the segment. This is almost never true in real life; it is not uncommon for a geocoded address to be off by several thousand feet.
- Segment Information (esp. from sources such as TIGER) includes a maximum upper bound for addresses and is interpolated as though the full address range is used. For example, a segment (block) might have a listed range of 100-199, but the last address at the end of the block is 110. In this case, address 110 would be geocoded to 10% of the distance down the segment rather than near the end.
- Most interpolation implementations will produce a point as their resulting "address" location. In reality, the physical address is distributed along the length of the segment, i.e. consider geocoding the address of a shopping mall - the physical lot may run quite some distance along the street segment (or could be thought of as a two-dimensional space-filling polygon which may front on several different streets - or worse, for cities with multi-level streets, a three-dimensional shape that meets different streets at several different levels) but the interpolation treats it as a singularity.

A very common error is to believe the accuracy ratings of a given map's geocodable attributes. Such "accuracy" currently touted by most vendors has no bearing on an address being attributed to the correct segment, being attributed to the correct "side" of the segment, nor resulting in an accurate position along that correct segment. With the geocoding process used for U.S. Census TIGER datasets, 5-7.5% of the addresses may be allocated to a different census tract, while 50% of the geocoded points might be located to a different property parcel.

Because of this, it is quite important to avoid using interpolated results except for non-critical applications, such as pizza delivery. Interpolated geocoding is usually not appropriate for making authoritative decisions, for example if life safety will be impacted by that decision. Emergency services, for example, do not make an authoritative decision based on their interpolations; an ambulance or fire truck will always be dispatched regardless of what the map says.

Other techniques

Other means of geocoding might include locating a point at the centroid (center) of a land parcel, if parcel (property) data is available in the geographic information system database. In rural areas or other places lacking high quality street network data and addressing, GPS is useful for mapping a location. For traffic accidents, geocoding to a street intersection or midpoint along a street centerline is a suitable technique. Most highways in developed countries have mile markers to aid in emergency response, maintenance, and navigation. It is also possible to use a combination of these geocoding techniques - using a particular technique for certain cases and situations and other techniques for other cases.

Uses

Geocoded locations are useful in many GIS analysis and cartography tasks.

Geocoding is common on the web, for services like finding driving directions to or from some address, or finding a list of the geographically nearest store or service locations.

Geocoding is one of several methods of obtaining geographic coordinates for geotagging media, such as photographs or RSS items.

Privacy concerns

The proliferation and ease of access to geocoding (and reverse-geocoding) services raises privacy concerns. For example, in mapping crime incidents, law enforcement agencies aim to balance the privacy rights of victims and offenders, with the public's right to know. Law enforcement agencies have experimented with alternative geocoding techniques that allow them to mask some of the locational detail (e.g., address specifics that would lead to identifying a victim or offender). As well, in providing online crime mapping to the public, they also place disclaimers regarding the locational accuracy of points on the map, acknowledging these location masking techniques, and impose terms of use for the information.

List of some geocoding systems

Web services:

- Google Maps Free up to 50,000 queries per day, but with numerous restrictions such as an obligation to display Google Maps pictures when using the service.
- Yahoo PlaceFinder Free up to 50,000 queries per day, but with numerous restrictions such as an obligation to display Yahoo Maps pictures when using the service. Does not include Australia or many Asian countries.
- Bing Maps (Microsoft) Free for "public-facing, non-password protected Web sites"; various commercial licence options.
- OpenStreetMaps Free. Poor coverage as at July 2010.
- USC Geocoder Free in batches of 2,500. US only. Uses various reference data sources.

Other systems (some of these code systems are free for use, others have different licences):

- ISO 6709 Standard Representation for Geographic Point Location by Coordinates
- C-squares - compact encoding of geographic coordinate bounds (latitude-longitude)
- FIPS country codes (FIPS 10-4), area code, administrative, free
- FIPS place codes (FIPS 55) U.S. only, free
- FIPS county codes (FIPS 6-4) US only, free
- FIPS state codes (FIPS 5-2) US only, free
- Canadian Location Code, encodes a weather forecast region
- Geohash, compact string encoding of a geographic coordinate with arbitrary precision, in public domain
- Georef, a military / air naviation coordinate system for point and area identification
- HASC (Hierarchical administrative subdivision codes)
- IATA airport codes, area /point codes, airports
- ICAO airport codes, area /point codes, airports

- IANA country codes similar to ISO 3166-1 alpha-2
- IOC country codes, area, worldwide
- ISO 3166 country and subdivision codes
- ITU-R country codes
- ITU-T country calling codes
- ITU-T mobile calling codes
- Maidenhead Locator System
- MapDot Protocol: world locations coded into a zone sequence, free
- MARC country codes
- Marsden Squares
- NAC, area codes (area can be indefinitely small)
- NUTS area code, partially administrative, worldwide: countries, Europe: country to community
- ONS code, UK only, administrative
- Postal codes, area, worldwide, country-codes by UPU, free
- Quarter Degree Grid Cells
- UN M.49 region codes, area code, continents, countries (like ISO 3166-1 numeric)
- SALB (Second Administrative Level Boundaries), by UN
- SGC codes, Canada only, statistical
- UN/LOCODE, area, administrative, cities
- UTM
- WMO squares

Reverse geocoding

Reverse geocoding is the process of back (reverse) coding of a point location (latitude, longitude) to a readable address or place name. This permits the identification of nearby street addresses, places, and/or areal subdivisions such as neighborhoods, county, state, or country. Combined with geocoding and routing services, reverse geocoding is a critical component of mobile location-based services and Enhanced 911 to convert a coordinate obtained by GPS to a readable street address which is easier to understand by the end user.

Reverse geocoding can be carried out systematically by services which process a coordinate similarly to the geocoding process. For example, when a GPS coordinate is entered the street address is interpolated from a range assigned to the road segment in a reference dataset that the point is nearest to. If the user provides a coordinate near the midpoint of a segment that starts with address 1 and ends with 100, the returned street address will be somewhere near 50. This approach to reverse geocoding does not return actual addresses, only estimates of what should be there based on the predetermined range. Alternatively, coordinates for reverse geocoding can also be selected on an interactive map, or extracted from static maps by georeferencing them in a GIS with predefined spatial layers to determine the coordinates of a displayed point. Many of the same limitations of geocoding are similar with reverse geocoding. The accuracy and timeliness of the reference layer used to reverse geocode a coordinate will have a significant impact on the accuracy of the results.

Reverse geocoding services have typically not been public due to the need for extensive computing resources and currently updated and large databases. However, public reverse geocoding services are becoming increasingly available through APIs and other web services as well as mobile phone applications. These services require manual input of a coordinate, capture from a GPS, or selection of a point on an interactive map; to look up a street address or neighboring places. Examples of these services include the GeoNames reverse geocoding web service which has tools to identify nearest street address, place names, articles, country, county subdivisions, neighborhoods, and other location data from a coordinate. GeoNames uses the United States Census Bureau's tiger line data set as the reference layer for reverse geocoding. Google has also published a reverse geocoding API which can be adapted for online reverse geocoding tools, which uses the same street reference layer as Google maps.

Privacy Concerns

Geocoding and reverse geocoding have raised potential privacy concerns, especially regarding the ability to reverse engineer street addresses from published static maps. By digitizing published maps it is possible to georeference them by overlaying with other spatial layers and then extract point locations which can be used to identify individuals or reverse geocoded to obtain a street address of the individual. This has potential implications to determine locations for patients and/or study participants from maps published in medical literature as well as potentially sensitive information published in other journalistic sources.

The ability to reverse engineer maps and obtain readable location information from static newspaper maps and hypothetical patient address maps has been examined. In one study a map of Hurricane Katrina mortality locations published in a Baton Rouge, Louisiana paper was examined. Using GPS locations obtained from houses where fatalities occurred, the authors were able to determine the relative error between the true house locations and the location determined by georeferencing the published map. The authors found that approximately 45% of the points extracted from the georeferenced map were within 10 meters of a households GPS obtained point. Another study found similar results examining hypothetical low and high resolution patient address maps similar to what might be found published in medical journals. They found approximately 26% of points obtained from a low resolution map and 79% from a high resolution map were matched precisely with the true location.

The findings from these studies raise concerns regarding the potential use of georeferencing and reverse geocoding of published maps to elucidate sensitive or private information on mapped individuals. Guidelines for the display and publication of potentially sensitive information are inconsistently applied and no uniform procedure has been identified. The use of blurring algorithms which shift the location of mapped points have been proposed as a solution. In addition, where direct reference to the geography of the area mapped is not required it may be possible to use abstract space on which to display spatial patterns.

Chapter 8

Web Mapping

Web mapping is the process of designing, implementing, generating and delivering maps on the World Wide Web and its product. While **web mapping** primarily deals with technological issues, **web cartography** additionally studies theoretic aspects: the use of web maps, the evaluation and optimization of techniques and workflows, the usability of web maps, social aspects, and more. **Web GIS** is similar to web mapping but with an emphasis on analysis, processing of project specific geodata and exploratory aspects. Often the terms web GIS and web mapping are used synonymously, even if they don't mean exactly the same. In fact, the border between web maps and web GIS is blurry. Web maps are often a presentation media in web GIS and web maps are increasingly gaining analytical capabilities. A special case of web maps are **mobile maps**, displayed on mobile computing devices, such as mobile phones, smart phones, PDAs, GPS and other devices. If the maps on these devices are displayed by a mobile web browser or web user agent, they can be regarded as **mobile web maps**. If the mobile web maps also display context and location sensitive information, such as **points of interest**, the term **Location-based services** is frequently used."

"The use of the web as a dissemination medium for maps can be regarded as a major advancement in cartography and opens many new opportunities, such as realtime maps, cheaper dissemination, more frequent and cheaper updates of data and software, personalized map content, distributed data sources and sharing of geographic information. It also implicates many challenges due to technical restrictions (low display resolution and limited bandwidth, in particular with mobile computing devices, many of which are physically small, and use slow wireless Internet connections), copyright and security issues, reliability issues and technical complexity. While the first web maps were primarily static, due to technical restrictions, today's web maps can be fully interactive and integrate multiple media. This means that both web mapping and web cartography also have to deal with interactivity, usability and multimedia issues."

A more general term is neogeography.

Development and implementation

The advent of web mapping can be regarded as a major new trend in cartography. Previously, cartography was restricted to a few companies, institutes and mapping agencies, requiring expensive and complex hardware and software as well as skilled cartographers and geomatics engineers. With web mapping, freely available mapping technologies and geodata potentially allow every skilled person to produce web maps, with expensive geodata and technical complexity (data harmonization, missing standards) being two of the remaining barriers preventing web mapping from fully going mainstream. The cheap and easy transfer of geodata across the internet allows the integration of distributed data sources, opening opportunities that go beyond the possibilities of disjoint data storage. Everyone with minimal knowhow and infrastructure can become a geodata provider. These facts can be regarded both as an advantage and a disadvantage. While it allows everyone to produce maps and considerably enlarges the audience, it also puts geodata in the hands of untrained people who potentially violate cartographic and geographic principles and introduce flaws during the preparation, analysis and presentation of geographic and cartographic data. Educating the general public about geographic analysis and cartographic methods and principles should therefore be a priority to the cartography community.

Types of web maps

A first classification of web maps has been made by Kraak. He distinguished *static* and *dynamic* web maps and further distinguished *interactive* and *view only* web maps. However, today in the light of an increased number of different web map types, this classification needs some revision. Today, there are additional possibilities regarding distributed data sources, collaborative maps, personalized maps, etc.

Analytic web maps

These web maps offer GIS analysis, either with geodata provided, or with geodata uploaded by the map user. As already mentioned, the borderline between analytic web maps and web GIS is blurry. Often, parts of the analysis are carried out by a serverside GIS and the client displays the result of the analysis. As web clients gain more and more capabilities, this task sharing may gradually shift.

Animated web maps

Animated Maps show changes in the map over time by animating one of the graphical or temporal variables. Various data and multimedia formats and technologies allow the display of animated web maps: SVG, Adobe Flash, Java, Quicktime, etc., also with varying degrees of interaction. Examples for animated web maps are weather maps, maps displaying dynamic natural or other phenomena (such as water currents, wind patterns, traffic flow, trade flow, communication patterns,social studies projects, and for college life, etc.).

Collaborative web maps

Collaborative maps are still new, immature and complex to implement, but show a lot of potential. Technically, an application allowing simultaneous editing across the web would have to ensure that geometric features being edited by one person are locked, so they can't be edited by other persons at the same time. Also, a minimal quality check would have to be made, before data goes public. Some collaborative map projects:

- OpenStreetMap
- (*Please add additional notes, references and examples here!*)

Customisable web maps

Web maps in this category are usually more complex web mapping systems that offer APIs for reuse in other people's web pages and products. Example for such a system with an API for reuse are the Open Layers Framework, Yahoo! Maps and Google Maps.

Distributed web maps

These are maps created from a distributed data source. The WMS protocol offers a standardised method to access maps on other servers. WMS servers can collect these different sources, reproject the map layers, if necessary, and send them back as a combined image containing all requested map layers. One server may offer a topographic base map, while other servers may offer thematic layers.

Dynamically created web maps

These maps are created on demand each time the user reloads the webpages, often from dynamic data sources, such as databases. The webserver generates the map using a web map server or a self written software. Some applications refer to depictions as **hyper maps**. One of the example is- Bhoosampada by Indian Space Research Organizations.

Hyper Maps

Any approach offering the planar presentation of a portion of an n-dimensional orthogonal web map structure with the option to chose the axes for depiction from the dimensions.

Interactive web maps

Interactivity is one of the major advantages of screen based maps and web maps. It helps to compensate for the disadvantages of screen and web maps. Interactivity helps to explore maps, change map parameters, navigate and interact with the map, reveal additional information, link to other resources, and much more. Technically, it is achieved through the combination of events, scripting and DOM manipulations.

Online atlases

Atlas projects often went through a renaissance when they made a transition to a web based project. In the past, atlas projects often suffered from expensive map production, small circulation and limited audience. Updates were expensive to produce and took a long time until they hit the public. Many atlas projects, after moving to the web, can now reach a wider audience, produce cheaper, provide a larger number of maps and map types and integrate with and benefit from other web resources. Some atlases even ceased their printed editions after going online, sometimes offering printing on demand features from the online edition. Some atlases (primarily from North America) also offer raw data downloads of the underlying geospatial data sources.

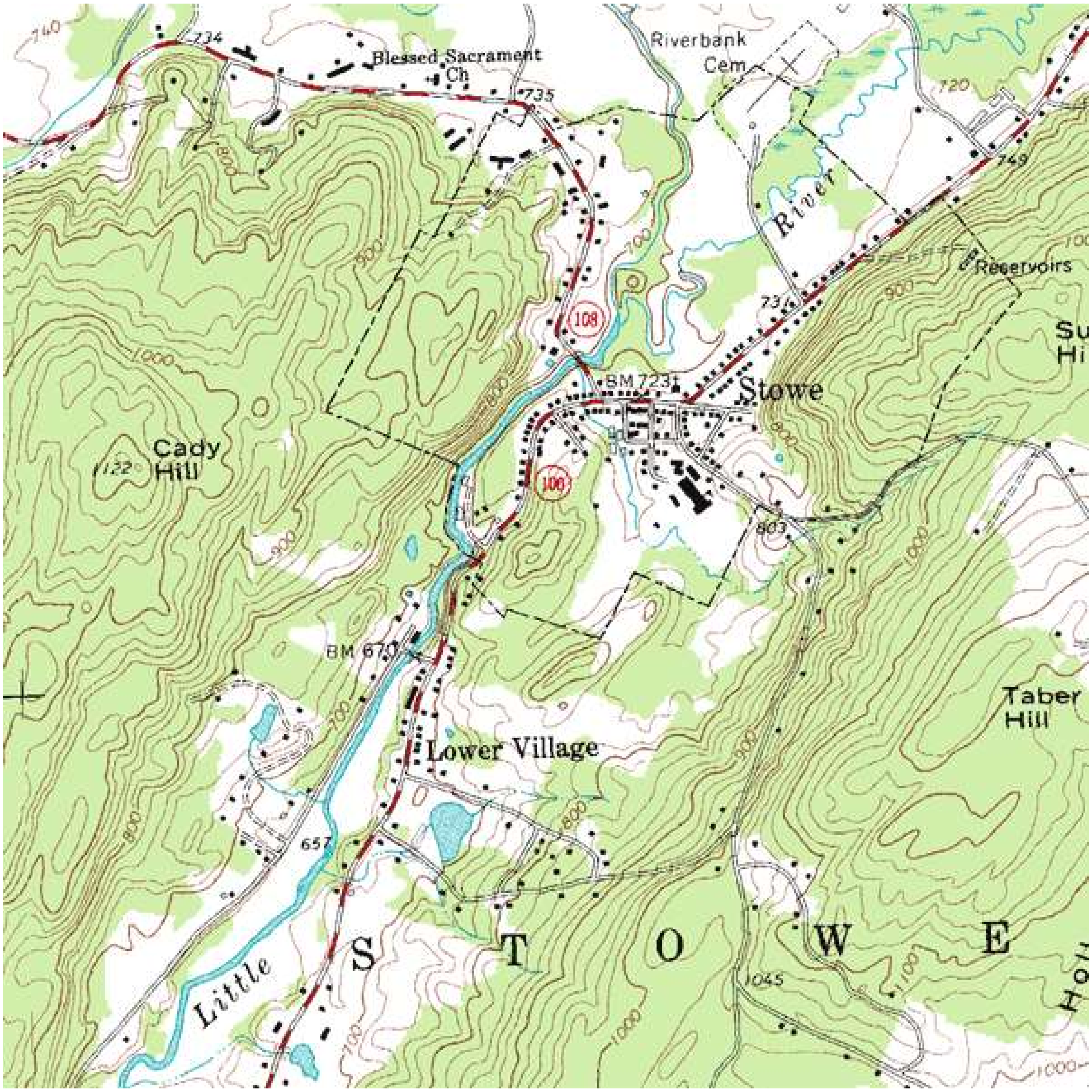
Personalized web maps

Personalized web maps allow the map user to apply his own data filtering, selective content and the application of personal styling and map symbolization. The OGC (Open Geospatial Consortium) provides the SLD standard (Styled Layer Description) that may be sent to a WMS server for the application of individual styles. This implies that the content and data structure of the remote WMS server is properly documented.

Realtime web maps

Realtime maps show the situation of a phenomenon in close to realtime (only a few seconds or minutes delay). Data is collected by sensors and the maps are generated or updated at regular intervals or immediately on demand. Examples are weather maps, traffic maps or vehicle monitoring systems.

Static web maps



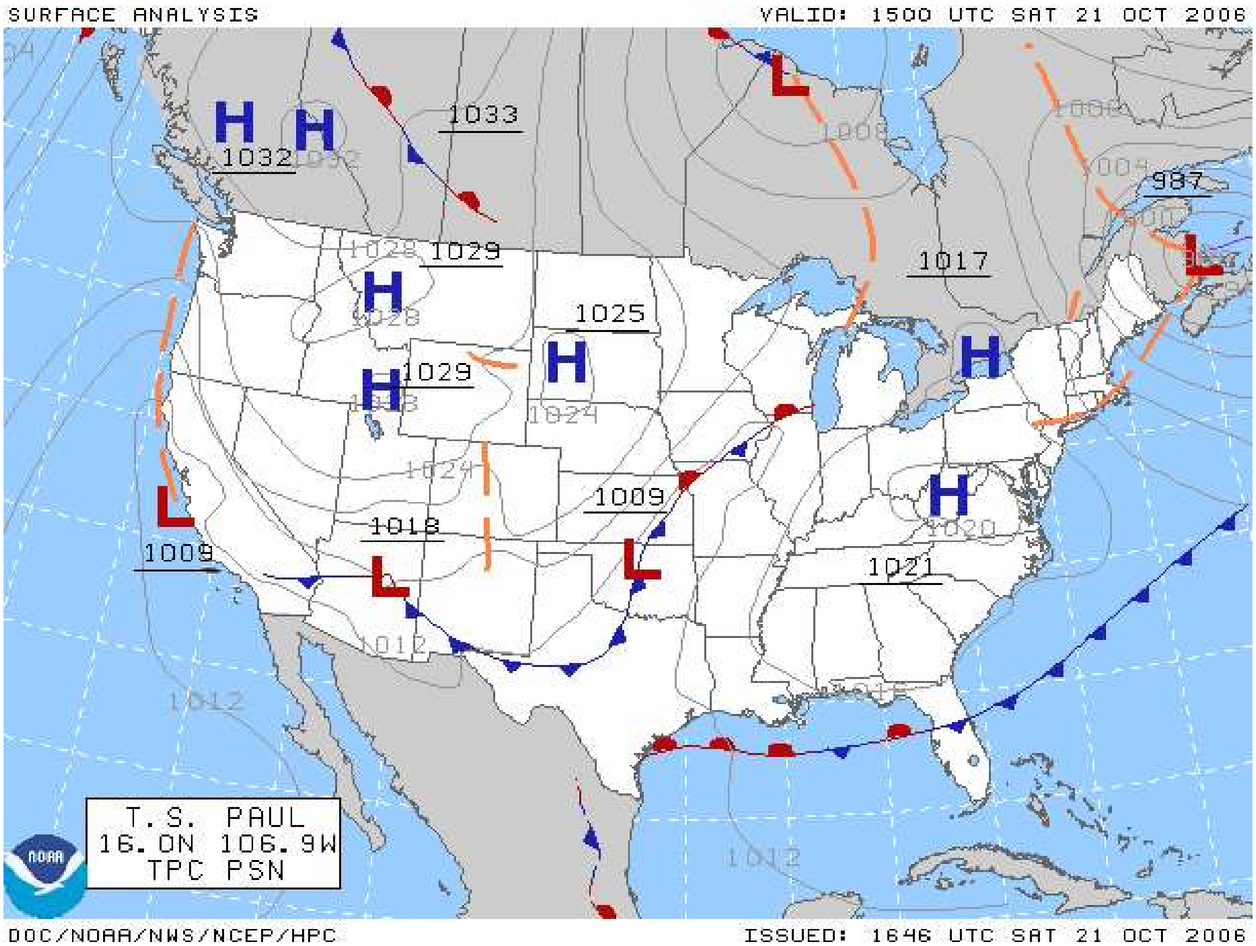
A USGS DRG - a static map

Static web pages are *view only* with no animation and interactivity. They are only created once, often manually and infrequently updated. Typical graphics formats for static web maps are PNG, JPEG, GIF, or TIFF (e.g., drg) for raster files, SVG, PDF or SWF for vector files. Often, these maps are scanned paper maps and had not been designed as screen maps. Paper maps have a much higher resolution and information density than typical computer displays of the same physical size, and might be unreadable when displayed on screens at the wrong resolution.

Temporal web maps

Any depiction of a portion of an n-dimensional orthogonal web map structure in a planar projection with time as one of the coordinate axes.

Advantages of web maps



A surface weather analysis for the United States on October 21, 2006.

- Web maps can easily *deliver up to date information*. If maps are generated automatically from databases, they can display information in almost realtime. They don't need to be printed, mastered and distributed. Examples:
  - A map displaying election results, as soon as the election results become available.
  - A map displaying the traffic situation near realtime by using traffic data collected by sensor networks.

    - A map showing the current locations of mass transit vehicles such as buses or trains, allowing patrons to minimize their waiting time at stops or stations, or be aware of delays in service.
    - Weather maps, such as NEXRAD.
- *Software and hardware infrastructure for web maps is cheap.* Web server hardware is cheaply available and many open source tools exist for producing web maps.
- *Product updates can easily be distributed*. Because web maps distribute both logic and data with each request or loading, product updates can happen every time the web user reloads the application. In traditional cartography, when dealing with printed maps or interactive maps distributed on offline media (CD, DVD, etc.), a map update caused serious efforts, triggering a reprint or remastering as well as a redistribution of the media. With web maps, data and product updates are easier, cheaper, and faster, and can occur more often.
- *They work across browsers and operating systems*. If web maps are implemented based on open standards, the underlying operating system and browser do not matter.
- *Web maps can combine distributed data sources*. Using open standards and documented APIs one can integrate (*mash up*) different data sources, if the projection system, map scale and data quality match. The use of centralized data sources removes the burden for individual organizations to maintain copies of the same data sets. The downside is that one has to rely on and trust the external data sources.
- *Web maps allow for personalization*. By using user profiles, personal filters and personal styling and symbolization, users can configure and design their own maps, if the web mapping systems supports personalization. Accessibility issues can be treated in the same way. If users can store their favourite colors and patterns they can avoid color combinations they can't easily distinguish (e.g. due to color blindness).
- *Web maps enable collaborative mapping*. Similar to the web mapping technologies, such as DHTML/Ajax, SVG, Java, Adobe Flash, etc. enable distributed data acquisition and collaborative efforts. Examples for such projects are the OpenStreetMap project or the Google Earth community. As with other open projects, quality assurance is very important, however!
- *Web maps support hyperlinking to other information on the web*. Just like any other web page, web maps can act like an index to other information on the web. Any sensitive area in a map, a label text, etc. can provide hyperlinks to additional information. As an example a map showing public transport options can directly link to the corresponding section in the online train time table.

- *It is easy to integrate multimedia in and with web maps*. Current web browsers support the playback of video, audio and animation (SVG, SWF, Quicktime, and other multimedia frameworks).

Disadvantages of web maps and problematic issues

- *Reliability issues* – the reliability of the internet and web server infrastructure is not yet good enough. Especially if a web map relies on external, distributed data sources, the original author often cannot guarantee the availability of the information.
- *Geodata is expensive* – Unlike in the US, where geodata collected by governmental institutions is usually available for free or cheap, geodata is usually very expensive in Europe or other parts of the world.
- *Bandwidth issues* – Web maps usually need a relatively high bandwidth.
- *Limited screen space* – Like with other screen based maps, web maps have the problem of limited screen space. This is in particular a problem for mobile web maps and location based services where maps have to be displayed in very small screens with resolutions as low as 100×100 pixels. Hopefully, technological advances will help to overcome these limitations.
- *Quality and accuracy issues* – Many web maps are of poor quality, both in symbolization, content and data accuracy.
- *Complex to develop* – Despite the increasing availability of free and commercial tools to create web mapping and web GIS applications, it is still a complex task to create interactive web maps. Many technologies, modules, services and data sources have to be mastered and integrated.
- *Immature development tools* – Compared to the development of standalone applications with integrated development tools, the development and debugging environments of a conglomerate of different web technologies is still awkward and uncomfortable.
- *Copyright issues* – Many people are still reluctant to publish geodata, especially in the light that geodata is expensive in some parts of the world. They fear copyright infringements of other people using their data without proper requests for permission.
- *Privacy issues* – With detailed information available and the combination of distributed data sources, it is possible to find out and combine a lot of private and personal information of individual persons. Properties and estates of individuals are now accessible through high resolution aerial and satellite images throughout the world to anyone.

History of web mapping

Event types

- ***Cartography-related events***

- **Technical events directly related to web mapping**
- *General technical events*
- Events relating to Web standards

This section contains some of the milestones of web mapping, online mapping services and atlases. Because web mapping depends on enabling technologies of the web, this section also includes a few milestones of the web.

- 1990-12: *First Web Browser and Web Server*, Tim Berners-Lee wrote first web browser and web server.
- 1991-04: HTTP 0.9 protocol, Initial design of the HTTP protocol for communication between browser and server.
- 1991-08: *WWW project announced in public newsgroup*, This is regarded as the debut date of the Web. Announced in newsgroup alt.hypertext.
- 1992-06: HTTP 1.0 protocol, Version 1.0 of the HTTP protocol. Introduces the POST method and persistent connections.
- 1993-04: *CERN announced web as free*, CERN announced that access to the web will be free for all. The web gained critical mass.
- 1993-06: HTML 1.0. The first version of HTML, published by T. Berners-Lee and Dan Connolly.
- 1993-07: ***Xerox PARC Map Viewer***, The first mapserver based on CGI/Perl, allowed reprojection styling and definition of map extent.
- 1994-06: ***The National Atlas of Canada***, The first version of the National Atlas of Canada was released. Can be regarded as the first online atlas.
- 1994-10: *Netscape Browser 0.9 (Mosaic)*, The first version of the highly popular browser Netscape Navigator.
- 1995-03: *Java 1.0*, The first public version of Java.
- 1995-11: HTML 2.0, Introduced forms, file upload, internationalization and client-side image maps.
- 1995-12: *Javascript 1.0*, Introduced first script based interactivity.
- 1995: *MapGuide*, First introduced as Argus MapGuide.
- 1996-01: *JDK 1.0*, First version of the Sun JDK.
- 1996-02: ***Mapquest***, The first popular online Address Matching and Routing Service with mapping output.
- 1996-06: ***MultiMap***, The UK-based MultiMap website launched offering online mapping, routing and location based services. Grew into one of the most popular UK web sites.
- 1996-11: **Geomedia WebMap 1.0**, First version of Geomedia WebMap, already supports vector graphics through the use of ActiveCGM.
- 1996-fall: *MapGuide*, Autodesk acquired Argus Technologies.and introduced Autodesk MapGuide 2.0.

- 1996-12: *Macromedia Flash 1.0*, First version of the Macromedia Flash plugin.
- 1997-01: HTML 3.2, Introduced tables, applets, script elements, multimedia elements, flowtext around images, etc.

National Atlas of the United States logo

- 1997-06: ***US Online National Atlas Initiative***, The USGS received the mandate to coordinate and create the online National Atlas of the United States of America .
- 1997-07: **UMN MapServer 1.0**, Developed as Part of the NASA ForNet Project. Grew out of the need to deliver remote sensing data across the web for foresters.
- 1997-12: HTML 4.0, Introduced styling with CSS, absolute and relative positioning of elements, frames, object element, etc.
- 1998-06: ***Terraserver USA***, A Web Map Service serving aerial images (mainly b+w) and USGS DRGs was released. One of the first popular WMS. This service is a joint effort of USGS, Microsoft and HP.
- 1998-07: **UMN MapServer 2.0**, Added reprojection support (PROJ.4).
- 1998-08: **MapObjects Internet Map Server**, ESRI's entry into the web mapping business.
- 1999-03: HTTP 1.1 protocol, Version 1.1 of the HTTP protocol. Introduces the request pipelining for multiple connections between server and client. This version is still in use as of 2007.
- 1999-08: ***National Atlas of Canada, 6th edition***, This new version was launched at the ICA 1999 conference in Ottawa. Introduced many new features and topics. Is being improved gradually, since then, and kept up-to-date with technical advancements.

- 2000-02: **ArcIMS 3.0**, The first public release of ESRI's ArcIMS.
- 2000-06: **ESRI Geography Network**, ESRI founded Geography Network to distribute data and web map services.
- 2000-06: **UMN MapServer 3.0**, Developed as part of the NASA TerraSIP Project. This is also the first public, open source release of UMN Mapserver. Added raster support and support for TrueType fonts (FreeType).
- 2000-08: *Flash Player 5*, This introduced ActionScript 1.0 (ECMAScript compatible).
- 2001-06: **MapScript 1.0 for UMN MapServer**, Adds a lot of flexibility to UMN MapServer solutions.
- 2001-09: SVG 1.0 W3C Recommendation, SVG (Scalable Vector Graphics) 1.0 became a W3C Recommendation.
- 2001-09: ***Tirolatlas***, A highly interactive online atlas, the first to be based on the SVG standard.
- 2002-06: **UMN MapServer 3.5**, Added support for PostGIS and ArcSDE. Version 3.6 adds initial OGC WMS support.
- 2002-07: **ArcIMS 4.0**, Version 4 of the ArcIMS web map server.
- 2003-01: SVG 1.1 W3C Recommendation, SVG 1.1 became a W3C Recommendation. This introduced the mobile profiles SVG Tiny and SVG Basic.

Screenshot from NASA World Wind

- 2003-06: ***NASA World Wind***, NASA World Wind Released. An open virtual globe that loads data from distributed resources across the internet. Terrain and buildings can be viewed 3 dimensionally. The (XML based) markup language allows users to integrate their own personal content. This virtual globe needs special software and doesn't run in a web browser.
- 2003-07: **UMN MapServer 4.0**, Adds 24bit raster output support and support for PDF and SWF.
- 2003-09: *Flash Player 7*, This introduced ActionScript 2.0 (ECMAScript 2.0 compatible (improved object orientation)). Also initial Video Playback support.
- 2004-07: **OpenStreetMap** was founded by Steve Coast. OSM is a web based collaborative project to create a world map under a free license.
- 2005-01: Nikolas Schiller creates the interactive "Inaugural Map" of downtown Washington, DC
- 2005-02: ***Google Maps***, The first version of Google Maps. Based on raster tiles organized in a quad tree scheme, data loading done with

XMLHttpRequests. This mapping application became highly popular on the web, also because it allowed other people to integrate google map services into their own website.

- 2005-04: **UMN MapServer 4.6**, Adds support for SVG.

- 2005-06: ***Google Earth***, The first version of Google Earth was released building on the virtual globe metaphor. Terrain and buildings can be viewed 3 dimensionally. The KML (XML based) markup language allows users to integrate their own personal content. This virtual globe needs special software and doesn't run in a web browser.
- 2005-11: *Firefox 1.5*, First Firefox release with native SVG support. Supports Scripting but no animation.
- 2006-06: *Opera 9*, Opera releases version 9 with extensive SVG support (including scripting and animation).
- 2006-08: SVG 1.2 Mobile Candidate Recommendation, This SVG Mobile Profile introduces improved multimedia support and many features required to build online Rich Internet Applications.
- 2009-01 Nokia makes Ovi Maps free on its smartphones.

Web mapping technologies

The potential number of technologies to implement web mapping projects is almost infinite. Any programming environment, programming language and serverside framework can be used to implement web mapping projects. In any case, both server and client side technologies have to be used. Following is a list of potential and popular server and client side technologies utilized for web mapping.

Server side technologies

- **Web server** – The webserver is responsible for handling http requests by web browsers and other user agents. In the simplest case they serve static files, such as HTML pages or static image files. Web servers also handle authentication, content negotiation, server side includes, URL rewriting and forward requests to dynamic resources, such as CGI applications or serverside scripting languages. The functionality of a webserver can usually be enhanced using modules or extensions. The most popular web server is Apache, followed by Microsoft Internet Information Server and others.
    - **CGI (common gateway interface)** applications are executables running on the webserver under the environment and user permissions of the webserver user. They may be written in any programming language (compiled) or scripting language (e.g. perl). A CGI application implements the common gateway interface

protocol, processes the information sent by the client, does whatever the application should do and sends the result back in a web-readable form to the client. As an example a web browser may send a request to a CGI application for getting a web map with a certain map extent, styling and map layer combination. The result is an image format, e.g. JPEG, PNG or SVG. For performance enhancements one can also install CGI applications such as FastCGI. This loads the application after the web server is started and keeps the application in memory, eliminating the need to spawn a separate process each time a request is being made.
    - Alternatively, one can use **scripting languages built into the webserver** as a module, such as PHP, Perl, Python, ASP, Ruby, etc. If built into the web server as a module, the scripting engine is already loaded and doesn't have to be loaded each time a request is being made.
- **Web application servers** are middleware which connects various software components with the web server and a programming language. As an example, a web application server can enable the communication between the API of a GIS and the webserver, a spatial database or other proprietary applications. Typical web application servers are written in Java, C, C++, C# or other scripting languages. Web application servers are also useful when developing complex realtime web mapping applications or Web GIS.
- **Spatial databases** are usually object relational databases enhanced with geographic data types, methods and properties. They are necessary whenever a web mapping application has to deal with dynamic data (that changes frequently) or with huge amount of geographic data. Spatial databases allow spatial queries, sub selects, reprojections, geometry manipulations and offer various import and export formats. A popular example for an open source spatial database is PostGIS. MySQL also implements some spatial features, although not as mature as PostGIS. Commercial alternatives are Oracle Spatial or spatial extensions of Microsoft SQL Server and IBM DB2. The OGC Simple Features for SQL Specification is a standard geometry data model and operator set for spatial databases. Most spatial databases implement this OGC standard.
- **WMS server** are specialized web mapping servers implemented as a CGI application, Java Servlet or other web application server. They either work as a standalone web server or in collaboration with existing web servers or web application servers (the general case). WMS Servers can generate maps on request, using parameters, such as map layer order, styling/symbolization, map extent, data format, projection, etc. The OGC Consortium defined the WMS standard to define the map requests and return data formats. Typical image formats for the map result are PNG, JPEG, GIF or SVG. There are open source WMS Servers such as UMN

Mapserver and Mapnik. Commercial alternatives exist from most commercial GIS vendors, such as ESRI ArcIMS, ArcGIS Server, GeoClip, Intergraph Geomedia WebMap, and others.

Client side technologies

- **Web browser** – In the simplest setup, only a web browser is required. All modern web browsers support the display of HTML and raster images (JPEG, PNG and GIF format). Some solutions require additional plugins (see below).
  - **ECMAScript support** – ECMAScript is the standardized version of JavaScript. It is necessary to implement client side interaction, refactoring of the DOM of a webpage and for doing network requests. ECMAScript is currently part of any modern web browser.
  - **Events support** – Various events are necessary to implement interactive client side maps. Events can trigger script execution or SMIL operations. We distinguish between:
    - **Mouse events** (mousedown, mouseup, mouseover, mousemove, click)
    - **Keyboard events** (keydown, keypress, keyup)
    - **State events** (load, unload, abort, error)
    - **Mutation events** (reacts on modifications of the DOM tree, e.g. DOMNodeInserted)
    - **SMIL animation events** (reacts on different states in SMIL animation, beginEvent, endEvent, repeatEvent)
    - **UI events** (focusin, focusout, activate)
    - **SVG specific events** (SVGZoom, SVGScroll, SVGResize)
  - **Network requests** – This is necessary to load additional data and content into a web page. Most modern browsers provide the XMLHttpRequest object which allows for get and post http requests and provides some feedback on the data loading state. The data received can be processed by ECMAScript and can be included into the current DOM tree of the web page / web map. SVG user agents alternatively provide the getURL() and postURL() methods for network requests. It is recommended to test for the existence of a network request method and provide alternatives if one method isn't present. As an example, a wrapper function could handle the network requests and test whether XMLHttpRequests or getURL() or alternative methods are available and choose the best one available. These network requests are also known under the term Ajax.
  - **DOM support** – The Document Object Model provides a language independent API for the manipulation of the document tree of the

webpage. It exposes properties of the individual nodes of the document tree, allows to insert new nodes, delete nodes, reorder nodes and change existing nodes. DOM support is included in any modern web browser. DOM support together with scripting is also known as DHTML or Dynamic HTML. Google Maps and many other web mapping sites use a combination of DHTML, Ajax, SVG and VML.

- **SVG support or SVG image support** – SVG is the abbreviation of "Scalable Vector Graphics" and integrates vector graphics, raster graphics and text. SVG also supports animation, internationalization, interactivity, scripting and XML based extension mechanisms. SVG is a huge step forward when it comes to delivering high quality, interactive maps. At the time of writing (2007–01), SVG is natively supported in Mozilla/Firefox >version 1.5, Opera >version 9 and the developer version of Safari/Webkit. Internet Explorer users still need the Adobe SVG viewer plugin provided by Adobe.
- **Java support** – some browsers still provide old versions of the Java virtual machine. An alternative is the use of the Sun Java Plugin. Java is a full featured programming language that can be used to create very sophisticated and interactive web maps. The Java2D and Java3D libraries provide 2d and 3d vector graphics support. The creation of Java based web maps requires a lot of programming know how. Adrian Herzog discusses the use of Java applets for the presentation of interactive choroplethe and cartogram maps.
- **Web browser plugins**
  - **Adobe Acrobat** – provides vector graphics and high quality printing support. Allows toggling of map layers, hyper links, multimedia embedding, some basic interactivity and scripting (ECMAScript).
  - **Adobe Flash** – provides vector graphics, animation and multimedia support. Allows the creation of sophisticated interactive maps, as with Java and SVG. Features a programming language (ActionScript) which is similar to ECMAScript. Supports Audio and Video.
  - **Apple Quicktime** – Adds support for additional image formats, video, audio and Quicktime VR (Panorama Images). Only available to Mac OS X and Windows.
  - **Adobe SVG viewer** – provide SVG 1.0 support for web browsers, only required for Internet Explorer Users, because it doesn't yet natively support SVG. The Adobe SVG viewer isn't developed any further and only fills the gap until Internet Explorer gains native SVG support.

- **Sun Java plugin** provides support for newer and advanced Java Features.

Chapter 9

Public Participation GIS

Public Participation Geographic Information Systems (PPGIS) was born, as a term, in 1996 at the meetings of the National Center for Geographic Information and Analysis (NCGIA). PPGIS is meant to bring the academic practices of GIS and mapping to the local level in order to promote knowledge production. The idea behind PPGIS is empowerment and inclusion of marginalized populations, who have little voice in the public arena, through geographic technology education and participation. PPGIS uses and produces digital maps, satellite imagery, sketch maps, and many other spatial and visual tools, to change geographic involvement and awareness on a local level.

Applications

Attendees to the *Mapping for Change International Conference on Participatory Spatial Information Management and Communication* conferred to at least three potential implications of PPGIS; it can: (1) enhance capacity in generating, managing, and communicating spatial information; (2) stimulate innovation; and ultimately; (3) encourage positive social change. This reflects on the rather nebulous definition of PPGIS as referenced in the *Encyclopedia of GIS* which describes PPGIS as having a definition problem.

There are a range of applications for PPGIS. The potential outcomes can be applied from community and neighborhood planning and development to environmental and natural resource management. Marginalized groups, be they grassroots organizations to indigenous populations could benefit from GIS technology.

Governments, non-government organizations and non-profit groups are a big force behind many programs. The current extent of PPGIS programs in the US has been evaluated by Sawicki and Peterman. They catalog over 60 PPGIS programs who aid in "public participation in community decision making by providing local-area data to community groups," in the United States (Craig et al., 2002:24). The organizations providing these programs are mostly universities, local chambers of commerce, non-profit foundations.

In general, neighborhood empowerment groups can form and gain access to information that is normally very easy for the official government and planning offices to obtain. It is easier for this to happen than for individuals of lower-income neighborhoods just working by themselves. There have been several projects where university students help implement GIS in neighborhoods and communities. It is believed that access to information is the doorway to more effective government for everybody and community empowerment. In a case study of a group in Milwaukee, residents of an inner-city neighborhood became active participants in building a community information system, learning to access public information and create and analyze new databases derived from their own surveys, all with the purpose of making these residents useful actors in city management and in the formation of public policy. In many cases, there are providers of data for community groups, but the groups may not know that such entities exist. Getting the word out would be beneficial.

Some of the spatial data that the neighborhood wanted was information on abandoned or boarded-up buildings and homes, vacant lots, and properties that contained garbage, rubbish and debris that contributed to health and safety issues in the area. They also appreciated being able to find landlords that were not keeping up the properties. The university team and the community were able to build databases and make maps that would help them find these areas and perform the spatial analysis that they needed. Community members learned how to use the computer resources, ArcView 1.0, and build a theme or land use map of the surrounding area. They were able to perform spatial queries and analyze neighborhood problems. Some of these problems included finding absentee landlords and finding code violations for the buildings on the maps (Ghose 2001).

The local, participatory management of urban neighborhoods usually follows on from 'claiming the territory', and has to be made compatible with national or local authority regulations on administering, managing and planning urban territory (McCall 2003). PPGIS applied to participatory Community/Neighborhood Planning has been examined by, among many others, [Howard (1999)], [Carver, Evans, Kingston, and Turton (1999)], [Leitner, McMaster, Elwood, McMaster, and Sheppard (2002)], and [Talen (1999)]. Specific attention has been given to applications such as housing issues (e.g. [Elwood (2002)]) or neighborhood revitalization (e.g. [Craig & Elwood (1998)]). Spatial databases along with the P-mapping are used to maintain a public records GIS or community land information systems (e.g. [Ventura, Niemann, Sutphin, & Chenoweth (2002)]). These are just a few of the uses of GIS in the community.

Approaches

There are two approaches to PPGIS use and application. These two perspectives, top-down and bottom-up, are the currently debated schism in PPGIS.

Top-down

According to Sieber (2006), PPGIS was first envisioned as a means of mapping individuals by many social and economic demographic factors in order to analyze the spatial differences in access to social services. She refers to this kind of PPGIS as *top-down*, being that it is less hands on for the public, but theoretically serves the public by making adjustments for the deficiencies, and improvements in public management.

Bottom-up

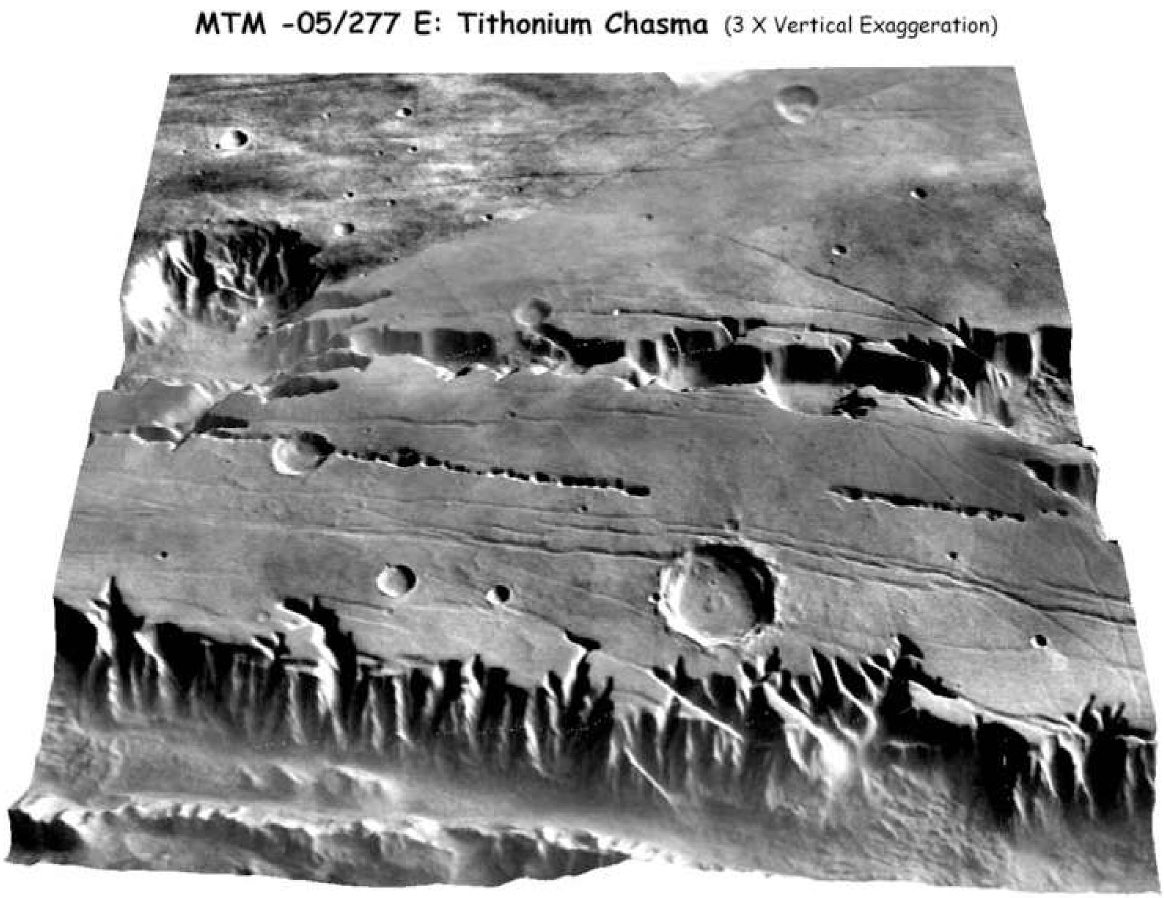
A current trend with academic involvement in PPGIS, is researching existing programs, and or starting programs in order to collect data on the effectiveness of PPGIS. Elwood (2006) in *The Professional Geographer*, talks in depth about the "everyday inclusions, exclusions, and contradictions of Participatory GIS research." The research is being conducted in order to evaluate if PPGIS is involving the public equally. In reference to Sieber's top-down PPGIS, this is a counter method of PPGIS, rightly referred to as *bottom-up* PPGIS. Its purpose is to work with the public to let them learn the technologies, then producing their own GIS.

Public Participation GIS is defined by Sieber as the use of geographic information systems to broaden public involvement in policymaking as well as to the value of GIS to promote the goals of nongovernmental organizations, grassroots groups and community based organizations (Sieber 2006). It would seem on the surface that PPGIS, as it is commonly referred to, in this sense would be of a beneficial nature to those in the community or area that is being represented. But in truth only certain groups or individuals will be able to obtain the technology and utilize it. Is PPGIS becoming more available to the underprivileged sector of the community? The question of "who benefits?" should always be asked, and does this harm a community or group of individuals.

The local, participatory management of urban neighborhoods usually follows on from 'claiming the territory', and has to be made compatible with national or local authority regulations on administering, managing and planning urban territory (McCall 2003). PPGIS applied to participatory Community/Neighborhood Planning has been examined by, among many others, [Howard (1999)], [Carver, Evans, Kingston, and Turton (1999)], [Leitner, McMaster, Elwood, McMaster, and Sheppard (2002)], and [Talen (1999)]. Specific attention has been given to applications such as housing issues (e.g. [Elwood (2002)]) or neighborhood revitalization (e.g. [Craig & Elwood (1998)]). Spatial databases along with the P-mapping are used to maintain a public records GIS or community land information systems (e.g. [Ventura, Niemann, Sutphin, & Chenoweth (2002)]). These are just a few of the uses of GIS in the community.

Chapter 10

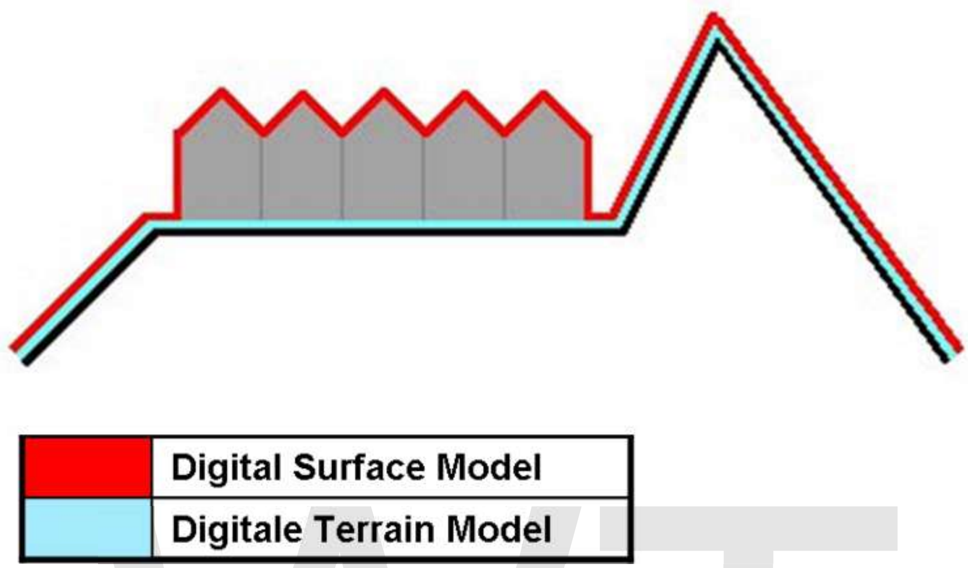
Digital Elevation Model



3D rendering of a DEM of Tithonium Chasma on Mars

A **digital elevation model** is a digital model or 3-D representation of a terrain's surface — commonly for a planet (including Earth), moon, or asteroid — created from terrain elevation data.

There is no common usage of the terms **digital elevation model (DEM)**, **digital terrain model (DTM)** and **digital surface model (DSM)** in scientific literature. In the most cases the term **digital surface model** represents the earth's surface and includes all objects on it. In contrast to a DSM, the **digital terrain model** represents the bare ground surface without any objects like plants and buildings (see Figure above).



Surfaces represented by a Digital Surface Model and Digital Terrain Model

The term Digital Elevation Model is often used as a generic term for DSMs and DTMs, only representing height information without any further definition about the surface. Other definitions equalise the terms DEM and DTM, or define the DEM as a subset of the DTM, which is also representing other morphological elements. There are also definitions which equalise the terms DEM and DSM. In the Web definitions can be found which define the DEM as a digital regularly spaced GRID and a DTM as a real three-dimensional model (TIN). Most of the data providers (USGS, ERSDAC, CGIAR) use the term DEM as a generic term for DSMs and DTMs. All datasets which are captured with satellites, airplanes or other flying platforms are originally DSMs (like SRTM or the ASTER GDEM). It is possible to compute a DTM from high resolution DSM datasets with complex algotithms (Li et al. (2005). In the following the term DEM is used as a generic term for DSMs and DTMs.

A DEM can be represented as a raster (a grid of squares, also known as a heightmap when representing elevation) or as a triangular irregular network (TIN). The TIN DEM dataset is also referred as a primary (measured) DEM, whereas the Raster DEM is referred as a secondary (computed) DEM. DEMs are commonly built using remote sensing techniques, but they may also be built from land surveying. DEMs are used often in geographic information systems, and are the most common basis for digitally-produced relief maps. The DEM could be acquired through techniques such as photogrammetry, LiDAR, IfSAR, land surveying, etc. (Li et al. 2005). While a DSM may be useful for landscape modeling, city modeling and visualization applications, a DTM is often required for flood or drainage modeling, land-use studies, geological applications, and much more.

Production

Relief map Sierra Nevada

Mappers may prepare digital elevation models in a number of ways, but they frequently use remote sensing rather than direct survey data. One powerful technique for generating digital elevation models is interferometric synthetic aperture radar: two passes of a radar satellite (such as RADARSAT-1 or TerraSAR-X), or a single pass if the satellite is equipped with two antennas (like the SRTM instrumentation), suffice to generate a digital elevation map tens of kilometers on a side with a resolution of around ten meters. Alternatively, other kinds of stereoscopic pairs can be employed using the digital image correlation method, where two optical images acquired with different angles taken from the same pass of an airplane or an Earth Observation Satellite (such as the HRS instrument of SPOT5 or the VNIR band of ASTER).

In 1986, the SPOT 1 satellite provided the first usable elevation data for a sizeable portion of the planet's landmass, using two-passes stereoscopic correlation. Later, further data were provided by the European Remote-Sensing Satellite (ERS) using the same method, the Shuttle Radar Topography Mission using single-pass SAR and the ASTER instrumentation on the Terra satellite using double-pass stereo pairs.

Older methods of generating DEMs often involve interpolating digital contour maps that may have been produced by direct survey of the land surface; this method is still used in mountain areas, where interferometry is not always satisfactory. Note that the contour line data or any other sampled elevation datasets (by GPS or ground survey) are not

DEMs, but may be considered digital terrain models. A DEM implies that elevation is available continuously at each location in the study area.

The quality of a DEM is a measure of how accurate elevation is at each pixel (absolute accuracy) and how accurately is the morphology presented (relative accuracy). Several factors play an important role for quality of DEM-derived products:

- terrain roughness;
- sampling density (elevation data collection method);
- grid resolution or pixel size;
- interpolation algorithm;
- vertical resolution;
- terrain analysis algorithm;

Methods for obtaining elevation data used to create DEMs

- LIDAR
- Stereo photogrammetry from aerial surveys
- Real Time Kinematic GPS
- Topographic maps
- Theodolite or total station
- Doppler radar
- Focus variation
- Inertial surveys

Uses

Bezmiechowa airfield 3D Digital Surface Model obtained using Pteryx UAV flying 200m above hilltop

Digital Surface Model of motorway interchange construction site. Note that tunnels are closed.

Common uses of DEMs include:

- Extracting terrain parameters
- Modeling water flow or mass movement (for example avalanches and landslides)
- Creation of relief maps
- Rendering of 3D visualizations.
- 3d flight planning
- Creation of physical models (including raised-relief maps)
- Rectification of aerial photography or satellite imagery.
- Reduction (terrain correction) of gravity measurements (gravimetry, physical geodesy).
- Terrain analyses in geomorphology and physical geography
- Geographic Information Systems (GIS)
- Engineering and infrastructure design
- Global positioning systems (GPS)
- Line-of-sight analysis
- Base mapping

- Flight simulation
- Precision farming and forestry
- Surface analysis
- Intelligent transportation systems (ITS)
- Auto safety / Advanced Driver Assistance Systems (ADAS)

Sources

A free DEM of the whole world called GTOPO30 (30 arcsecond resolution, approx. 1 km) is available, but its quality is variable and in some areas it is very poor. A much higher quality DEM from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument of the Terra satellite is also freely available for 99% of the globe, and represents elevation at a 30 meter resolution. A similarly high resolution was previously only available for the United States territory under the Shuttle Radar Topography Mission (SRTM) data, while most of the rest of the planet was only covered in a 3 arc-second resolution (around 90 meters). The limitation with the GTOPO30 and SRTM datasets is that they cover continental landmasses only, and SRTM does not cover the polar regions and has mountain and desert no data (void) areas. SRTM data, being derived from radar, represents the elevation of the first-reflected surface — quite often tree tops. So, the data are not necessarily representative of the ground surface, but the top of whatever is first encountered by the radar. Submarine elevation (known as bathymetry) data is generated using ship-mounted depth soundings. The SRTM30Plus dataset (used in NASA World Wind) attempts to combine GTOPO30, SRTM and bathymetric data to produce a truly global elevation model. A novel global DEM of postings lower than 12m and a height accuracy of less than 2m is expected being generated by the TanDEM-X satellite mission which started in July 2010.

The most usual grid (raster) is between 50 and 500 meters. In gravimetry e.g., the primary grid may be 50 m, but is switched to 100 or 500 meters in distances of about 5 or 10 kilometers.

Many national mapping agencies produce their own DEMs, often of a higher resolution and quality, but frequently these have to be purchased, and the cost is usually prohibitive to all except public authorities and large corporations. DEMs are often a product of National LIDAR Dataset programs.

Free DEMs are also available for Mars: the MEGDR, or Mission Experiment Gridded Data Record, from the Mars Global Surveyor's Mars Orbiter Laser Altimeter (MOLA) instrument; and NASA's Mars Digital Terrain Model (DTM).

United States

The US Geological Survey produces the National Elevation Dataset, a seamless DEM for the contiguous United States, Hawaii and Puerto Rico based on 7.5' topographic mapping. As of the beginning of 2006, this replaces the earlier DEM tiled format (one DEM per USGS topographic map).