

4.1 A Brief History of Navigation

Early man observed the sun and the stars, and presumably used these for navigation long before leaving any written records about it. As late as the Viking age (800-1100 AD), little further help was available for navigation on open seas. Rough estimates of the Polar star's height over the horizon, together with 'dead reckoning' (from 'deduced reckoning', estimating distances from course, currents, winds and speeds), did not always suffice to find the intended destinations. Innumerable marine disasters have been caused by navigational errors.

One particularly grizzly incident occurred on a foggy night in October, 1707 when a group of four British warships with about 2,000 men on board ran aground just off the English SW coast. Only two men reached shore. One happened to be the fleet commander, who was promptly murdered upon reaching shore by a local woman for a ring he was wearing. Maybe there was some justice in this. The fleet officers knew full well they were lost before the accident; nevertheless, a seaman who had kept a perfect log, and dared to very carefully and respectfully offer this to an officer the day before - knowing the risk but hoping to help avoid a disaster - was immediately hanged for insubordination.

The concept of longitudes and latitudes goes back at least to Ptolemy. All 27 sheets of his world atlas from 150 AD have such lines drawn, together with a separate list of coordinates for all its named locations. The equator was on his atlas marked as the zeroth parallel (latitude) and the Canary Islands defined the zero meridian (longitude). This latter choice was quite arbitrary, and indicative of the coming difficulties in determining the longitude at sea. Before settling at Greenwich, 'prime meridians' were at times placed at the Azores, Cape Verde Islands, Rome, Copenhagen, Jerusalem, St. Petersburg, Pisa, Paris, Philadelphia and many other places as well.

The big advances in navigation from the days of the Vikings have been

- discovery of the compass,
- finding the longitude,
(latitude can be read off easily from the height of stars - e.g. the pole star - over the horizon),
- navigation by radio beacons (LORAN), and
- GPS.

The first compasses were simple chunks of loadstone (magnetite, a common iron ore) which tend to orient themselves in a fixed direction, when suspended freely (by a string, or floated in a container of water). Their first use for navigation occurred in the Mediterranean during the 12th century. Magnetic compasses remain to this date indispensable on all ships, at the very least as a navigational back-up device. Gyro-compasses work by a completely different principle - a suitably suspended rapidly rotating disc will keep its axis aligned with that of the earth. These compasses will always point to true north, and are insensitive to variations in the magnetic field (which can be due to geological anomalies or electrical storms on the sun). Although far more complicated than magnetic compasses, they are nowadays used in most larger ships and aircraft, often in connection with 'inertial guidance' devices that compute changes in positions from sensed accelerations.

The lack of any reliable means for determining the longitude at sea caused great hazards for sea travels until the chronometer was developed in the second half of the 18th century. If it was not possible to simply follow coastlines (which can be dangerous, especially at night and in bad weather), it was common practice to try to reach ports by first finding the appropriate latitude, and then follow it until the destination. This procedure was not very satisfactory for several reasons

- it works less well for coastlines facing north or south (as opposed to east or west),
- when aiming for a small island, the approach to the desired latitude had to be quite far off to the east or to the west in order to leave no ambiguity about the direction to finally proceed in,
- it forced sailing ships to follow paths that might not be suitable with regard to shoals, winds and currents, and
- it offered opportunities for pirates to lie in wait at the latitudes of main harbors.

The main competing approaches for finding the longitude all required the local time (easily available by the position of the sun) to be compared to the simultaneous time at some fixed reference location (such as Greenwich). Ideas to determine this reference time included

- Observations of Jupiter's moons. Since their orbits could be tabulated accurately, the moons can serve as an accurate clock in the sky. Eclipses (a moon disappearing in the shadow of the planet - happening roughly every two days for each of the inner moons) are near-instantaneous events, allowing very accurate time readings. Although this worked well on land, it proved, even in good weather, to be utterly impractical at sea.
- Observing the position of our moon relative to the sun and the stars. Newton's law of gravity was discovered first in 1684, and the complicated orbit of the moon (a quite non-circular path influenced by both the earth and the sun) could not be predicted with enough precision until well after the whole approach had been made obsolete by the chronometer,
- The chronometer - basically an accurate clock, designed to be insensitive to motions and changes in temperature, humidity and gravity. This became the winner in the longitude competition. John Harrison's produced a series of increasingly accurate chronometers, culminating in 1760 with the pocket-sized "H-4". On its first sea trial - UK to Jamaica, arriving in January 1762 - it lost only 5 seconds. This corresponds to an error of only 2 km after 81 days at sea (however, this was somewhat lucky - an error of about one minute or 24 km could have been expected; even that a vast improvement over other methods). By 1780, chronometers were starting to come in wide use throughout the British (and other) navies (often privately purchased by the Captains, as official navy channels still were slow in providing them).

More exotic ideas at the time included

- Placing light-ships at known strategic locations. These would then every-so-often send up a rocket that exploded brightly - deemed to be visible at night for up to 60 to 100 miles, providing travelers within that range a time signal, and
- Mapping the vertical inclination of the earth's magnetic field. Lines of equal inclination would generally intersect the lines of constant latitude (or the angle could be mapped), thus together with the latitude providing complete positional information. However, not only does the earth's magnetic field change slowly with time, it can also fluctuate dramatically with solar activities (up to about 10 degrees - enough to cause positional uncertainties as wide as an ocean).

The history of how the longitude problem got resolved though the chronometer recounted in many books; a recent one being "Longitude" by Sobel (1995).

The first radio-based navigation technique amounted to determining the direction to a known transmitter by rotating a direction-sensitive antenna. Much higher precision was offered by a series of systems known as OMEGA, DECCA, GEE and LORAN (long range navigation). These were developed around the time of World War II. By the timing difference in arrivals of radio signals from a 'master' and a 'slave' transmitter (which re-transmitted the master signal the moment it received it), a ship could locate itself along a specific curve (in the 2-D plane case, a hyperbola). By also receiving signals from another transmitter pair, the ship could determine its location from the intersections of the two curves. This system gave a typical accuracy of around 1 km and a useful range of about 1000 km at daytime, and about double that at night time.

The GPS idea is to have a number of satellites in orbit, each transmitting both its orbital data and very accurate time pulses. A receiver can then clock the arrivals of the incoming time pulses. Knowing the speed of light ($c = 299,792,458$ m/s in vacuum), the distances to the satellites can be found. From knowing their orbits, the receiver's position can be found. With the high velocity of the satellites (and the high speed of light!), the demands on the precision of the equipment are extreme.

The cesium or rubidium clocks in the GPS satellites operate at 10.22999999545 MHz rather than the nominal 10.23 MHz to compensate for both the special relativity effect of a moving source and the general relativity effect of operating from a point of higher gravitational potential. The master clock at the GPS control center near Colorado Springs is set to run 16 ns a day fast to compensate for its location 1830 m above sea level.

The military's need for the system was also extreme - it was developed towards the end of the cold war as a means of accurately guiding ICBMs. Hence, it is hardly surprising that there today are two parallel fully operational systems in place, one created by the US Department of Defence and one by its Soviet counterpart. The cost for getting the GPS systems operational was staggering - at least 12 B\$ (i.e. $12 \cdot 10^9$ \$) for the US system. The fact that both systems now are available to the general public, without any charge, is almost as impressive as their technical capabilities. With low cost (< 200\$) handheld receivers, anyone can now determine his/her position to better than 100 meters at any time, in any weather, at any point on earth. With the best (and a lot more expensive) receiving equipment available, that can be improved to an amazing 1 mm (both horizontally and vertically). Surprisingly, GPS is still not used routinely in aviation - possibly because neither of the two signal providers is officially committed to providing uninterrupted

public service. For most civil and private usage, this concern is far outweighed by its practical advantages.