Chapter 1
Introduction to GPS

The NAVSTAR Global Positioning System (GPS) is a satellite-based radio-positioning and time-transfer system designed, financed, deployed, and operated by the U.S. Department of Defense. GPS has also demonstrated a significant benefit to the civilian community who are applying GPS to a rapidly expanding number of applications. What attracts us to GPS is:

- The relatively high positioning accuracies, from tens of metres down to the millimetre level.
- The capability of determining velocity and time, to an accuracy commensurate with position.
- The signals are available to users anywhere on the globe: in the air, on the ground, or at sea.
- Its is a positioning system with no user charges, that simply requires the use of relatively low cost hardware.
- It is an all-weather system, available 24 hours a day.
- The position information is in three dimensions, that is, vertical as well as horizontal information is provided.

The number of civilian users is already significantly greater than that of the military users. However, for the time being the U.S. military still operates several "levers" with which they control the performance of GPS (Section 1.2.3). Nevertheless, despite the handicap of GPS being a military system there continues to be tremendous product innovation within the civilian sector, and it is ironic that this innovative drive is partly directed to developing technology and procedures to overcome some of the constraints to GPS performance which have been applied by the system's military operators.

1.1 Introduction to the System Components

1.1.1 *System Design Considerations*

Development work on GPS commenced within the U.S. Department of Defense in 1973, the motivation being to develop an all-weather, 24-hour, global positioning system to support the positioning requirements for the armed forces of the U.S. and its allies. (For a background to the development of the GPS system the reader is referred to [1].) The system was therefore designed to replace the large variety of navigational systems already in use, and great emphasis was placed on the system's reliability and survivability. In short, a number of stringent conditions had to be met:

- suitable for all classes of platform: aircraft (jet to helicopter), ship, land (vehicle-mounted to handheld) and space (missiles and satellites),
- able to handle a wide variety of dynamics,
- real-time positioning, velocity and time determination capability to an appropriate accuracy,
- the positioning results were to be available on a single global geodetic datum,
- highest accuracy to be restricted to a certain class of user,
- resistant to jamming (intentional and unintentional),
- redundancy provisions to ensure the survivability of the system,
- passive positioning system that does not require the transmission of signals from the user to the satellite(s),
- able to provide the service to an unlimited number of users,
- low cost, low power, therefore as much complexity as possible should be built into the satellite segment, and
- total replacement of the Transit[1] satellite and other terrestrial navaid systems.

This led to a design based on the following essential concepts:

- A one-way ranging system, in which the satellites transmit signals, but are unaware of who is using the signal (no receiving function). As a result the user (or listener) cannot easily be: (a) detected by the enemy (military context), or (b) charged for using the system (civilian context).
- Use of the latest atomic clock and microwave transmission technology, including spread-spectrum techniques.
- A system that makes range-like measurements with the aid of pseudo-random binary codes

modulated on carrier signals.

- Satellite signals that are unaffected by cloud and rain.
- A multiple satellite system which ensures there is always a sufficient number of satellites visible simultaneously anywhere on the globe, and at any time.
- Positioning accuracy degradation that is graceful.

What was perhaps unforeseen by the system designers was the power of product innovation, which has added significantly to the versatility of the GPS as a system for precise positioning and navigation. For example, GPS is able to support a number of positioning and measurement modes in order to satisfy simultaneously a variety of users, from those satisfied with general navigational accuracies (tens of metres) to those demanding very high (sub-centimetre) relative positioning accuracies. *It has now so penetrated certain applications areas that it is difficult for us to imagine life without GPS!*

Rarely have so many seemingly unrelated technological advances been required to make a complex system such as GPS work. Briefly they are:

Space System Reliability: The U.S. space program had by 1973 demonstrated the reliability of space hardware. In particular, the Transit system had offered important lessons. The Transit satellites were originally designed to last 2-3 years in orbit, yet some of the satellites have operated well beyond their design life. In fact Transit continued to perform reliably for over 25 years.

Atomic Clock Technology: With the development of atomic clocks a new era of precise time-keeping had commenced. However, before the GPS program was launched these precise clocks had never been tested in space. The development of reliable, stable, compact, space-qualified atomic frequency oscillators (rubidium, and then cesium) was therefore a significant technological breakthrough. The advanced clocks now being used on the GPS satellites routinely achieve long-term frequency stability in the range of a few parts in 10^{14} per day (about 1 sec in 3,000,000 years!). This long-term stability is one of the keys to GPS, as it allows for the autonomous, synchronised generation and transmission of accurate timing signals by each of the GPS satellites without continuous monitoring from the ground.

Quartz Crystal Oscillator Technology: In order to keep the cost of user equipment down, quartz crystal oscillators were proposed (similar to those used in modern digital watches), rather than using atomic clocks as in the GPS satellites. Besides their low cost, quartz oscillators have excellent short-term stability. However, their long-term drift must be accounted for as part of the user position determination process.

Precise Satellite Tracking and Orbit Determination: Successful operation of GPS, as well as the Transit system, depends on the precise knowledge and prediction of a satellite's position with respect to an earth-fixed reference system. Tracking data collected by ground monitor stations is analysed to determine the satellite orbit over the period of tracking (typically one week). This reference ephemeris is extrapolated into the future and the data is then up-loaded to the satellites. Prediction accuracies of the satellite coordinates, for one day, at the few metre level have been demonstrated.

Spread-Spectrum Technology: The ability to track and obtain any selected GPS satellite signal (a receiver will be required to track a number of satellites at the same time), in the presence of considerable ambient noise is a critical technology. This is now possible using spread-spectrum and pseudo-random-noise coding techniques.

Large-Scale Integrated Circuit Technology: To realise the desired low cost, low power and small size necessary for much of the user equipment, the GPS program relies heavily on the successful application of VLSI circuits, and powerful computing capabilities built onto them.

The GPS system consists of three segments (figure 1.1). (Good general references on the GPS system are [2,3].):

- The **Space Segment**: comprising the satellites and the transmitted signals.

- The **Control Segment**: the ground facilities carrying out the task of satellite tracking, orbit computations, telemetry and supervision necessary for the daily control of the space segment.
- The **User Segment**: the entire spectrum of applications equipment and computational techniques that are available to the users.

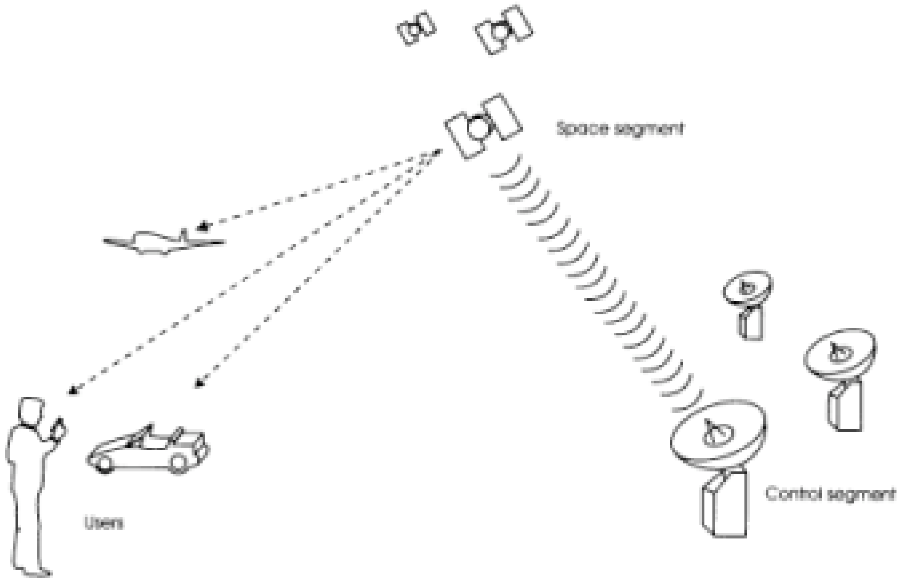

Figure 1.1: GPS System Elements.

1.1.2 *The Space Segment*

The Space Segment consists of the constellation of spacecraft and the signals broadcast by them which allow users to determine position, velocity and time. The basic functions of the satellites are to:

- Receive and store data transmitted by the Control Segment stations.
- Maintain accurate time by means of several onboard atomic clocks.
- Transmit information and signals to users on two L-band frequencies.
- Provide a stable platform and orbit for the L-band transmitters.

Several constellations of GPS satellites have been deployed, and several more are planned. The *experimental* satellites, the so-called "Block I" satellites, were built by the Rockwell Corporation. The first was launched in February 1978, and the last of the eleven satellite series (one exploded on the launchpad) was launched in 1985. The *operational* series of GPS satellites, the "Block II" and "Block IIA" satellites, were also built by the Rockwell Corporation. The 20 *replacement* "Block IIR" series of satellites, first launched in 1997, are built by the General Electric Corporation (now the Lockheed Martin Corporation). The "Block IIF" series are still in the design phase and may, for example, incorporate an additional civilian transmission frequency. They are planned for launch from 2005 onwards. The operational satellite I.D.s are separated into three space vehicle numbering series: SVN 13 through 21 for the Block II, SVN 22 through 40 for Block IIA, and SVN 41 and above for the Block IIR satellites.

The current status of the GPS constellation and such details as the launch and official commissioning date, the orbital plane and position within the plane, the satellite I.D. number(s), etc., can be obtained from several electronic GPS information sources on the Internet. Section 1.2.1 describes the general satellite orbit characteristics.

Each GPS satellite transmits a unique navigational signal centred on two L-band frequencies of the electromagnetic spectrum, permitting the ionospheric propagation effect on the signals to be eliminated. At these frequencies the signals are highly directional and so are easily reflected or blocked by solid objects. Clouds are easily penetrated, but the signals may be blocked by foliage (the extent of blockage is dependent on the type and density of the leaves and branches, and whether they are wet or dry). The signal is transmitted with enough power to ensure a minimum signal power level of -

160dBw at the earth's surface (the maximum it is likely to reach is about -153dBw, see [3]). The satellite signal consists of the following components (figure 1.2):

- The two L-band carrier waves.
- The ranging codes modulated on the carrier waves.
- The so-called "navigation message".

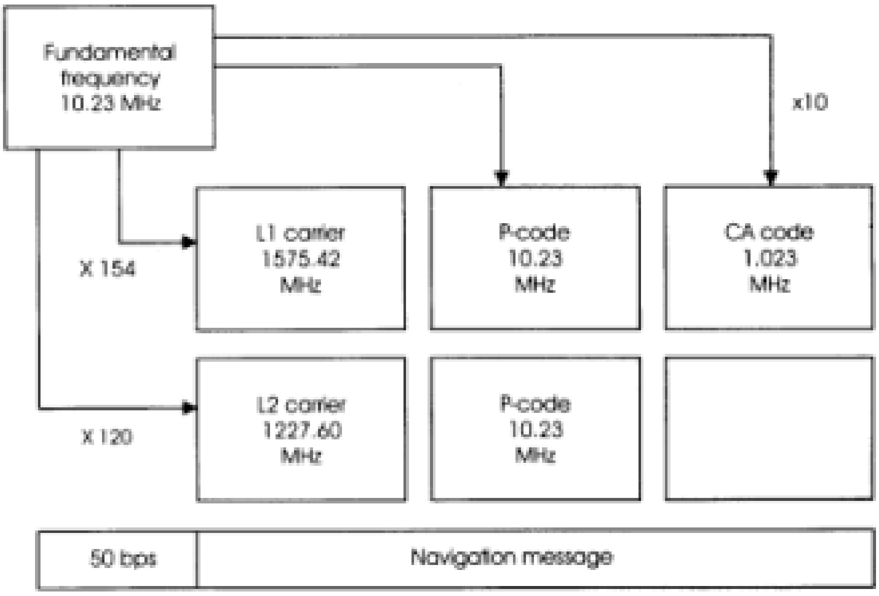

Figure 1.2: GPS Satellite Signal Components.

Modulated onto the carrier waves are the PRN ranging codes and navigation message for the user. The primary function of the ranging codes is to permit the *signal transit time* (from satellite to receiver) to be determined. The transit time when multiplied by the velocity of light then gives a measure of the receiver-satellite "range" (in reality the measurement process is considerably more complex). The navigation message contains the satellite orbit information, satellite clock parameters, and pertinent general system information necessary for real-time navigation to be performed. All signal components are derived from the output of a highly stable atomic clock. Each GPS satellite is equipped with several cesium and rubidium atomic clocks.

1.1.3 *The Control Segment*

The Control Segment consists of facilities necessary for satellite health monitoring, telemetry, tracking, command and control, satellite orbit and clock data computations, and data uplinking. There are five ground facility stations: Hawaii, Colorado Springs, Ascension Island, Diego Garcia and Kwajalein. All are owned and operated by the U.S. Department of Defense and perform the following functions:

- All five stations are *Monitor Stations*, equipped with GPS receivers to track the satellites. The resultant tracking data is sent to the Master Control Station.
- Colorado Springs is the *Master Control Station* (MCS), where the tracking data are processed in order to compute the satellite ephemerides and satellite clock corrections. It is also the station that initiates all operations of the space segment, such as spacecraft manoeuvring, signal encryption, satellite clock-keeping, etc.
- Three of the stations (Ascension Is., Diego Garcia, and Kwajalein) are *Upload Stations* allowing for the uplink of data to the satellites. The data includes the orbit and clock correction information transmitted within the navigation message, as well as command telemetry from the MCS.

Overall operation of the Control and Space Segments is the responsibility of the U.S. Air Force Space Command, Second Space Wing, Satellite Control Squadron at the Falcon Air Force Base, Colorado.

Each of the upload stations can view all the satellites once a day. All satellites are therefore in contact with an upload station three times a day, and new navigation messages as well as command telemetry can be transmitted to the GPS satellites every eight hours if necessary. The computation of: (a) the satellite orbits or "ephemerides", and (b) the determination of the satellite clock errors, are the most important function of the Control Segment. The first is necessary because the GPS satellites function as "orbiting control stations" and their coordinates must be known to a relatively high accuracy, while the latter permit a significant measurement bias to be reduced.

The GPS satellites travel at high velocity (of the order of 4 km/sec), but within a more or less regular orbit pattern. After a satellite has separated from its launch rocket and it begins orbiting the earth, it's orbit is defined by its initial position and velocity, and the various force fields acting on the satellite. In the case of the gravitational field for a spherically symmetric body (a reasonable approximation of the earth at the level of about 1 part in 10^3) this produces an elliptical orbit which is fixed in space -- the *Keplerian ellipse*. Due to the effects of the other, non-spherical gravitational components of the earth's gravity field, and non-gravitational forces, which perturb the orbit, the actual trajectory of the satellite departs from the ideal Keplerian ellipse (figure 1.3).

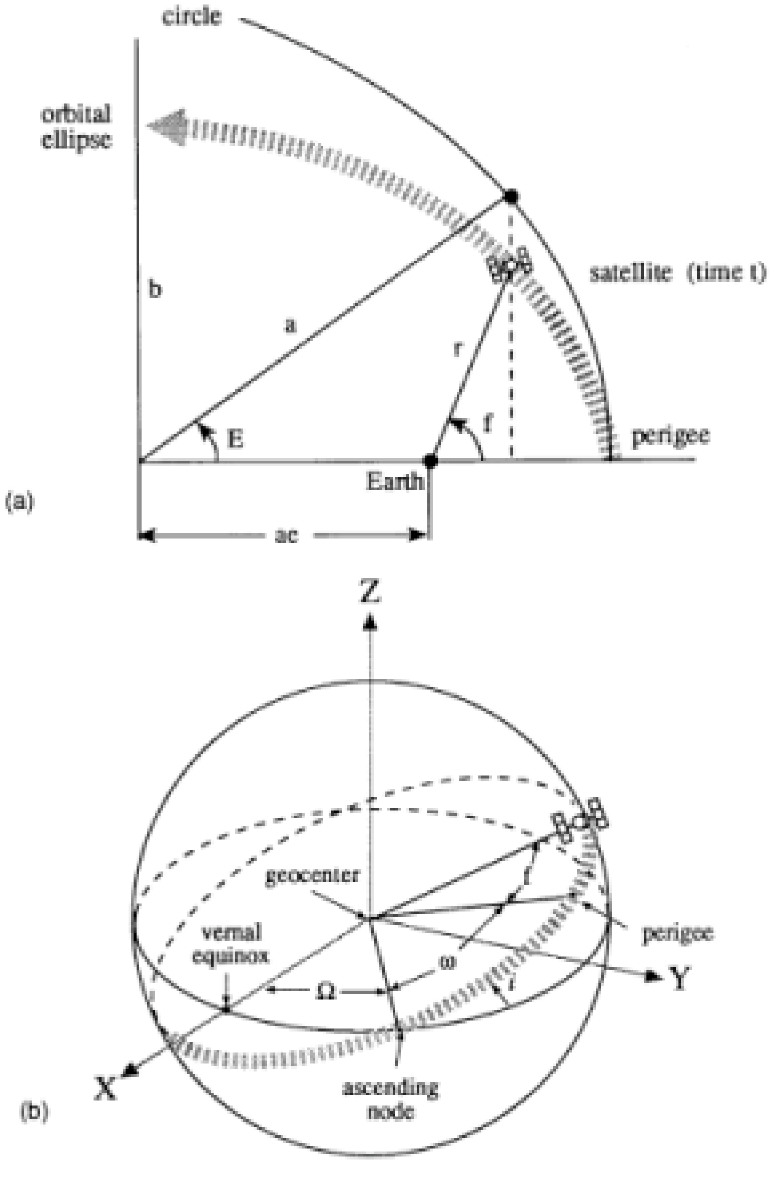

Figure 1.3: (a) The Keplerian Ellipse, and (b) Keplerian Orbital Elements.

The most significant forces that influence satellite motion are:

- the spherical and non-spherical gravitational attraction of the earth,
- the gravitational attractions of the sun, moon, and planets (the "third body" effects),
- atmospheric drag effects,
- solar radiation pressure (both direct and albedo effects), and
- the variable part of the earth's gravitational field arising from the solid earth and ocean tides.

To determine the motion of a satellite to a high precision these *perturbing* forces need to be modelled accurately. If these forces were known perfectly and the initial position and velocity of the satellite were given, then the integration of the Equations of Motion would give the satellite's position and velocity at any time in the future.

However, the perturbing forces are not known to sufficient precision. An "orbit computation" process is therefore performed, in which satellite observations obtained at tracking sites of known position (in the case of GPS, the monitor stations of the Control Segment) are analysed in order to produce an orbit that is a "best fit" to the available observations. This involves the adjustment of the appropriate parameters of the orbit, possibly together with several additional force model parameters (see [4]). Determining the orbit is a complex procedure, and in the case of GPS satellites this process occurs on a continuous, automatic basis.

The product of the orbit computation process at the MCS is the *satellite ephemeris* (or trajectory). A satellite ephemeris may be expressed in a number of forms:

- a list of 3-D coordinates, and velocities, at regular intervals of time,
- the Keplerian elements at some reference epoch, plus their rate-of-change with time,
- a polynomial representation of the trajectory in a suitable reference system, such as along-track, cross-track, or radial components, or
- satellite position and velocity at some reference time epoch, and requiring these values to be derived for subsequent times by integrating the Equations of Motion.

The GPS broadcast ephemeris, as represented in the navigation message, is actually a combination of all of the above orbit representations ([12]). The orbital ephemerides are expressed in the reference system most appropriate for positioning, which is an *earth-fixed* reference system such as WGS84. Hence the Control Segment has the function of propagating the satellite datum (Section 1.2.5), which users connect to via the transmitted satellite ephemerides.

The behaviour of each GPS satellite clock is monitored against GPS Time, as maintained by an ensemble of atomic clocks at the GPS Master Control Station. The satellite clock *bias, drift* and *drift-rate* relative to GPS Time are explicitly determined in the same procedure as the estimation of the satellite ephemeris. The clock behaviour so determined is made available to all GPS users via clock error coefficients (defining the mis-synchronization with GPS Time) in a polynomial form broadcast in the navigation message. However, what is available to users is really a *prediction* of the clock behaviour for some future time interval. Due to random deviations -- even cesium and rubidium oscillators are not entirely predictable -- the deterministic models of satellite clock error are only accurate to about 20 nanoseconds. This is not precise enough for accurate range measurement (see Section 1.3.6).

As the GPS system matures we can expect that the satellites will operate with greater independence from the ground-based Control Segment, without significant degradation in performance. For example the "Block IIR" and "Block IIF" satellites will have a crosslink capability enabling between-satellite communication and ranging. *The satellites will talk to each other!* This data will be processed to produce the ephemeris information within the space segment, with relatively little operator control having to be exercised.

1.1.4 *The User Segment -- The Applications*

This is the part of the GPS system with which we are most concerned -- the space and control segments being largely transparent to the operations of the navigation function. Of interest is the range of GPS:

- Applications,

- Equipment, and
- Positioning strategies.

The "engine" of commercial GPS product development is, without doubt, the *user applications*. Each day new applications are being identified, each with its unique requirements with regards to: accuracy of the results, reliability, operational constraints, user hardware, data processing algorithms, latency of the GPS results, etc. To make sense of the bewildering range of GPS applications it may be useful to classify them according to the following:

(1) Land, Sea and Air Navigation and Tracking, including enroute as well as precise navigation, collision avoidance, cargo monitoring, vehicle tracking, search and rescue operations, etc. *While the accuracy requirement may be modest and the user hardware is generally comparatively low cost, the reliability, integrity and speed with which the results are needed is generally high.*

(2) Surveying and Mapping, on land, at sea and from the air. Includes geophysical and resource surveys, GIS data capture surveys, etc. *The applications are of relatively high accuracy, for positioning in both the static and moving receiver mode, and generally require specialised hardware and data processing software.*

(3) Military Applications. *Although these are largely mirrored by civilian applications, the military GPS systems are generally developed to "military specifications" and a greater emphasis is placed on system reliability.*

(4) Recreational Uses, on land, at sea and in the air. *The primary requirement is for low cost instruments which are very easy to use.*

(5) Other specialised uses, such as time transfer, attitude determination, spacecraft operations, atmospheric studies, etc. *Obviously such applications require specially developed, high cost systems, often with additional demanding requirements such as real-time operation, etc.*

GPS user equipment has undergone an extensive program of development, both in the military and civilian area. In this context, GPS "equipment" refers to the combination of:

- hardware,
- software, and
- operational procedures or requirements.

While the military R&D programs have concentrated on achieving a high degree of miniaturisation, modularisation and reliability, the civilian user equipment manufacturers have, in addition, sought to bring down costs and to develop features that *enhance* the capabilities of the positioning system. The following general remarks can be made. Civilian users have, from the earliest days of GPS availability, demanded ever increasing levels of performance, in particular higher accuracy and improved reliability. This is particularly true of the survey user seeking levels of accuracy several orders of magnitude higher than that of the navigation user. In some respects the GPS user technology is being driven by the precise positioning market -- in much the same way that automotive technology often benefits from car racing. Yet another major influence on the development of GPS equipment has been the increasing variety of civilian applications. For although there may exist a similar positioning accuracy requirement across many user applications, to address a particular application in the most satisfactory manner, a specific combination of hardware and software features is often required.

It is expected that the worldwide market for GPS receiver equipment will grow from about US$1 billion at the present time, to over US$8 billion by the year 2000! Market surveys suggest that the greatest growth is expected to be in the commercial and consumer markets such as ITS applications, integration of GPS and cellular phones, and portable GPS for outdoor recreation and similar activities. These could account for more than 60% of the GPS market by the turn of the century. There are at present over 100 manufacturers of GPS instruments of varying kinds -- GPS instrumentation is discussed further in Section 1.4.

1.1.5 *The User Segment -- Positioning Principles*

The basic concept of GPS positioning is that of *positioning-by-ranges*. The geometrical principles of positioning can be demonstrated in terms of the intersection of locii. In the two-dimensional case, a measured range to a known point constrains the position to lie on circle with the measured range as radius. In three dimensions a measured range to a known point constrains the position in 3-D space to lie on the surface of a sphere centred at the known point, with radius being the measured distance (figure 1.4). In the case of GPS, the distance measurement is made to a satellite with known position (coordinates are obtained from the satellite ephemeris data transmitted within the navigation message), however the principle applies to any range measuring positioning system, terrestrial or satellite-based.

In two dimensions, position can be defined as the intersection of two circles, involving distances d_1 and d_2 to two known points, as shown in figure 1.5. Note that there are two possible solutions, only one of which is correct. In general one solution can be discarded rather easily through apriori knowledge of approximate position and velocity. Another possibility is to measure another range to a third point and if all ranges are measured without error the intersection of three LOPs is a single uniquely defined point.

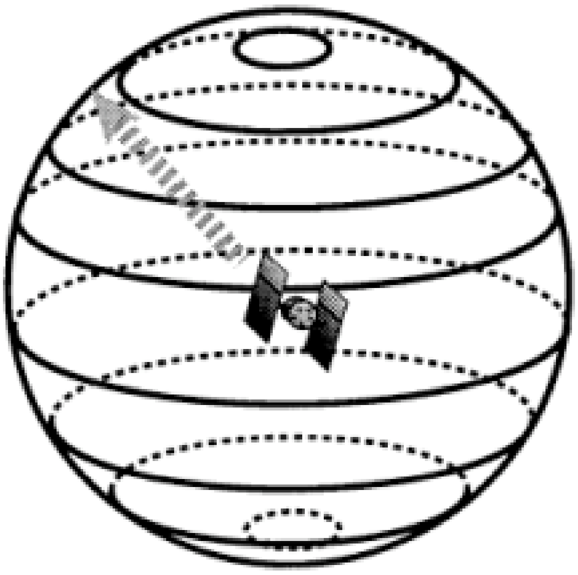

Figure 1.4: Surfaces of Position for Range Measurements.

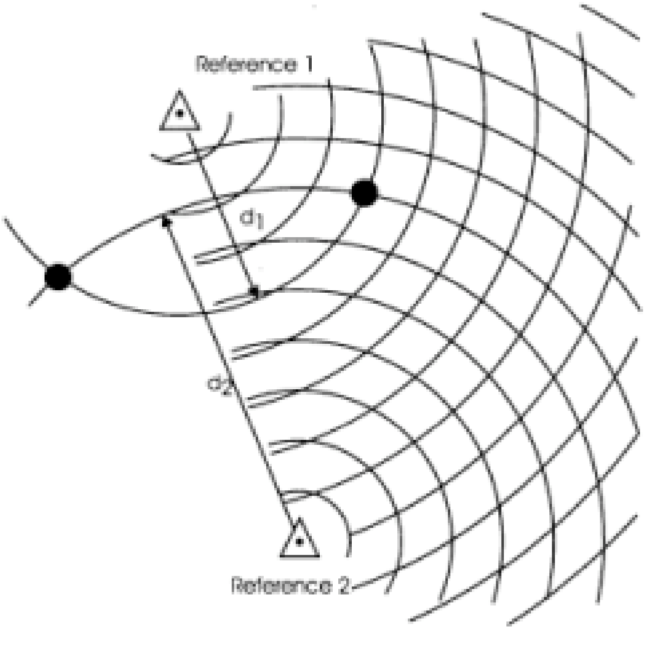

Figure 1.5: The Intersection of Circular Lines of Position for 2-D Positioning.

In the three-dimensional case, the intersection of three spheres describes two points in space, only one of which is correct (figure 1.6). Hence, a minimum of three ranges are required, to three separated known points, in order to solve the 3-D position problem. The quality of the positioning solution is dependent, amongst other things, on the accuracy with which the ranges can be measured and the geometry of the intersection.

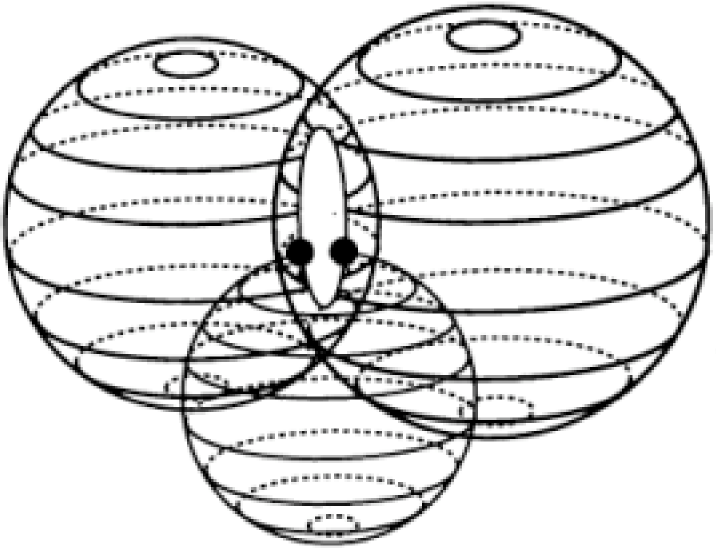

Figure 1.6: Intersection of Surfaces of Position Based on Range Measurements.

If the point being positioned is stationary, the two (or three) ranges do not need to be measured simultaneously. If the point is moving however, all ranges must be measured simultaneously (or over an interval of time during which the point has not moved by an amount greater than the uncertainty of the "fix"). Because the GPS constellation was designed to ensure at least four satellites are always visible anywhere on the earth, satellite positioning using simultaneously measured ranges is the basic positioning strategy for most navigation applications.

However, there is still the issue of how to account for *measurement biases*, as the technology used for making GPS range measurements does not give calibrated distance from the receiver to the satellite. Disturbing influences and errors in fact contaminate the range measurements to an unacceptable degree (in effect the radii of the spheres are incorrect -- Section 1.3.5), and hence the basic positioning principle is modified in several ways to satisfy the varying levels of accuracies required by different applications. (Chapter 2 describes some of the GPS enhancements that have been developed, or are under development.)

1.2 GPS Satellite Constellation and Signals

1.2.1 *GPS Constellation Design*

The operational Block II/IIA satellite constellation was to be fully deployed by the late 1980's. However, a number of factors, the main one being the Space Shuttle Challenger disaster (28 January 1986), has meant that the GPS system only became operational in the 1990's as far as most users were concerned. *Full Operational Capability* was declared on 17 July 1995 -- 24 Block II/IIA satellites operating satisfactorily. At an altitude of approximately 20,200km, a constellation of 24 functioning GPS satellites is sufficient to ensure that there will always be *at least four satellites visible*, at all unobstructed sites on the globe. Typically there are 6 to 10 satellites visible most of the day. The U.S. Department of Defense has undertaken to guarantee 24 satellite coverage 70% of the time, and 21 satellite coverage 98% of the time.

As the GPS satellites are in nearly circular orbits, at an altitude of approximately 20,200km above the earth (figure 1.7), this has a number of consequences:

- Their orbital period is approximately 11hrs 58mins, so that each satellite makes two revolutions in one sidereal day (the period taken for the earth to complete one rotation about its axis with respect to the stars).
- At the end of a sidereal day (23hrs 56mins in length) the satellites are again over the same position on earth.
- Reckoned in terms of a solar day (24hrs in length), the satellites are in the same position in the sky about four minutes earlier each day.
- The orbit groundtrack approximately repeats each day, except that there is a very small drift of the orbital plane to the west which is arrested by periodic manoeuvres.

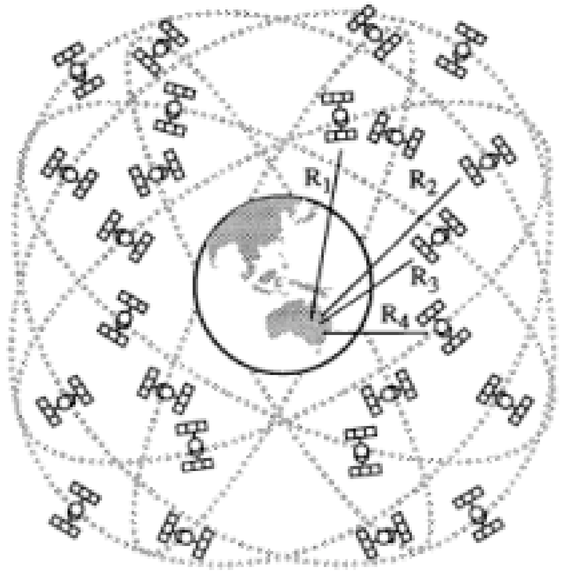

Figure 1.7: The GPS Constellation "Birdcage".

The following general remarks can be made with regard to satellite constellation design for navigation purposes:

- The higher a satellite, the longer it is visible above the horizon (the extreme case is the geostationary satellites).
- The higher a satellite, the better the coverage due to longer fly-over passes and extended visibility of the satellite across large areas of the earth.
- The higher a satellite, the less the rate-of-change of distance, and the lower the Doppler frequency of a transmitted signal.
- The greater the angle of inclination, the more northerly the track of the sub-satellite point across the surface of the earth.
- No satellite can be seen simultaneously from all locations on the earth.
- Depending on the positioning principles being employed, there may be a requirement for observations to be made to more than one satellite simultaneously from more than one ground station.

The Block II satellites are deployed in six orbital planes at 60^o intervals about the equator, with each containing four satellites. The satellites can be moved round their orbits if it becomes necessary to "cover" for a failed satellite. The orbital planes are inclined at an angle of 55^o to the equatorial plane (figure 1.3).

As the satellites are at an altitude of more than three times the earth's radius, a particular satellite may be above an observer's horizon for many hours, perhaps 6-7 hours or more in the one pass. At various times of the day, and at various locations on the surface of the earth, the number of satellites and the length of time they are above an observer's horizon will vary. Although at certain times of the day there may be up to 12 satellites visible simultaneously, there are nevertheless occasional periods of degraded satellite coverage (though naturally their frequency and duration will increase if satellites fail). "Degraded satellite coverage" is generally defined in terms of the magnitude of the Dilution of Precision (DOP) factor, a measure of the quality of satellite geometry. The highest the DOP value, the poorer the satellite geometry. For example, if all the visible satellites are located in the same part of the sky, the intersections of the SOPs will be very obtuse. (GPS DOPs are discussed further in Section 2.1.3.)

1.2.2 *GPS Signal Components*

The basis of the GPS signal are the two L-band carrier signals. These are generated by multiplying the fundamental frequency f_o (10.23MHz) by 154 and 120, yielding the two microwave L-band carrier waves L1 and L2 respectively (figure 1.2). The frequency of the two waves are: $f_{L1} = f_o$ x 154 = 1575.42 MHz, and $f_{L2} = f_o$ x 120 = 1227.6 MHz. These are radio frequency waves capable of transmission through the atmosphere over great distances, but which cannot penetrate solid objects. Note that all GPS satellites transmit carrier waves at the same two L-band frequencies (unlike the GLONASS system, where a different frequency is assigned to each satellite -- see Section 2.3.2). However, the L-band carrier waves themselves carry no information, and must be modified (or modulated) in some way. In the Global Positioning System the L-band carrier waves are modulated by two *ranging codes*, and the *navigation message*. The two distinct GPS ranging codes are:

- The **C/A code** (sometimes referred to as the "clear/access" or "coarse/acquisition" code), sometimes also referred to as the "S code".
- The **P code** (the "private" or "precise" code) was designed for use only by the military, and other authorised users.

The L1 carrier was designed to be modulated with both the P and C/A codes, whereas the L2 carrier would be modulated only with the P code. Under the policy of Anti-Spoofing (Section 1.2.3) the P code is encrypted through modulation by a further secret code (the "W code") to produce a new "Y code". Both carrier signals contain the navigation message.

The C/A and P (or Y) codes provide the means by which a GPS receiver can measure one-way ranges to the satellites. These codes have the characteristics of random noise, but are in fact binary codes generated by mathematical algorithms and are therefore referred to as "pseudo-random-noise" or PRN codes. Both the C/A and P code generating algorithms are known, and are based on a simple Tapped Feedback Shift Register scheme (see, for example [2,3]). One C/A code is assigned to each GPS satellite (the PRN code number is often used as the satellite I.D.). Each C/A code is a 1023 "chip" long binary sequence, generated at a rate of 1.023 million chips per second (that is, 1.023 MHz), thus the entire C/A code sequence repeats every millisecond.

The P code is a far more complex binary sequence of 0's and 1's, being some 267 days long with a chipping rate at the fundamental frequency f_o (10.23 MHz). The resolution of this code (length of the P code chip) is ten times the resolution of the C/A code. Instead of assigning each satellite a unique code, as is the case with the C/A code, the P code is allocated such that each satellite transmits a one week portion of the 267 day long PRN sequence, restarting the code sequence at the end of each week. Further details on how PRN codes are generated are given in, for example, [5,6].

To measure one-way range (from satellite to receiver), a knowledge of the ranging code(s) is required by the user's receiver. Knowing which PRN code is being transmitted by a satellite means that a receiver can generate a local replica of the same code sequence. These PRN codes possess a very important attribute: a given C/A (or P or Y) code will fully correlate with an exact replica of itself only when the two codes are aligned, and has a low degree of correlation with other alignments.

As the same P (or Y) code is synchronously modulated on both carrier waves, any difference in signal transit time of the same PRN sequence is due to the *retardation of the two L-band signals by a different amount* as they travel through the ionosphere. The effect of the ionosphere on signal propagation is essentially a function of signal frequency (Section 2.1.1), hence measurement on both frequencies is a very effective way of overcoming the ionospheric signal delay.

1.2.3 *The Civilian - Military Relationship*

Although GPS is a military navigation system the civilian sector represents an important (and rapidly growing) user group that has increasingly lobbied the U.S. Government in order to influence: (a) the direction of GPS system development; (b) official GPS policy concerning enhancement and control; and (c) the design of follow-on systems to GPS for the 21st century. Several policy decisions have already been made which impact on GPS performance. Some of these actions were agreed to during the early system design phase, while others were made only after much of the present system had been deployed:

- Two PRN ranging codes are implemented. The C/A code is intended for general, civilian use, while the P code was reserved for military and other authorised users. Because of the P code's higher measurement resolution it was expected that the accuracy of positioning using the P code would be much better than that possible using the C/A code. However, it was found that the performance of C/A code positioning was often no worse than that of P code positioning by a factor of two. (The latest C/A code tracking technology permits ranging quality almost as good as P code ranging and hence, all other things being equal, positioning accuracy close to that expected from the processing of P code ranges should be possible.)

- These two levels of positioning performance were designed into the GPS system from the very beginning. The positioning service based on using C/A code ranging data is referred to as the "Standard Positioning Service" (SPS), while the service based on P code ranging data is known as the "Precise Positioning Service" (PPS).

- As a result of the demonstration of a surprisingly good level of SPS accuracy, the policy of "Selective Availability" (SA) was endorsed in order to artificially widen the gap between the two levels of positioning ([7]). SA is an intentional degradation of the accuracy of GPS horizontal positioning to 100m and vertical positioning at the 160m (at the 95% confidence level) for SPS users. SA is implemented through an encryption of the navigation message whereby a part of the transmitted ephemeris and satellite clock data is falsified (the so-called "epsilon" effect) and the satellite clock is "dithered" (the so-called "delta" effect). SA therefore affects the precision of all measurements, code and carrier phase. SA does not affect PPS users who have user equipment able to decipher the correct ephemeris and clock error data.

- Any GPS hardware manufacturer is able to construct a P code ranging receiver (the P code PRN generation algorithm is published). The non-military market for P code receivers was always assumed to be very small, and one that could be controlled by the U.S. Government through the issuance of "export licences", etc. However, a significant demand by the land surveying market for dual-frequency phase tracking GPS receivers (code-correlating phase tracking receivers need a knowledge of the P code PRN in order to make a carrier phase measurement) led to an expansion in the production of P code capable positioning equipment.

- Under another policy known as "Anti-Spoofing" (AS), access is denied to the P code modulated on both L-band frequencies. AS was implemented on 31 January 1994, through the encryption of a secret "W code" onto the P code. The rationale behind this decision was that by keeping the military PRN code secret, an enemy of the U.S. could not jam the signal using a ground-based transmitter, nor "spoof" military GPS receivers by transmitting a false P code signal from a satellite. However, several GPS receiver manufacturers have developed proprietary techniques for making dual-frequency measurements even in the presence of AS.

- Dual-frequency observations will lead to more accurate positioning results than single frequency observations, because the ionospheric bias can be eliminated from the code range measurements. The fact that the C/A code is only modulated on the L1 carrier is therefore an intentional design decision to ensure that the SPS service cannot deliver the accuracy that the PPS service can, even with SA off.

It must be emphasised that the situation as far as the policies on SA and AS is under almost continuous review, particularly since the Presidential Decision Directive on the "U.S. National GPS Policy" was released in March 1996. The reader is referred to [8,9,10], which describe both the "official" policies and options on GPS, and report the debate and outcome of studies for alternative models of joint civilian-military GPS operation. The Vice President in early 1999 also announced “GPS Modernization Plans”, which include the transmission of a third frequency.

1.2.4 *Why is the GPS Signal is so Complicated?*

GPS was designed as a complex military navigation system, which would also be used by civilians, hence we can summarise the reasons why the signal is so complicated as:

(a) The requirement for GPS to be *multi-user system* is most easily satisfied if it designed as *a listen-only, or self-positioning system.* An unlimited number of users could be supported because no return signal was necessary, however, the measurements and positioning procedures must then be based on *one-way (satellite-to-receiver) signals.*

(b) The very important requirement for *real-time positioning* influenced a number of design criteria related to the satellite signals:
- Simultaneous measurements from many satellites, but all signals are at the same frequency (though Doppler-shifted) -- *need to identify signals by use of different codes*
- Unambiguous range measurements were required -- *need to determine signal delay*
- Satellite positions are needed -- *broadcast ephemerides*

Hence the PRN codes were used to distinguish the different satellite signals. The measurement of a "range" to a particular satellite required that a comparison be made of the PRN code generated within the receiver (corresponding to the satellite in question) to the PRN code sequence contained within the incoming satellite signal (see Section 1.3.2).

(c) The requirement for GPS to support *high accuracy positioning* also had implications for satellite signal design:
- High frequency modulation -- *P code at 10 MHz*
- Dual-frequency signal -- *permits ionospheric delay estimation*
- Microwave carrier frequency -- *1.2 to 1.6 GHz*

(d) The *anti-jamming* requirement could be partly met through the use of *spread-spectrum codes*. This also fulfilled the requirement for a low power signal that could still be detected even though it was below the ambient signal noise level. By ensuring that some of the PRN codes are secret (known only to military and a few authorised users), jamming or "spoofing" of signals was made much more difficult. (However, it is possible to jam a GPS receiver if the jamming source is within several metres of the receiver.)

(e) As the system was intended as a "dual-use" technology, satisfying both *military and civilian users*, this had several implications the most important of which was that the civilians were not to have the same level of accuracy as the military users:
- The use of two PRN codes -- *P and C/A code*
- Restriction on the dual-frequency signal for civilians -- *C/A code only on the L1 frequency*
- Microwave carrier frequency -- *1.2 to 1.6 GHz*

1.2.5 *GPS Satellite Ephemerides*

How are the GPS satellite ephemerides computed? As the forces of gravitational and non-gravitational origin perturb the motion of the GPS satellites, the coordinates of the satellites in relation to the WGS84 reference system (Section 3.1) must be continually determined through the analysis of tracking data. In the case of the GPS broadcast ephemeris, this procedure is a three-step process ([6]):

- An off-line orbit determination is performed through the analysis of tracking to generate a "reference orbit". This is an initial estimate of the satellite trajectories computed from about one week's of tracking from the five Control Segment monitor stations.
- An on-line daily updating of the reference orbit using a Kalman filter as new data are added.

This provides the current estimates of the satellite orbit which is used to predict the future orbit.

- The ephemeris is estimated for 1 to 14 days into the future. To obtain the necessary broadcast information, curve fits are made to 4 to 6 hour portions of the extrapolated ephemeris, and hourly orbit parameters determined.

Note that these parameters are not true Keplerian elements as they only describe the ephemeris over the interval of applicability and not for the whole orbit. (Although only intended for use during the transmission period, they do, however, adequately describe the orbit over intervals of 1.5 to 5 or more hours, with a graceful degradation in accuracy.). The user can derive the earth-centred, earth-fixed WGS84 Cartesian coordinates of the GPS satellite from the broadcast orbital parameters, using an algorithm described in [11], and implemented in every GPS receiver.

1.3 GPS MEASUREMENTS

1.3.1 *The Transmitted Signal*

The signal that actually leaves a GPS satellite antenna is a combination of the three components: carrier wave, ranging codes and navigation message. The generation of the signal to be transmitted is carried out in a number of steps, and relies on the fact that all the components are derived by multiplying or dividing the fundamental frequency (figure 1.2). There are two distinct procedures for the combination of signal components: (a) BPSK modulation of binary sequences onto the carrier waves, and (b) modulo-2 addition of binary sequences.

Biphase Shift Key Modulation (BPSK) is the technique used to added a binary signal to a sine wave carrier. This amounts to causing a 180° phase shift in the carrier each time the binary sequence undergoes a transition from "0" to "1", or "1" to "0". The P code plus navigation message is modulated on both the L1 and L2 carriers, while the C/A code plus navigation message is only modulated on the L1 carrier.

An example of binary-to-binary modification of codes is the modulo-2 addition of the navigation message data to the C/A code. Because of the difference in frequency between the C/A code and the message stream this has the effect of inverting 20 C/A code binary "states" whenever the data bit of the navigation message is equal to "1". Conversely, when the data bit is "0" the C/A code sequence remains unaffected. The same satellite message is also modulated onto the P code sequence using this modulo-2 addition procedure.

There are two main types of measurements that can be made on the GPS signals ([13]):

- range observations based on the PRN codes, sometimes referred to as "code range" or "code phase", and
- carrier phase observations, which are more precise range-type measurements, but which have a much higher degree of "ambiguity" than the code ranges (see Section 2.2.2).

1.3.2 *The GPS Range Measurements*

The PRN codes are accurate time marks that permit the receiver's navigation computer to determine the time of transmission for any portion of the satellite signal. Before considering how ranging is carried out we need to describe, in general terms, how the incoming satellite signal is processed within a GPS receiver. See references [2,3] for more details than are presented here!

Neglecting any discussion on extraneous noise (in relation to which PRN codes have special properties), and assuming that the only satellite signals received at the antenna originate from one GPS satellite, the following simplistic procedure is carried out within a receiver "channel" (Section 1.4.2). The L1 carrier modulated by the C/A code is converted to a signal of lower frequency and then mixed with a locally generated matching C/A code. The local C/A code is generated on a different time scale (due to non-synchronization of receiver clock to GPS Time and the travel time of the signal from the satellite to antenna) to that of the incoming C/A code. As soon as the incoming signal and the receiver C/A code sequences are aligned within the receiver the ones and zeros of the two codes cancel, leaving the received carrier modulated only by the binary navigation message.

The extraction of the code range, or more precisely the determination of the amount by which the receiver generated PRN code must be shifted to align it with the incoming signal, is carried out with the aid of a PRN code correlator in some form "delay-lock loop" scheme (see, for example [13]). *How accurately is this carried out?* The C/A code has a sequence rate of 1.023 Mbps, corresponding to a resolution of about 300m (speed of light divided by the frequency). The P (or Y) code, on the other hand, has a so-called "chip" rate of 10.23 Mbps, and hence an effective resolution of about 30m. As a *rule-of-thumb*, alignment of the incoming and receiver generated codes is generally possible to within about 1-2% of the chipping rate, hence the measurement precision of C/A code ranging is of the order of 3-5m, and for P code ranging it is of the order of 0.3-0.5m. *However, under the policy of "Anti-Spoofing", the Y code is encrypted and is therefore not available to civilian applications.*

It should be noted, however, that modern "narrow correlator" C/A code measurement technology as implemented in several "top-end" GPS receivers has demonstrated ten times better correlation performance for the C/A code correlation than that quoted above.

1.3.3 *The GPS Carrier Phase Measurements*

The wavelengths of the carrier waves are very short (approximately 19cm for L1 and 24cm for L2) compared to the C/A and P code wavelengths. Assuming a measurement resolution of 1-2% of the wavelength, this means that carrier phase can be measured to millimetre precision compared with a few metres for C/A code measurements. Unfortunately a phase measurement is *ambiguous* as it cannot discriminate one (L1 or L2) wavelength from another. In other words, time of transmission information for the L-band signal cannot be imprinted onto the carrier wave as is done using PRN codes (this would be possible only if the PRN code frequency was the same as the carrier wave, rather than 154 or 120 times lower in the case of the P code, and 1540 or 1200 times lower for the C/A code). The basic phase measurement is therefore an angle in the range 0^o to 360^o. *It is nevertheless the basis for high precision GPS positioning (Section 2.2).*

1.3.4 *Ranging Using PRN Codes*

Consider for a moment a perfect system, where all satellite clocks are synchronized to the same time system: the GPS Time (GPST). Furthermore, the ground receiver's clock also maintains the same synchronization, and none of the clocks drift from this GPST scale. Now suppose the satellite starts transmitting its L1 carrier (modulated with the combined C/A PRN code and navigation data). At the same instant, the receiver begins generating the C/A PRN code corresponding to that particular satellite (see figure 1.8). Under these circumstances the satellite and receiver generated C/A codes would be output in unison. When the satellite signal is received, however, it will be lagging the receiver generated code due to the *signal transit time*. Multiplying the time offset required to align the two codes (in effect determining the signal transit time) by the speed of light yields the satellite to receiver distance.

Measuring ranges simultaneously in this fashion to three satellites would fix one's position at the intersection of three spheres of known radii (the satellite ranges), centred at each satellite whose coordinates can be calculated from the navigation message. In reality the situation is more complex:

- GPS receivers are equipped with crystal clocks that do not keep the same time as the more stable satellite clocks (the satellite clocks can be nearly synchronised to GPST using the clock correction model transmitted in the navigation message). *Consequently each range is contaminated by the receiver clock error*. This range quantity is therefore referred to as **pseudo-range**, and in order for the user to derive position from pseudo-range data, the receiver equipment is required to track (a minimum of) four satellites, and solve for four unknown quantities: the three-dimensional position components and the receiver-clock offset (from GPST) -- see Section 1.3.5. *This is the basis of GPS real-time navigation*, and why GPS could be considered an example of a time-difference-of-arrival system.

- There is in fact a 300 km "ambiguity" in the C/A code pseudo-range measurements (300 km is the approximate length of the C/A code sequence). That is, all measured "distances" appear to have a range of 0 to 300 km. This ambiguity is resolved in a number of ways, but the easiest to assume that if the approximate receiver position is known to within say 100 km, the "missing" component of the distance can be determined, and hence the raw pseudo-range measurement can be corrected for this ambiguity to obtain the full satellite-receiver distance.

- Ranging (and hence receiver position determination) can be carried out using the C/A code or the P code. P code ranging can be done on the combination of the two frequencies, hence eliminating the bias due to ionospheric refraction. Furthermore, the C/A code is "coarser", and hence the C/A derived ranges are subject to greater measurement "noise". The absence of a C/A code on L2 is intentional, as one of the accuracy limitations of the GPS system for the general class of civilian users.

- As previously mentioned, this differentiation between ranging codes, and the formulation of policies for their use (in peacetime and in times of global emergencies), is responsible for the provision of two GPS services: The Precise Positioning Service based on P or Y code (dual-frequency) ranging, and the Standard Positioning Service based on single frequency C/A code ranging (Section 1.2.3).

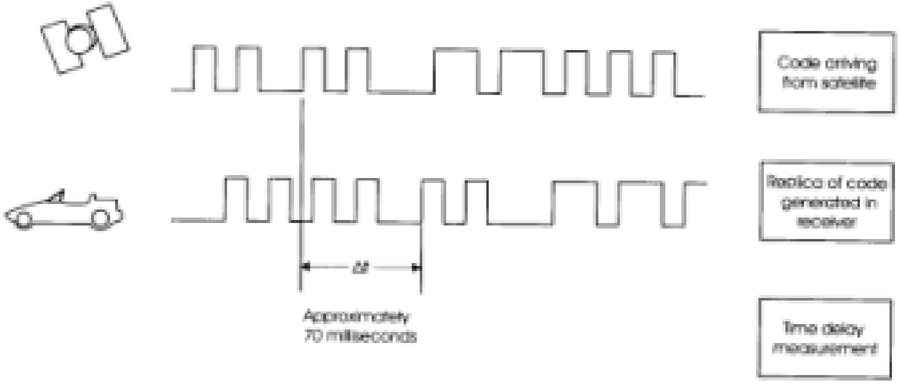
Code arriving from satellite
Replica of code generated in receiver
Δt
Approximately 70 milliseconds
Time delay measurement

Figure 1.8: One-Way Ranging Using PRN Codes.

1.3.5 *An Observation Model of the Pseudo-Range*

The observation equation for a receiver-clock-biased range is (see, for example [14]):

$$p = r + e_{rc}(t_r).c \qquad (1.1)$$

where c is the velocity of electromagnetic radiation in a vacuum (or simply the "velocity of light"), e_{rc} is the receiver clock error caused by the receiver oscillator (assume satellite clock time is "true" time) at time of reception t_r, p is the measured range and r is the true "geometric" range. Each observation made by the receiver can be parameterised as follows:

$$(x^s - x)^2 + (y^s - y)^2 + (z^s - z)^2 = (p - e_{rc}.c)^2 \qquad (1.2)$$

where x^s, y^s, z^s is the coordinate of the satellite and x, y, z is the coordinate of the receiver. Note that the time argument has been discarded (satellite coordinates are time-dependent, and so are the receiver coordinates if the receiver is moving).

As the 3-D coordinate of the satellite is known, then each measurement p contains four parameters which may be considered unknown: the 3-D coordinate of the receiver (x, y, z) and the receiver clock error (e_{rc}). By making four measurements, to four different satellites, the following system of equations is obtained:

$$\begin{aligned}
(x^{s1} - x)^2 + (y^{s1} - y)^2 + (z^{s1} - z)^2 &= (p^1 - e_{rc}.c)^2 \\
(x^{s2} - x)^2 + (y^{s2} - y)^2 + (z^{s2} - z)^2 &= (p^2 - e_{rc}.c)^2 \\
(x^{s3} - x)^2 + (y^{s3} - y)^2 + (z^{s3} - z)^2 &= (p^3 - e_{rc}.c)^2 \\
(x^{s4} - x)^2 + (y^{s4} - y)^2 + (z^{s4} - z)^2 &= (p^4 - e_{rc}.c)^2
\end{aligned} \qquad (1.3)$$

which has a unique solution (see, for example [15]). If more than four measurements are made the method of Least Squares, or other estimation techniques, can be used to determine the optimum solution. *Does the receiver clock error have to be estimated at each epoch?* That depends upon such factors as:

- How well the clock error is estimated.
- How often the position solution is carried out.
- The stability of the clock.

Under the assumptions that (a) there is a range measurement uncertainty of about 1 metre attributable to receiver "noise", and (b) the receiver is equipped with a quartz crystal clock (stability of 0.1 nanoseconds/second), then the uncertainty of the clock time after 30 seconds is as great as the range measurement error. Hence, once the clock error were determined, it would have to be independently re-estimated at least every 30 seconds, otherwise it would dominate the range error budget. This is one of the reasons why this procedure could not be used if pseudo-range measurements to different satellites were not made "simultaneously" (here, within 30 seconds of each other), so that the clock error could be assumed to be a constant.

This is the conventional strategy used for pseudo-range-based GPS positioning: the receiver clock error is simply treated as an additional unknown and the *navigation problem can be considered as 4-D estimation*. However, the GPS satellite clock error is assumed to be a known quantity, and parameters defining this clock error are transmitted in the navigation message. The ionospheric delay is also modelled using transmitted parameters. An estimate of the tropospheric delay may be obtained from a simple tropospheric model such as Hopfield. All other biases are assumed to be insignificant compared to the measurement noise (that is, their impact on the position solution quality is negligible -- see Section 2.1.1 for a discussion).

1.3.6 *Coping with the Satellite Clock Bias*

As the satellite clock error is the largest source of GPS measurement bias it deserves closer study. Under the assumption that the satellite clock error is an *unknown* quantity, the observation equation for such a satellite-biased range:

$$p = r + e^{sc}(T^s).c \qquad (1.4)$$

e^{sc} is the satellite clock error caused by the satellite oscillator not being synchronized to "true" time (GPST). p is the measured range, r is the true range and T^s is the time of transmission. Each observation made by the receiver can be parameterised as in equation (1.1), except for the replacement of the term e^{sc} for e_{rc}:

$$(x^s - x)^2 + (y^s - y)^2 + (z^s - z)^2 = (p - e^{sc}.c)^2 \qquad (1.5)$$

The 3-D coordinate of the satellite signal transmitter (x^s, y^s, z^s) is known, hence in the case of three range observations there are six unknowns in the system: the 3-D coordinate of the receiver (x_{r1}, y_{r1}, z_{r1}) and the three satellite clock error terms (e^{sc1}, e^{sc2}, e^{sc3}). Hence, at first glance, six satellite-biased range observations are required to solve this positioning problem. It is not, however, possible simply to make observations to more satellites as each new observation introduces a new satellite clock parameter. *There are two options for overcoming this dilemma.*

It is possible to take advantage of the fact that all observations made to a particular satellite are biased by the same amount (if made at the same time, or close enough together so that the satellite clock error can be assumed to have not changed by a significant amount). If three range observations are made from another station, whose coordinate is known (x_{r2}, y_{r2}, z_{r2}), then it is possible to obtain a system of six equations in six unknowns:

$$\begin{aligned}
(x^{s1} - x_{r1})^2 + (y^{s1} - y_{r1})^2 + (z^{s1} - z_{r1})^2 &= (p_{r1}{}^{s1} - e^{sc1}.c)^2 \\
(x^{s2} - x_{r1})^2 + (y^{s2} - y_{r1})^2 + (z^{s2} - z_{r1})^2 &= (p_{r1}{}^{s2} - e^{sc2}.c)^2 \\
(x^{s3} - x_{r1})^2 + (y^{s3} - y_{r1})^2 + (z^{s3} - z_{r1})^2 &= (p_{r1}{}^{s3} - e^{sc3}.c)^2 \\
(x^{s1} - x_{r2})^2 + (y^{s1} - y_{r2})^2 + (z^{s1} - z_{r2})^2 &= (p_{r2}{}^{s1} - e^{sc1}.c)^2 \\
(x^{s2} - x_{r2})^2 + (y^{s2} - y_{r2})^2 + (z^{s2} - z_{r2})^2 &= (p_{r2}{}^{s2} - e^{sc2}.c)^2 \\
(x^{s3} - x_{r2})^2 + (y^{s3} - y_{r2})^2 + (z^{s3} - z_{r2})^2 &= (p_{r2}{}^{s3} - e^{sc3}.c)^2
\end{aligned} \qquad (1.6)$$

for which a unique solution can be obtained. In a conceptual sense this is the basis of *differential GPS positioning*, although in reality there are several possible implementations (Section 2.1.2 and 2.2).

The other strategy for accounting for satellite clock error is for the GPS operators to periodically determine the clock error. As the satellite clocks have significantly better long-term drift characteristics than the receiver clocks, a suitable clock error model could be a time polynomial:

$$e^{sc} = a_0 + a_1\ (t - t_{oc}) + a_2\ (t - t_{oc})^2 \qquad (1.7)$$

where: a_0 is the clock bias term
a_1 is the clock drift term
a_2 is the clock drift-rate
t is satellite clock time
t_{oc} is some reference epoch for the definition of the coefficients

What is actually available to users via the navigation message is a *prediction* of the satellite clock error behaviour for some time into the future (24 hours or more). Such deterministic models of satellite clock error are accurate to about 20 nanoseconds, or about six metres in equivalent range, depending upon the time since last navigation message update. Hence the measured pseudo-ranges may be corrected for satellite clock error, and then the observation model in equation (1.1) can be used.

Selective Availability complicates matters because it is a further artificial "dithering" of the satellite clock causing an additional several tens of metres error in the pseudo-range measurement (Section 1.2.3).

1.4 GPS Instrumentation

The following components of a generic GPS receiver can be identified (figure 1.9):

- *Antenna* and *Preamplifier*: Antennas used for GPS receivers have broadbeam characteristics, thus they do not have to be pointed to the signal source like satellite TV receiving dishes. The antennas are compact and a variety of designs are possible. There is a trend to integrating the antenna assembly with the receiver electronics.

- *Radio Frequency Section* and *Computer Processor*: The RF section contains the signal processing electronics. Different receiver types use somewhat different techniques to process the signal. There is a powerful processor onboard not only to carry out computations such as extracting the ephemerides and determining the elevation/azimuth of the satellites, etc., but also to control the tracking and measurement function within modern digital circuits, and in some cases to carry out digital signal processing.

- *Control Unit Interface*: The control unit enables the operator to interact with the microprocessor. Its size and type varies greatly for different receivers, ranging from a handheld unit to soft keys surrounding an LCD screen fixed to the receiver "box".

- *Recording Device*: in the case of GPS receivers intended for specialised uses such as the surveying the measured data must be stored in some way for later data processing. In the case of ITS applications such as the logging of vehicle movement, only the GPS-derived coordinates and velocity may be recorded. A variety of storage devices were utilised in the past, including cassette and tape recorders, floppy disks and computer tapes, etc., but these days almost all receivers utilise solid state (RAM) memory or removable memory "cards".

- *Power Supply*: Transportable GPS receivers these days need low voltage DC power. The trend towards more energy efficient instrumentation is a strong one and most GPS receivers operate from a number of power sources, including internal NiCad or Lithium batteries, external batteries such as wet cell car batteries, or from mains power.

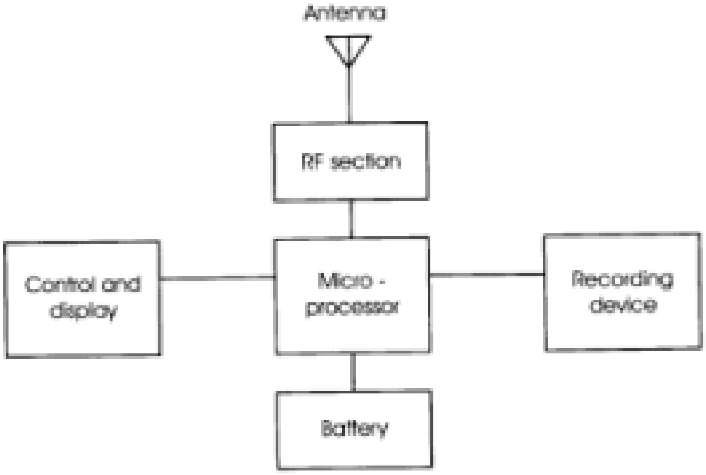

Figure 1.9: The Generic GPS Receiver.

The antenna and RF technology components are briefly discussed below. For further details the reader is referred to [6,16].

1.4.1 *Antennas*

The task of the antenna is to convert the energy of the arriving electromagnetic waves into an electric current that can be processed by the receiver electronics. There are a number of special considerations as far as antenna design is concerned:

- the antenna must be able to pick and discriminate very weak signals,
- the antenna may need to operate at just the L1 frequency, or at both the L1 and L2 frequencies,
- as the signals are right-hand circularly polarised, the GPS antenna must also be right-hand circularly polarised,
- antenna gain pattern that enhances the ability of the RF section to discriminate against multipath signals (such as left-hand circularly polarised signals),
- a stable electrical centre,
- low cost, and
- reliable.

There have been several types of GPS antennas used:

- monopole or dipole configurations,

- quadrifilar helices,
- spiral helices,
- microstrip,
- choke ring, and other multipath resistant designs

In short, the antennas are required to be rugged, simple in construction, have stable phase centres, be resistant to multipath, have good gain and pattern coverage characteristics. The microstrip antenna is almost universally used for navigation applications.

1.4.2 *Signal Processing Principles for Code-Correlating Receivers*

The RF section of a GPS receiver converts the incoming (preamplified) signal to a signal of a more manageable frequency. This intermediate frequency is obtained by mixing the incoming signal with a pure sinusoidal signal generated by the local oscillator (the quartz "clock"). The frequency of this beat frequency is the difference between the original (Doppler-shifted) received carrier frequency and the local oscillator. The intermediate or beat frequency is then processed by the signal tracking circuitry.

There are several classes of signal processing techniques that can be employed to make observations, as well as several proprietary implementations of tracking technologies. This is particularly the case for instruments which track carrier phase. It is beyond the scope of this chapter to discuss in detail the electronic circuitry and the reader is referred to [2,3]. As our interest is the pseudo-range measuring instruments, we need only consider the issues of the "code-correlating" signal processing technique, and some brief details of the processing *architecture* within the receiver, and in particular the form of the tracking "channel(s)".

Code-correlating receivers employ tracking loops to extract the useful measurements from the beat signal. A typical GPS receiver contains two types of tracking loops:

- the *delay-lock*, or code-tracking, loop, and
- the *phase-lock*, or carrier tracking, loop.

The delay-lock loop is used to make the alignment of the PRN code sequence (C/A or P code) that is contained in the signal from a satellite with an identical PRN code generated within the receiver. A correlator in the delay-lock loop continuously cross-correlates the two code streams, time shifting the receiver generated stream until alignment is obtained. The time shift is then converted to a pseudo-range measurement. Once the code-tracking loop is aligned, the PRN code can be stripped from the satellite signal. The resulting signal then passes to the phase-lock loop where the satellite message is extracted. Once the local oscillator is locked onto the satellite signal it will continue to follow the variations in the phase of the carrier as the satellite-receiver distance changes.

A GPS receiver may be described as *continuous* or *switching* depending on the type of channel(s) it has. A continuous-tracking receiver has dedicated hardware channels, and each channel tracks a single satellite, maintaining continuous code and/or phase lock on the signal. Each channel is controlled and sampled by the receiver's microprocessor with input/output operations being performed fast enough so that tracking is not disturbed. Continuous-tracking receivers may enjoy a signal-to-noise advantage over switching receivers in that the satellite signal is continuously available and may be more frequently sampled. A further advantage is a potential redundancy capability. Should one of the hardware channels fail, it may still be possible to obtain sufficient data to determine a position. One disadvantage of a multi-channel receiver is that the differences in signal path delay in the channels, the so-called inter-channel biases, must be well calibrated. This is the most common architecture used in GPS receivers today..

A *switching receiver* has hardware channels which sequentially sample the incoming signal from more than one satellite. There may be a single channel for all satellites, or each channel may track, say, two satellites. Code and/or carrier phase tracking for the individual signals is controlled by software (or, more usually, the "firmware") within the microprocessor. As a result, greater demands are placed on the microprocessor in a switching receiver and its programming is necessarily more complex -- in effect hardware complexity is exchanged for software complexity. If the cost of hardware components is a significant factor in determining the selling price of a receiver then in principle, the cost of a switching receiver should be cheaper than a continuous receiver (this is certainly the case for the low cost GPS navigation units).

There are three basic kinds of switching receivers, distinguished by the time required to sequence through the signals tracked by a particular channel. A *multiplexing channel* is one for which the sequencing time to sample all satellites assigned to the channel is equal to 20 milliseconds, the period of one bit in the satellite navigation message. The sampling can be arranged so that no message bit boundary is spanned by any tracking interval. In this way, the messages from all satellites phase tracked by the channel can be read simultaneously. A multiplexing channel can be used to obtain both L1 and L2 data by alternating between the frequencies every 20 milliseconds. If a channel switches between signals at a rate which is asynchronous with the message bit rate, the channel is referred to as a *sequencing channel*. A fast sequencing channel is one which takes the order of a second or so to sequence through the signals. A slow sequencing channel may take several seconds or even minutes. A single sequencing channel would lose bits in a particular satellite message during those intervals spent sampling the signals from other satellites. Consequently, sequencing receivers may have an extra hardware channel just for message decoding. Alternatively, the navigation message must be decoded before the receiver starts the tracking cycle for real-time positioning (note that the message only changes once an hour).

The configuration of channels is selected so that, for example in the case of a navigation receiver, a minimum of four satellites can be tracked at the same time. This may call for a single switching channel, or two switching (between two satellite signals) channels, or even five channels (one used for calibrating the other four channels, and decoding the navigation message).

1.4.3 *Trends In GPS Instrumentation*

It is impossible to precisely predict trends in GPS instrumentation. The task is no easier if we focus our attention only to a specific market segment, such as positioning hardware for ITS applications. Nevertheless, speculation based on R&D activities being undertaken at present gives us some clues:

- The third generation of GPS receivers presently available already exhibit significant gains in miniaturisation, reduction in power consumption, and portability, over the earlier models. *We can expect this trend to continue.*

- The choice confronting GPS manufacturers is whether to maintain a broad-based development program, and market a product that is versatile enough to satisfy many applications (including the provision of interchangeable components such as antennas, RF units, memory, etc., to accomplish this), or to focus on specialist applications (military, navigation, differential navigation, surveying, kinematic, etc.). To date, most manufacturers are split between those with products for the high precision surveying market, and those focussing on the low cost, high volume navigation market. *Only a few address all markets.*

- There have been many predictions of low cost GPS receivers. However there is an enormous range of prices from the "top-of-the-line" surveying receivers to the "bare-bones" GPS "engines" implemented on one or more chips. GPS boardsets with basic navigation functionality are available at less than one hundred U.S. dollars each (when purchased in high volumes).

- There is a trend towards product differentiation, with many different configurations of tracking channels, data recording options, and, in particular, software options. Although some of these trends may be due to manufacturers wanting to give their product "an edge" in the marketplace, it is equally valid to suggest that this is in response to the different demands (some very specialist) of the market. *However, even in the case of complex ITS systems, it is possible to identify the "basic GPS component", and distinguish it from the "add-ons".*

- An exciting marriage is possible between satellite navigation and satellite communication, combining real-time positioning with instantaneous transmission of position. GPS receivers may be fitted to many different platforms (some unmanned, for example rail rolling stock and ship cargo containers), and their locations remotely monitored at a central site via a satcom link. In addition, for land navigation these could include electronic map displays to aid the driver. *What is clear is that for many applications the navigation data provided by a GPS receiver will* *merely be the first link in a complex information system.*

- Market surveys suggest that the greatest growth is expected to be in the commercial and consumer markets. Consumer applications such as ITS, integration of GPS and cellular phones, and portable GPS for outdoor recreation and similar activities will account for more than 60% of the market by the year 2000..

- New tracking electronics, such as "narrow-correlator" technology, would improve pseudo-range measurement precisions, as would a combination of phase and pseudo-range measurement. Both are, however, still relatively expensive technologies and are not found on "bare-bones" GPS boardsets. Further electronic tracking refinements will lead to more multipath-resistant receivers.

References

[1] PARKINSON, B.W., 1994. GPS eyewitness: the early years. **GPS World, 5(9)**, 32-45.

[2] KAPLAN, E. (ed.), 1996. **Understanding GPS: Principles & Applications**. Artech House Publishers, Boston London, 554pp.

[3] SPILKER Jr., J.J. & PARKINSON, B.W. (eds.), 1995. **Global Positioning Systems: Theory & Applications**. American Institute of Aeronautics & Astronautics (AIAA), 1995, Vol.1(694pp), Vol.2(601pp).

[4] SEEBER, G., 1993. **Satellite Geodesy: Foundations, Methods & Applications**. Walter de Gruyter, Berlin New York, 531pp.

[5] SPILKER Jr., J.J., 1980. GPS signal structure and performance characteristics. In: Global Positioning System, papers published in **Navigation,** reprinted by the (U.S.) Inst. of Navigation, Vol.1, 29-54.

[6] WELLS, D.E., BECK, N., DELIKARAOGLOU, D., KLEUSBERG, A., KRAKIWSKY, E.J., LACHAPELKLE, G., LANGLEY, R.B., NAKIBOGLU, M., SCHWARZ, K.P., TRANQUILLA, J.M. & VANICEK, P., 1987. **Guide to GPS Positioning**. 2nd. ed. Canadian GPS Associates, Fredericton, New Brunswick, Canada, 600pp.

[7] GEORGIADOU, Y. & DOUCET, K.D., 1990. The issue of Selective Availability. **GPS World, 1(5)**, 53-56.

[8] N.R.C., 1995. The Global Positioning System: a shared national asset. Rept. by National Research Council, National Academy Press, 264pp.

[9] N.A.P.A, 1995. The Global Positioning System: charting the future. Rept. by National Academy of Public Administration & the National Research Council, National Academy Press, 332pp.

[10] GIBBONS, G., 1996. A national GPS policy. **GPS World, 7(5)**, 48-50.

[11] VAN DIERENDONCK, A.J., RUSSELL, S.S., KOPTIZKE, E.R. & BIRNBAUM, M., 1980. The GPS navigation message. In: Global Positioning System, papers published in **Navigation,** reprinted by the (U.S.) Inst. of Navigation, Vol.1, 55-73.

[12] LANGLEY, R.B., 1991b. The orbits of GPS satellites. **GPS World, 2(3)**, 50-53.

[13] LANGLEY, R.B., 1993. The GPS observables. **GPS World, 4(4)**, 52-59.

[14] LANGLEY, R.B., 1991d. Time, clocks, and GPS. **GPS World, 2(10)**, 38-42.

[15] LANGLEY, R.B., 1991c. The mathematics of GPS. **GPS World, 2(7)**, 45-50.

[16] LANGLEY, R.B., 1991a. The GPS receiver - An introduction. **GPS World, 2(1)**, 50-53.

Footnotes:

[1] Transit is an early satellite-based navigation system based on Doppler measurements (see [4]).

Chapter 2
GPS Enhancements

There are several aspects to GPS *performance*:

- **Accuracy**, at a certain level when the appropriate hardware, software and operational procedures are used.
- **Availability**, the extent to which the system is available to all users, anywhere on the earth, and at any time of the day.
- **Continuity**, the degree to which a certain level of accuracy is maintained on a continuous basis.
- **Reliability** of the system and results, often evidenced by a certain "repeatability" of the positioning accuracy.
- **Integrity**, the capacity to monitor performance and warn users when accuracy falls below a certain level.
- **Cost**, of hardware and software as well as indirect operational costs.
- **Competitive technologies**, do they exist? what do they offer in terms of superior accuracy, etc.?

However, the most important performance measure for most users is *accuracy*. Hence we will need to examine the main factors influencing GPS positioning accuracy:

- Measurement errors and biases, as they affect observations.
- Absolute or differential positioning mode.
- Satellite-receiver geometry.
- Processing algorithms, operational mode and other enhancements.

2.1 Factors Influencing GPS Accuracy

All GPS measurements, be they pseudo-range, carrier phase or Doppler frequency, are affected by **biases** and **errors** (figure 2.1). Their combined magnitudes will affect the accuracy of the positioning results. Errors may be considered synonymous to "internal instrument noise" or *random errors*, as well as any unmodelled or residual biases. Biases may therefore be defined as being those errors of the measurements that cause *true ranges* to be different from *measured ranges* by a "systematic amount", such as, for example, all distances being measured too short, or too long, over a significant period of time.

Biases must somehow accounted for in the measurement model used for data processing if high accuracy is sought. There are several sources of bias with varying characteristics of magnitude, periodicity, satellite or receiver dependency, etc. Biases may have physical bases, such as the atmosphere effects on signal propagation, but may also enter at the data processing stage through imperfect knowledge of constants, for example any "fixed" parameters such as the satellite orbit, station coordinates, velocity of light, etc. *Residual biases* may therefore arise from incorrect or incomplete observation modelling and hence it is useful to assemble under the heading of "errors" all unaccounted for random measurement process effects, as well as any unmodelled or residual biases that remain.

A useful way for us to define biases is as errors which are *correlated* in space and/or time, best illustrated by an examination of the basic GPS pseudo-range measurement model ([1]):

$$p_i^s = \rho_i^s + \varepsilon_{rc(i)} + \varepsilon^{sc(s)} + \varepsilon^{orbit(i,s)} + \varepsilon^{atmos(i,s)} + \nu_i^s \qquad (2.1)$$

The subscript in brackets refers to the GPS receiver "i", the superscript in brackets refers to the satellite "s", and the time argument has been omitted. p is the measured pseudo-range, ρ is the true geometric range from one receiver to one satellite, ε_{rc} is the receiver clock error, ε^{sc} is the satellite clock error, $\varepsilon^{orbit(i,s)}$ is the 3-D satellite orbit error mapped into the 1-D range, $\varepsilon^{atmos(i,s)}$ is the atmospheric refraction error, and ν_i^s are any other errors not explicitly accounted for in the above observation model. All error and bias terms in equation (2.1) are expressed in units of metres. Although the time argument has been dropped for the sake of clarity, all quantities in equation (2.1) vary with time, and hence the equation represents a "snapshot" of a GPS pseudo-range measurement at a single *epoch*, or instant of time.

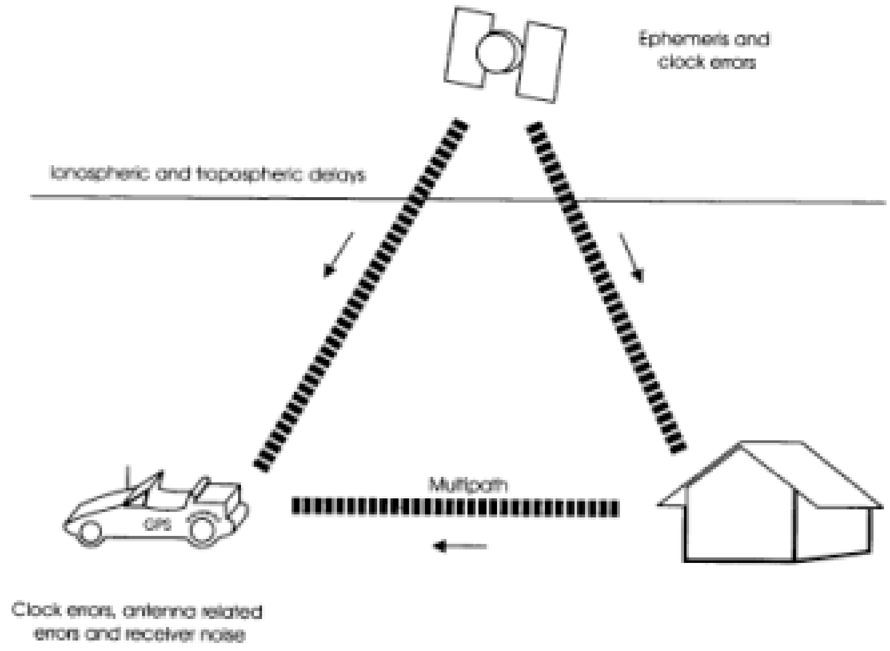
Ephemeris and clock errors
Ionospheric and tropospheric delays
GPS
Multipath
Clock errors, antenna related errors and receiver noise

Figure 2.1: Some GPS Biases and Errors Affecting Observations.

The spatial correlation of errors is obvious when an observation from another GPS receiver "k" to the same satellite, *at the same epoch*, is modelled as in equation (2.1):

$$p_k^s = \rho_k^s + \varepsilon_{rc(k)} + \varepsilon^{sc(s)} + \varepsilon^{orbit(k,s)} + \varepsilon^{atmos(k,s)} + \nu_k^s \qquad (2.2)$$

The following comments can be made with regard to equations (2.1) and (2.2):

- The receiver clock error $\varepsilon_{rc(i)}$ (length units) systematically affects all measurements made at receiver "i" to all satellites by exactly the same amount. It is completely unrelated to the value of $\varepsilon_{rc(k)}$ at the same measurement epoch. $\varepsilon_{rc(i)}$ across different epochs may also be considered uncorrelated.
- The satellite clock error $\varepsilon^{sc(s)}$ (length units) systematically affects all measurements made to satellite "s" by any GPS receiver making a measurement at the same time(-of-transmission). Hence $\varepsilon^{sc(s)}$ is spatially correlated (across different receivers) at an epoch, and this property can be exploited to overcome the effect of this bias. *One component of Selective Availability (Section 1.2.3), the so-called "clock dither", can be modelled in terms of the satellite clock bias*.
- The satellite orbit error $\varepsilon^{orbit(i,s)}$ (length units), although explicitly associated with satellite "s", is mapped differently as a range error in the case of measurements made by receiver "i" compared to those made by receiver "k" (hence the use of the double index denoting both the receiver and satellite involved in the measurement). However, if the two receivers are close together, the residual bias ($\varepsilon^{orbit(i,s)} - \varepsilon^{orbit(k,s)}$) will be very small. $\varepsilon^{orbit(i,s)}$ across different epochs will change comparatively slowly, hence there is a high temporal correlation.
- The atmospheric refraction error $\varepsilon^{atmos(i,s)}$ (length units) expressly tags the measurement made to satellite "s" by receiver "i". However, if the two receivers are close together the atmospheric conditions along the direction of line-of-sight to the satellite can be expected to be very similar,

and hence the residual bias ($\varepsilon^{atmos(i,s)} - \varepsilon^{atmos(k,s)}$) will be quite small in magnitude. The atmospheric refraction effect has an ionospheric and tropospheric component, each with their own spatial and temporal characteristics. $\varepsilon^{atmos(i,s)}$ across different epochs will change comparatively slowly, that is, there is high temporal correlation.

- The error term ν_i^s contains the random effects of measurement noise (which will vary according to satellite, receiver and measurement epoch), disturbing biases which are not spatially correlated (their effects are too dissimilar at different receivers), and any other biases not explicitly included in equations (2.1) and (2.2), such as signal multipath.

Clearly the use of two GPS receivers, simultaneously tracking the same satellites, is an effective means of overcoming the effect of spatially correlated biases. The method of positioning based on the use of two receivers is *differential GPS positioning* (or DGPS for short). There are different implementations of the differential positioning procedures, but all share the characteristic that the position of the GPS receiver of interest is derived *relative* to another fixed or reference receiver. DGPS is discussed further in Section 2.1.2 and 2.2. However, let's first take a closer look at GPS biases.

2.1.1 *GPS Measurement Biases and Errors*

As we've already mentioned, a useful "rule-of-thumb" for GPS measurement states that it is possible to make a measurement with a basic resolution at the level of 1%, or better, of the wavelength of the signal. Hence the level of measurement resolution is generally considered synonymous with measurement "noise". In the case of pseudo-range measurement "noise":

- The C/A code resolution is approximately 300m, hence there is a 3m range precision. (Noting that there is a trend to instrumentation with sub-metre C/A code resolution.)
- The P code resolution is approximately 30m, hence there is a tenfold improvement in precision, down to the 0.3m range level.

Carrier phase measurement "noise", on the other hand, is significantly less. The L1 carrier resolution is approximately 0.19m (the signal wavelength), while the L2 carrier resolution (or wavelength) is approximately 0.24m, implying millimetre accuracy for carrier phase measurements!

The level of measurement noise has a considerable influence on the precision attainable with GPS. Low measurement noise would be expected to result in comparatively high accuracy. This would only be true if there were no systematic biases contaminating the measurements, and hence the aim would be to remove as many of the biases as feasible so that only the smallest possible measurement errors remain. Dealing with the various biases in GPS data is a considerable problem and various strategies have been developed to account for them:

- They can be *estimated* as explicit (additional) parameters.
- Those biases linearly correlated across different datasets can be *eliminated by differencing*.
- The biases can be *directly measured*.
- The biases can be considered known or adequately *modelled*.
- They can be *ignored*.

It must be emphasised that biases larger than the noise level of the measurement, and likely to cause the quality of positioning to fall below some specified level of accuracy, should be accounted for somehow. In the case of pseudo-range measurements, the magnitude of most of the biases is below the noise level of the observations and can therefore be simply ignored (they become absorbed into the "error"). However, in the case of carrier phase measurements all biases are potentially a matter of concern. In other words, *different GPS applications require different levels of GPS accuracy, hence there is the possibility of a different partitioning of "biases" and "errors"*. At one extreme, as in the case of GPS pseudo-range absolute positioning, all biases with the exception of the receiver and satellite clock "uncertainty" are treated as errors, their effects ignored and hence these biases would be expected to distort the positioning results. At the other extreme, involving precise GPS position determination for geodesy applications such as crustal motion surveys, all measurement biases are explicitly accounted for in any solution in order that the results are highly accurate and reliable. Many ITS and mapping applications fall somewhere in between, requiring the elimination of some of the biases, particularly Selective Availability.

Fortunately, *differential positioning* is the most effective means of accounting for many of the troublesome GPS measurement biases, and hence is the basis for all high precision GPS positioning techniques. Although differential positioning does place additional operational demands (two GPS receivers are required, with a data link between them when DGPS is implemented in real-time), no explicit measurement or modelling of the spatially correlated biases is required. They are simply lumped together and eliminated, to a lesser or greater extent, from the relative position results.

In general, the biases can be considered to belong to one of three classes: satellite-dependent biases, receiver-dependent biases, and signal propagation biases.

Satellite-Dependent Biases

The ephemeris information used to calculate the GPS satellite positions is generated from the tracking data collected by the five monitor stations of the Control Segment (Section 1.1.3). The data is processed at the Master Control Station and the satellite navigation message information is uploaded to every satellite, and are available to GPS users at the time of observation. The *satellite orbit bias* is therefore the discrepancy between the "true" position (and velocity) of a satellite and its broadcast ephemeris. With regard to accuracy, there are (in principle) several distinct effects:

- There is the effect arising from the accuracy of the orbit *computation* procedure itself. The data used are P code pseudo-ranges, and although the tracking geometry is not strong (most of the tracking stations are in the equatorial belt), accuracies better than 5 metres are achievable.
- There are errors resulting from unpredictable orbital motion during the period since upload. These are essentially the *prediction* errors. Their magnitude can vary from a few metres (close to the time of navigation message upload) to several tens of metres.
- There is the effect due to Selective Availability (SA), which involves the deliberate degradation of the broadcast ephemeris parameters. The resulting orbit error may be highly variable (the degradation algorithm is classified), but we may assume it could be as high as 100 metres or more. (Though it is unclear whether SA also degrades the orbit information at present.)

What is the effect of the satellite orbit bias on GPS positioning? Expressed another way, if there is no other option than to assume the available orbit data is correct, what is the effect of an error in the satellite orbit? The following comments can be made:

- The height is a weakly determined component because there are no satellites below the horizon. This component is usually of the order of 2 or 3 times less accurate than the horizontal components.
- The effect on single receiver operation is to propagate the orbit error into the position results. A position result error is amplified.
- When two receivers are being positioned, both will be in error by nearly the same amount (the extent to which this is true is a function of the distance between the two receivers -- the closer they are, the more similar the error due to orbital bias). Relative positioning is therefore an effective strategy for *minimising* the effect of this bias.

One option for overcoming satellite bias error is to use a precise ephemeris as generated by the International GPS Service (the IGS -- Section 3.1.5). These ephemerides are accurate to the sub-metre level and are computed after global tracking data is collected from the IGS stations. Hence they are only available "post-mission" (unlike the broadcast ephemerides which are predicted into the future from the computed orbit and which can be used in real-time applications*). Plans are underway to generate predicted IGS orbits of sub-metre accuracy*.

Although GPS satellites use high quality cesium or rubidium atomic clocks for time-keeping and signal synchronization, there are unavoidable clock errors which change with time. These *satellite clock errors* cannot be ignored, hence they are a significant bias which are monitored by the control segment during tracking data analysis. The only way they can be accounted for in single receiver positioning is by using the broadcast clock error model defined by the polynomial coefficients (equation (2.8)). The three polynomial coefficients are known well enough to match the basic pseudo-range accuracy, that is to an accuracy of a few metres. However, under the policy of SA the coefficients are no longer able to adequately model satellite clock error. The range error resulting from the residual satellite clock bias (after correcting the range using the broadcast error model) can be more than 30m.

As all observations made at an instant, to a particular satellite, by all GPS receivers, are contaminated by the same satellite clock error, then the possibility exists for eliminating this bias through the principles of differential positioning.

Receiver-Dependent Biases

GPS receivers are equipped with relatively inexpensive quartz crystal oscillators. Although the time defined by individual receiver clocks have essentially arbitrary origins, they can be tied to a well established time scale, such as GPS Time (GPST), in a number of ways. The offset between the receiver clock time and GPST is the *receiver clock error* that contaminates all satellite-receiver ranges made at that instant by that receiver, and leads to these quantities being referred to as "pseudo-ranges" (Section 1.3.5). Typically, the solution to this problem is to treat the clock bias as an additional parameter in the pseudo-range navigation estimation procedure, requiring that four or more pseudo-range measurements are available. An alternative strategy is to take differences between data collected to the different satellites so that the common bias is eliminated. The subsequent time scale defined by the corrected receiver clock is then nominally that of GPST because:

(1) The synchronisation at some epoch (that is, the process of defining the time origin) is susceptible to error. *Generally, it can be carried out only at the few metre level.*
(2) The stability of the time scale is directly related to the quality of oscillator used, and how often the current clock time is synchronised using GPS pseudo-range observations.

Signal Propagation Biases

The ionosphere is the band of the atmosphere from around 50 to 1000 km above the surface of the earth. In this region, free electrons are released as a result of the gas molecules being excited by solar radiation. When the electromagnetic GPS signals propagate through this medium dispersion occurs, changing the velocity of the propagated signal ([2]). The *ionospheric propagation delay* of the code signals will cause the measured range to be longer than the true range. (On the other hand, the delay to the carrier phase signals is negative, and hence will cause the measured phase-range to be shorter than the true range.)

Ionospheric delay can range from about 50m for signals at the zenith to as much as 150m for observations made at the receiver's horizon. To reduce the ionospheric effect, coefficients of a correction formula are transmitted within the satellite navigation message. The correction can be applied to the measured data. However, the accuracy of the correction is very much dependent on the reliability of the estimate of Total Electron Content (TEC) along the signal path, which varies as a function of: the latitude of the receiver, the season, the time of day the observation of a satellite's signal is being made, and the level of solar activity at the time of observation.

For example, at night the ionospheric delay is approximately five times less than for day time observations. TEC is a maximum at mid to low latitudes, and is a minimum at the poles. A diurnal cycle for TEC occurs with a maximum occurring two hours after solar noon and is a minimum before dawn. Ionospheric disturbances, which can occur suddenly and be very severe, also affect the value of TEC. As the TEC is difficult to accurately determine, applying the correction formulae cannot effectively remove this effect. It is generally conceded that the broadcast correction model can be used to remove up to about 50% of the ionospheric delay at mid-latitude regions. For single frequency receivers the use of the correction model parameters is often the only option for point positioning. However, the ionospheric bias is spatially correlated (it is approximately the same for receivers up to a few tens of kilometres apart), and effectively is eliminated in differential positioning.

The ionospheric delay on a signal is a function of the signal frequency, hence if dual-frequency receivers are available this factor can be used to remove almost all of the ionospheric effect by making measurements on L1 and L2 signals and combining them in a special linear combination. However, all civilian GPS navigation receivers likely to be installed in vehicles are the single frequency (that is, the L1 tracking) variety.

The troposphere extends from the surface of the earth to about 8 km. GPS signals travelling through this medium will experience a *tropospheric refraction delay* that is a function of elevation and the altitude of the receiver, and is dependent on the atmospheric pressure, temperature, and water vapour pressure. The bias ranges from approximately 2m for signals at the zenith to about 20m for signals at an elevation angle of 10^o ([3]). The propagation of GPS signals in this medium is frequency independent (the troposphere is sometimes referred to as the "neutral atmosphere") therefore this effect cannot be removed by combining observations made on two frequencies. There are several options:

- Many models are available for this correction, the commonly used ones are the Hopfield model, the Black model, and the Saastamoinen model ([1]). About 90% of the delay stems from the dry gas component of the troposphere which can be modelled accurately without too much difficulty. The remaining 10% owing to the water vapour content is much more difficult to accurately model.
- For high precision applications the residual tropospheric bias has to be parameterised in the final position solution. *This is not an option that is exercised for GPS navigation applications.*
- Avoid tracking low elevation satellites. Generally satellites below 20^o have much greater problems with the tropospheric delay than high elevation satellites. *However, for navigation applications tracking of satellites down to the horizon is usually necessary.*
- As with the ionospheric bias, the fact that the bias is spatially correlated over distances up to several tens of kilometres means that differential positioning is an effective strategy for mitigating the effect of the tropospheric bias on positioning results. *However, the correlation times and distances for ionospheric delays and tropospheric delays are different, with tropospheric delays being typically a more local phenomenon.*

Another signal propagation bias does not have its origin in the physics of the atmosphere. The satellite to receiver distance, if measured using the carrier wave signal can be expressed as N cycles of the signal wavelength (approximately 19cm for the L1 signal, and 24cm for the L2 signal) plus the fraction of a cycle. For example, a measurement of -1993673.239 L1 cycles is a valid range measurement for while the fractional part of this data (0.239 cycle) is accurately measured, the integer cycle part (-1993673) is only arbitrarily assigned at the beginning of the satellite tracking process. The count of cycles since the beginning are precisely recorded and assuming that there is no loss of signal lock between the beginning of the tracking process t_{OC} and time t, the actual distance at time t will be equal to the phase measurement at time t, plus an integer bias. This bias is usually termed the *initial carrier phase cycle ambiguity*. It has the following characteristics:

- The ambiguity is an integer number (a multiple of the carrier wavelength).
- The ambiguity is different for L1 and L2 phase observations.
- The ambiguity is different for each satellite-receiver pair.
- The ambiguity is a constant for a satellite-receiver pair for all epochs of continuous tracking.

Once this bias is estimated the *ambiguous* carrier phase measurement can be converted into a standard (unambiguous) range measurement similar to the pseudo-range -- it is affected by the same biases mentioned earlier -- but with much higher measurement precision (that is, lower measurement noise ν') and significantly lower multipath error. Equation (6.2) may be extended to model the carrier phase measurement as:

$$\Phi_k{}^s = \rho_k{}^s + \varepsilon_{rc(k)} + \varepsilon^{sc(s)} + \varepsilon^{orbit(k,s)} + \varepsilon^{atmos(k,s)} + \lambda.N_k{}^s + \nu'_k{}^s \qquad (2.3)$$

where $\lambda$is the wavelength (L1 or L2), and N is a constant representing the unknown integer ambiguity. Carrier phase measurement is the basis of high precision GPS positioning.

GPS Measurement Errors

Apart from the measurement noise and the residual biases referred to earlier there are several other errors.

Multipath effects are propagation errors arising from the interference of the direct signal by reflected signals from water and metallic surfaces and nearby buildings. The combined direct and reflected signals will give rise to incorrect pseudo-range or phase measurements. The maximum multipath error that can occur in the case of pseudo-range data is one half the chip length (or resolution) of the code, that is, about 300m for the C/A code measurements, and 30m for P code measurements. This is a very large error which must be guarded against, especially in land navigation applications where the GPS receiver is located on the metal surface of a vehicle. (Carrier phase multipath on the other hand does not exceed one quarter of the wavelength; that is, 19cm on L1.) Effective ways to reduce this effect include the use of specially designed antennas and careful antenna mounting (for example, avoiding reflective surfaces -- often very difficult in many mapping applications). New receiver technology is being developed to effectively filter out multipath effects using advanced signal processing.

Another important error affects only carrier phase measurements (equation (2.3)). If satellite signals are obstructed by objects, or interfered by other signals, a *loss of lock* on the satellite signal will occur. On the resumption of lock to the satellite(s), the accurate fractional part of the phase observable can again be measured, however the integer part will be re-initialised and the initial integer ambiguity will no longer be a valid connection between the ambiguous fractional cycle measurement and the satellite-receiver range. For this reason there is a "jump" in the measurement data just before and immediately after the epoch at which the loss of lock occurred, and all measurements beyond this epoch are shifted by the same integer number of cycles. This "jump" is known as a *cycle slip*, and can occur independently on L1 and L2. The detection and repair of cycle slips is therefore an important carrier phase data pre-processing step.

2.1.2 *Absolute and Relative Positioning*

The GPS system was designed to provide users with two levels of performance. The highest accuracy was reserved for military users, who had access to the coded signals on both L-band frequencies. This is referred to as the Precise Positioning Service (PPS), and is intended to give absolute accuracies of the order of 10-20 metres. A lower level of accuracy was to be provided to the general (civilian) user by the Standard Positioning Service (SPS). The accuracy was intended to be an order of magnitude worse. However, after extensive testing in the early to mid 1980's, it was found that the SPS proved to be far more accurate than expected, achieving accuracies in the few dekametre range rather than the expected 100m level of accuracy.

GPS positioning accuracy can be expressed in a number of ways. Firstly in an *absolute* sense, with respect to a coordinate system such as WGS84 (Section 3.1.4). This coordinate system is realised first through the coordinates of the monitor stations (of the Control Segment), and subsequently transferred to users via the (changing) coordinates of the GPS satellites. As the satellite coordinates are essential for the computation of user position, any error in these values will directly affect the quality of the position determination. This sets a lower bound for the magnitude of GPS absolute positioning error. The Precise Positioning Service can deliver positioning accuracies of the order of 10-20 metres, mainly because of the lower level of measurement noise and the data is unaffected by unmodelled ionospheric delay error. A significantly lower level of accuracy was to have been provided to the general (civilian) user by the Standard Positioning Service. However, after extensive testing it was found that the SPS proved to be far more accurate than expected, achieving accuracies in the few dekametre range. As a consequence SA was implemented on the SPS in 1990.

Higher accuracies are possible if relative position is computed instead. Because many errors will affect the absolute position of two or more GPS users to almost the same extent, these errors largely cancel when differential positioning is carried out. This is the standard mode for GPS surveying, which essentially measures the *baseline components* (Δx, Δy, Δz) between simultaneously observing receivers.

In differential or relative positioning, one receiver is at location A, whose absolute coordinates are already known (x_A, y_A, z_A), and another receiver is at point B, whose position is to be determined. Both receivers observe the same satellites, and the observation data collected at both sites is then used to compute position B, but *relative* to A. As the coordinates of A are known, the absolute position of B will simply be the addition of the coordinates of A to the baseline components Δx, Δy, Δz. This technique of relative positioning can remove or reduce most of the biases common to two receiver sites, thus resulting in a higher positioning accuracy. An example of a bias that is significantly reduced in DGPS positioning is the satellite orbit bias (Section 2.1.1). The following approximate relations can be used:

Effect of orbit error on point positioning:

Position error = PDOP . Orbit error

- Example:

  If the Position Dilution Of Precision = 2 (Section 2.1.3), and

  orbit error = 20m

  Then *position error* = 40m
- Position error is therefore an amplification of the measurement error, where the amplification factor is a function of satellite-receiver geometry.
- Complications in the above simplistic relation due to the orbit error varying for different satellites, and with time.

Effect of orbit error on relative positioning:

$$\textbf{\textit{Baseline error}} = \frac{d}{20000} \; . \; \textbf{\textit{Orbit error}}$$

d = baseline length in kms

- Example:

  If d = 10km, and *orbit error* = 20m

  Then *baseline error* = 1cm
- Baseline error is therefore a function of baseline length.

There are several comments which can be made:

- Point B may be stationary or moving, but point A is stationary.
- There are essentially two strategies for data processing in DGPS: (a) data differencing so that the mathematical model contains the baseline components Δx, Δy, Δz explicitly, or (b) correction of data at B using the known position of A and the raw tracking data from A. The former is usually implemented in carrier phase processing software. The latter is the normal mode for DGPS using pseudo-range data (there are two implementations possible -- see Section 2.2.1).
- If B is stationary, then data can be collected over an "observation session" (with duration that may range from several minutes to many hours), and a more precise solution is possible.
- The accuracy of the relative position is a function of the distance between the two receivers A and B.
- The relative position can be determined in real-time if the data (or data corrections) from the reference receiver A are transmitted to receiver B, where they are combined with B's raw measurement data before processing.

2.1.3 *Satellite - Receiver Geometry*

The effect of satellite geometry on receiver positioning can be considered in relation to the intersection angles of the "lines of position" (or in the 3-D case, "surfaces of position"). For example, figure 1.5 (Chapter 1) illustrates the "cut" of two LOPs. If the angle of the "cut" is less than say 30^o (or greater than 150^o), the quality of the intersection is lowered. Another way of describing this effect is *mathematically*, using the pseudo-range measurements and one of the products of the navigation solution, the standard deviations of the position components.

Consider the case of four simultaneous measurements to four different satellites, a system of equations as in equation (2.1) can be constructed. While this 4-D estimation problem has a unique solution (four independent measurements, and four unknown quantities), if more than four measurements are made then the method of Least Squares can be applied to obtain the optimal solution. This solution, in addition to providing the values of the parameters, gives the variance-covariance matrix containing the standard deviations of the estimated parameters.

The basic steps to a Least Squares computation are outlined below ([1]):

(1) **Set up the solution**: compute the elements of the design matrix A, containing the partial derivatives of the range observations with respect to the parameters x, y, z, t (based on the parameterisation in equation (2.1)):

$$\frac{\partial\rho}{\partial x} = -\frac{x^s - x}{\rho}$$

$$\frac{\partial\rho}{\partial y} = -\frac{y^s - y}{\rho}$$

$$\frac{\partial\rho}{\partial z} = -\frac{z^s - z}{\rho} \qquad (2.4)$$

$$\frac{\partial\rho}{\partial t} = 1$$

(2) **Define approximate, or apriori, values of the parameters**: in particular the coordinate parameters $\overset{o}{\mathbf{x}}$ used for the computation of the partial derivatives and the residual quantities, or difference between the actual and calculated observations (the "sum of squares" of which are to be minimised):

$$\overset{o}{\mathbf{v}} = (\boldsymbol{l} - \boldsymbol{f}(\overset{o}{\mathbf{x}})) \qquad (2.5)$$

where $\boldsymbol{l}$ is the vector of actual observations and $\boldsymbol{f}(\mathbf{x})$ is the functional model for the observations (equation (6.1)).

(3) **Specify the quality of the observations**: by constructing the variance-covariance (VCV) matrix $Q_{\boldsymbol{l}} = P^{-1}$, where P is the "weight" matrix of the observations. Usually the VCV matrix contains only diagonal elements, the square of the standard deviation of the random measurement error for each satellite $\sigma_{URE}{}^2$.

(4) **Form the normal matrix**: $N = A^TPA$, and solve the system of equations:

$$\delta\hat{\mathbf{x}} = N^{-1}A^TP\,\overset{o}{\mathbf{v}} \qquad (2.6)$$

where $\delta\hat{\mathbf{x}}$ are corrections to the apriori values of the parameters $\overset{o}{\mathbf{x}}$. The quality of the estimated parameters can be extracted from the VCV matrix of the parameters $Q_{\hat{x}\hat{y}\hat{z}} = N^{-1}$.

This is the standard mode of pseudo-range positioning used in GPS navigation, in which the receiver clock error is treated as an additional unknown. All other biases are assumed to be insignificant (that is, their impact on the position solution accuracy is negligible).

The accuracy with which position can be determined is therefore not just a function of the measurement precision σ_{URE}, and the correct modelling of the significant biases, it is also a function of the *satellite-receiver geometry*.

How is this geometry indicator characterised? The variance-covariance matrix $Q_{\hat{x}\hat{y}\hat{z}}$ contains both the contribution to positioning error of the geometry and the random measurement error. Traditionally, for navigation applications, the components of the VCV matrix of the parameters are transformed into the Dilution of Precision (DOP) factor ([4]). DOP is simply the ratio of the positioning precision to the measurement error:

$$\sigma = DOP.\sigma_{URE} \qquad (2.7)$$

where σ_{URE} is the measurement precision (root-mean-square of the random errors)
σ is the position precision (root-mean-square of the position error)

Note that DOP is always a number greater than unity when there are no redundant observations. In the presence of unmodelled Selective Availability errors in the satellite clock, the value of σ_{URE} is estimated as being 25-35 metres.

There are a number of different definitions of DOP factors, depending on the coordinate component, or combination of coordinate components, of interest. For example, the two factors often used in GPS positioning is PDOP (Position DOP):

$$\text{PDOP} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_h^2} = \sqrt{\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2} \qquad (2.8a)$$

and GDOP (Geometric DOP):

$$\text{GDOP} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_h^2 + \sigma_T^2} = \sqrt{\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2 + \sigma_T^2} \qquad (2.8b)$$

where σ_E^2 , σ_N^2 , σ_h^2 variances of the east, north and height error components
σ_X^2 , σ_Y^2 , σ_Z^2 variances of the X, Y and Z error components
σ_T^2 is the variance of the error of the estimated receiver clock offset parameter

are the diagonal elements of the VCV matrix of the Least Squares position solution $Q_{\hat{x}\hat{y}\hat{z}}$. (The corresponding VCV for the local geographic components E, N, h can be obtained by transforming the $Q_{\hat{x}\hat{y}\hat{z}}$ matrix.) Figure 2.2 illustrates the situations of good and poor GDOP.

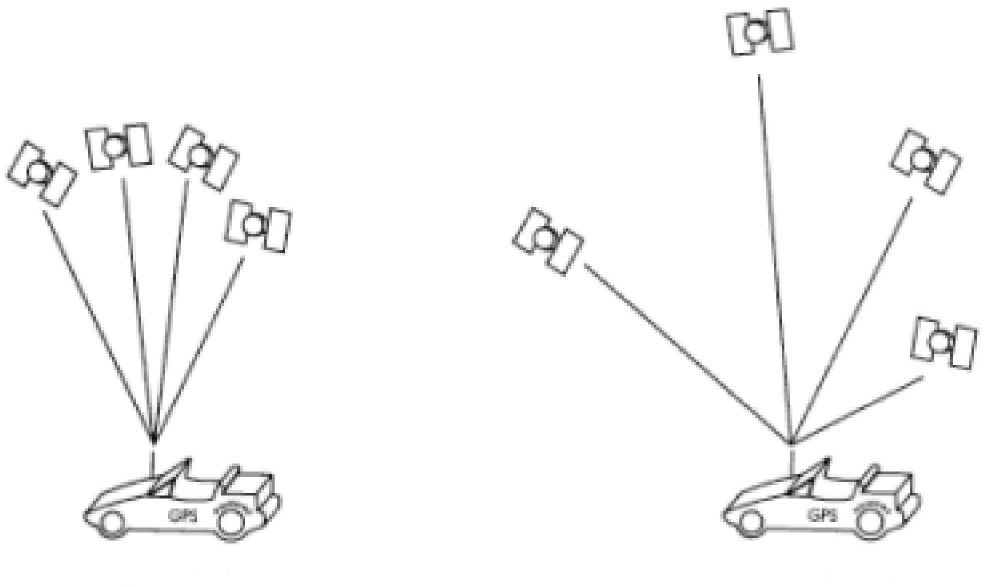

Figure 2.2: The Relationship Between Satellite-Receiver Geometry and GDOP (bad GDOP left; good GDOP right).

The following comments can be made regarding DOPs:

- The smaller the value of DOP, the higher the accuracy of the position results -- *the* *measurement errors are not as strongly amplified*.

- The DOP value for the vertical component estimate (the "VDOP") is consistently larger than the DOP for the horizontal positioning problem (the "HDOP") -- *confirming the well known fact that vertical error in GPS-based position determination is larger than horizontal error by a factor of 2 to 3*.
- DOP is usually greater than unity, however, if many satellites are observed (say >8), the value of DOP can be less than unity.
- DOP can be used as the basis of selecting satellites for solution -- *if GPS receiver cannot track all satellites that are in view*.
- A high DOP (say >10) indicates an "outage" -- *a situation where the position solution is too unreliable*.
- DOP varies with time of day and geographic location -- *but the pattern of DOP at a location repeats itself each day because the constellation is unchanged from day-to-day (except that it rising approximately four minutes earlier each day), hence it is highly predictable*.
- DOP varies with number of satellites considered-- *due to such factors as elevation cutoff angle used, number of satellites used by receiver to give "fix", etc.*

To decrease the occurrences of outages, and to improve the system availability, a number of technological solutions have been proposed (see Section 2.3), including:

- The most promising is the integration of GPS and GLONASS so that receivers can track both types of signals ([5,6]). This will significantly increase the number of simultaneously visible satellites.
- The use of geostationary, and other, communication satellites to transmit GPS-look-alike signals.
- The use of "pseudolites" -- ground stations transmitting a GPS-look-alike signal.
- Additional "pseudo-observations" such as provided by precise receiver clocks (permitting so-called "clock-aiding") and barometers (permitting "height-aiding"), see Section 2.3.3.

2.1.4 *Processing Algorithms, Operational Mode and Other Enhancements*

Finally, GPS accuracy is also dependent on a host of other *operational, algorithmic and other factors* such as:

- Whether the user is moving or stationary. Obviously repeat observations at a stationary station would permit an improvement in precision due to the effect of averaging down random errors over time. A moving GPS receiver does not offer this possibility.

- Whether the results are required in real-time, or if post-processing of the data is possible. Real-time positioning requires a "robust" but less precise technique to be used. The luxury of post-processing the data permits more sophisticated modelling and processing of GPS data which minimises the magnitude and impact of residual biases and errors. *Post-processing is not an option for most ITS applications*.

- The level of measurement noise has a considerable influence on the precision attainable with GPS. Low measurement noise would be expected to result in comparatively high accuracy. Hence carrier phase measurements are the basis for high accuracy techniques, while pseudo-range measurements are used for low accuracy applications. In addition, carrier phase data can be used to "smooth" the relatively noisy pseudo-range measurements prior to their use in the positioning algorithm.

- The degree of redundancy in the measurements. For example, such factors as the number of tracked satellites (dependent upon the elevation cutoff angle, the number of receiver tracking channels, satellites apart from GPS such as GLONASS, the use of pseudolites, etc.), the number of observations (dual-frequency carrier phase, dual-frequency pseudo-range data). This permits more sophisticated quality control procedures to be implemented that "trap" (and delete or down weight) bad quality data which would otherwise bias solutions.

- The algorithm type may also impact on GPS accuracy. For example, "exotic" data combinations are possible (carrier phase plus pseudo-range), Kalman filter solution algorithms, more sophisticated phase processing algorithms, etc.

- Techniques of data enhancements and aiding may be employed. For example, the use of carrier phase smoothed pseudo-range data, external data such as from inertial navigation systems (and other such devices which can be used to navigate by "dead reckoning" when satellite positioning is not possible), additional constraints, etc.

2.2 Relative GPS Techniques

As discussed in Section 2.1, the use of two GPS receivers, simultaneously tracking the same satellites, is an effective means of overcoming the effect of spatially correlated biases. There are essentially two ways in which measurements from two receivers are used to account for biases, and hence improve accuracy:

(a) Each set of measurements at a receiver are independently used to derive a position which is in error by more or less the same amount. *This is the DGPS procedure implemented for precise navigation applications using pseudo-range data*.

(b) Differencing measurements between receivers leads to an observable that is essentially free of biases (or at least substantially reduced if the receivers are not too far apart). *This is the GPS surveying mode of differential positioning using carrier phase data*.

DGPS is discussed further in the following sections.

2.2.1 *Differential GPS Navigation*

The distinguishing characteristic of navigation is the *urgency* with which positioning information is required. Relative positioning information is derived from two separate navigation "fixes" based on the processing of pseudo-range observations. Differential GPS (or DGPS for short) requires that one of the receivers is located at a station of known position (the "base station"), while the other is at an unknown location (the so-called "rover receiver"). Two implementations are possible:

(a) Differential positioning can be accomplished by the continuous transmission of the coordinate solution from the "base station" to the "rover receiver" as illustrated in figure 2.3. This block shift technique, or *position domain differential strategy*, is the easiest to implement (although it does have certain serious limitations):

- Base receiver at known point -- *WGS84 or local datum coordinates*.
- Compare known position and instantaneously computed position.
- Generate correction ΔX, ΔY, ΔZ.
- Transmit correction to rover receiver for immediate correction of "raw" point coordinates.

It is important that both the rover and base receivers use the *same* satellite constellation to generate their point solutions, otherwise severe errors can result, possibly worse than those of the (uncorrected) point positioning. This is a significant limitation as it is rarely the case that the same constellation of satellites are simultaneously visible if the two receivers are a long distance apart, or when the rover receiver is operating in cities, as is the case for ITS applications, where the effect of urban "canyons" causes significant occlusion of the satellite signals.

(b) A popular real-time DGPS strategy is the method of range corrections. Rather than making corrections to the coordinates, the ranges *before* computation of the receiver position are corrected (figure 2.4), this is also known as the *measurement domain differential strategy*,. This is achieved by a process similar in many respects to that of the block shift method:

- Base station at known point --> *WGS84 or local datum coordinates*.
- Using known position compute "true" range.
- Generate corrections to all pseudo-range data by comparing "true" to "observed" range.
- Transmit correction data to the rover receiver for correction of ranges before solution is carried out.

The technique is far more flexible because the correction is made to the pseudo-ranges and hence the rover GPS receiver can use any combination of corrected ranges to obtain a solution, and not just the satellite constellation used at the base station.

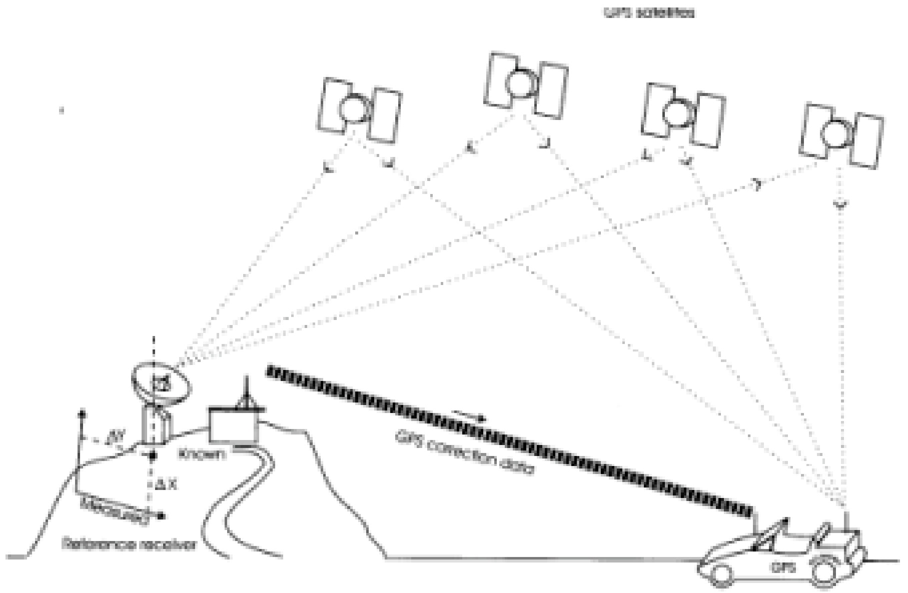

Figure 2.3: Principle of DGPS Using the Block Shift Method.

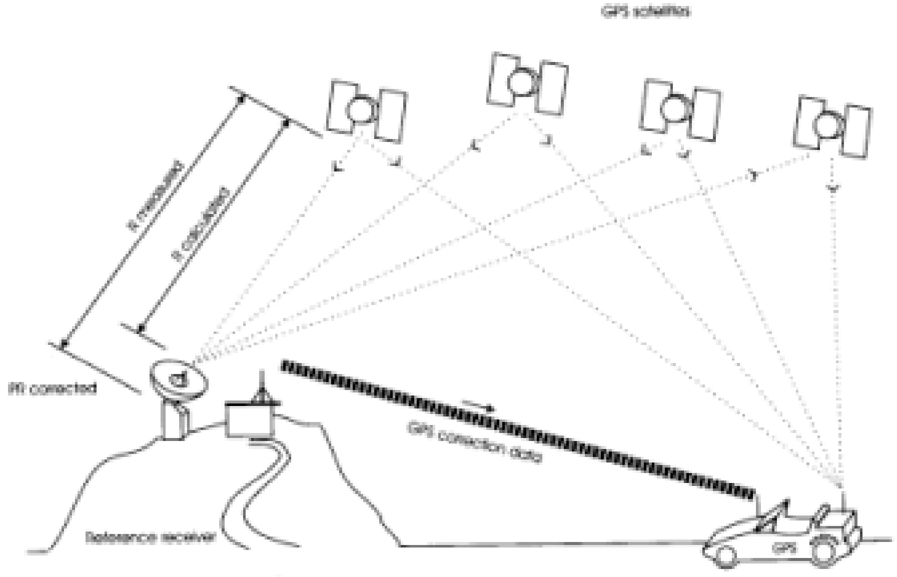

Figure 2.4: Principle of DGPS Using the Range Corrections Method.

The relative accuracies expected in such schemes are of the order of 1-10m, though enhancements have been introduced that appear to deliver sub-metre accuracy.

There are several DGPS scenarios possible:

- DGPS may be implemented in **real-time** or **post-processed** mode. The range corrections method is preferred for real-time implementations because only the set of corrections for all visible satellites need to be transmitted. On the other hand, if implemented in the post-processed mode, all the raw observations made by the rover receiver would have to be written to file as well as the differential corrections computed at the base receiver. An alternative that doesn't require the outputting of raw (pseudo-range) measurements by the rover receiver is based on the block shift method.
- DGPS in support of **autonomous navigation** or **vehicle tracking** applications. The former

provides precise GPS position to the rover, for its own use, but to no-one else. The latter is a means by which a rover receiver's location can be monitored at some central facility. Different DGPS scenarios are possible depending upon whether the operation is implemented in real-time or in a post-processed mode, as discussed below.

Real-time DGPS or post-processed DGPS? The following comments can be made:

Some advantages of the Real-Time DGPS Implementation:

- No data archiving required and no post-processing is necessary.
- The rover receiver equipment can therefore be small and lightweight.
- Transmission of the industry-standard DGPS correction message format means that real-time DGPS capability is built into all receivers at low (additional) cost.
- When DGPS is in broadcast (or "open service") mode, all receivers operate independently.
- Can take advantage of communication link to transmit other (non-positional) data to and from the base facility.

Some disadvantages of the Real-Time DGPS Implementation:

- The requirements of a communication channel leads to greater infrastructure complexity, and associated problems such as signal coverage, fade, etc.
- Real-time tracking has capacity limitations.
- Real-time tracking is likely to be a "closed" system.
- Quality assurance within real-time system is more difficult than in the case of data post-processing.

Some advantages of the Post-Processed DGPS Implementation:

- No additional instrumentation (such as communications equipment) is required.
- Quality assurance measures can be applied.
- There may be no urgency for DGPS positions.

Some disadvantages of the Post-Processed DGPS Implementation:

- Post-processing software is likely to be instrument specific.
- The operation requires coordination of data capture at both rover and base receivers.
- Cannot be used for real-time positioning, as in the case of most ITS applications.

An important consideration for post-processed GPS schemes is data file formats. There are two file format options available for data to be post-processed. The first are the receiver-specific data formats, only useful if the same make of receiver is operated as both the base and rover receiver; and the universally recognised standard format for GPS data known as the RINEX (Receiver INdependent EXchange) format. Privately owned and public base stations generally only make data available to users in the RINEX format, for example via floppy disk, via Internet, or by email. If the data differencing mode cannot be used, for example because the pseudo-range tracking data is not available from both the rover and base receiver, then some form of modified Block Shift method would have to be used (see above). The standard post-processing procedure is for the rover receiver to track and compute coordinate positions, and to record the position together with the time and the set of satellite I.D.s used to determine those positions. The base receiver tracks all available satellites and stores the information in the ASCII RINEX format, at some regular epoch rate. During post-processing, time-tags and satellites I.D.s are compared (between the rover and base files) and the necessary tracking data required for coordinate computation is extracted from the base RINEX file, the computation made with this subset of satellite data, the Block Shifts determined and then applied to the uncorrected rover coordinates.

The use of Range Corrections is the preferred mode for real-time autonomous DGPS navigation. The Range Corrections are calculated at the base receiver and then transmitted to one or more rover receivers. The base and rover(s) operate independently of each other. The rover receiver corrects the pseudo-range tracking information it has collected using the transmitted corrections. The corrected data is then used by the rover receiver's computer to determine the position. This corrected position information need not be stored for most real-time navigation applications.

If the raw data (position or pseudo-ranges) is sent from the rover receiver to a central facility (say the base station), it can be corrected, and then the base immediately has the rover's corrected position or tracking information. In this real-time scenario only a limited number of rover receivers can be tracked by the base station, simply as a result of the limit to how much information can be received and processed. Nevertheless, such a system is ideally suited for such applications as the monitoring of precious goods while in transit. The system can be extended further by combining the tracking and navigation modes, through a two-way communications link, to provide the ideal all-round navigation/tracking system.

2.2.2 *Carrier Phase-Based Relative Positioning*

The basis of high precision relative positioning is carrier phase measurements because of the very low level of "noise" on these measurements (section 1.3.3). Carrier phase measurements are used in two ways:

- They may be used to "smooth" the much noisier pseudo-range observations prior to being used to compute absolute, or relative, positions (section 2.3.3).
- They may be used on their own, in preference to the pseudo-range observations. Initially developed to cater for static surveying applications requiring centimetre level accuracy, modern instrumentation can deliver this level of accuracy even when the rover receiver is in motion.

Carrier phase-based positioning relies on the ability to eliminate, or significantly reduce, the common measurement biases across observations made simultaneously by ground receivers to the ensemble of visible GPS satellites. An inspection of eqns (2.1) and (2.3), and the accompanying discussion, reveals that if GPS observations made by a receiver to several satellites are differenced, the **receiver dependent biases** are eliminated (principally the receiver clock bias -- section 2.1.1). Further, if a number of GPS observations are made by several receivers simultaneously to the same satellite, the **satellite dependent biases** are eliminated (or their effects significantly mitigated). At any epoch of measurement, differencing pseudo-range observations made simultaneously by receivers k and i, to satellites s and q, yields the following "double-difference" observable:

$$\begin{aligned} \nabla\Delta p_{ki}^{pq} &= p_k^s - p_k^q - p_i^s - p_i^q \\ &= \rho_k^s - \rho_k^q - \rho_i^s - \rho_i^q + \nabla\Delta\nu_{ki}^{sq} \qquad (2.9) \end{aligned}$$

Hence four pseudo-range measurements are combined to create a new "observable", with a noise level that is double that of a single one-way pseudo-range measurement (therefore of the order of several metres). Typically, the pseudo-range measurements are made using the C/A code on the L1 carrier.

The same procedure may be applied to carrier phase measurements (which can be made on either the L1 or L2 carrier), to yield in metric units, a similar relation:

$$\begin{aligned} \nabla\Delta\Phi_{ki}^{sq} &= \Phi_k^s - \Phi_k^q - \Phi_i^s - \Phi_i^q \\ &= \rho_k^s - \rho_k^q - \rho_i^s - \rho_i^q + \lambda.\nabla\Delta N_{ki}^{sq} + \nabla\Delta\nu'_{ki}{}^{sq} \qquad (2.10) \end{aligned}$$

Note that $\nabla\Delta N_{ki}^{sq}$, the double-differenced ambiguity term, the only new term in eqn (2.10) (compared with eqn (2.9)), is an integer. The noise of the double-differenced phase observable is less than one centimetre. The use of (double-differenced) carrier phase is problematic because any data processing scheme must ensure the estimation of two classes of parameters: (a) the baseline parameters Δx_{ki}, Δy_{ki}, Δz_{ki} (contained within the geometric range term $\nabla\Delta\rho_{ki}^{sq}$ -- see eqn (2.3)), and (b) the ambiguity parameter $\nabla\Delta N_{ki}^{sq}$ (for each pair of independent satellites s, q). The following comments may be made with regard to the processing of carrier phase data:

- The problem of processing "ambiguous" carrier phase data was overcome for static GPS surveying applications by the collection of sufficient epochs of data (up to one hour or more) to ensure the "separability" of the position and ambiguity parameters within a Least Squares adjustment scheme (that is, the reliable and accurate determination of both sets of parameters).

- Although the estimation of the ambiguity parameters are real-valued quantities at the same time as the baseline components is a one option, advantage can be taken of the fact that these parameters should be integers in order to strengthen the solution further.
- If, on the basis of the real-valued ambiguity estimates, and their precisions, it is possible to "resolve" what the likeliest integer values of the ambiguities are, they may be eliminated from the estimable parameter set in eqn (2.10) -- in effect the ambiguous double-differenced phase observable $\nabla\Delta\Phi$ is converted to the unambiguous double-differenced pseudo-range observable $\nabla\Delta p$ -- and the position estimation problem becomes one involving very precise "range" observables.
- This process of estimating ambiguity parameters, and then selecting the likeliest integer values, is known as "ambiguity resolution", and its reliability is a function of:
  - baseline length (typically the receivers are not separated by more than about 20km),
  - the number of satellites (the more the better),
  - the satellite-receiver geometry,
  - whether observations are made on both frequencies (it is much easier to resolve ambiguities when dual-frequency observations are available), and
  - the length of the observation session (the longer the better).
- Although the above remarks imply that ambiguity resolution is a process which can only be applied to static GPS carrier phase data, in recent years algorithms have been developed to carry out ambiguity resolution "on-the-fly", that is, while one of the receivers is in motion.

Note that, once the ambiguities have been resolved, all carrier phase data can be converted to a range-like observable which can be used for precise (centimetre-level) instantaneous positioning. There are an increasing number of applications for high precision carrier-phase based positioning, for machine and vehicle guidance and control.

Nowadays kinematic carrier phase-based positioning can be carried out in real-time if an appropriate communications link is provided over which the carrier phase data collected at a static base receiver can be made available to the rover receiver's onboard computer; to generate the double-differences, resolve the ambiguities and perform the position calculations. Although many such systems are now available, the formats of the transmitted carrier phase data are mostly of a proprietary nature, and hence it is not generally possible for Brand "X" base receiver to "talk" to a Brand "Y" rover receiver. A broadcast carrier phase service is being offered in the U.S. and parts of Europe and Asia, based on a message format that is compatible with a number of GPS receivers. On the other hand, in the case of pseudo-range corrections, an industry standard transmission format is available.

2.2.3 *Real-Time DGPS: Data Correction Transmissions*

The United States body, the Radio Technical Commission for Maritime (RTCM) Services, is a group concerned with the communication issues as they pertain to the maritime industry. Special Committee 104 was formed to draft a standard format for the correction messages necessary to ensure an open real-time DGPS system (see [7]). The format has become known as *RTCM 104*, and has recently been updated to version 2.2.

According to these recommendations, the pseudo-range correction message transmission consists of a selection from a large number of different message types. Not all message types are required to be broadcast in each transmission, some of the messages require a high update rate while others require only occasional transmission. Provision has also been made for carrier phase data transmission, to support carrier phase-based real-time kinematic (usually referred to as "RTK") positioning using the RTCM message protocol. GLONASS differential corrections can also be transmitted within this protocol. Many message types are still undefined, providing for considerable flexibility.

The greatest consideration for the DGPS data link is the rate of update of the range corrections. Selective Availability errors vary more quickly than any other bias (such as orbit error, atmospheric refraction, etc.), hence they are the primary concern and the major constraint for real-time DGPS communications options. The correction to the pseudo-range and the rate-of-change of this correction are determined and transmitted for each satellite. If the message "latency" (or age) is too great then temporal decorrelation occurs, and the benefit of the DGPS corrections is diminished.

The DGPS correction message format is patterned on the satellite navigation message, and was originally designed to operate with communication links with as low a data rate as 50 bps (bits per second). Most navigation-type GPS receivers are "RTCM-capable", meaning that they are designed to accept RTCM messages through an input port, and hence output a differentially corrected position. RTCM is not instrument-specific, hence Brand "X" rover receiver can apply the corrections even though they were generated by a Brand "Y" base receiver.

2.2.4 *Real-Time DGPS: Communication Link Considerations*

The following considerations must be addressed by DGPS communication links:

- **Coverage:** This is generally dependent on the frequency of the radio transmission that is used, the distribution and spacing of transmitters, the transmission power, susceptibility to fade, interference, etc.
- **Type of Service:** For example, whether the real-time DGPS service is a "closed" one available only to selected users, whether it is a subscriber service, or an open broadcast service.
- **Functionality:** This includes such link characteristics as whether it is a one-way or two-way communications link, the duty period, whether it is continuous or intermittent, whether other data is also transmitted, etc.
- **Reliability:** Does the communications link provide a "reasonable" service? For example, what are the temporal coverage characteristics? Is there gradual degradation of the link? What about short term interruptions?
- **Integrity:** This is an important consideration for critical applications, hence any errors in transmitted messages need to be detected with a high probability, and users alerted accordingly.
- **Cost:** This includes the capital as well as ongoing expenses, for both the DGPS service provider as well as users.
- **Data rate:** In general the faster the data rate, the higher the update rate for range corrections, and hence better positioning accuracy. Typically a set of correction messages every few seconds is acceptable.
- **Latency:** Refers to the time lag between computation of correction messages and the reception of message at the rover receiver. Obviously this should be kept as short as possible, and typically a latency of less than 5 seconds is suggested.

In this age of communication technology and information transfer, there are a number of communication options available for DGPS operation including:

General Systems: Two-way communications suitable for full DGPS operation include:

- HF/VHF/UHF radio systems -- dedicated frequencies, as well as open citizen bands.
- Satellite communications -- via geostationary or low-earth-orbiting (LEO) satellites.
- Cellular (or mobile) phone network -- a growing number of options including digital/analogue systems, packet based systems, etc.

Broadcast Options: These are one-way systems suitable for carrying the RTCM messages to the user:

- Bulletin Board Services.
- Paging services.
- FM radio ancillary channels.
- LF/MF frequency transmissions via navigation beacons, AM radios, etc.
- Television Blanking Interval transmissions.
- Pseudolites.

Some of the two-way communication systems may already be in place, or can easily be established in virtually any location. The infrastructure for their operation may therefore provide immediate and effective coverage. Satellite communications is a particularly attractive option because of its wide coverage, and hence is very commonly used for offshore DGPS applications. It is, however, still a relatively expensive option and is generally used only if there are no cheaper alternatives.

Some of the broadcast options may be considered to be "piggyback" systems that take advantage of an communication infrastructure that is already in place. Direct or dedicated radio systems are the alternative to piggyback systems. However these normally would require the establishment of an independent licensed radio transmitting station. Pseudolite (or "pseudo-satellite") services, on the other hand, are ground-based stations that transmit a signal that is similar to that of a GPS satellite. Such systems are still in the experimental stage, but show great promise for specific applications when the number of visible satellites is low due to significant shading. GPS-like signals will increasingly be transmitted in future by non-GPS satellites. These signals will, in addition to carrying correction messages, provide alternative range measurements for position fixing. *One thing is certain, the variety of satellite and ground-based communication systems is likely to grow rapidly over the next ten years*.

2.2.5 *Real-Time DGPS Services*

It is possible to differentiate between *Local Area DGPS* (LADGPS) and *Wide Area DGPS* (WADGPS). The assumption made when a pseudo-range correction message is generated at a base receiver is that it is a valid calibration quantity representing the "lumped" satellite and propagation link biases as monitored at the base receiver: satellite clock error, satellite orbit bias, troposphere and ionosphere refraction. However, apart from the first bias quantity, this assumption breaks down as the separation of base and rover receivers increases. In addition, the chances that the constellation of visible satellites at both the base and rover receivers are the same diminishes as inter-station distances grow. Typically, a range of 100 km or so is considered the limit beyond which it is unreasonable to assume that biases will cancel when the pseudo-range corrections are applied. (In the case of carrier phase data this assumption generally becomes invalid at distances of around 30 km.) *Hence, LADGPS refers to real-time differential positioning typically over distances up to a few hundred kilometres using the DGPS corrections generated by a single base station*. The DGPS correction is generally delivered by some form of short-range terrestrial-based communications system.

Wide Area DGPS, as the name implies, is a DGPS technique that distributes the accuracy advantages of DGPS across a very wide region. This may be over a continental extent or, in the extreme case, could represent a global service. Although there are a number of different implementations of WADGPS (see, for example [8]), all rely on a *network of base stations distributed across the region of interest* and a communications system with the requisite coverage and availability characteristics. In its crudest form WADGPS can be considered a means by which *multiple* RTCM messages are received at the rover receiver (from each of the base stations within the WADGPS network) and the corrections are, in effect, "averaged" and input through the receivers I/O port as a "synthetic" RTCM message (this obviates the need for elaborate new software having to be embedded within the GPS receiver). This implementation is sometimes referred to as "Network DGPS". More sophisticated implementations model (in real-time) the spatial variation of errors due to atmospheric refraction and orbital bias so that the WADGPS message contains the values of model parameters and a special algorithm computes the rover receiver corrections on the basis of geographic location.

A variety of real-time DGPS services have been established over the last few years to address precise navigation and positioning applications on land, at sea and in the air. Such services may be characterised according to the following:

- LADGPS or WADGPS implementation.
- The type of communications link, whether it is terrestrial or satellite-based.
- Whether the service addresses a specific group of users (for example, marine), or is a general service.
- The nature of the organisation providing the service, is it a government agency, an academic institution or a private company?
- Whether the service is freely available, or whether it is operated as a commercial activity.
- Whether it is restricted to RTCM pseudo-range correction broadcasts, or the transmission of carrier phase data, or both.
- Whether the system supports post-processed DGPS by archiving the base station data.
- Whether the service uses a single base station, or is part of a network of DGPS base stations.
- The sophistication of the Quality Control measures that are in place.

WADGPS is appropriate for those applications where the delivery of differential corrections by satellite communications link is the only feasible option. This is unlikely to be useful for most ITS applications, for the foreseeable future, unless the new low-earth-orbiting communications satellite systems lead to dramatic reductions in DGPS service charges. Most ITS applications are based in cities, where a range of communication systems are available to carry DGPS corrections. LADGPS is at present the most common implementation of real-time DGPS.

It is instructive to review the situation in Australia with regard to DGPS services, cost structures, user profiles and dominant applications. Australia may be considered a representative example of what is happening in many other countries, because the majority of system developers and service providers are global companies operating across different geographic regions. The four DGPS services are all multi-site systems which between them effectively address almost 100% of all DGPS needs in the Australian region. Some are WADGPS, others LADGPS. Some are commercial systems, however one is a free "public service". Several of them are mature systems, while others will expand as additional base stations are established.

DGPS by satellite communications link offered by the huge Dutch conglomerate Fugro. They have a number of GPS base stations across Australia which are connected by landline to their head office, Fugro Starfix, in Perth. The DGPS corrections are uplinked to the Optus B1 satellite from where they are broadcast all over Australia and parts of S.E.Asia. The procedure is identical to that used by Fugro in other parts of the world except that instead of using the Inmarsat satellites (which require a gimbal-mounted directional antenna), they use Australia's L-band mobile satellite communication system which uses small whip omnidirectional antennas (a similar service is available in North America and Europe). The DGPS service, known as Omnistar, is a strictly commercial operation and offers several levels of service. In its simplest (and cheapest) configuration it operates as a LADGPS (RTCM messages from one base station), which can deliver a few metres accuracy up to a few hundred kilometres from a base station, with accuracy degradation being a function of the distance from the nearest base station. A WADGPS option uses a proprietary Fugro format to combine the correction messages from several base stations, delivering sub-metre accuracy. The advantages of the service is that it is available anywhere in Australia (and offshore) using relatively compact hardware. It is, however, comparatively expensive on a day-rate basis, though it is possible to be charged on an on-time basis.

DGPS by satellite communications link offered by the large British company, Racal Survey. As with the Fugro system, they have a number of GPS base stations across Australia which are connected by landline to their head office, also located in Perth. In every other respect the systems are direct competitors. The DGPS corrections are also transmitted by the Optus B1 satellite in the Australian region, and by Inmarsat satellites over the rest of the world. The DGPS service, known as Landstar, is also a commercial operation and offers both a LADGPS and WADGPS option. Despite competition with the Fugro service, it also is a comparatively expensive service with user charges a factor of ten or more higher than LADGPS services based on terrestrial communications systems.

DGPS by Radio Data Service (RDS) link offered by the AusNav company in Canberra. They have established a large number of GPS base stations across Australia, mostly located in the capital cities or where there is a large (generally niche) local market for DGPS services. The DGPS corrections generated at each base station are sent to a local Australian Broadcasting Corporation FM radio station by landline, where they are encrypted and modulated on the sideband RDS signal. (RDS is a protocol for encrypting digital data on FM sidelobe signals.) Obviously this service takes good advantage of existing radio service infrastructure and allows for the use of very small low-cost FM receivers no larger than a pager to receive and decode the RTCM message. The DGPS range is limited to that of the FM signal reception range. The LADGPS service in Australia is known as AusNav Service, and is a strictly commercial operation aimed at small volume users, and users who cannot justify an expensive DGPS service such as that provided by the Racal or Fugro systems. The transmission of RTCM messages (and even carrier phase data) using the RDS principle is now being offered in many countries by two companies: Differential Corrections Inc. (as in the AusNav network) and PinPoint (or Accqpoint Communications Corporation as it is known in the U.S.).

DGPS for marine users is a unique service because it is intended for only one class of user. It is patterned on a similar service offered in North America and Europe (where there are over a hundred DGPS base stations), and is in fact compatible with these international systems because the marine users travel from country to country, and must be able to acquire and use transmitted messages wherever they go. In Australia the government agency tasked to provide this navigation service is the Australian Maritime Safety Authority (AMSA). It is a LADGPS service offered for free to users who can pick up an unused marine frequency -- 304 KHz. AMSA is rapidly expanding the service and has transmitting stations in Sydney, Melbourne, and several stations up the east coast of Australia (particularly the Great Barrier Reef) and the north-west shelf area. The network will eventually cover all coastal waters, and out to about 100-200km from the coast. The signal can also give limited inland coverage, though the extent to which it could be used as an alternative to AusNav, in captial city environments, is not yet known.

A useful discussion on the future of "commercial" DGPS services can be found in [9].

2.3 GPS Augmentation Options

"Augmentation" refers to those enhancements to the system, the algorithms or the hardware, designed to improve the performance of GPS in some way. (Some of these were mentioned in Section 2.1.4.) The improvement(s) may be measurable in terms of some global performance variables such as accuracy, or reliability, or availability. On the other hand, some enhancements may be introduced in order to make GPS a more attractive technology for addressing a specific application. In this section we mention briefly only some representative examples of:

- Satellite system enahncements,
- Data processing algorithm or software enhancements, and
- Hardware enhancements.

2.3.1 *Global GPS Augmentation*

This is a system augmentation designed to improve: (a) accuracy, (b) integrity, and (c) availability. The impetus for this has come from civil aviation authorities who wish to replace traditional navaids by technology which is less expensive, more reliable and more versatile. The International Civil Aviation Organisation (ICAO) has developed the concept of the Future Air Navigation System (FANS). An important component of FANS is the Global Navigational Satellite System (GNSS) based on GPS, but enhanced in order to satisfy the varying requirements for accuracy, availability and integrity for the different phases of flight: en route navigation, airport approach and landing, and surface movement. Although motivated by aviation concerns, GNSS is important for other land-based applications because it provides the first model of *extension* of the GPS satellite-based positioning technology. The augmentation that ICAO has in mind for the GNSS consists of three components:

(a) Transmission of *differential corrections* by satellite to users over large areas.
(b) Transmission of *integrity information* by satellite.
(c) Transmission of *additional GPS-like signals* by other satellites.

The enhancements:

- improve accuracy, integrity, availability and continuity,
- represent a significant investment in additional space and ground infrastructure,
- are intended to service users across large areas of the world,
- will require agreement on global specifications for all components,
- will require modifications to GPS user hardware,
- are primarily intended for one applications sector,
- indicate a possible system architecture for the successor to the GPS system, and
- are expected to be operational in the U.S.A. over the next few years, with other countries following suit with some or all of the components, according to an agreed timetable.

Although there has been international agreement that GPS alone cannot fulfil all the requirements for a sole navigation aid for civil aviation (and hence GNSS is not just a synonym for GPS), there is not yet unanimous agreement on the details of the augmentation. As a consequence, there will be different regional implementations of the augmentation, with the U.S.A. promoting its Wide Area Augmentation System (WAAS), the European nations have their European Geostationary Navigation Overlay Service (EGNOS), and more recently unveiled plans for their own navigation system known as “Galileo”, and the Japanese have proposed the Multi-Functional Transport Satellite (MT-SAT). A description of a "generic" version of the ICAO augmentation model is presented below.

Component 1: WADGPS Corrections
Wide Area DGPS corrections (see Section 2.2.5) provide the means of delivering high navigation accuracy (better than 10m) to civil aviation users over large areas, by means of universal communication systems such as geostationary satellites. The WADGPS corrections would be generated under contract to the national aviation control authorities, who would make them available to suitably equipped users. The preferred communications service provider is Inmarsat, an international not-for-profit organisation specialising in the provision of marine satellite communications, though other options may be used.

As each country implements a WAAS-like system within their own airspace, a network of reference receivers will be established to gather the tracking data and a Master Station will calculate the differential corrections and integrity data (see below). This data will be put into the appropriate message format and sent to a Ground Earth Station for uplink to a geostationary satellite such as those of the Inmarsat series, which will, in turn, broadcast the information to all users. *It is unclear whether the delivery of the WADGPS correction messages will be provided at a competitive rate to present commercial services -- or even if it will be free (just as the marine navigation industry benefits from the free DGPS correction service indirectly financed from "lighthouse dues").*

Component 2: Integrity Warnings
This is intended to address the concerns of system "integrity", a Quality Control (QC) issue. Critical users require timely warning when <u>not</u> to use GPS for navigation. There are several options.

The Receiver Autonomous Integrity Monitoring (RAIM) is a GPS receiver hardware-based QC technique which takes advantage of redundant satellites (the excess over the minimum four required for GPS pseudo-range positioning). In principle, different sub-sets of four satellites are used to generate solutions, and the results are continuously intercompared. In this way if five satellites are tracked, RAIM will be capable of *detecting* if one satellite is malfunctioning. If six or more satellites are tracked, RAIM will be capable of *identifying* malfunctioning satellite (but only if one is malfunctioning). *RAIM-capable receivers will be authorised for aircraft operations, however it is unclear as to whether such receivers will be used extensively outside the narrow civil aviation applications.*

GPS Integrity Channel techniques take advantage of ground facilities. What is required is one or more ground GPS receivers collecting data to the GPS satellites, a Master Station to monitor the quality of the GPS system, and a communications infrastructure to alert users when operations are degraded. The most complex component is the Master Station and its interface to the communications system, for example the Inmarsat geostationary satellites, as the maximum delay between the detection of a problem and the alert message being received by an aircraft must be of the order of a few seconds for airport landing applications!

Civil aviation users have several means of acquiring integrity information, however, the provision of a satellite communications channel to deliver WADGPS corrections makes the additional provision of ground-generated integrity information a relatively straightforward procedure.

Component 3: GPS Signals from Other Satellites and Pseudolites
A minimum of four simultaneously visible satellites are required for 3-D GPS positioning (Section 1.3.5), however, more satellites would be required for reliable RAIM operations. Further, the U.S. Department of Defense has guaranteed 24 satellite coverage for 70% of the time. Consequently, by having access to additional GPS ranging signals the issue of "availability" (for positioning and RAIM) can be addressed. As only 37 C/A PRN codes are reserved for the GPS satellites, and there are 1024 possible C/A codes, these unassigned codes can be used by other transmitters. Inmarsat has already made available transponders on its third generation communications satellites for the transmission of navigation signals. Other satellites could also transmit GPS-like signals. Different PRN codes would be assigned to these satellites, and as far as the user hardware is concerned, they would be indistinguishable from the GPS satellite signals. Accuracy and integrity are expected to be enhanced as well.

A specialised development that could benefit aviation users is the *pseudolite* (or "pseudo-satellite").

The concept has now been developed to such a degree that pseudolites are being aggressively promoted as the best means to satisfy the requirements of precision automatic aircraft landings. However, there are other applications of pseudolites that are of greater interest to land navigation. For example, a pseudolite could be used to improve coverage in urban areas. Of added utility is the ability of including differential correction data within the broadcast messages, thus supporting DGPS operations (or even RTK, if pseudolite carrier phase data is included in the transmission) without the need for dedicated communications links or RTCM-capable GPS receivers.

2.3.2 *Non-GPS Satellite-Based Navigation Systems*

The Russian Federation's Global Navigation Satellite System (GLONASS) was developed for the Russian military, and is at present the only satellite-based positioning system which is a natural competitor to GPS. GLONASS has the following characteristics ([5,6,7]):

- 21 satellites + 3 active spares (figure 2.5).
- 3 planes, 8 satellites per plane.
- 64.8° inclination, 19100 km altitude (11hr 15min period).
- Dual-frequency (L1 in the range: 1597-1617 MHz; L2 in the range: 1240-1260 MHz).
- Each satellite transmits a different frequency on L1 (=1602 + Kx0.5625 MHz; $K\in[-7,24]$) and L2 (=1246 + Kx0.4375 MHz; $K\in[-7,24]$).
- Spread-spectrum PRN code signal structure.
- Global coverage for navigation based on simultaneous pseudo-ranges, with an autonomous positioning accuracy of better than 20m horizontal, 95% of the time.
- A different datum and time reference system to GPS.
- There is a PPS and a SPS, as in the case of GPS.
- No Selective Availability is implemented.

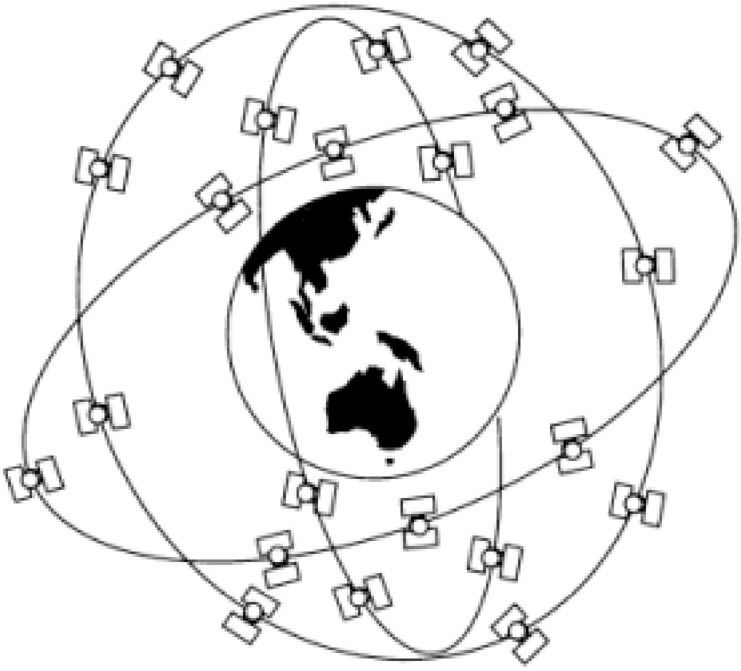

Figure 2.5: The GLONASS Satellite Constellation.

Although some of the characteristics of GLONASS are very similar to GPS, there are nevertheless significant technical differences. In addition, the level of maturity of the user receiver technology and the institutional capability necessary to support the GLONASS space and control segment are significantly less than in the case of GPS. GLONASS will continue to be viewed by many user communities as a technically inferior system to GPS, a system concerning which there are many question-marks regarding its long-term viability. This uncertainty is stifling much needed market investment in new generation receiver hardware. Yet to dismiss GLONASS as a serious candidate for a 21st century satellite positioning technology because it cannot *compete* with GPS technology is too simplistic an analysis. Although GLONASS has the potential to rival GPS in coverage and accuracy, this potential is unlikely to be realised in the medium term, and hence for the foreseeable future GLONASS should be considered a *complementary* system to GPS. GLONASS was declared operational (with 24 satellites in orbit) in 1996.

The combination of GPS and GLONASS as part of a Global Navigational Satellite System (GNSS) is being actively promoted in version GNSS-1 of the European EGNOS. Although there are GLONASS-only receivers available on the market, these are generally inferior to GPS products. However, there is a distinct trend to develop receivers that can track and process signals from *both* the GPS and GLONASS satellites. One of the first commercial GPS+GLONASS systems is the Ashtech GG24 receiver. A combined GPS+GLONASS receiver can track signals from a 48 satellite constellation, twice as many as the GPS-only constellation and therefore significantly improving *availability*. For example, simulation studies have shown that with a 45° obstruction to half the sky (as would be caused by a tall building), five or more GPS satellites are only available for about 33% of the day, and four or more satellites for about 85% of the day. However, there is 100% availability of five or more satellites when both GPS and GLONASS satellites are considered.

GLONASS positioning results are of higher accuracy than GPS because no Selective Availability is implemented. The improvements in positioning accuracy based on GLONASS pseudo-range data compared with GPS (with SA on) is evident from figure 2.6.

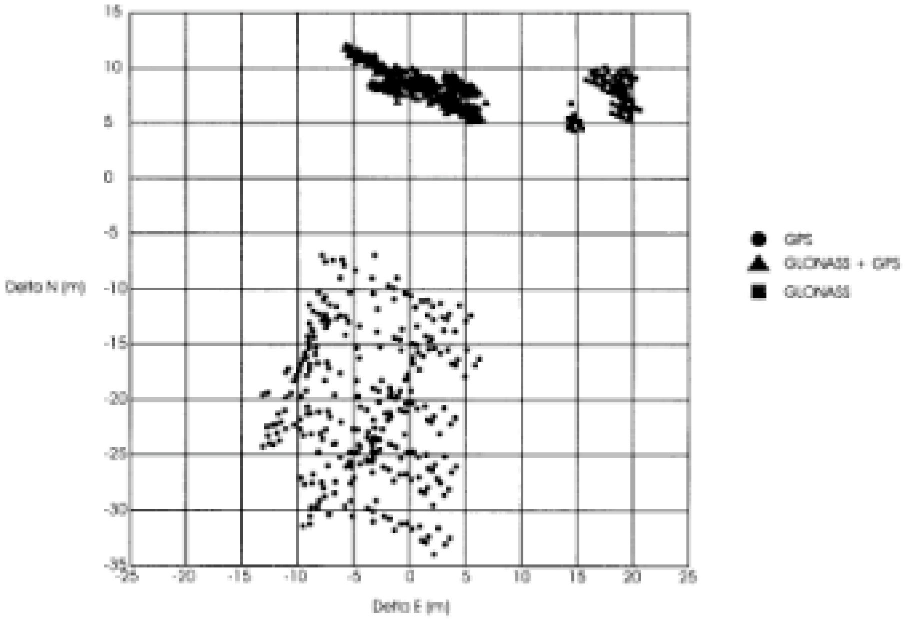

Figure 2.6: Typical Positioning Results of GLONASS and GPS.

Autonomous (single receiver) horizontal accuracy at the 95% confidence level is therefore significantly enhanced in the case of GLONASS-only (20m) and GPS+GLONASS (16m) receivers, compared to GPS-only systems (100m). Differential accuracies are quoted as being of the order of 1m and 75cm for differential GLONASS (DGLONASS) and DGPS+DGLONASS receivers, respectively. DGLONASS is implemented through special messages within the RTCM differential correction transmission format.

With falling prices for GPS+GLONASS receivers (though they are never likely to be as inexpensive as GPS-only receivers), many more applications will be addressed without the need to implement differential positioning techniques. Remember, in urban "canyon" environments satellite signals are severely restricted, and the ability to track both GPS and GLONASS satellite signals means that the chances that valid positioning can be carried out (based on the availability of four or more satellites) are increased. Hence, in the context of many positioning applications, GLONASS can be viewed as a GPS augmentation, rather than as an alternative positioning technology.

Although the first GNSS is likely to be based on some combination of GPS and GLONASS, the issues of who develops, who controls, and who finances subsequent generations of GNSS are still to be addressed in detail though it is expected that several scenarios will emerge over the next few years. It seems likely that satellite communications companies will be in the best position to also address the positioning-navigation market. *One scenario has been suggested by Inmarsat* ([10]).

The Inmarsat navigation transponders referred to earlier, although intended to provide an augmentation to GPS, could be the first step on a road leading eventually to a fully independent, internationally owned, civilian satellite navigation system. Such a development will be important in many countries that do not have a close relationship with the U.S.A. For such countries, the issue of being dependent on a foreign-controlled system, is an issue that has to be addressed.

A four-step evolution from Inmarsat’s current supplementary role in satellite navigation (as is proposed in European GNSS-1) to the provision of a completely independent, civil GNSS has been defined. The navigation payload to be carried on the Inmarsat-3 satellites is just the first step. The Inmarsat-3 satellites will augment GPS by providing ground-derived system integrity information, additional GPS-lookalike ranging signals and WADGPS corrections, that is, the full WAAS-like service (although this is being primarily championed by the European Union). Despite the improvements Inmarsat’s overlay will bring, the organisation is counting on further augmentation being required in the longer term. This may involve additional navigation payloads to its future generation(s) of communications satellites, both the traditional geostationary satellites and the new low-earth-orbiting systems.

Such a civilian system is feasible from a technical point of view, but as Inmarsat concedes, the really big questions are institutional (who would run such a system?) and financial (how would it be financed?). If the current misgivings concerning sole-nation ownership that are felt about GPS are to be addressed, the answer to the first question will involve some sort of international cooperative body such as Inmarsat. Clearly the capital and running costs of a GNSS will be enormous, and this must remain the primary hurdle to the establishment of a civil GNSS, not only for the Inmarsat organisation, but any other navigation service provider competing against a GPS service that levies no user charges. Nevertheless, it is possible that a satellite communications company may see opportunities for generating income from value-added services such as positioning, and move to satisfy the market before Inmarsat can obtain permission from its constituent members to proceed with its own plans. *It must be emphasised that such developments will not only impact civil aviation users, but will affect ALL satellite navigation applications.*

2.3.3 *Examples of Software Enhancements*

There are many enhancements that can be made to GPS data processing, some relatively minor, some significant (and perhaps patentable), though most enhancements can be considered in a hierarchical manner: from the straightforward processing of pseudo-range data, through levels of complexity for the estimation algorithms which may involve the introduction of "exotic" additional data types. Four examples which may be relevant to many applications are:

- Clock or height-aided position solution.
- Carrier phase smoothing of pseudo-range data.
- Processing of carrier phase data in kinematic mode.
- Kalman filter algorithms for GPS data processing.

It must be emphasised, however, that to ensure high quality GPS positioning results using the techniques referred to above, it may be necessary to also upgrade the GPS hardware (for example, to use carrier phase tracking receivers), as well as to implement stringent data collection procedures (for example, the most effective way of improving accuracy is for the GPS antenna to be static for periods of several minutes or more!).

Clock-Aiding and Height-Aiding

Clock-aiding refers to the process by which we assume that the receiver clock offset from GPS Time is not an entirely unknown parameter. *Height-aiding* refers to the technique by which we can assume that the height of the receiver is known. There are in fact two ways in which these enhancements can be implemented: with or without extra hardware.

With no extra hardware (and hence extra observations), information on the receiver clock offset can be included as an extra constraint within the pseudo-range solution. This constraint can be considered a "pseudo-observation", that is, the clock offset can be input as an observable (let us refer to this as "option 1"), or it can be considered a known quantity which does not need to be estimated ("option 2") in the standard four satellite 4-D estimation procedure (Section 1.3.5). The solution becomes stronger because: (a) there is one extra observation to estimate the same number of parameters (option 1), or (b) there is one less parameter to be estimated using the same number of pseudo-range observations (option 2). Where does the receiver clock error estimate come from? If the receiver clock were of good enough quality such that the clock bias were highly predictable for a short time into the future, then once the bias and bias-rate were determined in the conventional way, the estimated clock bias at some future time could be assumed accurate enough not to warrant estimation on an epoch-by-epoch basis. However, according to the investigations by [11], the standard quartz crystal clock cannot satisfy this role, but an oven-controlled crystal oscillator would.

A similar approach can be applied to height. *Most navigation applications involve 2-D (that is, horizontal) positioning*. If the hardware, such as a barometer, were available an extra observation of height could be added to the solution. (This would have a similar effect to adding an extra satellite to the constellation being tracked.) Alternatively, once the height had been estimated, it could be assumed to not change in value for some time into the future, and hence removed from the estimable parameter set. In fact, many GPS receivers have this option when less than four satellites are visible (particularly when there is significant signal shading, as would be experienced in urban "canyons").

Using Carrier Phase Data to Smooth Pseudo-Range Data

One way of overcoming the two main problems associated with pseudo-range data, that is: (a) the high measurement noise, and (b) the greater multipath disturbance in comparison to carrier phase data, is to create a pseudo-range / carrier phase combination that, in effect, "smooths" the pseudo-range data. The basis of all data smoothing techniques is to derive the rate-of-change of range from the carrier phase data, and to combine this with the absolute measurement of range provided by the pseudo-range data.

An early implementation of a GPS data smoothing technique was described by [12], making use of dual-frequency phase and pseudo-range data. Alternative smoothing algorithms have been developed which use Doppler data in place of carrier phase data. Furthermore, all smoothing algorithms are also applicable to single frequency data, though of course such techniques are inferior to the dual-frequency techniques because they cannot account for the ionospheric bias. Many GPS receivers nowadays use such carrier-smoothed data in the standard navigation solution..

Ambiguity Resolution "On-the-Fly"

The basis of modern precise GPS positioning techniques is the ability to convert the data processing problem from one that addresses the carrier phase observation model in eqn (2.10) to one that solves the much simpler problem posed by pseudo-range data (eqn 2.9). This is achieved through the mathematical process of **ambiguity resolution**, or the determination of the integer values of the double-differenced ambiguity parameters. Ambiguity resolution is therefore the process by which a precise, but ambiguous, carrier phase observation is converted to an unambiguous range quantity, having all the advantages of a pseudo-range observation but with much lower measurement noise.

Techniques for achieving this have been progressively refined until it is now possible to do this even while the receiver is in motion (that is, in the so-called "on-the-fly" mode), within a period generally of the order of seconds to minutes. The ambiguity resolution procedure consists of several stages:

(1) Definition of the apriori values of the ambiguity parameters.

(2) Operation of a search algorithm to identify likely integer values.

(3) Implementation of a decision-making algorithm to select the "best" set of integer values.

(4) Application of the ambiguities to the data to create unambiguous range measurements.

Ideally the first three steps can occur transparently to the user, with minor delay and with high reliability. Several techniques have been developed to address steps (2) and (3), including:

- The Fast Ambiguity Resolution Approach (FARA).
- The Cholesky Decomposition based search technique.
- Spectral Decomposition based search technique.
- The Least Squares Ambiguity Search Technique.
- The Fast Ambiguity Search Filter (FASF) technique.
- The Least-squares AMBiguity Decorrelation Adjustment (LAMBDA) technique.
- The Ambiguity Function Method.
- Direct estimation from a combination of phase and pseudo-range data.

None of the above procedures are 100% foolproof, and hence it is still necessary to have favourable operational conditions in order to maximise the chances of achieving the correct results:

- Keep the distance between the two receivers short! *Generally less than 20km.*
- Track as many satellites as possible and ensure good satellite geometry. *Preferably five or more.*
- Use dual-frequency observations. *Can use clever combinations of data on the two frequencies.*
- Use precise pseudo-range data. *Helps give very good apriori positioning information.*
- Improve ambiguity search, selection and testing algorithms. *It is here that much research is being concentrated.*

The "holy grail" of precise GPS positioning is to achieve instantaneous ambiguity resolution (that is, with just one epoch of data). Although significant progress has been made, it is still not a routine and reliable process. *Furthermore, it requires "state-of-the-art" GPS hardware and hence cannot be considered viable for many "standard" GPS positioning applications.*

Kalman Filter Algorithms

The use of Kalman filters for GPS data processing is a growing trend, *however is it just a market gimmick?*

The standard Least Squares estimation technique is typically used when the estimation problem is "over-determined" or, in other words, when there are more observations than required to estimate the position parameters (Section 2.1.3). In kinematic applications Least Squares procedures can be applied to data on an "epoch-by-epoch" basis. However, the parameters of interest (the position), and the dominant system errors (for example, the clock or atmospheric refraction errors), are time-varying quantities. In addition, the time variation is more or less predictable. For such applications, the data processing techniques that are the most efficient and optimal, and therefore the most appropriate, are those based on the extension of the principles of Least Squares to encompass the concepts of *prediction, filtering* and *smoothing*.

The three concepts of prediction, filtering and smoothing are closely related and are best illustrated through an example, in this case a moving vehicle for which the parameters of interest are its instantaneous position at some time t. The process of computing the vehicle's position in real-time (that is, observations are taken at time t_k, and position results are required at t_k) is referred to as *filtering*. The computation of the expected position of the vehicle at some subsequent time t_k, based on the last measurements at t_{k-1} is properly termed *prediction*, while the estimation of where the vehicle was (say at time t_k), once all the measurements are post-processed to time t_{k+1}, is referred to as *smoothing*.

Although the three procedures are separate, and can be applied independently, they may also be applied *sequentially* :

- The **prediction step**: based on *past positioning information together with a kinematic model*, the expected position and its precision at the next epoch of measurement is computed. The kinematic model is composed, as is the measurement model, of functional and stochastic components. Thus *four* models must be considered and, given a particular application and a certain data type, the filter design process is therefore one of selecting the appropriate models.
- The **adjustment** or **filtering step**: is a classical adjustment, except that a fairly good apriori estimate of the parameters is already provided from the prediction step. Basically, the resulting parameter estimates are *weighted combinations of predicted quantities and measurement data*. The Kalman filter is a particular form of the generalised Least Squares filter.
- The **smoothing step**: by which *all the measurements are reprocessed* after the last measurement has been made and the filtering step has been completed.

As indicated above, the implementation of the filter requires the specification of the stochastic and mathematical models for both the *measurement system* and the *system dynamics*. Once the mathematical and stochastic models have been defined, the implementation within a Kalman filter is, in principle, relatively straightforward, although there exist several different implementations which may have different advantages from the computational, numerical stability or quality control point-of-view. The reader is referred to such classic texts as [13,14] for details.

Kalman filtering techniques are particularly suited for GPS navigation because:

- Standard Least Squares procedures treats each measurement epoch independently, and hence does not use information on the system dynamics, such as the motion of the vehicle to which the GPS receiver is attached.
- Permits the rigorous computation of precision and reliability measures such as error ellipses and "marginally detectable errors".
- The Kalman filter is also central to many "quality control" or "fault detection" procedures which can be implemented in real-time in order to *detect* "failure" (where poor quality data is introduced into the process, or where there is an error in the measurement or system dynamics models), to then *identify* the source of error, and to then *adapt* (or *recover*) the system to ensure that the results are not biased due to this system "failure".
- Estimate small biases that affect the data over many epochs. For example, many measurement biases in modern navigation technology have the signature of *drifts* which are not apparent at the single epoch level -- because they appear as "noise" in standard epoch-by-epoch Least Squares solutions.
- By taking into account information on system dynamics, such as the regular motion of the GPS receiver, it is possible to carry out position estimation even if there is insufficient data -- for example, when only two satellites are visible.
- Is an optimal linear estimator in the presence of Gaussian white noise.
- A Kalman filter can accept data as and when it is measured, and does not have to be "reduced" to some specified epoch.
- The Kalman filter is well suited to the mixing, or "fusion", of various data types (including from non-GPS sensors).

Some GPS receivers incorporate Kalman filters as the navigation computing algorithm, but their real utility is generally only obvious when the positioning system involves several sensors such as when GPS is integrated with Dead Reckoning sensors. However, Kalman filters are not "magical" procedures, because if the input data is of questionable quality, or there is an error in the assumptions regarding the model for the system dynamics, the positioning results will still be seriously biased.

2.3.4 *Hardware Augmentation: GPS and Other Sensors*

For many vehicle navigation applications GPS is insufficient as a stand-alone positioning system, particularly in urban environments, because of satellite signal degradation and obstruction. It is for such reasons that many positioning systems are based on a combination of several technologies. The integration of GPS with Dead Reckoning (DR) sensors would appear to be ideal for supporting vehicle positioning in ITS applications, because they are complementary systems and can output continuous position information to the required accuracies of urban vehicle navigation ([15,16]). However, the DR sensors do add to the overall costs of the navigation hardware, and hence such systems are unlikely to be installed in *all* vehicles.

The principle of a DR system is the relative position fixing method, which requires knowledge of the location of a vehicle and its subsequent speed and direction (for example, the last position and velocity determination before GPS signal interruption) in order to calculate its present position. A typical DR system therefore comprises distance and heading sensors. Such a system can only give the two-dimensional position of a vehicle (although more sophisticated DR systems may include altitude sensors or inclinometers which can provide the three-dimensional position of the vehicle). However, because of unfavourable error accumulation (a small error in heading, grows over time into a large error in position), frequent calibration is required. It is in this context that GPS is integrated with DR systems. That is, the DR sensors provide information on relative position (relative to a starting location), but GPS-derived coordinates are used to determine the DR sensor errors, which may be fed back into the navigation computer.

The sensors that are favoured for GPS-DR systems in ITS applications may consist of some or all of the following ([15]):

- An **odometer** is a distance sensor which may be mounted singly or in pairs onto either the wheel or the transmission of the vehicle. Odometers are prone to errors due to wheel slippage and changes in wheel circumference due to tyre pressure and velocity changes. Their accuracy is typically of the order of 0.3-2% of the distance travelled.
- A **magnetic compass** that measures the heading of the vehicle. The most popular electronic compass technology for land vehicle applications uses the *fluxgate* principle. Empirical tests with fluxgate compasses in urban environments has shown that, because of their sensitivity to external magnetic field disturbances such as bridges, railway tracks, overpasses, etc., a standard error of about 2-4° is expected.
- A **tilt sensor** that gives information about the pitch and roll angles of the vehicle may involve one or more inclinometers. Such a sensor is comparatively expensive, but its accuracy is of the order of 0.1°.
- A **gyroscope** measures the rate-of-change of heading of the vehicle. The *fibre optic gyros* have a drift rate of the order of 1-10°/hour, but the lower cost (and more widely used) *vibrational and solid-state gyros* exhibit poor bias and scale factor stability, and hence require almost continuous calibration.
- **Digital maps** can be used to relate mathematical coordinates to locations on street segments and intersections ([16]). In turn, the stored coordinates of the map features provide a means of navigating in coordinate space, and hence allow the digital map to contribute to the navigation function. The accuracy requirements for digital maps must be high enough to ensure that the vehicle does not appear to be navigating a neighbouring street!

2.3.5 *DR Sensor Selection for ITS Applications*

A generic vehicle navigation system architecture which incorporates GPS and DR sensors is illustrated in Figure 2.7.

There is a trade-off in accuracy requirements of the DR sensors relative to the primary navigation device, the GPS receiver ([17]). The more accurate the DR sub-systems, the longer the duration of the maximum tolerable GPS outage. The sensors that comprise the DR sub-system can be classified as either sources of heading or velocity information. The vehicle's odometer is the most attractive option for obtaining velocity information because it's free! The odometer, once calibrated, provides good long-term stability, and any effects due to tire pressure variation can be accounted for by real-time calibration from GPS. The only other velocity option is the use of low-cost accelerometers, however, the accelerometer bias must be removed initially and periodically recalibrated.

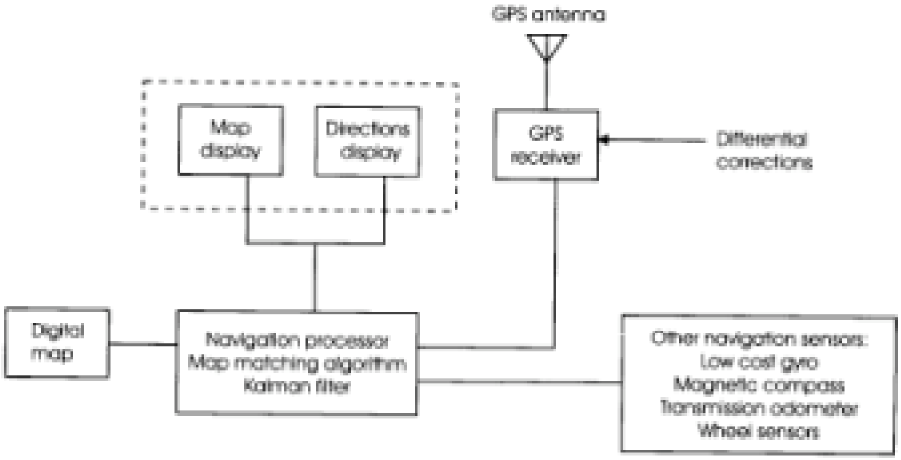

Figure 2.7: Generic GPS-DR Vehicle Navigation System (after [10]).

Potential sources of heading information include magnetic compasses, accelerometers, two-wheel odometers, and gyros. Magnetic compasses are an attractive low-cost sensor but require calibration to remove the effects of local magnetic disturbances. These require that the raw heading information be filtered against other sensors, in particular GPS. Gyros provide only heading-rate information, hence the other sensor(s) is (are) needed for heading initialisation. Several different types of gyros are possible, but the most common are the vibrational gyros, though fibre optic gyros promise to become competitive as their cost decreases.

The central problem in integrating GPS and DR sensors is the design of the data processing algorithm. This is invariably a form of Kalman filter. There are two options: (a) a loose integration or coupling where some prior processing is carried out in sensor-specific filters, or (b) a tightly integrated implementation in which all observations are processed simultaneously. The most common filter used for GPS-DR products is the loosely coupled filter. Because of the non-homogeneous types of sensors (they all invariably come from different manufacturers) and their relatively low cost, each sensor will usually have its own filter. *Fusion* of the outputs from each sensor is then performed within a master filter. Note that this means that the GPS "observations" as passed to the master filter are position, velocity and time (PVT) (not the pseudo-range or carrier phase measurements). This type of approach calibrates the local sensor biases and scale factors, as well as yielding a globally optimal solution for the vehicle's PVT and heading, along with accuracy and reliability measures for quality control.

New inertial sensors (gyroscopes and accelerometers) are predominantly developed by the defence and aerospace industry, with R&D being market driven. The tendency is not towards more accurate systems, but cheaper and smaller systems which can be easily integrated with GPS.

References

[1] WELLS, D.E., BECK, N., DELIKARAOGLOU, D., KLEUSBERG, A., KRAKIWSKY, E.J., LACHAPELKLE, G., LANGLEY, R.B., NAKIBOGLU, M., SCHWARZ, K.P., TRANQUILLA, J.M. & VANICEK, P., 1987. **Guide to GPS Positioning**. 2nd. ed. Canadian GPS Associates, Fredericton, New Brunswick, Canada, 600pp.
[2] KLOBUCHAR, J.A., 1991. Ionospheric effects on GPS. **GPS World, 2(4)**, 48-51.
[3] BRUNNER, F.K. & WELSCH, W.M., 1993. Effect of the troposphere on GPS measurements. **GPS World, 4(1)**, 42-51.
[4] LANGLEY, R.B., 1991c. The mathematics of GPS. **GPS World, 2(7)**, 45-50.
[5] KLEUSBERG, A., 1990. Comparing GPS and GLONASS. **GPS World, 1(6)**, 52-54.

[6] IVANOV, N.E. & SALISTCHEV, V., 1991. GLONASS and GPS: Prospects for a partnership. **GPS World, 2(4)**, 36-40.

[7] LANGLEY, R.B., 1994. RTCM SC-104 DGPS standards. **GPS World, 5(5)**, 48-53.

[8] MUELLER, T., 1994. Wide Area Differential GPS. **GPS World, 5(6)**, 36-44.

[9] THOMSON, S., 1996. The future for commercial DGPS. **The Hydrographic Journal, 82**, 3-8.

[10] KAPLAN, E. (ed.), 1996. **Understanding GPS: Principles & Applications**. Artech House Publishers, Boston London, 554pp.

[11] MISRA, P.N., 1996. The role of the clock in a GPS receiver. **GPS World, 7(4)**, 60-66.

[12] HATCH, R.R., 1982. The synergism of GPS code and carrier measurements. Proc. 3rd Int. Symp. on "Satellite Doppler Positioning", New Mexico, 8-12 February, 1982, 1213-1231.

[13] GELB, A., (ed.), 1974. **Applied Optimal Estimation**. MIT Press, Cambridge, Mass., 374pp.

[14] MINKLER, G. & MINKLER, J., 1993. **Theory and Application of Kalman Filtering**. Magellan Book Company, Palm Bay, Florida, USA.

[15] KRAKIWSKY, E.J. & McLELLAN, J.F., 1995. Making GPS even better with auxiliary devices. **GPS World, 6(3)**, 46-53.

[16] KRAKIWSKY, E.J. & BULLOCK, J.B., 1994. Digital road data: putting GPS on the map. ***GPS World***, 5(5), 43-46.

[17] GEIER, G.J., HESHMATI, A., McLAIN, P., JOHNSON, K. & MURPHY, M., 1993. Integration of GPS with Dead Reckoning for vehicle tracking applications. Proc. 49th Annual Meeting of the U.S. Inst. of Navigation, Cambridge, MA, June 21-23, 75-82.

Chapter 3
Mapping Issues

Central to almost all navigation and positioning tasks is the notion that a *map coordinate* is of little value without reference to recognisable features on the surface of the earth. For example, a mariner may know where the vessel is, and have the coordinates of the journey's end, but in order to navigate the route safely, the location of potential dangers must also be known, as well as of ports of haven, designated shipping lanes, restricted waters, etc. In the case of a land vehicle the coordinates of the vehicle's present location or of the destination may be of little use to the average driver who is not expected to be trained in traditional navigation skills, and may probably be largely ignorant of mathematical coordinate systems. The driver would prefer locations to be expressed in terms of an "address", such as a house number and street name. Furthermore, because the vehicle is constrained to travel along roads, the route to be followed is most appropriately depicted in terms of a turn-by-turn reference to a road map.

The road map is a compact, graphical representation of the essential spatial information that a driver needs to negotiate a journey to a new location and, in the context of Intelligent Transport Systems (ITS), is the *interface* between the driver and the positioning technology being used. But there are many other ways in which maps can be used within ITS. For example, although the in-vehicle map may assist in answering such questions as 'where am I?', a vehicle tracking application will have a map at the central tracking facility to answer questions such as 'where are you?', 'how do you get to location A?', 'which vehicle is in the best position to render aid at location B?', etc. A compact, informative road map is only the end product of a long and complex map-making process whose principles and procedures have been refined over centuries. In this chapter we introduce the mapping issues which are relevant to, and underpin, the geographic framework within which spatial applications must operate. These map issues include:

- The definition of the fundamental reference datum to which position is referred.
- The variety of ways in which coordinates can be expressed.
- The transformation procedures between different coordinate systems and datum frameworks.
- The characteristics of modern satellite datums.
- The processes by which the real world is mapped, and the related mapping accuracy standards.
- The issue of map data update and maintenance, and the expanded range of map information now required to support a variety of user applications.
- The display of spatial data in various projection systems.
- The variety of ways in which map data is made available, on demand, in electronic form to ITS users.
- How map data can be used to aid navigation and other functions.

The chapter therefore touches on the disciplines of geodesy, surveying and cartography. Although no "catch-all" word has been universally agreed to which describes all the modern map production and display operations, in the last few years the word *Geomatics* has been coined to cover all spatially related activities, the traditional disciplines mentioned above as well as those that have emerged recently and which are closely identified with the geospatial technologies such as GPS, GIS (Geographic Information Systems) and Remote Sensing. In fact, the positioning aspects of ITS would be included under the discipline of "geomatics" (or “geoinformatics” as it is also known), and many "geomatics" departments at universities do carry out research into GPS technology and its applications, and therefore are making contributions to ITS.

3.1 Datums and Coordinate Systems

3.1.1 *Introduction to Geodesy: The Figure of the Earth*

As is conceded in [1,2], the meaning and relevance of "geodesy" is largely a mystery to the general public. 'What is geodesy?' 'Who needs it and why?' Are therefore questions which must be answered.

Geodesy, from the Greek, literally means "dividing the earth", and has as its first practical objective the provision of an accurate framework for the control of national topographical surveys, and hence is the foundation of a nation's maps. To do this, geodesy must first define the basic geometrical and physical properties of the figure of the earth. The scientific objective of geodesy has therefore always been to determine the *size, shape and gravitational field of the earth.* Geodesy has traced the development of our understanding of the earth from the times of the flat earth concept, through the sphere and spheroid, to the geoid. In the process, the technology of *measurement* and *computation* of position has undergone tremendous change, from the knotted rope for the measurement of distance, and the telescope to measure angles, to presentday electronic systems and orbiting satellites.

During the last quarter of the 20th century, geodesy has become closely associated with *accurate positioning*. So many applications now involve accurate position determination -- whether it be oil platforms located many kilometres offshore, support of space missions, measurement and monitoring of manmade and natural structures, laser bathymetry for chart production, inertial measurement systems, gravity mapping, seismic surveys, etc. -- that geodesy is now an indispensable tool of the scientist and engineer. If one considers the entire spectrum of positioning accuracy for *real-world* applications, most geographic/mapping applications would be clustered around the low to medium accuracy end of the scale. On the other hand, "geodetic positioning" would be at the top-end of the scale. "GPS geodesy", for example, is the application of the GPS technology for the definition of sub-centimetre accurate positions, on continental and global scales, to support such applications as the measurement of crustal motion!

However, "geodesy" is not just an esoteric activity concerned with measuring the snail's pace motion of the continents. Geodetic principles underpin our system of map coordinates, hence it is necessary to review our knowledge of the shape of the earth and consider the impact that it has had on the definition of the mathematical framework in which we refer *latitude* and *longitude*.

The "figure of the earth" has been a central concern of philosophers and mathematicians since the dawn of civilisation. Early humans were increasing their knowledge of the planet simply by observing nature, and the motion of the sun and the stars in the sky. The first clues as to the shape of the earth would have been rather obvious. For example, an observer of maritime activity would have noticed that a distant ship disappears from view lower part first, with the top of the mast the last part to vanish ([2]). Stargazers in the northern hemisphere would have noticed that as they travelled in an east or west direction the star now recognised as the Pole Star (and around which the signs of the Zodiac revolved) stayed more or less at the same elevation angle in the sky, whereas if they went in a north or south direction the elevation angle would change. Travellers would have noticed that the length of their shadow at midday changed as they travelled north or south, whereas (over periods of a few days) the shadows were of constant length. Gradually these clues contributed to the development of ideas of the earth's shape, from flat, through a disc or cylinder, to a number of variations on these. All textbooks on geodesy (for example [3,4]), as well as introductory geodetic monographs (such as [1,2]), describe the evolution in our understanding of the earth's shape.

By the time of Pythagoras (ca 580-500 B.C.) the earth was considered to be a spherical shape -- and hence a most perfect figure! First attempts at determining the dimension of the sphere are credited to Aristotle (ca 384-322 B.C.) with 400000 stades (a figure which could vary from 84000 to 63000 km depending on the choice of scale factor), but his method is unknown. A century later Archimedes estimated the circumference as 300000 stades (63000 - 47000 km), though he may have used a different length of the "stade". Although several other "guesstimates" were made by various Greek philosophers, the first account of an explicit experiment to deduce the circumference of the earth from actual measurements is credited to be that used by Eratosthenes, by a method which is still valid today (figure 3.1). Eratosthenes's estimate of 250000 stades (52500 - 39400 km) is close to the present day accepted value of around 40000 km.

For an approach to the determination of the size of the earth which has the greatest scientific significance we must turn to ancient Egypt, and the Greek philosopher Eratosthenes, librarian at the famous library of Alexandria. The size of a sphere can be found if two quantities are known: (a) the distance between two points that lie on the same meridian of longitude, and (b) the angle subtended by these two points at the centre of the earth. The circumference is 360°s/α, where s and α are the two measured quantities. *But how to measure α?*

The problem was solved rather elegantly. Eratosthenes observed that on the day of the summer solstice, the midday sunshone to the bottom of a well in the town of Syene (now Aswan) (Figure 3.1). At the same time he noted that the sun was not overhead at Alexandria, but cast a shadow equivalent to 1/50th of a circle (or about 7°12'). He then made several fortuitous assumptions: (a) that Syene was on the Tropic of Cancer (in order for the sun to shine directly into a vertical shaft on the summer solstice), (b) the linear distance between Syene and Alexandria was 5000 stades, and (c) Alexandria and Syene lay on the same meridian.

Although the experiment design is sound, the observations and assumptions are full of errors: Syene is not on the Tropic of Cancer, but around 60km to the north of it; Syene and Alexandria are several degrees from being on the same meridian; their distance apart was some 10% in error; the difference in latitude between Alexandria and Syene is 7°5' (not 7°12'); and there is an uncertainty in the calibration of the "stade". It is remarkable that the derived value for the circumference (52500 - 39500km) is so close to the accepted value of about 40000km.

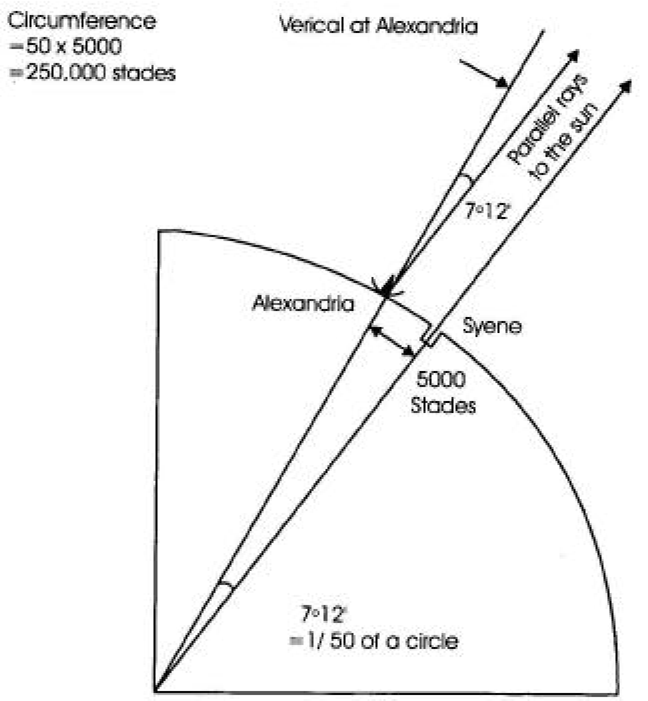

Figure 3.1: Eratosthenes' Method for Determining the Size of the Earth.

Another ancient measurement of the size of the earth was made, a century later, by the Greek, Poseidonius. His value of 240000 stades, obtained using an independent method, was comparable to that of Eratosthenes. However, a revised value of 180000 stades was the one promulgated by the great geographer Ptolemy (100-178 A.D.). The world maps of Ptolemy strongly influenced cartographers up to the middle ages. Furthermore, this too-small a value of the earth's circumference had an unexpected impact on world history! By using the wrong conversion factor, it appeared to Columbus that Asia was only some 4000 miles (6400km) west of Europe! It was not until the 15th century that Ptolemy's figure for the shape of the earth was revised upwards.

The 16th and 17th centuries were the turning point in the understanding and application of many aspects of science, including geodesy. Developments in instrumentation (such as the telescope, vernier, thermometer and barometer), and in computing techniques (with the aid of logarithmic and trigonometric tables) provided new tools for studying the shape and size of the earth. Of interest is the method of arc measurements using the technique of *triangulation* (Section 3.2.1). In this method, the distance between two terminal points could be deduced indirectly, rather than by direct linear measurement. All that was needed was the measurement of a precise baseline distance, and then the coordinates of points could be determined through computation, by using telescope measurements of the angles of a series of triangles formed by the geodetic control points which run from one terminal point of a *network* to the other terminal point.

By the end of the 16th and in the early years of the 17th century, several arcs were measured in France. Under the guidance of the Cassini family, a continuous north-south arc of triangles was measured from Dunkirk south to the Spanish border. The measured arc was divided into two parts, one northward of Paris, another southward. When Cassini computed the length of a degree of arc from both chains, he found that the length of one degree in the northern part of the chain was shorter than that in the southern part. This result could only be caused by an egg-shaped earth (figure 3.2). Almost at the same time, scientific expeditions were measuring the oscillations of pendulums at various places around the (then known) world. These results, together with the theories of Newton and Huygens, suggested that the earth must be flattened at the poles. The results started an intense controversy between French and English scientists which was finally settled in favour of the "figure of the earth" predicted by Newton. *Since all the computations involved in a geodetic survey technique such as triangulation are carried out on a mathematical surface that must closely resemble the shape of the earth, the findings were very important*.

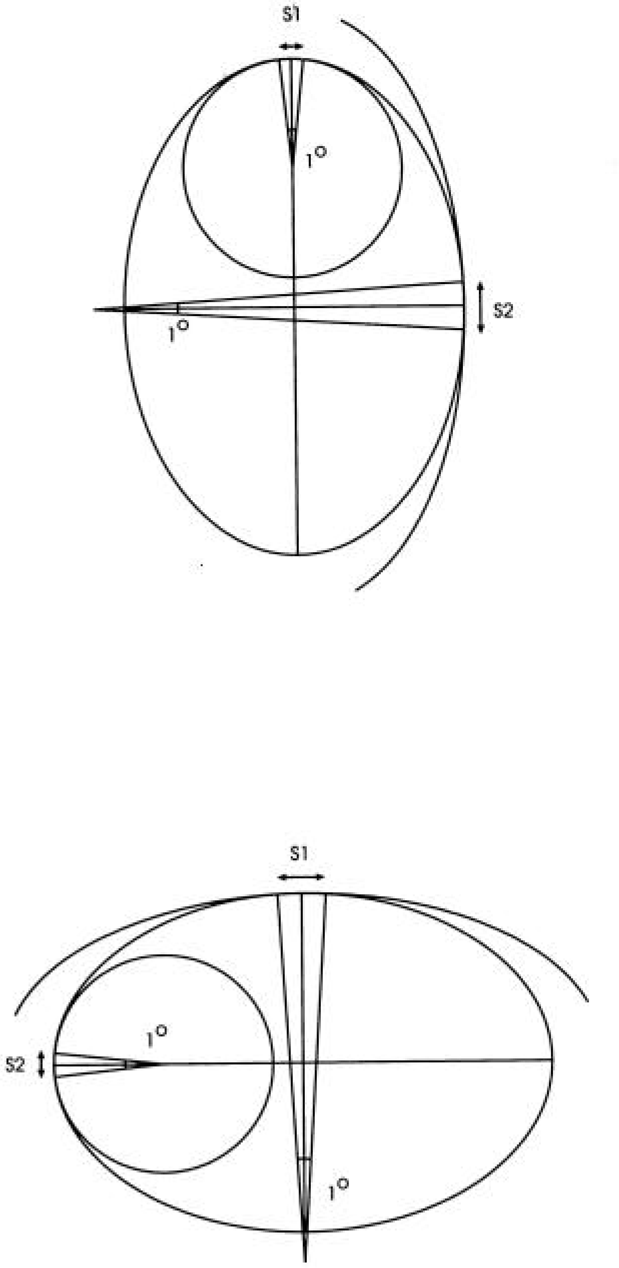

Figure 3.2: Oblate and Prolate Spheroids or Ellipsoids.

The contest between the French and English models of the earth was a scientific controversy par excellence, or as described by some, the pumpkin versus the egg contest. The technical names given to the rival shapes are "oblate" spheroid for flattening at the poles, and "þrolate" spheroid for flattening at the equator (Figure 3.2). For any given angular value the equivalent arc length will be longer at the equator for a prolate spheroid and longer at the poles for an oblate spheroid.

To settle the controversy, once and for all, the French Academy of Science sent a geodetic expedition to Peru in 1735 to measure the length of a meridian degree close to the equator, and another to Lapland to make a similar measurement near the Arctic Circle. The measurements conclusively proved the earth to be flattened -- an oblate spheroid -- as Newton had predicted. The two mean radii computed from the experimental results were 6376.45km and 6355.88km, or in terms of the length of 1° of arc, a difference of only 350m in 111km. It is not surprising that the errors in the equipment of the Cassinis swamped the small amount they were trying to resolve.

In the two hundred years since these classic expeditions there have been many other attempts to determine the difference in radii of curvature (and hence the "flattening" of the spheroid of best-fit to the earth) using essentially the same "meridian-arc" procedure. The bewildering variety of reference spheroids that have been used since the 18th century is due to distance and angle measurement error. For example, in mountainous areas, the direction of the plumbline (from which the astronomic latitude is determined) will not be in the direction of the centre of mass of the earth, but will deviate because of the nearby mountain masses. This will result in different apparent flattening values for the best-fitting earth spheroid.

The expression "figure of the earth" has various meanings in geodesy, according to the purpose to which it is used and the level of precision sought for expressing relative positions. Hence although the oblate spheroid is one such figure, the actual topographic surface is another candidate -- it is the surface on which actual measurements are made. However, the surface is highly variable and is therefore not suitable for mathematical computations without adopting certain generalisations. The idea of a flat earth is still acceptable for small areas where the curvature of the earth can be neglected. For example, a map of a city could be produced with high accuracy assuming that the earth were a plane surface the size of the city. A spherical approximation may suffice for larger areas. However, for the highest precision an oblate spheroid is used.

The oblate spheroid has its north-south radius slightly less (some 22 km, compared to an earth radius of almost 6400 km) than that in the east-west direction at the equator. Figure 3.3 is a depiction of a section of the earth spheroid, with the flattening grossly exaggerated. An alternative term used to define this reference surface is the "ellipsoid of revolution". The spheroid is generated by rotating an ellipse about its semi-minor axis, so that all meridianal sections are ellipses, and all sections taken perpendicular to the axis of rotation are circles. It remains to define the two parameters of the spheroid (or ellipsoid): (a) the length of the semi-major axis, and (b) the length of the semi-minor axis or the flattening or the eccentricity.

So how big should the spheroid be? Those parts which have the most regular surfaces are the 70% of the earth which is covered by oceans -- in practice the *mean sea level*, to average out the time-varying component of the vertical motion of the ocean tides. If it were imagined that we could continue this mean sea level surface under the continents we arrive at another surface of fundamental importance to geodesy, the **geoid**. The geoid is an equipotential surface of the earth's gravitational field which on average coincides with mean sea level. It is therefore another "figure of the earth", whose surface departs from a "best-fitting" (in the geometric sense) spheroid by up to 100m. However, this surface is too irregular to be used as the surface upon which geodetic computations are made. But it is the generally accepted datum surface for heights, that is, *height above sea level!*

Many spheroids have been used over the last few centuries to support geodetic work in various parts of the world ([3,4]). Variations in the size and shape of the spheroid can be attributed to the results of different experiments to determine the best-fitting spheroid surface to a portion of the earth's surface. International or global spheroids are relatively recent concepts.

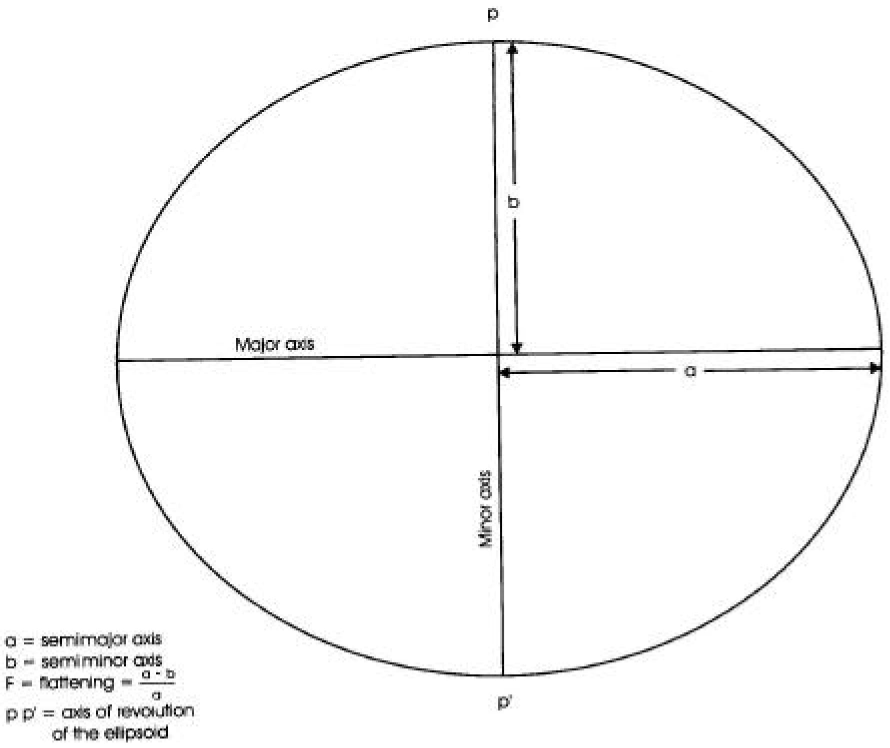

Figure 3.3: Elements of an Ellipse.

3.1.2 *Geodetic Datums*

The oblate spheroid may be constrained so that its centre is located at the earth's centre of mass -- the so-called **geocentre** -- and the semi-minor axis is, for all intents and purposes, coincident with the earth's rotation axis. The only parameters which may vary are the two which define the size (the semi-major axis) and the shape (semi-minor axis, flattening, or eccentricity), and these are selected to be a best-fit to the geoid *on a global basis*. In 1979 the spheroid known as the Geodetic Reference System 1980 (GRS80) was approved and adopted at the congress of the International Union of Geodesy and Geophysics as the global "figure of the earth". GRS80 is also the basis of the World Geodetic System 1984 (WGS84) and the International Terrestrial Reference System (ITRS) datums (see Sections 3.1.4 and 3.1.5). Its semi-major axis is 6378137m, and its flattening is 1/298.257223563[1]. *In this chapter we will use the terminology "reference ellipsoid" to designate that spheroid which is the basis of a nation's geodetic datum*.

The constraining of the location and orientation of the reference ellipsoid is, in fact, a form of *datum definition* ([3]). In practice, national datums in the past were defined in a more complex manner. Because there was no geodetic technique that could locate the earth's centre of mass, the centre of the reference ellipsoid is generally not at the geocentre but located arbitrarily so that it is a best-fit to the geoid across the region of interest (typically the land surface across which the datum is to be used). Instead of defining the datum origin in terms of the amount the centre of the reference ellipsoid is offset from the geocentre, the datum origin was often defined in terms of the geodetic coordinates of one special ground station -- the so-called *origin* or *datum station*. (The reader is referred to standard geodesy textbooks such as [3] and [4] for further information on datum definition.)

The launch of the first artificial earth-orbiting satellite ushered in a new era in which: (a) datums could be defined on a global basis, and (b) the relationship between different geodetic datums could be determined. GPS is merely the latest in a series of space-based positioning technologies which have been used for this purpose. The differences between satellite datums and traditional local geodetic datums are:

- The origin of satellite datums is nominally the geocentre (the point about which all earth satellites orbit), while for local geodetic datum it is the origin station. *Consequently the local geodetic datum is not geocentric*.
- The satellite datum is realised by a sparse network of tracking stations, and is accessed through satellite ephemerides (time-varying coordinates of the satellite). The local geodetic datums are realised by a dense network of control marks, and the datum is accessed by making measurements to/from these control marks.
- In the case of local geodetic datums the horizontal datum is typically independently defined from the vertical datum. *Satellite datums are three-dimensional in nature*.
- Coordinates in the satellite datum are expressed in the Cartesian system, while geodetic (or ellipsoid) coordinates are the preferred system for local geodetic datums.
- Satellite datums have global relevance, while local geodetic datums are only used over a relatively small portion of the globe.

Traditional local datums in many countries are presently undergoing a process of redefinition, to make them compatible to new positioning technologies such as the Global Positioning System. GPS allows us to determine our position anywhere on or near the earth's surface in a single global satellite datum; that of WGS84.

Specifying the size, shape, location and orientation of the reference ellipsoid suffices as a definition of the datum, but this is not sufficient to make it *accessible* to users. The primary purpose of a datum is to provide the means by which the horizontal locations of points can be defined in both mathematical terms (by coordinates) as well as graphically (on maps). This is accomplished by establishing connections to the datum at many points within a *geodetic control network*. This requirement is the same whether we consider traditional local geodetic datums or geocentric satellite datums.

In the traditional (pre-satellite) methodology the datum was propagated out from the origin station, to the geodetic control network using geodetic surveying techniques (Section 3.2.1). These techniques actually relate to horizontal position on the ellipsoid, the latitude and longitude of a point. A national datum is therefore *realised* by thousands of physical control marks (of varying accuracy). To *access* the datum, a surveyor or navigator merely has to go to the nearest control mark, define the reference azimuth and then propagate this datum information to any other points of interest through a combination of linear and angular measurement. In many respects the same procedure is used for a satellite datum. However, there may not be a single origin station, but several *fundamental* reference stations whose coordinates in the satellite datum are known to very high accuracy, having been established, for example, by "GPS geodesy" techniques. From this small number of precisely located physical benchmarks, denser networks are established using geodetic surveying techniques, which nowadays are almost exclusively based on GPS surveys.

3.1.3 *Datum and Coordinate Transformations*

The coordinate of a point in relation to a geodetic datum is typically expressed in one of two systems: (a) the ellipsoidal coordinate system, or (b) the Cartesian coordinate system. The ellipsoidal coordinate components are latitude (φ), longitude (λ) and height above the reference ellipsoid (h). Although the latitude and longitude are curvilinear components, they relate to horizontal position on the surface of an earth-model, while the height component is measured in relation to the normal to the ellipsoid at the point. It is a straightforward matter to convert the ellipsoidal coordinate into its Cartesian equivalent form (see figure 3.4).

(There are similar formulae for the conversion of ellipsoidal and Cartesian coordinates to other coordinate representations such as topocentric and projection coordinate systems.)

In addition to converting coordinates from one mathematical system to another, there is often the need to relate two geodetic datums. This was a rare occurrence in the past because local geodetic datums did not overlap, and hence there was no need to transform coordinates from one datum to another. With a global datum such as that used for GPS positions there is now often a need to relate GPS-derived coordinates to the local geodetic datum (for example, the system in which the map data is referenced), as the two may not be coincident. There are two options:

- Either change the geodetic datum so that it is coincident with the GPS datum (Section 3.1.4), or
- use a transformation model to relate coordinates in one datum to those on another datum.

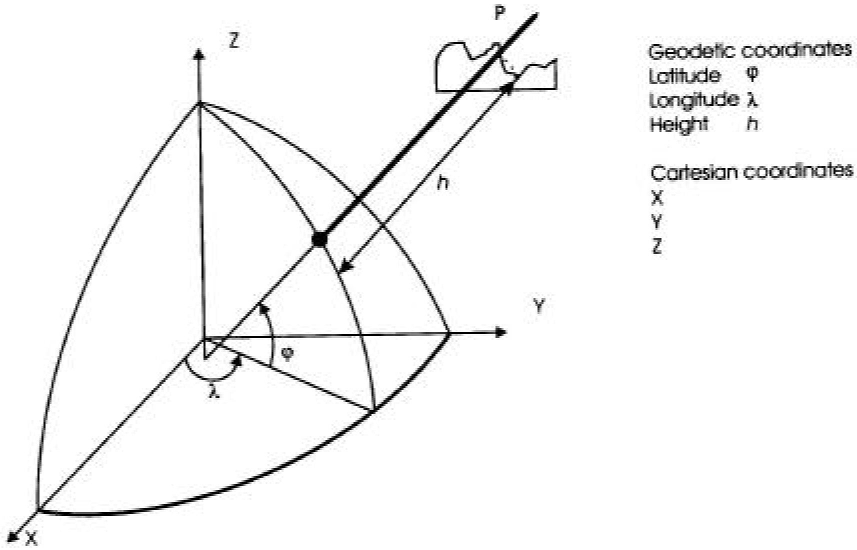

Figure 3.4: The Cartesian and Ellipsoidal Coordinate Systems.

The first option is being exercised by many countries which see that an expansion of the use of GPS, for a host of applications, results in greater efficiencies if there was no need to transform GPS coordinates *before* putting them to use. For example, the Australian datum is being redefined so that coordinates expressed in terms of this datum will be within one metre of WGS84 (and hence for most navigation users there will be no requirement for any transformation). The new datum is known as the Geocentric Datum of Australia ([5]). North America, and many other countries, have already adopted geocentric datums. The main obstacle to such a datum redefinition is when a large body of coordinates have already been generated on an earlier non-geocentric datum -- for example, the coordinates of blocks of land -- and to change the datum would introduce massive confusion. In this case the only option is to use a transformation model.

There are a number of ways of defining the relationship between one reference system and another. The most general of the transformations is the *affine transformation*. An affine transformation transforms straight lines to straight lines and parallel lines remain parallel. Generally the size, shape, position, and orientation of lines will change. The scale factor depends on the orientation of the line but not on its position. Hence the lengths of all lines in a certain direction are multiplied by the same scalar. Alternatively it is possible to define a *projection transformation* where the scale factor is also a function of position.

A transformation in which the scale factor is the same in all directions is called a *similarity transformation*, and is by far the most widely used of the transformation models. A similarity transformation preserves shape, so angles will not change, but the lengths of lines and the position of points may change. An *orthogonal transformation* is a similarity transformation in which the scale factor is unity. In this case the angles and distances within the network will not change, but the positions of points do change on transformation.

The similarity transformation model operates on 3-D Cartesian coordinates, and requires the definition of seven parameters: one scale factor, three translation parameters (representing the offset of the centre of one datum origin with respect to that of the other datum), and three rotation parameters (representing the orientation angles relating the Cartesian axes of one datum to those of the other datum). *The parameters relating the WGS84 datum to most of the world's geodetic datums are stored within GPS receivers, so that the user has the option of presenting the coordinate results in the datum of choice.* Transformations are discussed further in Section 3.3.2.

The Similarity Transformation Model

The 3-D similarity transformation model relating coordinates of a network of points in the $X_BY_BZ_B$ datum to coordinates in the $X_AY_AZ_A$ datum is (figure 3.5):

$$\begin{pmatrix} X_B \\ Y_B \\ Z_B \end{pmatrix} = s\ .\ \mathbf{R}\ .\ \begin{pmatrix} X_A \\ Y_A \\ Z_A \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} \qquad (3.1)$$

where s is the scale factor and $\mathbf{R}$ is a $3x3$ orthogonal rotation matrix (equation (3.3)). Note that there are seven parameters: three rotation angles, three translation components and one scale factor. The translation terms T_x, T_y, T_z are the coordinates of the origin of the $X_AY_AZ_A$ network in the frame of the $X_BY_BZ_B$ network.

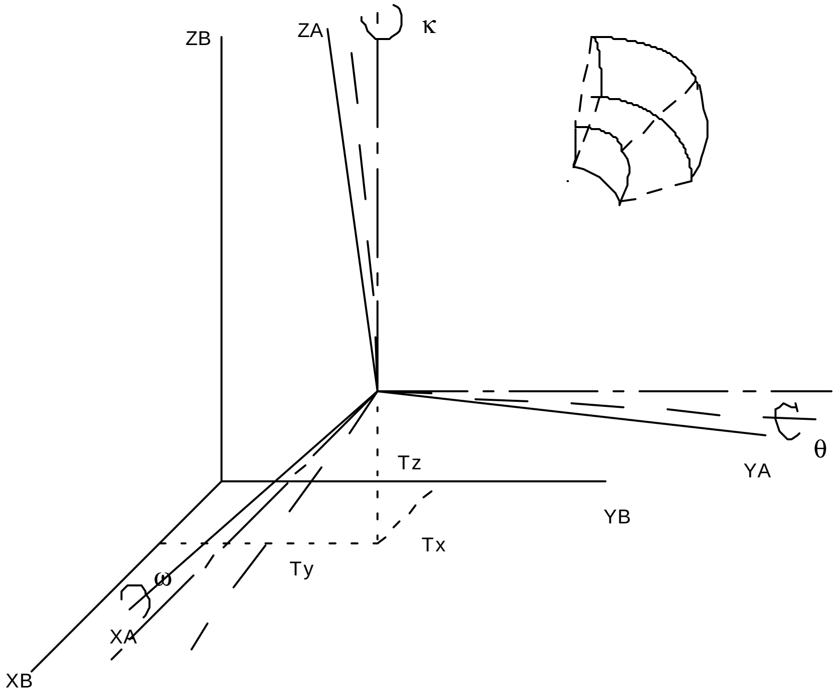

Figure 3.5: The Seven Parameter 3-D Similarity Transformation Model.

The transformation model defined in equation (3.1) is often referred to as the *Bursa-Wolf model*. When this model is invoked for small areas the rotation parameters are highly correlated with the translation parameters. (The reader can convince themselves of this by considering, for example, a rotation about the Z-axis of a point on the Greenwich meridian, the effect of which is almost indistinguishable from a translation of the area along the Y-axis.) An alternative formulation that avoids this correlation "problem" is the *Molodensky-Badekas model* ([3,4]).

The rotation matrices about the X-, Y-, and Z-axes are:

$$\mathbf{R_z}(\kappa) = \begin{pmatrix} \cos\kappa & \sin\kappa & 0 \\ -\sin\kappa & \cos\kappa & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbf{R_y}(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix} \quad \mathbf{R_x}(\omega) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\omega & \sin\omega \\ 0 & -\sin\omega & \cos\omega \end{pmatrix} \tag{3.2}$$

The most common combined rotation matrix is: $\mathbf{R} = \mathbf{R_z}(\kappa)\,.\mathbf{R_y}(\theta)\,.\mathbf{R_x}(\omega)$, leading to:

$$\mathbf{R} = \begin{pmatrix} \cos\kappa\cos\theta & \cos\kappa\sin\theta\sin\omega+\sin\kappa\cos\omega & \sin\kappa\sin\omega-\cos\kappa\sin\theta\cos\omega \\ -\sin\kappa\cos\theta & \cos\kappa\cos\omega-\sin\kappa\sin\theta\sin\omega & \sin\kappa\sin\theta\cos\omega+\cos\kappa\sin\omega \\ \sin\theta & -\cos\theta\sin\omega & \cos\theta\cos\omega \end{pmatrix} \tag{3.3}$$

For small rotations this matrix may be approximated by:

$$\mathbf{R} \approx \begin{pmatrix} 1 & \kappa & -\theta \\ -\kappa & 1 & \omega \\ \theta & -\omega & 1 \end{pmatrix} \tag{3.4}$$

where ω, θ, and κ are the rotation angles in radians about the X-, Y-, and Z-axes respectively. The small angles assumption is usually valid for rotation angles up to 10". The rotation angles depend on the baseline vectors (that is, the relative positions) and not on the absolute coordinates.

A scale factor can be visualised as follows. Imagine a network of points drawn on the surface of an inflatable sphere. As the sphere is inflated, the points of the network spread apart from each other, and from the centre of the sphere (figure 3.6). The inflation of the sphere is equivalent to the application of a scale factor greater than unity. Multiplication of a set of rectangular Cartesian coordinates by a scale factor is identical to multiplying the corresponding baseline lengths by the same scale factor. Hence the scale factor can be determined from either the 3-D site coordinates or from the baseline lengths. Thus, as with the rotation angles, the origin of the coordinates has no affect on the results. In the case of ellipsoidal coordinates the longitude is not affected by a scale change but the geodetic latitude does change slightly. For example, a one part-per-million (ppm) scale change will change geodetic latitudes by less than about 0.0007" (2cm). However, the effect on ellipsoidal height is significant. For example, a 1ppm scale change will produce a change in height of about 6.4 metres.

Examples of seven-parameter similarity transformation models are given in table 3.1. Row 1 is an example of a transformation model between a local geodetic datum (in this case the Australian Geodetic Datum) and the WGS84 satellite datum. Note the large values of the origin offset parameters. Row 2 are the parameters relating two satellite datums, defined using different satellite tracking technology and data processing procedures: GPS and Satellite Laser Ranging. Note the very close agreement. Row 3 are the parameters relating two realisations of the International Terrestrial Reference System (Section 3.1.5).

Table 3.1: Transformation Parameters Between Various Geodetic Datums.

Coordinate system (datum)	T_x[a] (m)	T_y[a] (m)	T_z[a] (m)	ω[b] (")	θ[b] (")	κ[b] (")	ds[c] (ppm)
WGS84 -> AGD84	116.00	50.47	-141.69	+0.23	+0.39	+0.34	-0.098
BTS87 -> WGS84	+0.071	-0.509	-0.166	-0.0179	+0.0005	+0.0067	-0.017
ITRF93 -> ITRF94	+0.021	+0.001	-0.001	+0.0013	+0.0009	+0.0005	+0.0002

[a] origin offsets [b] rotations about the x-, y-, z-axes [c] scale difference from unity

Although the seven-parameter similarity transformation model is the most common of the transformation models, there are simplified models such as defining only the three "block shift" origin shift parameters T_x, T_y, T_z. This may be adequate for lower accuracy applications. Alternatively, the shift parameters for the horizontal ellipsoid components may be used ($\Delta\phi$, $\Delta\lambda$). This is discussed further in Section 3.3.2.

Eqn (3.1) relates sets of Cartesian coordinates. There are also formulae that relate two sets of ellipsoidal coordinates. In addition to accounting for the differences in scale, origin and orientation between the two datums, they must also relate the results on different spheroids (due to possible differences in the size and shape of the spheroids). These formulae are collectively referred to as the **Molodensky Formulae**. They are particularly useful for 2-D transformations as only the effect on latitude and longitude need be considered. The *abridged* Molodensky Formulae (accounting for only differences in origin and differences in the size/shape of the reference ellipsoids), giving the corrections, in arcseconds, to the ellipsoidal coordinates of points on datum A, to transform them to datum B, are:

$$\Delta\phi = (-T_x.\sin\phi.\cos\lambda - T_y.\sin\phi.\sin\lambda + T_z.\cos\phi + (a.\Delta f + \Delta a.f).\sin 2\phi)$$
$$/ \ (R_M.\sin(1 \text{ second}))$$

$$(3.5)$$

$$\Delta\lambda = (-T_x.\sin\lambda + T_y.\cos\lambda) / (R_N.\cos\phi.\sin(1 \text{ second}))$$

where: a is the semi-major axis of the datum A ellipsoid
b is the semi-minor axis of the datum A ellipsoid
T_x, T_y, T_z are the offsets of the datum A ellipsoid relative to the datum B ellipsoid origin
Δa is the change in semi-major axis (from datum A ellipsoid to datum B ellipsoid)
Δf is the change in flattening (from datum A ellipsoid to datum B ellipsoid)
$R_N = \dfrac{a}{\sqrt{1-e^2\sin^2\phi}}$ where e is the eccentricity of the datum A ellipsoid
$R_M = \dfrac{a(1-e^2)}{(1-e^2\sin^2\phi)^{1.5}}$

For example, in the case of the transformation from the Australian datum AGD84 to WGS84, the values of T_x, T_y, T_z are -116, -50, 142 metres (Table 3.1). The values of Δa, Δf are -23m and -0.000000081204 (see sections 3.1.2 and 3.1.4)

3.1.4 *The WGS84 Satellite Datum*

Unlike local geodetic datums, which are essentially defined by parameters associated with a single "origin" terrestrial station, satellite datums are defined by a combination of: (a) *physical models* such as the adopted model of the earth's gravity field, gravitational constant of the earth, the rotation rate of the earth, the velocity of light, etc.; and (b) *geometric models* such as the adopted coordinates of the satellite tracking stations used in the orbit determination procedure, and the models for precession, nutation, polar motion, and earth rotation, that relate the celestial reference system (in which the satellite's ephemeris is computed) to the earth-fixed reference system (in which the tracking station coordinates are expressed). Such a datum has the following characteristics:

- It is geocentric, because the geocentre is the physical point about which the satellite orbits.
- It is generally defined as a Cartesian system (although a reference ellipsoid is often also defined), with axes oriented close to the principle axes of rotation ("z-axis") and the intersection of the Greenwich Meridian plane and the equatorial plane ("x-axis"), with the "y-axis" forming a right-handed system.
- There are a number of different satellite datums, each associated with different satellite tracking technology (for example, Satellite Laser Ranging, Transit Doppler, etc.) and different combinations of gravity field model, earth orientation model, and tracking station coordinates used for orbit computation.

The World Geodetic System 84 (WGS84) is one such satellite datum, defined and maintained by the U.S. National Imagery and Mapping Agency (NIMA) (formerly the Defense Mapping Agency) as a *global geodetic datum* ([6]). It is the datum to which all GPS positioning information is referred by virtue of being the reference system of the broadcast ephemeris. The realisation of the WGS84 satellite datum is the catalogue of coordinates of over 1500 geodetic stations (most of them either active or past tracking stations) around the world. They fulfil the same function as national geodetic stations do, that is, they provide the means by which a position can be related to a datum. WGS84 is an earth-centred Cartesian coordinate system fixed to the surface of the earth such that (see figure 3.5):

- The Z-axis is aligned parallel to the direction of the Conventional Terrestrial Pole (CTP)[2] for polar motion, as defined by the International Earth Rotation Service (IERS).
- The X-axis being the intersection of the WGS84 Reference Meridian Plane and the plane of the CTP Equator (the Reference Meridian being parallel to the Zero Meridian defined by the IERS).

The defining parameters of the WGS84 reference ellipsoid are:

- Semi-major axis: 6378137m
- Ellipsoid flattening: 1/298.257223563[3]
- Angular velocity of the earth: 7292115 x 10^{-11} rad/sec
- The earth's gravitational constant (atmosphere included): 3986005 x 10^{-8} m^3/sec^2

It's relationship to other global datums and local geodetic datums has been determined empirically, and transformation parameters of varying quality have been derived (row 1, table 3.1). Reference systems are periodically redefined, for a number of reasons, such as when the primary tracking technology improves (for example when the Transit Doppler system was superseded by GPS), or if the configuration of ground stations alters radically enough to justify a recomputation of the global datum coordinates. The result is generally a small refinement in the datum definition, and a change in the numerical values of the coordinates. For example, prior to January 1987, the satellite datum in use for GPS, and other navigation systems, was WGS72. In 1994 the GPS reference system underwent a subtle change to WGS84(G730) to bring it into the same system as used by the International GPS Service to produce precise GPS satellite ephemerides (see Section 3.1.5 and [7]).

However, with dramatically increasing tracking accuracies another phenomenon impacts on datum definition and maintenance: the motion of the tectonic plates across the earth's surface, or "continental drift" as it is commonly referred. (It is assumed there is comparatively little vertical motion.) This motion is measured in centimetres per year, with the fastest rates being over 10cm/year.

3.1.5 *The International Terrestrial Reference System*

The WGS84 system is the most widely used global reference system because it is the system in which the GPS coordinate results are expressed. Other satellite reference systems have been defined but these have mostly been for "scientific" purposes. However, since the mid 1980's, geodesists have been using GPS to measure crustal motion, and to define more precise satellite datums. The latter were essentially by-products of the sophisticated data processing, which included the re-computation of the GPS satellite orbits. These GPS surveys required coordinated tracking by GPS receivers spread over a wide region during the period of the GPS "campaign". Little interest was shown in these alternative datums until:

- The network of tracking stations spread across the global, rather than being concentrated about the region of interest of the GPS campaign.
- The network of tracking stations was maintained on a permanent basis, rather than operated intermittently.
- The scientific community initiated a project to define and maintain a datum at the highest level of accuracy.
- The policy of Selective Availability was announced (Chapter 2), whereby the WGS84 datum as realised by the broadcast GPS orbit could be corrupted.

In 1991, the International Association of Geodesy decided to establish the International GPS Service (IGS) to promote and support activities such as the maintenance of a permanent network of GPS tracking stations, and the continuous computation of the satellite orbits and ground station coordinates ([7]). Both of these were preconditions to the definition and maintenance of a new satellite datum independently of WGSS84, defined by the U.S. National Imagery and Mapping Agency network and the Control Segment monitor station network (used to provide the data for the operational computation of the GPS broadcast ephemerides). After a test campaign in 1992, routine activities commenced at the beginning of 1994. The network is an international collaborative activity consisting of about 50 core tracking stations located around the world, supplemented by more than 200 other stations (some continuously operating, others only tracking on an intermittent basis). The precise orbits of the GPS satellites are available over the Internet, from the IGS.

The definition of the reference system in which the coordinates of the tracking stations are expressed, and periodically re-determined, is the responsibility of the International Earth Rotation Service. The reference system realisation is known as the International Terrestrial Reference Frame (ITRF), and its definition and maintenance is dependent on a suitable combination of Satellite Laser Ranging, Very Long Baseline Interferometry and GPS coordinate results. (However, increasingly it is the GPS system that is providing most of the data.) Each year a new combination of precise tracking results is performed, and the resulting datum is referred to as ITRFxx, where "xx" is the year. A further characteristic that sets the ITRF series of datums apart from the WGS, is the definition not only of the station coordinates, but also their *velocities* due to continental and regional tectonic motion. Hence, it is possible to determine station coordinates within the datum, say ITRF98, at some "epoch" such as January 1st 1999, by applying the velocity information and predicting the coordinates of the station at any time into the future (or the past).

Such ITRF datums, initially dedicated to geodynamical applications requiring the highest possible precision, have been used increasingly as the fundamental basis for the redefinition of many national geodetic datums. For example, the new Australian datum, known as the Geocentric Datum of Australia is a realisation of *ITRF92 at epoch 1994.0* at a large number of control stations ([5]). Of course other countries are free to chose any of the ITRF datums (it is usually the latest), and define any epoch for their national datum (the year of the GPS survey, or some date in the future, such as the year 2000.0). *Only if both the ITRF datum and epoch are the same, can it be claimed that two countries have the same geodetic datum*. However, differences in such datums can still be accommodated through similarity transformation models (Section 3.1.3).

3.2 Spatial Data Capture

Once we have adopted a mathematical model for the "figure of the earth", the next problem is to answer the question 'where am I on the earth?' or 'where on earth am I?' Of course these questions could be answered simply with reference to some obvious physical object or feature, for example, 'so many kilometres, in such a direction, or along the river', or 'the 2nd house, on Hillsdale Street, after the church', and so on. However, we will assume that the answer is required to be given in some coordinate form. In that case, where you are, or where you want to go, where some point of interest is located, must be defined with reference to a datum and a coordinate system. Yet for most people, the map is the most widely used device for answering the aforementioned questions, because it permits both *qualitative* (through the graphical representation) as well as *quantitative* (through coordinates) interpretation.

Maps have always played a special role in history. The earliest maps are likely to have been very ephemeral in nature. They may have been simply a series of lines scratched in the sand. However, early civilisations were using maps for many of the same reasons we use them today: to depict the topographic form of territory, as an aid to navigation, to mark dangers, as an inventory of national assets, as evidence of ownership of land areas, to wage war, for the collection of taxes, to assist in development, to display spatial relationships, etc. What has varied over time has been the accuracy of the maps, the methods of compilation, and the format of the map display.

3.2.1 *The Survey Operation*

The ingredients needed to make a map are: (a) a datum, (b) a projection, (c) data capture techniques, (d) standards and specifications. The first has already been dealt with in earlier sections. Projections are discussed in Section 3.3.1. In this section, the data capture techniques are discussed, while the issue of standards and specifications relevant to map-making is touched on in the next section.

What do we mean by spatial data? "Spatial data" can be simply defined as the *geometric parameters* that define the position of points on the earth's surface, or near it (below ground, on the ocean bottom, or in the atmosphere or beyond). Spatial data capture is therefore concerned with the determination of the coordinates of points, and has traditionally been the preserve of surveyors. If we restrict ourselves to horizontal position determination, there are a number of techniques which can be used ([1,2]): Positional astronomy, Triangulation, Traversing, Trilateration, Satellite positioning techniques, Inertial Positioning Systems, and Photogrammetry.

Until the last few centuries latitude was far easier to determine than was longitude. While the latter had to await the invention of good chronometers, its was relatively straightforward to get latitude by simple stellar observations. Although used for marine navigation, and later by explorers in uncharted regions of the world, the single isolated fixes by astronomical methods were of little value to the surveyor because they could not be inter-related. However, when these astronomically determined positions were connected by other means they were very valuable. Geodesists used astronomic methods to define local geodetic datums (section 3.1.2), as well as to control the growth of relative position errors in networks.

The propagation of the datum, and the primary means of determining precise relative coordinates of a network of control points, for several centuries was by means of *triangulation* (figure 3.7). A network of points was established on hilltops, and their coordinates determined by the principles of trigonometry using the following data: (a) the given coordinate of one point (the origin station), (b) the azimuth of one line (one end of which was the known point), (c) the distance between two points in the network, and (d) measurement of the included angles within the triangles formed by the points in the network. The angles were measured using a theodolite, and because the points had to be visible from at least two other points, the spacing of these points was likely to be of the order of a few tens of kilometres at most. Once these primary points were established, a similar technique could be used to "intersect" any other point of interest.

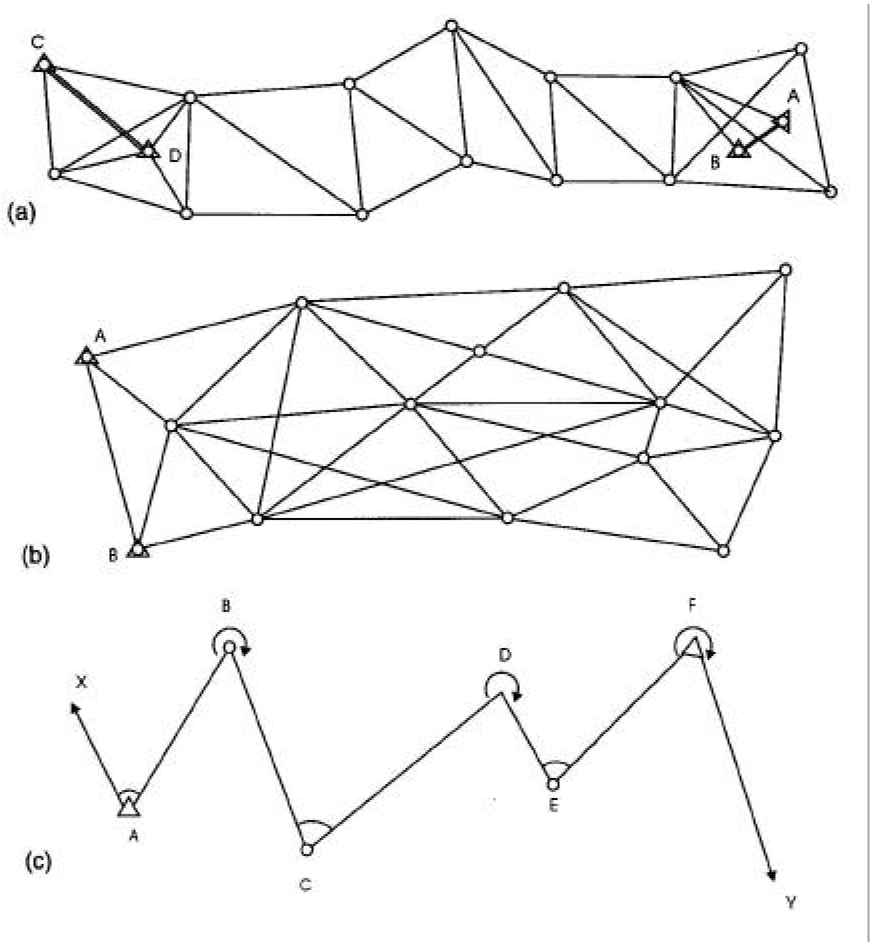

Figure 3.6: The Principles of (a) Triangulation, (b) Trilateration and (c) Traversing.

Triangulation techniques were favoured by surveyors because angular measurement using a theodolite was a far easier (and more accurate) measurement to make than distance. Herculean efforts were expended simply to measure occasional "baselines" of about 10 km length using specially manufactured wooden or metal bars. These bars (around 2m in length) had to be carefully handled, and laboriously placed end-to-end in order to span the distances required. In the 1960's, with the advent of both the Tellurometer and Geodimeter instruments, the measurement technology changed radically ([3,4]). Both of these were first generation Electronic Distance Measurement (EDM) instruments -- the Tellurometer was a microwave instrument, and the Geodimeter was an optical (laser) instrument -- which measured a distance by observing the time elapsed for the electro-magnetic signal to travel from the transmitter out to a reflector station (which could be tens of kilometres away), and back. Nowadays EDM instruments are the mainstay of the surveyors' toolkit, and are integrated within the theodolite to provide a highly refined instrument for positioning.

Almost overnight the technique of triangulation was superseded by the techniques of traversing and trilateration (figure 3.6). *Trilateration* is similar to triangulation except that instead of measuring the angles within the chain of triangles, the distances are measured. The technique of *Traversing*, on the other hand, requires that both distances and angles are measured. A characteristic of all three of these terrestrial methods is that the stations whose coordinates are being determined must be visible from the measuring stations, and therefore the coordinates which are so derived are all *relative* quantities.

In 1967 the first satellite-based positioning system, the *Transit Doppler system*, was made available to the general public. In addition to providing a navigation service, special observing and data processing techniques were developed so that it could provide geodetic positioning accuracies. Transit was able to provide *absolute* (that is, with respect to the geocentre) as well as relative coordinates. Absolute positioning accuracies at the sub-metre level, and relative positioning accuracies at the few decimetre level were possible. Transit has now been superseded[4] by the GPS system (Chapter 1 and 2), *which is superior to it in almost every respect except one: GPS absolute positioning accuracy is at the several dekametre (ten-metre) level at best*. GPS when used in the relative positioning mode -- requiring the determination of the coordinates of one GPS receiver in relation to another GPS receiver located at a point of known location -- is capable of high accuracy: at the few metre level when implemented using pseudo-range data, to sub-centimetre accuracy when carrier phase data is processed (see Section 2.2.2).

A technique which does not require inter-station visibility, and does not rely on measurements of external signals, but which can still determine relative position to a high accuracy is *inertial surveying*. The principle is simple, and makes use of the gyroscopic principle to sense the change in direction of the inertial platform as it moves from point to point. Once it has been initialised on a point of known coordinates, the trajectory is measured continuously through the integration of the readings on three gyroscopes and three accelerometers, fixed orthogonal to each other. The technology is of course well known in navigation as "dead reckoning". The differences are that the instrumentation is more accurate and that special observation procedures are employed to control the effect of sensor drift errors. However, the instrumentation is very expensive and not used extensively these days.

A technique which is quite like no other positioning technology is *photogrammetry*. This technique was originally developed for the purpose of making maps from aerial photographs and differs from all other "spatial data capture" methods in that it can coordinate, in one process, a large number of ground points. Nowadays photogrammetric techniques are used for virtually all map-making, from very small scales (likely to involve satellite remote sensed images) to very large scales. Almost all road maps are produced by photogrammetric means.

Photogrammetry is defined by the American Society of Photogrammetry ([8,9]) as the "art, science and technology of obtaining reliable information about physical objects and the environment through processes of recording, measuring, and interpreting photographic images and pattern of recorded radiant electromagnetic energy and other phenomena". Photographs are still the principal source of information, and included within the domain of photogrammetry are two distinct processes: (1) *metric* photogrammetry, and (2) *interpretative* photogrammetry. Metric photogrammetry consists of making precise measurements from photographs to determine the relative location of points. This enables distances, angles, areas, elevations, sizes and shapes of objects to be determined. The most common application of metric photogrammetry is the preparation of planimetric or topographic maps from aerial photographs.

The airplane was first used in 1913 for obtaining photographs for mapping purposes. In the period between the two World Wars, aerial photogrammetry for mapping evolved rapidly, and in the last 50 years improvements in instrumentation and techniques have made photogrammetry so accurate, efficient, and advantageous that at the present time, except for mapping small land parcels, very little mapping is done by other means. The principle of metric photogrammetry (figure 3.7) is to carry out measurements on a pair of photographs which have an overlap area which appears in both photographs. Through a mechanical process (but nowadays through computation) the location and orientation of the two camera locations are determined in a model, which includes the terrain in its correct 3-D (scaled) visualisation. The horizontal (planimetric) and vertical coordinates can be scaled off this virtual model (viewed stereoscopically through special optics).

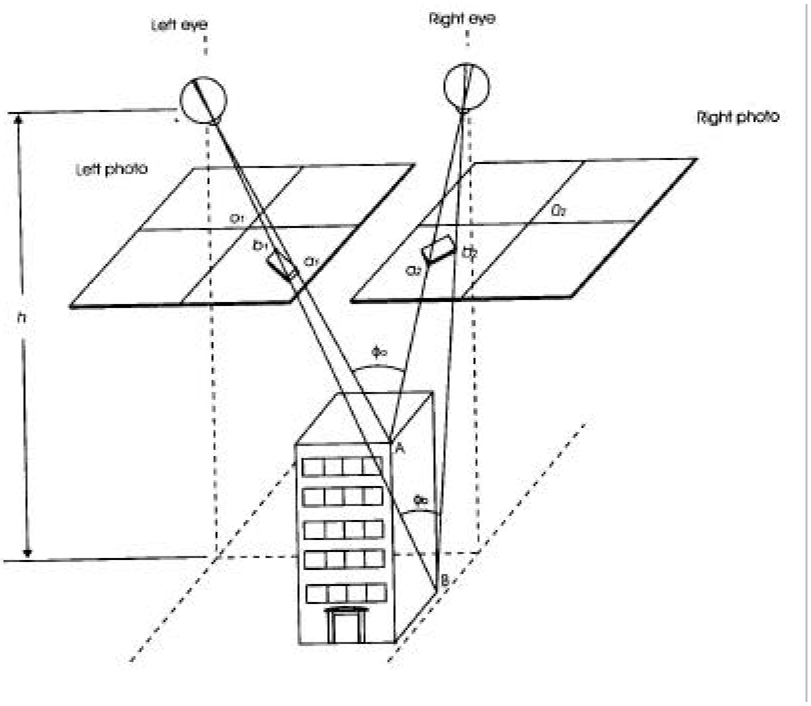

Figure 3.7: The Principle of Metric Photogrammetry (after [8]).

Some of the most significant advances in photogrammetry have been the transformation of the process of making a map from the labour-intensive analogue photogrammetric techniques, to the modern automatic digital photogrammetric procedures in which the photographic image is stored within a computer rather than in the form of a plate or negative. Although airborne cameras are still used to make the initial image, in the next few years high resolution satellite images will be increasingly used for small scale maps. These images have a resolution as fine as one metre.

3.2.2 *Accuracy Standards and Specifications*

One of the most important functions of geodesy is the determination of the precise position of points in relation to a well-defined reference system or datum. Hence geodesy is both the science of determining the "figure of the earth" <u>and</u> the set of practical tools for capturing spatial data. However, it is customary to distinguish between those techniques which are "geodetic", capable of the most precise position determination, from those which are not able to satisfy the same stringent accuracy requirements. There is, of course, a continuous spectrum of positioning accuracies from high to low. This has long been recognised by the hierarchical system used in *control surveying*. Geodesy provides the most fundamental control or geodetic network (or framework of physical benchmarks), which is then progressively "broken down" or "densified" using less accurate techniques.

It must be emphasised that until the advent of satellite techniques such as the Transit Doppler system and GPS, positional astronomy was the only means by which "absolute" position, that is, latitude and longitude, could be determined in relation to coordinate axes or a datum. At best, the accuracy of latitude and longitude that can be determined using several nights of astronomical observations in the field is about a few tenths of an arcsecond (about 10 metres), and the datum was a geocentric *celestial* datum based on the coordinates of visible stars. Sub-metre accuracy was possible from several days of Transit observations. However, GPS absolute accuracy is typically only at the few tens of metres level because of "Selective Availability" (see Section 2.1.2).

Accuracy in telative terms means the size of the average or maximum error in the coordinate of one object relative to another, expressed as a ratio of error magnitude to point separation. For example, a one centimetre error in the location of two objects one metre apart is large, hence the technique of fixing position is one of low accuracy, but if the two objects were 100 km apart, then the position fixing technique is a high accuracy one. Hence, a "relative" measure of positioning accuracy would be to express the ratio in terms of "parts per million", or simply "ppm". We could define a "geodetic positioning technique" as being anything that was able to consistently (or with some high probability) determine relative position to an accuracy of 1ppm -- this is a 1cm error in the relative coordinates of points 10 km apart. There are position fixing techniques, such as those of positional astronomy, which determine "absolute" position, independently of any other nearby point. In such circumstances the coordinate is determined relative to the *origin* of the reference system, which for a point on the surface of the earth would be at least 6000000 km away! An error 6m (equivalent to 0.2 arcseconds error in astronomic latitude or longitude), would be expressed as 1ppm -- a very high accuracy indeed.

National or state geodetic authorities are responsible for the establishment, maintenance and densification of the control framework. The greatest of care and the most precise positioning technology is used for this purpose. As we have seen, the control framework is the physical realisation of the geodetic datum, and hence by making measurements to/from the control stations we ensure that all our spatial information is in a consistent coordinate frame. Such a framework of points may be given a certain quality flag, for example "zero order" or "first order", implying a certain relative accuracy (measured in terms of "ppm"). Points established by connection to this framework using any of the survey techniques referred to earlier, may be labelled "second order", "third order", etc., depending on the quality of the (relative) coordinates. (Different authorities may adopt different "ppm" criteria for the different orders of accuracy.)

In general, the highest orders are used to control the spatial errors across the distances usually represented by map sheets. For example, a map sheet 0.5m across, at a scale of 1:100000, should have relative errors between two well defined points on the map which are less than about 70 metres if standard plotting quality criteria are used -- that 90% of points are within 0.5 millimetre of their correct position on the map. In the case of a map at a scale of 1:10000, the relative accuracy requirement is ten times more stringent. However, the maximum separation of points may be of the order of 70 km in the former case (1:100000 scale) and 7 km in the latter case (1:10000 scale), but both imply a relative accuracy of only 1 part in 1000, well inside the level of geodetic accuracy. Hence, only a few points on the map might have had their coordinates determined to geodetic accuracy, but these points are used in the photogrammetric technique to ensure that the scale, orientation and overall spatial integrity of the map are to the standard required.

The geodetic network is therefore the very foundation of a map, can be considered as a series of layers of spatial information overlain one on the other (figure 3.8). A map may serve a certain purpose by including only those layers of spatial information which are relevant to that application. Hence we may have topographic maps, tourist maps, air navigation maps, geological maps, census and political maps, etc. Nowadays the basic layers of data (topography, natural watercourses, vegetation, roads and railways, public buildings, etc.) can be electronically stored and maps produced which are highly customisable using anything from simple desktop mapping programs to sophisticated computer mapping software.

The maps which will be used within vehicles for ITS applications will be generated from the raw (layered) spatial datasets that already exist. Additional spatial data capture may have to be carried out in order to "geo-code" or locate points-of-interest such as restaurants, parking areas, theatres and cinemas, banks, post offices, police stations, etc. However, spatial data is not the only information which may be contained within a map. The role of the "legend" on a paper map is to provide basic information regarding the features on the map. In the case of electronic maps, this "intelligence" is provided by attribute data linked to the spatial data.

3.2.3 *Map Attribute Data*

Modern computer-based spatial data storage, management, analysis and display technologies such as Geographic Information Systems (GIS) do more than simply draw maps on demand, they are able to support decision making because of the ability to match the spatial characteristics of the data (such as position, and topology of lines and polygons -- that is, how points are *connected* to each other) with other textural data. This auxiliary data about points lines and polygons on a map is part of the *attribute data* (figure 3.8).

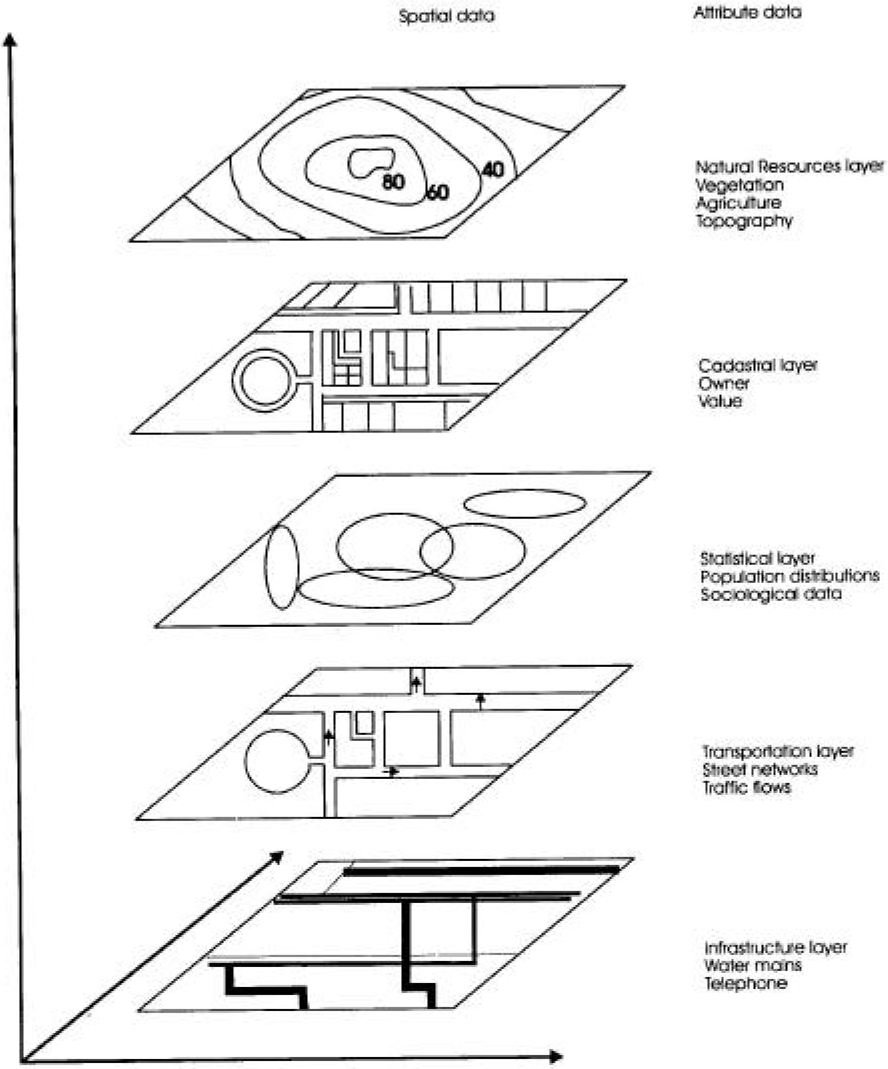

Figure 3.8: Spatial and Attribute Data.

Each layer of a map may have a set of global attribute data specifying, for example, the origin of the data set, when it was last updated, the purported accuracy of the data, how it was captured, etc. This is sometimes referred to as 'metadata". However, attribute data for specific features or points in a map (or map layer) generally is in the form of a text table. The definition of the text table is generally different for different features. For example, a tree may have attribute data linked to it concerning species, height, girth, etc., while the attribute data linked to a road segment will have different entries such as width, date of construction, surface quality, etc. The ability to access this additional data (hidden away in attribute data tables) simply by pointing to the map feature on the computer screen makes GIS a powerful query tool.

Other queries could be 'display where such-and-such is located?'. More complex questions such as 'where does a road cross a river south of a line of latitude xx?' require the interrogation of more than one map layer.

How does this impact on ITS applications? In a number of ways. For example, given an "intelligent" spatial dataset comprising of many layers (road network, topography, watercourses, property boundaries, etc.), with associated attribute data, *it is possible to create the specialised maps which are central to the in-vehicle navigation system*. Firstly, only those layers are selected which are relevant to road users, and those features of significance to road users may be highlighted in some way. These could be petrol stations, traffic lights, car parking stations, restaurants, etc. Such a customised map may then be converted into some convenient form for display within a vehicle, for example, raster or vector maps stored on CD-ROM. Such datasets are the electronic form of the humble *street directory*, and many of them have been developed to have almost the same "look and feel" of the paper equivalent.

A more sophisticated approach to compiling the in-vehicle map is to go beyond just producing an electronic version of a paper map, but to also make the attribute data available to the driver. Hence, in addition to compiling the spatial data in the manner described earlier, the attribute data is compressed into succinct descriptive information. For example, the name of the restaurant is included, the car park entrance is described, the brand of gasoline sold at the petrol station is included, etc. If the basic map data (spatial and attribute) has been carefully collected, and it is both current and comprehensive, then this process is not too onerous. However, there may be attribute data which is useful to drivers that has not been collected, such as which streets carry one-way traffic (and in which direction), the weight restrictions on roads, speed restrictions, where right turns (or left turns) are prohibited, which streets have median strips, and so on. If it is not available, this data will have to be obtained in some way, and then regularly updated.

The requirement for "intelligent" road maps is even more stringent for ITS applications such as fleet management and emergency vehicle dispatch. In such applications the algorithms which must advise the driver on how to travel from location A (present position) to location B (point of cargo pickup, site of emergency, etc.) would require not only information on the road *geometry* (the spatial data which describes how roads connect to each other), but also traffic restrictions and conditions, etc., in order to perform optimum route guidance computations. The ultimate attribute data would be up-to-the-minute information on traffic conditions, so that route guidance will be sensitive to congestion and other ephemeral traffic restrictions.

A variation on the above is *map-aided navigation* (Section 3.3.4). In this case, the road vector data (coordinates of the start and end of road segments) is used to generate route options, which require certain attribute data to be available in order to eliminate any routes that are not suitable (for example, suggesting that the vehicle travel the wrong way up a one-way street).

3.2.4 *Data Storage, Maintenance and Distribution*

The following comments may be made with regards to "traditional" maps:

- In the era of paper maps, the mapping authority was responsible for all phases of map production, from the raw data capture, through the various steps of the map compilation process, to map production and distribution.
- Map updates were comparatively infrequent, being dependent on such factors as the degree to which the mapped features changed, mapping program priorities, budget constraints, impact of new technology, etc.
- The sale price of paper maps is significantly less than the true costs of making the map.

- Almost all maps require the use of aerial photographs, and the application of the principles of photogrammetry.
- Apart from the standardised topographic map series developed by the federal or state mapping authority, various forms of thematic maps were produced by other agencies, such as geological maps, hydrographic/navigation charts, road maps/atlases, cadastral maps, tourist maps, etc.
- There was some coordination of mapping activities across several agencies, such as the use of the same aerial photographs, or the use of one contour map overlay.
- There is now a recognition that the true cost of producing a map must somehow be recouped or accounted for.
- There is now a trend to the rationalisation of map databases so that there is a minimum of duplication.

"Digital maps", or maps that are either stored in electronic form or displayed on a computer screen, are the latest metamorphosis of the humble paper map. However, many so-called "digital" or "electronic" maps are merely derived by scanning existing paper maps (or the cartographic material used in the printing process). The scanned image may be processed and "vectorised", which is essentially a process of automatically digitising the line and point features. Hence, by this process the map coordinates of all points of interest on a map can be determined in a form of "reverse engineering". Digital maps are increasingly also being generated from the raw data (the photographic or satellite images).

There are certain unique features of such maps from the point of view of storage, distribution and maintenance:

- Digital map database(s) offer tremendous opportunities for customisation of the final map display -- the storage of the data is of no concern to the user, it is a set of files resident on some mass storage device.
- The map database(s), or the final electronic map form, can be easily distributed to users on computer disk or CD-ROM.
- The original sources of the map data (or "layers") may no longer be known to the user -- the 'right to use', or 'right to publish' may have been sold or licensed several times.
- Many of the digital map databases have been established simply by digitising paper maps -- sometimes with no regard to issues of "copyright" and ownership
- The format in which the map database may be stored, as well as the specifications of what is stored, is not a universally recognised standard.

The last issue is particularly important. Certain companies have established strong positions as digital map data providers through strategic alliances with companies (software developers, hardware vendors, GPS or GIS service providers, etc.), and while many countries struggle with formulating "standards" for spatial data storage, these companies have been defining de facto standards for digital maps. [10] claims that most digital road map data in North America, Europe and Japan is supplied by just five companies or associations. This digital data is "intelligent" in that it can be used in ITS products that require such functions as (Section 3.3.4): address matching, map-matching, best-route calculation, and route guidance. The following information is largely drawn from [10].

The Japan Digital Road Map Association (JDRMA) is an 82 member consortium of companies involved in vehicle navigation (including car and electronics manufacturers), and in 1988 released the first Digital Road Map (DRM) of Japan. This was derived from the 1:50000 and 1:25000 topographical maps, and virtually covers the entire country.

There are two United States companies providing DRM data for vehicle navigation systems. Etak, based in Menlo Park, California, has been generating and distributing DRMs for over 10 years. Etak's maps provide extensive coverage in the U.S. and other countries, including France, Germany, Japan, Canada, Hong Kong, and The Netherlands. City DRMs are produced at an equivalent scale of about 1:25000, but in rural areas the accuracy is equivalent to 1:100000. The other U.S. company is Navigation Technologies (NavTech), based in Sunnyvale, California, and released its first databases in 1991. NavTech has focussed on developing DRM data that can support all ITS applications, and guarantees that its databases are 97% complete and accurate both in position (better than 15m) and in the correctness of the restrictions and geometry of the road network.

In Europe, the development of DRMs has also been driven by the requirements of vehicle navigation. At the heart of European road databases is the Geographic Data File (GDF) format, an extensive set of information on road geometry, conditions, traffic restrictions, etc. So massive is this dataset that vehicle navigation system manufacturers generally only use a sub-set of the available GDF data for their own systems. GDF is therefore an archive and exchange format, and data suppliers contribute digital road map data in the GDF format. The European Digital Road Map Association (EDRA) is a consortium which has created a pool of DRM data for all of Europe. Another group, EGT of The Netherlands, opted to map Europe on its own, probably because EDRA's Etak and EGT's partner NavTech are competitors in the United States.

3.3 Map Display

An inspection of a standard paper map sheet reveals several constant features of maps which may have to be modified for digital maps intended for in-vehicle display:

- Graticules which permit coordinates to be scaled off, or simple navigational calculations to be performed. The grid may be composed of straight lines marking off "east" and "north" values, or they may be curved and designate parallels of latitude or meridians of longitude.
- Statement of map projection used, from which it is possible to infer whether the scale of the map is a constant or not, whether it is possible to measure off bearings, what the graticule values correspond to, etc.
- Scale of the map, in the form of a statement, or as a scale bar.
- North point indicator, which may be in the direction of true north, magnetic north, or grid north.
- Legend table, which describes the codes or symbols used for different features.
- Statements concerning origin of raw spatial data, authority responsible for producing and printing the map, date of survey/printing, statements concerning accuracy, etc.

However, there are also differences between electronic maps and their paper equivalents:

- Electronic maps are usually "tiled" so that the image on a relatively small computer screen is not displayed at too small a scale -- hence the whole map sheet may not be visible, including the legend table, scale bar, etc., and some other means must be used to access this information.
- The "up" direction of the map display is north, the direction of travel, or some other arbitrary direction fixed by the navigation computer.
- The scale of the map display can be changed by zooming in, or zooming out.
- The level of detail displayed may be user-selectable.
- Data in general may not be "entered" on the map -- certainly the case for "closed file" formats of in-vehicle electronic maps.
- The ability to display auxiliary data such as the present vehicle's location (if a positioning device is fitted).
- The ability to make "queries", such as 'what is the distance between two points on the map?', or 'where is the nearest petrol station?'.
- Electronic maps may have to use colours, symbols, labels and notations which are optimised for viewing on computer screens.

In the sections below we introduce some basic material on map projections, and make some comments regarding map accuracy and map-aided positioning.

3.3.1 *Map Projections*

Cartographers have been struggling for centuries with the problem of representing the surface of our (almost) spherically-shaped planet on a flat piece of paper. A map is an attempt to represent the curved surface of the earth on a flat piece of paper or on a computer screen. As anyone who has ever cut an orange in half, removed the fruity interior, and then tried to flatten the orange peel onto a flat surface knows, accurately completing such a task is quite a challenge! The orange skin tears and we must stretch or compress the skin, which distorts its original shape.

This distortion dilemma has challenged cartographers, mathematicians and geodesists for two millennia. The ideal projection would portray the features of the earth in their true relationship to each other; that is, directions would be true and distances would be represented at a constant scale over the entire map, resulting in equality of area and true shape of all parcels of land. Because an exact solution has not yet been found, the only workable solution has been to design map projections with prescribed distortion characteristics. More than 250 map projections have been developed and proposed through the years. The characteristics most commonly desired in a map projection are:

- Conformality.
- Constant scale.
- Equal area.
- Great circles[5] portrayed as straight lines.
- Rhumb lines[6] portrayed as straight lines.
- True azimuth or bearing.
- Geographic position easily located.

In general, conformal, or equal angle, projections are most commonly used because: (a) scale at any point is independent of azimuth, (b) the outline of small areas on the map conform to the shape of the feature, and (c) the longitude and latitude gridlines interest at right angles. Common *classes* of conformal projections are ([11]):

Conical projection is formed by considering a cone tangential to the ellipsoid at some *standard parallel of latitude*. After the cone is flattened out, the meridians of longitude are straight lines converging to an apex, which is also the centre of circles representing the projected parallels. The *Lambert projection* is one example of this type of projection.

Azimuthal projection is a special case of the conical projection where the apex is the pole and the cone degenerates to a plane tangential at the pole. The pole is therefore the centre of circles representing the parallels and of the straight lines representing the meridians. Examples of this projection include the *stereographic projection* (projection point is located on the surface of the earth opposite the point of the tangent plane), the *gnomonic projection* (projection point is located at the centre of the earth), and the *orthographic projection* (projection point is located at infinity).

Cylindrical projection is a special case of the conical projection where the apex is moved to infinity so that the cone becomes a cylinder which is tangential to the equator. After the cylinder is unrolled, the equator is mapped without distortion. The *Mercator projection* is one of the most widely used of all projection systems.

The Mercator projection is a conformal, *non-perspective* projection -- that is, it is constructed by means of mathematical formulae and cannot be obtained by graphical means. The distinguishing feature of the Mercator projection is that at any latitude the ratio of expansion of both meridians of longitude and parallels of latitude is the same, hence all directions and all distances are correctly represented. The Mercator projection is the only projection to depict rhumb lines as straight lines, and their directions can be measured directly on the map. Distances can also be measured directly, but not by a single distance scale on the entire map (unless the spread of latitude is small). Great circles appear as curved lines, convex to the equator (but in each hemisphere, concave to the pole). The shapes of small areas are nearly correct, but are of an increased size unless they are near the equator. Hence the Mercator projection is not an equal area projection, and is useless in polar regions above 80°N or below 80°S. The *Transverse Mercator projection* is designed for areas not covered by the equatorial Mercator.

In the conventional Transverse Mercator (TM) projection the *standard meridian* (or "central meridian") is mapped without distortion as it is the line of tangency of the spherical approximation of the ellipsoid with the cylinder. The central meridian is the y-axis (north direction) of the projection, while the x-axis is the mapping of the equator. In the case of the TM projection the property of straight meridians and parallels is lost, and the rhumb line is no longer represented by a straight line -- all are complex curves, making the TM difficult to use as a plotting map. On the TM projection a fictitious graticule similar to, but offset from, the familiar meridians and parallels is used. The fictitious "meridians" and "parallels" are straight lines perpendicular to each other, and a straight line on the TM projection makes the same angle with all the fictitious "meridians", in effect creating a fictitious rhumb line.

A modification of the TM is the *Universal Transverse Mercator* (UTM) system. Imagine the earth as an orange with parallels of latitude and meridians of longitude drawn upon it. Using a knife, it is possible to make a series of straight north-south cuts in the skin at equal intervals of 6° completely around the "orange" until 60 identical strips have been detached.

Each segment forms the basis of a separate map projection, and because each zone is relatively narrow, there is minimal distortion of the features shown on the surface. By international convention, all of the UTM grid zones are numbered from west to east, 1 to 60, beginning at the *International Dateline* (180°E longitude). Hence UTM grid zone 1 is a vertical area running between meridians located at 180°E and 186°E longitude, with its *central meridian* at 183°E.

3.3.2 *Coordinates, Datums and Maps*

The UTM zones across Australia are numbered 49 to 56, while the UTM zones of the contiguous United States are numbered, from west to east, 10 to 19. Unlike the standard TM projection, where the scale factor on the central meridian is unity, the scale factor at the central meridian of the UTM projection is 0.9996 to reduce the large distortions on the fringes of the zone.

All UTM grid zones' central meridians (for example, 111°E, 117°E, 123°E, etc.) are arbitrarily labelled 500000mE to create a west to east coordinate numbering system within the grid zone. In the northern hemisphere the equator serves as the 0mN grid line, and for the southern hemisphere the equator serves as the 10000000mN grid line. It is therefore possible to have the same coordinates for many different features, for example 733675mE, 6291250mN, located in different parts of the world because in addition to the coordinates the distinguishing data is the UTM zone number.

What other possibilities are there for "coordinate confusion"? As mentioned previously, before the age of satellite positioning technology, differences in geodetic datums did not pose any problems because each country or region was using its own reference system. The GPS datum -- WGS84 -- does not always coincide with the local geodetic datum. Hence, the same latitude and longitude values from two different geodetic datums does not identify a unique point. This is also true if the same datum is considered, but different reference ellipsoids are used. Finally, the map coordinates of a point will vary depending on which map projection is selected. As a result, the same map coordinate from two different map projections do not necessarily refer to the same point on the surface of the earth.

The procedures of transformation and projection are extremely important because the difference between WGS84 coordinates and local map coordinates can exceed one kilometre in some parts of the world. For example, the difference between the UTM eastings and northings of points referenced to the geodetic datum used in Japan and WGS84 is almost 1500m (figure 3.9) ([13]). Although this is an extreme case, differences between local geodetic datums and WGS84 (or ITRF) of several hundred metres are common. Hence care must be taken to ensure that datum and projection incompatibilities are not present when relating GPS-derived coordinates to features on a paper or electronic map.

The process of converting between the different datums and map projections used throughout the world is relatively straightforward. The U.S. National Imagery and Mapping Agency (NIMA) (formerly the DMA) has published a technical report that contains the numerical relationships between more than 100 geodetic datums and WGS84 ([14]). Most GPS receivers intended for handheld and in-vehicle applications offer transformation and projection options, but they must be configured to transform to the right datum and to project to the correct map system using the appropriate parameter values.

Several procedures can be used for transformations: simple block shifts, similarity transformations, and projective transformations. The block shift is the simplest transformation procedure, involving the application of a bias in latitude and longitude to the WGS84 coordinates to make them compatible with the local datum. However it is the least accurate method and is suited only for small areas such as a city. (The similarity transformation was discussed in Section 3.1.3.) Values of the standard seven similarity transformation parameters are usually published by the national geodetic authority or the NIMA/DMA, however, to achieve higher accuracy "tailored" parameters sets may be used for small areas. This requires GPS measurements to be made at local geodetic control stations in order to provide common points in both datums. With sufficient common points (three is the minimum for a seven parameter transformation model) users can derive their own transformation parameters.

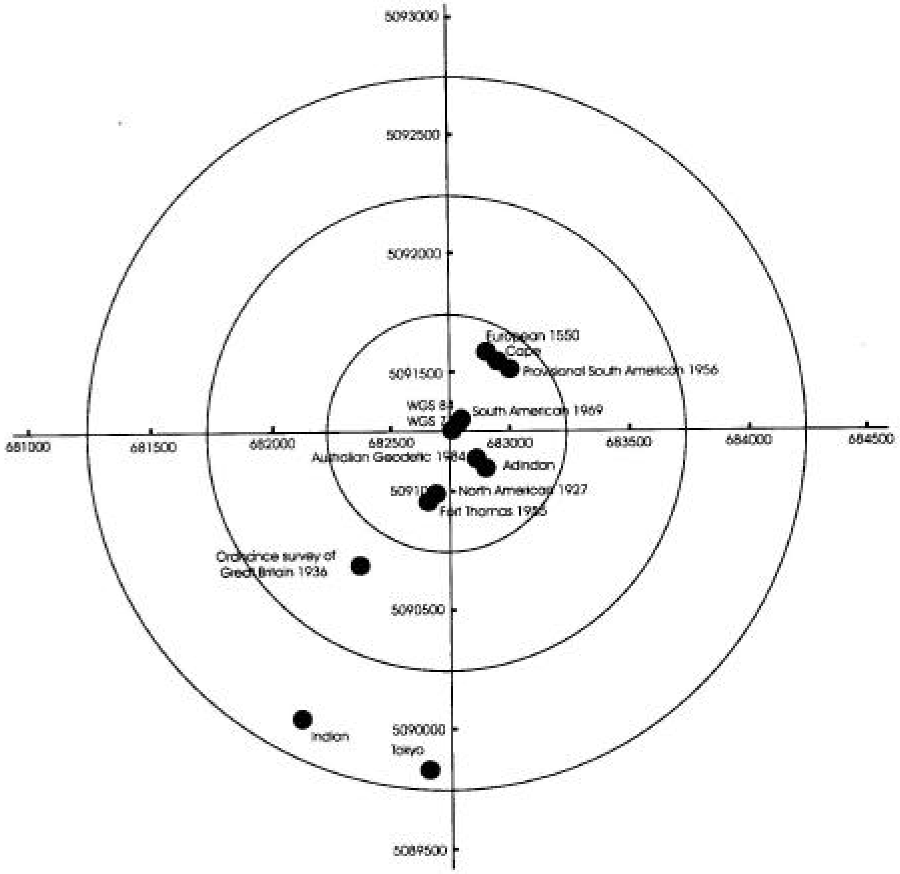

Figure 3.9: Plot of UTM Eastings and Northings (in metres) of a Point with WGS84 Geographical Coordinates 45°57'0.96"N, 66°38'32.22"W, for a Variety of Geodetic Datums. Circles have radii of 500, 1000 and 1500 metres (after [13]).

An alternative approach is to use a *projective transformation*. Such a model is especially useful for adsorbing any errors or distortions in the original local geodetic datum which can be detected using more precise GPS survey techniques. The transformation model accounts for both the internal errors in the local datum and the change in datum. NIMA/DMA has generated such transformation parameters, used in so-called multiple-regression equations, for seven continent-sized regions of the world -- Australia, Brazil, Argentina, Western Europe, Canada, the United States, and the rest of South America.

Algorithms required to convert latitude and longitude (or Cartesian components) to easting and northing values can be found in textbooks on map projections (see, for example [11]).

Electronic map products which use GPS-derived position for indexing the vehicle to an electronic map generally perform datum and projection coordinate changes in a seamless manner. If the plotted location of a vehicle appears to be off the roadway, as depicted on an electronic map, it is generally because either: (a) the position is not determined accurately enough (for example, because GPS is used in the absolute positioning single-receiver mode), or (b) the road feature is not mapped accurately enough for the scale at which it is displayed.

3.3.3 *Map Scales and Accuracies*

Map scales may be confusing to those who are not trained in the basics of cartography, or simply cannot "read" a map. Scale is usually expressed in the form of a statement such as '1 in xxxxxx', for example 1 in 100000, or simply 1:100000. This implies that 1 unit on the map represents 100000 units "on the ground", hence at this scale 1mm on the map is equivalent to 100m in the real world. This would be generally classed as a "small scale" map. *How is this so?* One's first reaction is that 100000 is a "large" number! Or that a small scale map must depict "small" details. This is, of course, the opposite to what really happens. The words "small" or "large", as they relate to map scale, refer to the *fraction* $\frac{1}{100000}$. Hence, a 1:1000 map is at a large scale, and could depict features down to 1m in size, while 1:1000000 is a very small scale map. Many countries use (or have used in the past) scales which are less convenient, often based on imperial units, such as '4 inches to the mile', etc.

Maps are generally produced at different scales to satisfy a diverse user population. Some countries are comparatively small, and hence their maps may all be at large scales. On the other hand, a country such as Australia or the U.S.A. cannot afford to map the entire continent at all scales, from the smallest to the largest. For example, the entire Australian continent is mapped at scales 1:250000 and 1:100000. However, maps at scales of 1:25000, 1:10000, and larger, are only available in areas where they are most likely to benefit the community, such as along the coast, around towns and cities, etc. In cities a great variety of maps have been produced for all sorts of purposes, however the raw data (or map layers) may, in most cases, have come from the same survey or photogrammetric procedure. Maps for ITS applications may be at scales of about 1:5000 to 1:10000 in cities (similar to those used in "street directories"), and at smaller scales along the major roads outside the metropolitan areas.

There are several aspects to "accuracy": (a) the accuracy with which the coordinates of a feature can be determined, and (b) the accuracy with which the coordinates of a feature must be known if it is to be correctly displayed on a map at a certain scale. Examples of varying accuracies with which point features can be surveyed:

Control mark	1cm
Corner of building	10cm
Street intersection	1m
River boundary	2-5m
Forest boundary	10m
Soil type demarcation boundary	50m

Hence, depending on the map scale used to depict this data, these accuracies may be either higher, or lower, than the plotting accuracy. Map plotting accuracy is usually specified by requirements such as that 90% of "well defined" points should be within 0.5mm of their correct position at map scale. This translates to the following accuracies (of the smallest plottable feature):

1:1000	0.5m
1:10000	5m
1:25000	12.5m
1:100000	50m

Another factor that impacts on the accuracy with which a feature is shown on a map is the size or bulk of the feature, and whether, relative to the scale of the map, it should be drawn at an exaggerated scale. The best example of this is the depiction of a roadway by a pair of lines the apparent separation of which many be much larger than the actual width of the road. This is particularly common in the case of "street directory" type maps.

The accuracy with which points must therefore be determined or defined is related to the map scale and the exaggeration factor that is applied, as well as to the "sharpness" or quality of definition of the feature.

However, electronic "vector" maps rely on databases of object coordinates, and must be drawn on the computer screen "on the fly". The map scale may vary considerably through the process of "zooming", and when an area originally mapped at a relatively small scale is zoomed and displayed at a larger scale, then the accuracy of the feature's position (based on the quality of the original survey, and the "sharpness" of the feature) may no longer satisfy the plotting accuracy requirements.

3.3.4 Map-Aided Positioning

Digital maps are the basis of many other ITS functions besides locating the vehicle in a map reference frame ([10]). However, vehicle navigation systems are limited by the map data they use, and are crippled when inexpensive, complete, and "seamless" Digital Road Map (DRM) data is not available. It is possible to define several functions that DRM data can perform in the context of ITS applications:

Address Matching -- in which the function is to transform a given latitude and longitude (provided, for example, by a GPS system) into a street address, or vice versa. People know the address of their destination rather than its coordinates! Hence to *navigate* to an address requires that the positioning system be able to unambiguously convert the address to a coordinate. This would require that the navigation system accuracy be better than 20 metres -- a requirement that under Selective Availability only differential GPS can satisfy.

Map-Matching -- is based on the premise that the navigating vehicle is on a road. Hence when a positioning system outputs coordinates that are not on a road segment as defined by the DRM, the map-matching algorithm finds the nearest segment and "snaps" the vehicle onto the road segment. Obviously map-matching requires maps with high positional accuracy to minimise incorrect road segment selections. The map-matching algorithm can also be used within a GPS-Dead Reckoning system (Section 2.3). The DR sensors are used to measure distance and heading (or heading change) in order to compute relative changes in position, for example, since the last GPS "fix", while map-matching contributes information on "absolute position".

Best-Route Calculation -- this supports driver planning by providing assistance on selecting the optimal travel route. A DRM coupled with a best-route calculation algorithm provides an optimal route based on travel time, travel distance, or some other specified criterion. The result is a turn-by-turn description of a journey. Best-route calculation requires a high level of map information, for example, the DRM must include traffic and turn restrictions so that the route selected is not illegal or dangerous!

Route Guidance -- supports drivers as they navigate along a route (selected by the driver, or the best-route algorithm). Route guidance includes turn-by-turn instructions, street names, distances, intersections, and landmarks. This is particularly challenging in real-time because the algorithm must process position information, perform address and map-matching, and display the DRM to the driver. If the driver misses a turn, the system must be able to compute a new best-route "on-the-fly" and provide new guidance information.

Each of the four functions relies on specific features in the DRM database, however the most demanding are the last two: best-route calculation and route guidance. If a DRM supports these functions it is said to be *navigable* ([10]). The DRM database accuracy requirement would be similar for all the map-aided applications listed above, however, navigable DRM databases are more complex because of the extensive additional information that must be stored. Hence, navigable DRM databases are significantly more expensive to produce and maintain. The most sophisticated ITS map functions can therefore be supported only when accurate, complete and seamless DRM databases are available.

References

[1] NOAA, 1985. **Geodesy for the Layman**. 5th Ed., 96pp.

[2] SMITH, J.R., 1988. **Basic Geodesy**. Landmark Enterprises, 151pp.

[3] VANICEK, 1995. **Global Positioning Systems: Theory & Applications**. American Institute of Aeronautics & Astronautics (AIAA), 1995, Vol.1(694pp), Vol.2(601pp).

[4] TORGE, G., 1993. **Geodesy**. Walter de Gruyter, Berlin New York, 531pp.

[5] MANNING, J. & HARVEY, B., 1994. Status of the Australian geocentric datum. ***Aust. Surveyor***, 39(1), 28-33.

[6] D.M.A., 1991. Department of Defense World Geodetic System 1984. Technical Report 8350.2, 2nd ed. U.S. Defense Mapping Agency.

[7] DIXON, K., 1995. Global reference frames with time. ***Surveying World***, September 1995.

[8] WOLF, P.R., 1983. **Elements of Photogrammetry**. McGraw Hill, Inc., 2nd ed., 628pp.

[9] THOMPSON, M.M. (ed.), 1966. **The Manual of Photogrammetry**. American Society of Photogrammetry, 3rd ed., Vol.1 & 2.

[10] KRAKIWSKY, E.J. & BULLOCK, J.B., 1994. Digital road data: putting GPS on the map. ***GPS World***, 5(5), 43-46.

[11] SNYDER, J.P., 1993. Map projections -- a working manual. USGS professional paper 1395, U.S. Government Printing Office, Washington, D.C., 383pp.

[12] HOTCHKISS, N.J., 1994. **A Comprehensive Guide to Land Navigation with GPS**. Alexis Publishing, USA, 187pp.

[13] FEATHERSTONE, W. & LANGLEY, R.B., 1997. Coordinates and datums and maps! Oh my! ***GPS World***, 8(1), 34-41.

[14] DoD World Geodetic System 1984 - Its definition and relationships with local geodetic systems, 1991. NIMA Technical Report 8350.2, 2nd ed., National Imagery and Mapping Agency, Washington, D.C.

Footnotes:

[1] The semi-minor axis is therefore about 20km shorter than the semi-major axis.

[2] The line joining the north and south CTP is an imaginary axis assumed fixed to the earth's crust, against which the instantaneous rotation axis of the earth is tracked. The rotation axis of the earth moves relative to the earth's crust in a complex "polar motion" which is monitored by the IERS, which publishes the "coordinates" of the point where the rotation axis of the earth pierces the crust near the North Pole relative to the CTP .

[3] Derived from the value of the normalised second degree zonal harmonic coefficient of the gravitational field: -484.16685 x 10^{-6}.

[4] At the start of 1997 the TRANSIT Doppler System was closed down by the U.S. government.

[5] Lines on the globe formed by intersecting a plane with a sphere such that the plane passes through the centre f the sphere. The equator and all the lines of longitude are great circles.

[6] Lines of constant bearing.