INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
34100 TRIESTE (ITALY) - P.O.B. 586 - MIRAMARE - STRADA COSTIERA 11 - TELEPHONES: 224281/2/3/4/5/6
CABLE: CENTRATOM - TELEX 460392-I

SMR/98 - 6

AUTUMN COURSE ON GEOMAGNETISM, THE IONOSPHERE
AND MAGNETOSPHERE

(21 September - 12 November 1982)

GEOMAGNETISM

S.R. MALIN

Institute of Geological Sciences
Murchison House
West Mains Road
Edinburgh EH9 3LA
U.K.

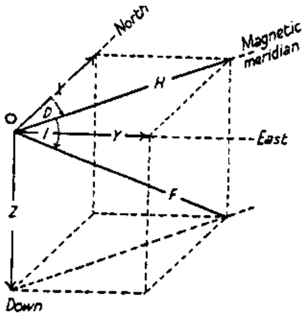
These are preliminary lecture notes, intended only for distribution to participants.
Missing or extra copies are available from Room 230.

Geomagnetism 1 (S R C Malin)

MAGNETIC MEASUREMENT AND OBSERVATORIES

Definitions

At any point on the Earth, the magnetic field is a vector, having amplitude and direction. It is usual to describe it in geographical components: X, the north component; Y, the East component and Z, the vertically downward component. Other elements are H, the horizontal component ($H^2 = X^2 + Y^2$); F, the total intensity ($F^2 = X^2 + Y^2 + Z^2$); D the angle of declination ($\tan D = Y/X$); I the angle of dip or inclination ($\tan I = Z/H$)



From Chapman & Bartels, Vol. 1, p. 2

The intensity elements (X, Y, Z, H, F) are usually measured in nano Tesla (nT); this unit used to be called the gamma (γ).

10^5 nT = 1 Gauss; 10^9 nT = 1 Tesla

The angular elements are measured in degrees, D positive to the East I positive downwards.

Measurement of D

The measurement of declination (D) requires a determination of true north and of magnetic north. Until very recently true north was invariably determined *via* an astronomical observation, usually of the Sun or the pole star. When at sea, with a clear horizon, the magnetic azimuth of the rising or setting sun could be compared with the true azimuth given in an ephemeris such as the Caroline tables, for the appropriate latitude and date. This was the method used by Halley, who also took account of atmospheric refraction.

When there were no satisfactory tables an admirable method of determining the north from the Sun was that used by Borough (1581). He noted the azimuth of the Sun as it ascended and descended through a chosen almacantur, using a quadrant to define the altitude. The mean of these two azimuths gives true north, and is independent of refraction. A simpler version of this method, using the shadow of a vertical wire, is described by Bourne (1574). If only a single observation is made of the altitude and azimuth of the Sun, true north can still be determined with the aid of an ephemeris. This method was commonly used in the 17th and 18th centuries.

In the 19th and 20th centuries true north was still commonly determined from a solar observation, particularly when doing field work. However, the altitude measurement has been replaced with a time measurement. A theodolite is clamped at a fixed azimuth and the time of transit of the Sun across this azimuth is noted. The readily attained accuracy of 0.4 s is sufficient to define the north to 0.1'.

At observatories, true north is usually obtained by observing polaris with a theodolite, both directly and reflected in mercury. These observations are reduced with the aid of pole star tables and a knowledge of the time of observation which is much less critical than for a solar observation.

Since the direction of true north does not change with time, at least to the accuracy required for declination measurements, it need be determined only very infrequently for each site. It can be recorded either by inscribing a meridian line on a fixed object (as was done by Gunter (1624) on the dial in the King's gardens at Whitehall), or by noting the true azimuth of a distant reference mark, as is the practice at observatories and survey stations.

Over the past few years an instrument has been developed that gives true north by sensing the Earth's rotation without any reference to an astronomical object. This gravity-influenced gyro fits on a theodolite, is battery powered, and permits true north to be determined to within 20 seconds of an arc in less than half an hour. Its disadvantages are its cost and the fact that it allows survey work to continue in bad weather.

Magnetic North

For the early determinations of declination, the magnetic azimuth was determined by means of magnetized needles fixed to a pivoted card, as described and illustrated by Moore (1681). In better instruments the card was dispersed with and the compass needle was fitted with an agate cup which rested directly on the pin. Various refinements were added, such as a reversible agate cup to allow the needle to be mounted either side up (and hence remove errors due to a difference between the magnetic and geometrical axes of the needle), mirrors to avoid parallax when reading the needle, verniers and a microscope to improve the reading of the circle, etc. Cavendish (1776) gives a good illustrated description of the Royal Society instrument, which was a fine example of such a compass. Some observers used needles several feet long; a few inches to a foot was more usual.

The main problems with pivoted compass needles are the friction in the bearings and the difficulty in persuading them to stay horizontal. Indeed, it was this latter problem that led Norman to his discovery of dip. These difficulties are overcome with a suspended magnet, as in the Kew-pattern magnetometer, which can also be used for measuring horizontal intensity. A detailed description of this instrument and its use is given by Stewart & Gee (1903). It is essentially the same instrument that is used for absolute measurements of declination in present day observatories.

Another instrument that can be used to find the magnetic north is the fluxgate. This is a pair of ferrite rods with coils of wire round them connected to electronic equipment that can be used to indicate the magnetic field intensity along the rods. The field indicated will be zero when the rods are magnetic East-West, so this direction can be found by holding the sensor horizontal and rotating about a vertical axis until a null reading is obtained. With suitable precautions, this can be as accurate as a suspended magnet, though probably no better.

A similar null-detector is the turbomag. This consists of a single-turn coil rotated at great speed by compressed air. Only when the rotation axis is aligned with the magnetic vector will the magnetic flux through the coil be constant, and no alternating current will be generated. The orientation is detected optically, using reflection from a mirror on the coil. The turbomag shows great promise as a survey instrument, but is still being assessed.

Measurement of Dip

For dip, the reference datum is the horizontal, which is readily defined by a spirit level, mercury dish, or via a plumb line.

Until the beginning of this century dip was invariably measured with a dip circle, consisting of a magnetic needle free to turn about a horizontal axis at the centre of a verical graduated circle. It was usual to observe with the plane of rotation of the needle in the magnetic meridian (defined to sufficient accuracy by a simple compass), though dip can readily be deduced from observations made in two perpendicular azimuths. Eccentricity of the axis is overcome by reading both ends of the needle and taking the mean, and misalignment of the magnetic and geometrical axes of the needle is allowed for by observing with one face of the needle first east, then west. A more serious problem is that the axis of rotation does not, in general, pass through the centre of gravity of the needle. It can be made to do so by means of counterweights, adjusted with the plane of rotation of the needle perpendicular to the magnetic meridian, but it is easier to allow for it by making two sets of measurements with the sense of magnetization of the needle changed in between, by stroking the needle with a magnet. This method is strictly valid only if the magnetic moment is the same for each direction, when the tangent of the true dip is the mean of the tangents of the two measures.

Another serious problem with the dip circles is mechanical resistance to rotation. Various means were tried to overcome this, such as roller bearing as described by Nairne (1772), or a cylindrical axle rolling on agate flats or knife edges, but the problem was never completely solved. Despite numerous detailed refinements, the dip circle changed remarkably little in its essentials from that which Norman built in 1586 to the Airy apparatus of 1861.

The dip circle remained in use as a field instrument until quite recently, but at observatories it was replaced by the dip inductor from 1914. The dip inductor consists of a coil of wire connected *via* a commutator to a galvanometer, and mounted within a verical circle on a firm horizontal base. When the coil is rotated (by means of a band or cable drive) it generates a current except when its axis of rotation is aligned with the magnetic vector. The null point is found by adjusting the orientation of the axis, and the dip is read off. This instrument is still in use at some observatories, though its use has generally lapsed. Its main drawbacks are the vibrations caused by the rotation, and the difficulty of adjusting the lignum vitae bearings to allow free rotation without slackness.

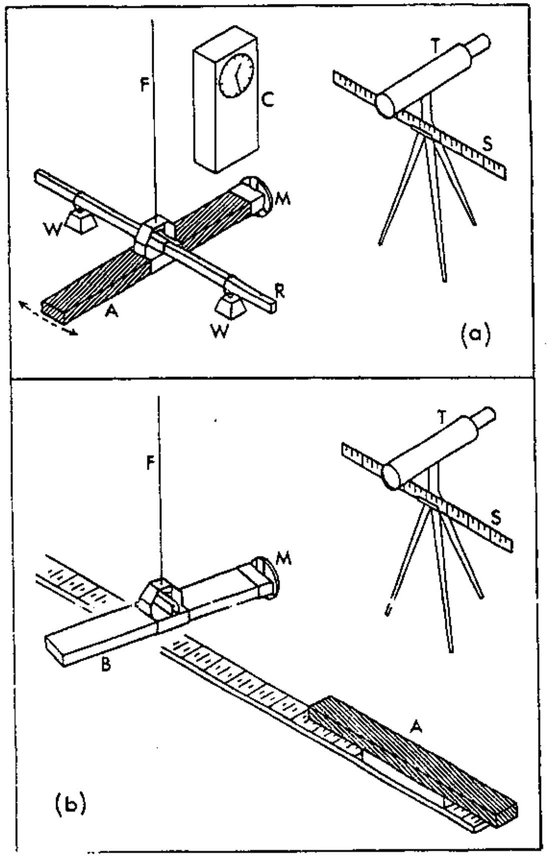
The turbomag (mentioned under 'declination') is essentially a modern version of the dip inductor, but with the drawbacks overcome. The turbomag and the fluxgate can both be used as null-detectors to determine dip in the same way as with a dip inductor. The fluxgate has the advantage of no moving parts.

Instead of measuring dip directly, it is not uncommon nowadays to deduce it from measurements of *H* and *Z*, or of *H* and *F*.

<u>Measurement of *H*</u>

The early measurements of magnetic intensity were relative rather than absolute. For example, Humboldt compared the magnetic intensity in many parts of the world by comparing the rate of oscillation of his standard dip-needle at the different sites. In this way, he showed that the field strength increases towards the poles, but he could not have measured changes with time at the same site since he could not have known if such changes were in his magnet or in the Earth's field strength.

Gauss's experiment.



(a) The vibration experiment. The magnet, A, suspended by a silk thread, F, oscillates in a horizontal plane. Its moment of inertia can be varied by hanging weights, W, from the rod, R. The period of oscillation is obtained by observing the reflection of the scale, S, in mirror M, through the telescope, T, and timing an integral number of swings with the clock, C. (b) The deflection experiment. Magnet B is suspended and is deflected from magnetic north by placing magnet A at a known distance. The deflection is measured by observing the reflection of the scale, S, in mirror M, through the telescope, T.

The first absolute measurement was by Gauss in 1832. His experiment was in 2 parts, the first to measure *MH*, where *M* is the magnetic moment of his magnet, and the second to measure *M/H*. From these 2 results, *M* can be eliminated to give *H* in nT. In the first experiment the period of oscillation, *T*, and the movement of inertia, *I*, were measured. There $MH = 4\pi^2 I/T^{-2}$. In the second experiment, the ratio of *M* to *H* is measured by noting the deflection of a compass from the magnetic meridian when the experimental magnet is at a known distance from the compass.

For observatory work, the highest accuracy was required, and apparatus for Gauss's experiment was refined at the Kew Observatory, Richmond, to attain this. The resulting Kew unifilar magnetometer was a beautiful instrument which permitted the absolute determination of *D* as well as *H*. From about 1860, this instrument came into almost universal use at magnetic observatories, and is still in use at some.

Gauss's method for the absolute measurement of magnetic intensity was not improved upon until the introduction of the coil magnetometer in the 1920's. This instrument was developed by F E Smith at the National Physical Laboratory, based on principles set out by Sir Arthur Schuster. If a known current is passed through a coil of known dimensions, the magnetic field it produces at the centre can be calculated. The current is increased until the artificial field just balances the horizontal component of the Earth's magnetic field, when a magnet suspended at the centre will swing round. The accuracy of the instrument depends on the precision with which the current and coil dimensions can be measured.

The coil magnetometer is excellent for observatory use, but is not suitable for field work. The Kew magnetometer *can* be used in the field, but is slow, demanding and unwieldy. For these reasons, a portable instrument, the Quartz Horizontal Magnetometer or QHM was developed in Denmark. This is not really an absolute instrument (though it is used as such by several observatories) and needs to be calibrated at an observatory both initially and at intervals of a few years to confirm that its constants have not changed. It consists of a magnet suspended from a quartz thread. When 360° of torsion is introduced into the thread (simply by rotating the instrument while the magnet remains pointing north), the magnet is deflected from the meridian until the turning moment of *H* on the magnet exactly equals the torsion. The measurement consists simply of measuring the angle of deflection and noting the temperature, since the torsion is temperature-sensitive. The reduction takes a matter of minutes and gives *H* to within a few nT.

Measurement of Other Elements

F: Nowadays, absolute measurements of the intensity of the Earth's magnetic field are made quickly and easily with a proton magnetometer. Protons precess around magnetic lines of force at a rate that depends on the strength of the magnetic field. The precession of randomly oriented protons cannot be detected, but if their spin axes are initially aligned by a strong magnetic field, when it is removed they will precess around the geomagnetic field in unison, producing a signal which can readily be detected. A proton magnetometer can measure the magnetic field to $5/10^6$ and is an excellent survey instrument.

Z: By backing-off the *H* with an artificial field supplied by a Helmholz coil, a proton magnetometer can be used to measure *Z*. It is not necessary to know the value of *H*; provided the coil has its axis horizontal in the magnetic meridian, it is sufficient to adjust the backing-off field until the residual field is a minimum. Similarly, by backing-off *Z* with a vertical coil, *H* can be measured. A proton magnetometer equipped with coils for use in this way is known as a proton vector magnetometer. It is the best observatory absolute instrument for measuring *H* and *Z*.

For field instrument for measuring *Z* (the BMZ) was developed at the same time as the QHM, though it was never so accurate or successful. An artificial field is used to cancel *Z* and a balanced magnet is used to detect when the cancellation is exact. Most of the artificial field comes from a permanent magnet which is screwed onto the BMZ, with fine adjustment provided by a weaker magnet than can be rotated in a verical plane. The measurement consists of finding the angle of the weaker magnet that causes the balanced magnet to rest horizontal (ie no vertical field). It is difficult to attain an accuracy of anything better than 10 nT. In general, for field work it is better and simpler to deduce *Z* from *F* measured by proton magnetometer and *H* measured by QHM.

Just as *H* could be measured with a Schuster-Smith coil, so *Z* could be measured with a vertical version of the same instrument, called a Dye coil. It was never widely used and has now been overtaken by the proton vector magnetometer.

Variometers

The instruments described so far give instantaneous values of the magnetic elements. For a complete description of the magnetic field at an observatory, it is necessary to know how it varies in the intervals between the spot observations. This information is provided by variometers, of which the La Cour variometer is a typical example.

It consists of 3 sensors - one each for measuring changes in D, H and Z - and a recorder. The D sensor is simply a suspended magnet fitted with a mirror that reflects a spot of light from a fixed source. As the magnet moves to and fro, following the small changes in the direction of magnetic north, the reflected spot moves left and right. The H sensor is essentially similar, except that torsion is introduced into the suspension until the magnet hangs E-W, with the torsion balancing the turning moment of H. If H increases, the magnet turns slightly towards magnetic north; if H decreases the magnet moves away from magnetic north. These movements cause a reflected light spot to move left and right. The Z sensor consists of a magnet balanced on knife edges so that it can turn in a vertical E-W plane. If pivoted at its centre of gravity, the turning moment of Z would cause it to stand vertical, but it is pivoted off-centre so that it comes to rest horizontally, with the Z turning moment compensated by the gravitational couple. Stability is achieved by having the knife edges above the magnet's centre of gravity. If Z increases, the magnet turns slightly towards the vertical, and *vice versa*. These movements are translated into left and right movements of a reflected spot by means of a prism and a polished upper surface on the magnet. The recorder is a horizontal cylinder which rotates once a day about its axis and which is covered with a sheet of photographic paper. The moving light spots are focussed on the drum by cylindrical lenses. Time signals are introduced either by interrupting the light source or by switching on an additional one at regular intervals. Baselines are provided by reflections from fixed mirrors. When developed, the *magnetogram* shows 3 wiggly lines that indicate the changes in D, H and Z throughout the day, time marks and straight baselines.

Before it can be used, the magnetogram must be calibrated. First, it is essential to know how to convert 1 mm of movement on the paper into nT. This is done for H and Z by imposing a known field on the sensor, and noting the deflection this produces on the photographic trace. This gives the *scale value* in nT/mm. For D, the scale value is in arcmin/mm, and can be deduced from the distance of the sensor from the recorder, since it is a simple optical lever. Secondly, we need to know the value of at least one point on the trace before the other values can be deduced. This is provided by the absolute observation.

In the well-controlled and constant temperature environment available at an observatory, the La Cour variometer is an excellent instrument, but it is certainly not portable. Special variometers have been developed for use in the field which are portable, robust and compact. One problem in the field is the lack of temperature control. In the Gough-Reitzel variometer, this is overcome by designing the instrument for installation in a shallow borehole.

Another disadvantage of the La Cour is the photographic recording and associated optics. It would be preferable to use electronic means of detecting and recording, so that the output could be immediate, *via* pen recorders, and machine readable, by recording digitally onto magnetic tape. Fluxgate sensors would appear to be suitable for this, but at present they appear to be less stable than suspended magnets. Another sensor with excellent accuracy and stability is the rubidium vapour magnetometer. This has been used with great success in the field, but has not been reliable enough for observatory use, where continuity of record is vital. Probably the real answer will be the SQUID magnetometer, but the price and the requirement for liquid helium will put it beyond the range of most observatories for at least a decade.

Automatic Instruments

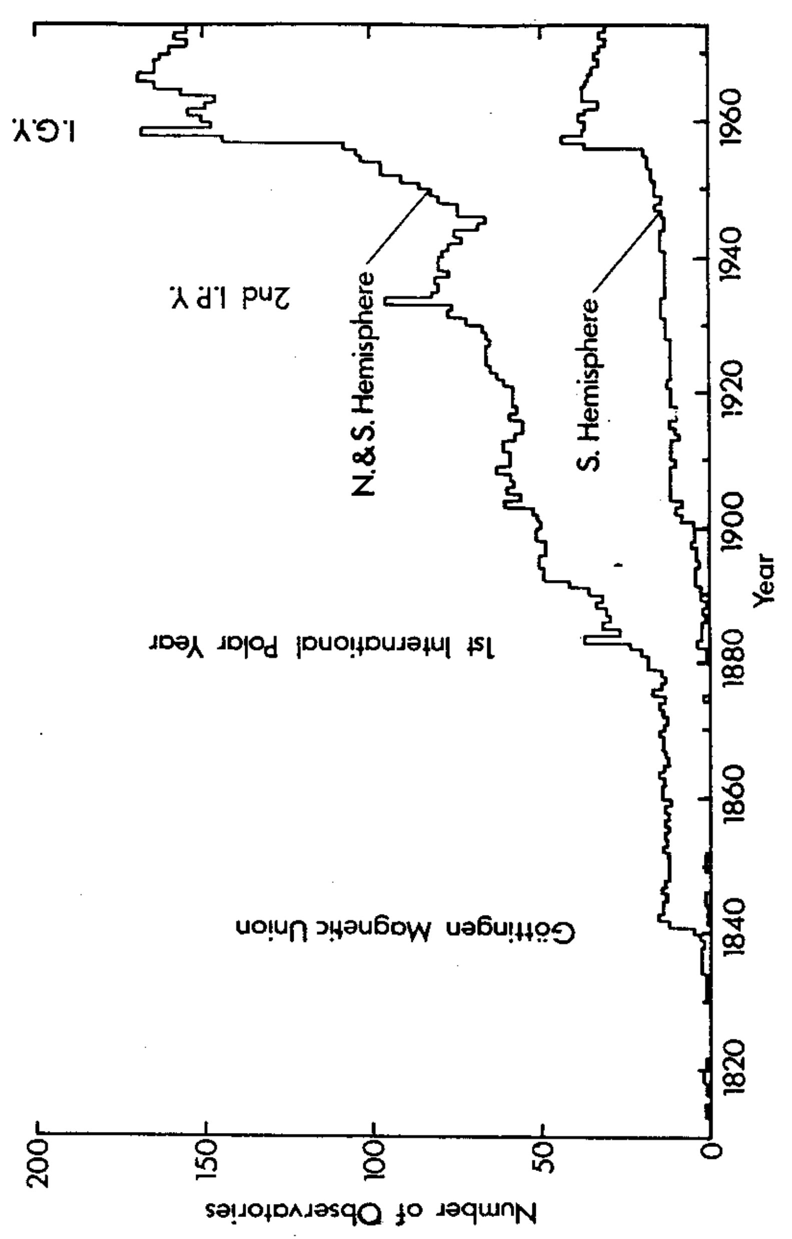
Some 2/3 of the Earth's surface is ocean. For a realistic global coverage it would be desirable to have ocean bed observatories, and this requirement prompted several groups to attempt to develop these. None met with any great success at sea, but they have proved useful at some land sites. One system is the Digitally Recording Proton Vector Magnetometer, or DRPVM. A proton magnetometer sensor is placed between 2 sets of Helmholtz coils, one with its axis magnetic East-West and the other perpendicular to both the magnetic vector and the other set of coils, in a direction which we will denote PQ. A sequence of 5 observations gives data from which absolute values of X, Y and Z can be deduced, and this is repeated every 30 s. The sequency is as follows: (i) no bias field, (ii) a bias field in the E direction, (iii) the same bias field in the W direction, (iv) a bias field in the P direction, (v) the same bias field in the Q direction. The unravelling of these five field intensity measures to give X, Y and Z is left as an easy exercise for the reader!

The Automatic Magnetic Observatory System (AMOS) is an instrument developed in Canada that has been widely used. Three orthogonal fluxgates measure X, Y and Z, and overall absolute control comes from a proton magnetometer measuring F and the constraint that $F^2 = X^2 + Y^2 + Z^2$. This is essentially the same system as was used on the recent very successful satellite MAGSAT.

Observatories

Much has been said of individual instruments such as might be used at observatories, but what of the observatories themselves? A magnetic observatory is a fixed site at which regular (preferably continuous) observations are made of the geomagnetic field. The earliest observatories were concerned only with D, such as that of George Graham, the famous clockmaker, who made thousands of observations of D at his house in Fleet Street between 1742 and 1748, and discovered the diurnal variation. Mark Beaufoy ran similar observatories at Hackney Wick and Bushey Heath from 1813 - 1822. However, the present worldwide system of observatories has grown from the Gottingen Magnetic Union of 1839.

In an attempt to discover the geographical extent of magnetic disturbances, Gauss and Humboldt solicited the collaboration of scientists from many countries in a project to observe the magnetic field at 5 minute intervals throughout 4 selected "term days" each year. With time, the number of collaborating observatories increased (see figure) and the frequency of observation increased from the 4 term days to every day and (with the introduction of photographic recording in the late 1840's) from 5 or 15 minute intervals to continuous. Various other projects, such as the International Polar Years of 1882 and 1932 and the International Geophysical Year of 1957, have given impetus to the observatory network which has continued



long after the specific project has finished. At present there are some 150 observatories in operation, of which less than a quarter are in the southern hemisphere.

A typical observatory comprises an office building, which houses the staff, workshop, darkroom, electronics laboratory and clocks; an absolute building in which the absolute instruments (eg proton vector magnetometer and declinometer) are installed on custom-built piers; a variometer building, usually windowless and commonly underground or double walled for thermal insulation, and miscellaneous other small buildings for batteries, instrument calibration, visitors experiments, etc. The absolute and variometer buildings are constructed throughout from non-magnetic materials, and should be remote from any artificial disturbance. It is the encroachment of artificial disturbances in the form of electric trains, motorways, carparks, factories, and housing estates that is the most common reason for the closure of observatories.

Routine Observations

At a typical observatory the routine is as follows:

Daily: change and develop magnetograms, check variograph lights on and drums rotating.

Weekly: make absolute observations (this should ideally be daily, but we are considering what is typical).

Monthly: make scale-value determinations.

Annually: check azimuth of mark against an astronomical observation.

Routine Reductions

Daily: estimation of C and K indicies from magnetograms.

Weekly: determine baseline values on magnetograms using absolute observations.

Monthly: adopt final baseline values representing a smooth fit to the individual determination; adopt scale values similarly, measure hourly mean ordinates in mm, convert these to nT using the adopted baselines and scale values. Note phenomenon.

Annually: complete annual volume by taking means for all days, quiet days and disturbed days for hours, days, months and year.

Use of Observatories

Observatories provide a means of monitoring magnetic activity, which is of importance in communications both by radio and telephone, and in power transmission. They provide a base for surveys where instruments can be calibrated and to which survey data can be reduced. They provide information on the secular variation which is essential for cartographers.

These are a few of the uses which are of direct commerical importance. However, the real reason for running an observatory is to contribute to the body of geomagnetic data which is used for purely scientific investigations relating to properties of all parts of the Earth from core to magnetosphere.

DESCRIPTION AND ANALYSIS OF MAIN FIELD

Data

The main sources of data on the geomagnetic field are as follows:

Observatories: These data have been described in some detail. They are of the highest accuracy, but it should be recognised that they refer only to the observatory site which may be anomalous and not representative of the surrounding area. Also, their number distribution is poor, with only 150 in total mostly concentrated in Europe.

Survey data: Surveys are carried out on land, at sea and in the air. Land survey data are usually of the highest quality, but are available for relatively few sites and only in restricted areas. Ocean surveys are now almost exclusively of F, usually measured with a towed proton magnetometer. The earlier ocean surveys by specially designed ships such as the Galilee, the Carnegil and the Zarya were of all elements, and more widely spread than recent ones, but they are now badly out of date. Air surveys blossomed in the 1950's and 60's, with the US "Project Magnet" and Canadian surveys from the Pacific to Scandinavia. While Project Magnet is still going on, it is at a reduced level except for regions near the US coast. The quantity of air data is not, in general, matched by its quality.

Satellite data: Until 1979, nearly all satellite magnetic data was for F only, though there was a vast quantity of it. The main sources were the OGO (Orbiting Geophysical Observatory) series of satellites, the even numbered ones being in nearly circular polar orbits, giving a detailed coverage of the whole Earth, though with greater density of observations near the poles. The Cosmos 49 satellite did not go to higher latitude (N or S) than 50°. The recent, highly successful MAGSAT satellite was in a polar orbit at very low altitude and measured the complete magnetic vector. The MAGSAT data are still being actively exploited by many groups.

Charts

The simplest way of summarizing a body of magnetic data is by means of a chart showing isolines of one of the elements. The first map of isomagnetic lines was probably that produced in about 1640 by the Jesuit Christoforo Borri and Martin Martino, though it was not uncommon for earler charts to give some form of compass information. This chart is referred to by Kircher (1643) and is probably also that alluded to by Whiston (1721). Unfortunately, no copies of the chart have survived, so credit for the first isomagnetic chart usually goes to Edmond Halley for his Atlantic chart of declination, published c. 1701. Certainly this chart, which was based on Halley's own observations, and its worldwide successor, which incorporated data from other sources and appeared a year later, were the first to gain general acceptance and to be of practical use to navigators.

As Halley himself was well aware, the magnetic field pattern does not remain constant, and this secular change gradually rendered his charts obsolete. Over the next 150 years, various individuals produced charts when the need arose, collecting data in an *ad hoc* manner, reducing them to epoch where they had information on the secular change, and contouring the results. In 1858 the British Hydrographic Office assumed responsibility for the systematic collection of magnetic data and the preparation and publication of world magnetic charts. At first, the preparation was by naval officers, but was later passed on to the Magnetical and Meteorological Department of the Royal Greenwich Observatory, which department, though now absorbed into the Geomagnetism Unit of the Institute of Geological Sciences, still retains the responsibility. Similar arrangements exist in the USA and the USSR. World declination charts are published every 5 years, and charts of the other magnetic elements every 10 years.

In the early days, when data were sparse, magnetic cartography was more art than science, with considerable scope for imaginative contouring in the spaces between observations. As more data became available, the overall pattern became more clearly defined, but artistic judgement was still required in drawing smooth contours through scattered values. Occassionally, charts have been produced with the contours contorted to provide an exact fit to all the data (eg Bock 1948), but experience has shown that smoother charts usually provide better fits to subsequent data as well as being easier to read.

Besides fitting the data a model may be expected to satisfy physical constraints. In the case of charts, some of the constraints are geometrical. As well as the obvious ones, such as the requirement that contours should not cross one another, there are more subtle ones, particularly regarding the behaviour of contours near the geographical and geomagnetic poles. Where charts of several different elements are produced they should, of course, be mutually compatible. A full discussion of the constraints that isomagnetic maps should satisfy has been given in a series of papers by Chapman. These papers are mainly theoretical. The detailed techniques for the practical production of magnetic charts have mostly been passed on by word of mouth rather than by publication, but some of them have been recorded by Sucksdorff (1981).

The greatest revolution in magnetic cartography, as in many branches of science, has been brought about by electronic computers. Not only do they permit the direct processing of large quantities of data to produce the parameters of a mathematical model of the magnetic field, they can also be coaxed into drawing the actual contours with only the minimum of human intervension. This is not all gain; there are those of us who miss the satisfaction of refining a set of contours for maximum sensual appeal while still not departing too far from the truth. Also, computers may make magnetic charts obsolete. Already it is common in aircraft for magnetic declination data to be generated in a small computer rather than read from a chart.

Spherical Harmonic Analysis

The method of spherical harmonic analysis was devised by Gauss (1839) specifically for the purpose of modelling the geomagnetic field. He showed that, if the geomagnetic field intensity can be represented as the gradient of a potential, it can be written as a linear combination of an infinite series of spherical harmonic coefficients, some of which represent the part of the potential of internal origin and others the part of external origin. Here, internal and external are relative to the surface of a reference sphere, which can be chosen to be the surface of the Earth.

The potential, V, at a point with spherical coordinates (θ,ϕ,r) may be represented in spherical harmonic form by

$$V = K + R \sum_n \sum_m \{(c_n^m \cos m\phi + s_n^m \sin m\phi)\ (R/r)^{n+1} + (Y_n^m \cos m\phi$$
$$+ \sigma_n^m \sin m\phi)\ (r/R)^n\}\ P_n^m\ (\cos\theta).$$

Here, θ denotes the colatitude (North polar distance),

ϕ denotes East longitude,

r denotes the radial distance from the centre of the Earth,

K is an arbitrary constant,

R denotes the radius of a reference sphere, whose centre coincides with that of the Earth,

c_n^m, s_n^m denote the spherical harmonic coefficients associated with the part of V that originates within the reference sphere,

Y_n^m, σ_n^m denote the spherical harmonic coefficients associated with the part of V that originates outside the reference sphere

and P_n^m (cos θ) denotes the associated Legendre polynomial of degree m and order n.

Following the recommendations of the International Association of Terrestrial Magnetism and Atmospheric Electricity (Goldie 1940), Schmidt quasi-normalization is used, so that

$$P_n^m\ (x) = \frac{1}{2^n n!} \left\{ \frac{\varepsilon_m (n-m)!\ (1-x^2)^m}{(n+m)!} \right\}^{\frac{1}{2}} \frac{d^{n+m}}{dx^{n+m}}\ (x^2-1)^n,$$

Where $\varepsilon_m = 1$ for $m = 0$; $\varepsilon_m = 2$ for $m = 1, 2, 3, \ldots..$

The North (X), East (Y) and vertically downward (Z) components of magnetic intensity are derived from V as follows:

$$X = \frac{1}{r}\frac{\partial V}{\partial\theta},\quad Y = \frac{1}{r\sin\theta}\ \frac{\partial V}{\partial\phi}\ \text{and}\ Z = \frac{\partial V}{\partial r}.$$

That is

$$X = \sum_n \sum_m \{(c_n^m \cos m\phi + s_n^m \sin m\phi)\ (R/r)^{n+2} + (Y_n^m \cos m\phi + \sigma_n^m \sin m\phi)$$

$$(r/R)^{n-1}\}\ nX_n^m\ (\cos\theta),$$

$$Y = \sum_n \sum_m \{(c_n^m \sin m\phi - s_n^m \cos m\phi)\ (R/r)^{n+2} + (Y_n^m \sin m\phi - \sigma_n^m \cos m\phi)$$

$$(r/R)^{n-1}\}\ nY_n^m\ (\cos\theta),$$

$$Z = -\sum_n \sum_m \{(c_n^m \cos m\phi + s_n^m \sin m\phi)\ (n + 1)\ (R/r)^{n+2} - (Y_n^m \cos m\phi$$

$$+ \sigma_n^m \sin m\phi)\ n\ (r/R)^{n-1}\}\ P_n^m\ (\cos\theta);$$

where

$$nX_n^m\ (\cos\theta) = \frac{d}{d\theta}\{P_n^m\ (\cos\theta)\} \text{ and } nY_n^m\ (\cos\theta) = \frac{m}{\sin\theta}\ P_n^m\ (\cos\theta).$$

A spherical harmonic analysis is the process of determining the numerical values of the spherical harmonic coefficients from a set of geomagnetic data. Clearly, it is impossible to determine an infinite set of coefficients, so in practice the series is truncated at some point. The level of truncation determines the complexity of the model, corresponding to degree of smoothing in the case of a chart. Provided the data are in the form of the orthogonal elements X, Y, or Z, each datum can be expressed as a linear combination of the required coefficients, as shown above, giving an equation of condition. There are usually many more equations of condition than there are coefficients to be determined, so an exact solution is not possible. Instead, another legacy from Gauss, the method of least squares, is generally used to determine the set of coefficients that most nearly fits the data.

It is common practice to solve the X and Y equations for g_n^m and h_n^m, where

$$g_n^m = \left(c_n^m - \frac{n}{n+1}\,Y_n^m\right) \text{and} \quad h_n^m = \left(s_n^m - \frac{n}{n+1}\,\sigma_n^m\right)$$

The internal and external spherical harmonic coefficients are then separated as follows:

$$\left.\begin{aligned} c_n^m &= ng_n^m + (n + 1)\ g'^m_n\ /\ (2n + 1), \\ s_n^m &= nh_n^m + (n + 1)\ h'^m_n\ /\ (2n + 1), \\ Y_n^m &= (g_n^m - g'^m_n)\ (n + 1)\ /\ (2n + 1), \\ \sigma_n^m &= (h_n^m - h'^m_n)\ (n + 1)\ /\ (2n + 1). \end{aligned}\right\}$$

If, as appears to be true for the main geomagnetic field, the external part is negligible, then $g_n^m = g'^m_n$ and $h_n^m = h'^m_n$. In this case, the equations for X, Y and Z may be solved simultaneously to give c_n^m and s_n^m directly.

If the data are read from charts, a number of simplifications are possible. For example, with data on a uniform geographical grid, the analysis can be separated into two distinct parts. The first is concerned with the dependence of the data on longitude. This dependence can be described by the first few terms of a Fourier series, with a separate set of terms for each value of latitude. Since the data are uniformly spaced in longitude, the Fourier coefficients are simply the sums of the products of the magnetic field values and constant factors. In the second part of the analysis, latitude dependence is considered. Each of the Fourier coefficients can be considered separately as a function of latitude involving not more than n spherical harmonic coefficients. For example, the first (constant) term in the Fourier series is a function of the spherical harmonic coefficients $g_1^o, g_2^o, \ldots g_n^o$. The n coefficients could be directly determined by least squares, involving the solution of simultaneous equations in n unknowns, but a further simplification is possible. Since the data are uniformly spaced in latitude and spherical harmonics are alternately symmetrical and antisymmetrical about the equator, the sum of the northern and southern hemispheres involves only half of the coefficients, and the difference of the hemispheres involves only the other half. Thus, there is never need to solve for more than $n/2$ unknowns.

In practice, the data are not sufficiently well distributed for a uniform grid to represent their distribution, so it is better to do a direct solution of the equations of condition, with the data weighted according to their accuracy.

It has been implicitly assumed that the reference sphere corresponds to the surface of the Earth, but this can be only an approximation since the Earth is more nearly an oblate spheroid than a sphere. It is possible to make a sphroidal harmonic analysis to allow for this (eg Jones & Melotte 1953), but it is more usual to stick to a spherical reference surface, and take account of the fact that observations on the surface of the Earth are not on the reference sphere, and that geographical north and verical are slightly off the directions of θ and r. The only complication is that 'internal' and 'external' are relative to the reference sphere rather than the Earth's surface.

The equations of condition for the non-orthogonal elements D, I, H, F are not linear in the spherical harmonic coefficients, so their direct solution is not possible. With the advent of satellite data which was initially exclusively F, it became urgent to find a way of solving the F equations. After some false starts, the solutions proved quite simple. When the equations were differentiated they became linear in Δg_n^m, Δh_n^m, which denotes small corrections to g_n^m, h_n^m. Thus, one specifies an initial set of g_n^m, h_n^m and solves for the corrections to the true values. In principle, the solution should be iterative, but the rate of convergence is so rapid that only one iteration is needed, even when the initial model is poor.

The numerical techniques for solving large numbers of equations in many unknowns have been greatly modified over the years, particularly since the introduction of electronic computers. The early solutions were *via* normal equations using Gaussian elimination. This was replaced by matrix inversion methods, though still *via* normal equations (see Appendix). There are many ways of inverting matrices, but the best is probably that of Gauss and Jordan, with pivotal searching. The most recent methods of solving simultaneous equations completely by-pass the normal equations stage, and operate directly on the equation of condition matrix. These 'QR decomposition' methods are much more stable and not much slower than matrix inversion.

Results

The spherical harmonic coefficients are themselves a summary of a vast body of geomagnetic data, so it is difficult to summarize them further. However, some of the low order coefficients lend themselves to physical interpretation.

The g_1^0 term corresponds to a dipole at the centre of the Earth with its axis pointing north. This *axial dipole* is the first approximation to the geomagnetic field. The next approximation is the *centre dipole*, in which the axis is allowed to point in the optimum direction (towards north Greenland at present). Its movement, M, is given by $MR^3 = \left[(g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2\right]^{\frac{1}{2}}$. The next approximation, the *eccentric dipole*, requires higher order coefficients for its specification. There are several ways of defining the eccentric dipole, depending on the number of coefficients one takes into account, but they all end up by specifying the position of the dipole (3 parameters), the direction of its axis (2 parameters) and its moment (1 parameter). The simplest version is deduced directly from the first 6 spherical harmonic coefficients: g_1^0, g_1^1, h_1^1, g_2^0, g_2^1, h_2^1.

One can go no further with a dipole; better approximations require more dipoles or alternatively, centred multipoles. Centred multipoles are more simply related to spherical harmonic coefficients, since there is one for each n. The dipole is specified by the 3 $n = 1$ terms, the quadrupole by the 5 $n = 2$ terms, the octupole by $n = 3$, and so on. Going backwards, the single term g_0^0 corresponds to a monopole at the centre of the Earth. In the few analyses where g^0, has been sought, it has never been found to differ significantly from zero.

SECULAR VARIATION

Discovery

The discovery of secular variation was a fine example of teamwork, starting with an accurate measure of declination in London in 1580 by William Borough, who had served under Drake. In 1622 Edmund Gunter (professor of astronomy and inventor of the surveyor's chain and the slide rule) made a new measurement at the same place, and found a value of declination 5 degrees less than that of Borough. Being a cautious man, he did not leap to any conclusions in case the earlier observation was wrong. It was left to his successor, Henry Gellibrand, to find that the declination had decreased by a further 2 degrees by 1634 and thus show that the Earth's magnetic field really does change with time.

The secular variation is of particular geophysical importance since it originates in the Earth's core and is one of the very few clues to dynamical processes deep within the Earth. It also gives information on the conductivity of the mantle, through which it has to pass before being observed at the surface.

Data

By far the most reliable sources of secular variation (sv) data are observatory annual means, but we have already noted the unsatisfactory geographical distribution of these.

Other sv data come from *repeat stations*. These are marked sites (the mark is commonly a buried tile or a brass stud, which can be found with the aid of a 'treasure map' held by the observer) at which a set of absolute, or near absolute, observations are made at intervals of a few years. Provided the observations are good and the site has not been poluted by magnetic material in the interval between observations, the change in field should be a reasonable measure of secular variation. In practice, it is surprising how seldom the provisos are satisfied. However, repeat stations do give some additional sv information which can be of value in regions remote from observatories.

Another source of sv data, though still less reliable, is the difference between successive surveys of the same region. Unless the survey points are the same on both occasions (ie repeat stations), the comparison must be made *via* maps, or some other form of local model.

It is anticipated that really good sv data will come from a comparison of the MAGSAT results with those from a subsequent, similar satellite, that has yet to be launched. While this is undoubtedly true, this will only give the mean sv between the dates of the 2 satellites, and observatory data will still be required for the higher temporal resolution required to reveal such important features as jerks (see below).

Morphology

Charts of sv have been produced since about 1858 for use in correcting world chart data. They are still used for this purpose, but are now also used for geophysical research. The greatest rates of change occur in Z, the isolines of which shows 3 deep foci, one positive in Antarctica and two negative, one in the Atlantic and the other in the Indian ocean. In the first 2 foci, the rate of change exceeds 140 nT yr^{-1}. The X and Y charts both show zero contours passing through these foci. Those for X pass through from E to W, those for Y from N to S, as would be expected from potential theory. Spherical harmonic analysis of the sv potential shows that it is less dominantly dipolar than the main field and that the convergence of the harmonic series is less rapid. The slow convergence is an embarrassement. It suggests that a large number of terms are required for an adequate description, but the inadequacy of the data prevents this. It also suggests that the higher (undetermined) coefficients are of great importance at the core-mantle interface, where the field originates, and to which level we would wish to extrapolate it.

Higher Derivatives

The sv itself changes with time, and its rate of change is called the *secular acceleration*. It has been shown that secular acceleration is sufficiently stable and well-determined to be of value in extrapolating data for the production of navigation charts.

The third time derivative in the main field is called the *jerk*, by analogy with mechanics. It is usually nearly zero, but a pulse occurred in about 1970 which was of surprisingly short duration. The pulse has been shown to originate within the Earth, and it would be expected that such a short lived phenomenon would be attentuated to insignificance when passing through the conducting mantle. The fact that it is detectable at the surface of the Earth suggests that the conductivity of the mantle is a lot less than had been previously suspected.

Other Implications

If the Earth's core is (to a good approximation) a perfect conductor, the total number of lines of force passing through its surface should be constant. This proposition can be tested using spherical harmonic models of the geomagnetic field and its sv. The departures of the flux from constancy are within the uncertainty of the models. The procedure can be reversed, and the constraint of constant flux be applied when determining the coefficients, and this should lead to better models.

There are other consequences of perfect core conductivity that suggest that more complicated functions of the main field and sv should be invariant. These are being examined.

Again assuming constant core conductivity, the lines of force are 'frozen in' to the core, so movement of one implies the same movement of the other. We cannot see lines of force, but the main field and sv tell us some of their properties and movements, so we can obtain some (limited) information about motions at the surface of the liquid core.

Appendix

Least Square Solutions

Consider a set of equations of condition:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \;\cdots\cdots + a_{1n}x_n = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \;\cdots\cdots + a_{2n}x_n = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \;\cdots\cdots + a_{3n}x_n = c_3 \\ \\ a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 \;\cdots\cdots + a_{mn}x_n = c_m \end{array}\right\} \quad (1)$$

here $m \geqslant n$, the number of unknowns.

We wish to determine the values of x_j, $j = 1$ to n, that minimises the sum of the squares of residuals, R , where

$$R = \sum_{k=1}^{n} (\epsilon_k)^2$$

and

$$\epsilon_k = a_{k1}x_1 + a_{k2}x_2 + a_{k3}x_3 \cdots\cdots + a_{kn}x_n - c_k \qquad (2)$$

For R to be a minimum,

$$\frac{dR}{dx_j} = 0. \quad \text{But } \frac{dR}{dx_j} = \sum_{k=1}^{m} (2a_{kj}\,\epsilon_k), \qquad (3)$$

so

$$\sum_{k=1}^{m} (a_{k1}.a_{kj}\,x_1 + a_{k2}.a_{kj}\,x_2 + a_{k3}.a_{kj}\,x_3 \;\cdots\cdots\; a_{kn}.a_{kj}\,x_n - c_k.a_{kj}) = 0 \qquad (4)$$

Since j takes all values from 1 to n, this provides n such equations in n unknowns.

Adopting the notation

$$\sum_{k=1}^{m} (a_{ki}\cdot a_{kj}) \equiv [a_i a_j] ,$$

these *normal equations* may be written in the form

$$\begin{bmatrix} [a_1a_1] & [a_2a_1] & [a_3a_1] & \cdots\cdots & [a_na_1] \\ [a_1a_2] & [a_2a_2] & [a_3a_2] & \cdots\cdots & [a_na_2] \\ [a_1a_3] & [a_2a_3] & [a_3a_3] & \cdots\cdots & [a_na_3] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [a_1a_n] & [a_2a_n] & [a_3a_n] & \cdots\cdots & [a_na_n] \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} [c\ a_1] \\ [c\ a_2] \\ [c\ a_3] \\ \vdots \\ [c\ a_n] \end{bmatrix} \qquad (5)$$

In matrix form: AX = C. Hence X = $A^{-1}C$, where A^{-1} is the inverse of A.

Having thus obtained the n values x_j, we still require their standard deviations, s_j, where

$$s_j = \left(\frac{W_j R}{m-n}\right)^{\frac{1}{2}}$$

and w_j is the jth element of the leading diagonal of the inverse matrix. R may be obtained rather laboriously by substituting the values of x_j into equations (2) and summing ε_k^2 for k = 1 to m (though this method has the advantage of permitting an examination of the individual residuals, which frequently indicates erroneous data). Alternatively, R may be obtained from:

$$R = \sum_{k=1}^{m} (c_k)^2 - \sum_{j=1}^{m} x_j\ [c\ a_j] \qquad (6)$$

The proof of (6) is as follows: from (2)

$$R = \sum_k \varepsilon_k\cdot\varepsilon_k = x_1\sum_k a_{k1}\,\varepsilon_k + x_2\sum_k a_{k2}\,\varepsilon_k + x_3\sum_k a_{k3}\,\varepsilon_k \cdots$$

$$\cdots + x_n\sum_k a_{kn}\,\varepsilon_k - \sum_k c_k\,\varepsilon_k$$

but from (3), all terms except the last are zero when R is a minimum. Thus

$$R = -\sum_k c_k\,\varepsilon_k = -x_1\sum_k a_{k1}\,c_k - x_2\sum_k a_{k2}\,c_k - x_3\sum_k a_{k3}\,c_k \cdots$$

$$\cdots - x_n\sum_k a_{kn}\,c_k + \sum_k (c_k)^2$$

which is the same as (6).

Weighting If we give the kth equation of condition weight N_k, this means that we consider its importance to be equivalent to N_k equations of unit weight. We wish its contribution to the normal equations to be the same as that which would be obtained by entering the same equation N_k times with unit weight; that is, we wish to replace $a_{ki}.a_{kj}$ with $N_k a_{ki}.a_{kj}$. This is most simply done by multiplying the whole equation k by $\sqrt{N_k}$, and then proceding *exactly* as in the un-weighted case. (NB. N_k is not constrained to be an integer; m is not affected by the weighting; no different treatment of the residuals is required after the above-mentioned multiplication has been carried out).

Appendix to Appendix

SOLUTION TO NORMAL EQUATIONS

1 Gaussian Elimination

Probably the most basic method of solving *n* equations in *n* unknowns. Multiples of one equation, the pivotal equation, are subtracted from the other equations to eliminate one of the variables from all of them. This process is repeated until a triangular system is obtained and solution is by back substitution.

This may be executed with or without 'pivotal searching'. Loss of significant figures may result if an equation with a relatively small coefficient for the variable to be eliminated is chosen as the pivotal equation. It is usual to take the variables in order and use the equation with the largest coefficient at that stage for the elimination of a given variable. Full pivotal searching involves looking for the largest coefficient in the whole array and eliminating the corresponding variable with that equation; however this may be counter-productive by giving very small pivots in the later stages of elimination. (The extra computing time is not worth the effort.)

An example is given below:

$$\begin{array}{rcrcrcrll}
x & + & 1/2y & + & z & = & -5 & \text{i} & \text{Pivotal eq}^{n}\text{ for stage 1} \\
1/2x & + & 3/2y & + & z & = & 6 & \text{ii} & \\
x & + & y & + & 3/2z & = & -1 & \text{iii} & \\
\hline
 & & 5/4y & + & 1/2z & = & 8\ 1/2 & \text{ii} & -1/2 \text{ x i Pivotal eq}^{n}\text{ for stage 2} \\
 & & 1/2y & + & 1/2z & = & 4 & \text{iii} & -1 \text{ x i} \\
\hline
 & & & & 3/10z & = & 3/5 & \text{iii} & -2/5 \text{ x ii}
\end{array}$$

$z = 2,\ y = 6,\ x = -10$

This may also be written omitting the variable names:

$$\begin{array}{ccc|c}
1 & 1/2 & 1 & -5 \\
1/2 & 3/2 & 1 & 6 \\
1 & 1 & 3/2 & -1 \\
\hline
 & 5/4 & 1/2 & 8\ 1/2 \\
 & 1/2 & 1/2 & 4 \\
\hline
 & & 3/10 & 3/5
\end{array}$$

The sum of squares of residuals must be calculated by substituting into the equations of condition.

2 Matrix Inversion

Start with the same set of normal equations, which may be written in matrix form:

$$\begin{bmatrix} 1.0 & 0.5 & 1.0 \\ 0.5 & 1.5 & 1.0 \\ 1.0 & 1.5 & 1.0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -5.0 \\ 6.0 \\ -1.0 \end{bmatrix} \qquad \begin{matrix} (1) \\ (2) \\ (3) \end{matrix}$$

The unit matrix on the right is not necessary, but will illustrate the method which is a systematic way of converting the left hand matrix to a unit matrix while the right hand matrix becomes the inverse.

A(i) Divide the first row by its first element, so that the first element becomes 1.0. (In this example it is already 1.0, so we merely rewrite the line:

$$\begin{array}{ccccccc} 1.0 & 0.5 & 1.0 & 1 & 0 & 0 & \qquad (1A) \end{array}$$

A(ii) Reduce the first element of the second row to zero by subtracting a suitable multiple of (1A) from it. In this case the multiple is 0.5:

$$\begin{array}{ccccccc} 0 & 1.25 & 0.5 & -0.5 & 1 & 0 & \qquad (2A) \end{array}$$

A(iii) Reduce the first element of the third row to zero by subtracting a suitable multiple of (1A) from it. In this case the multiple is 1.0:

0 0.5 0.5 -1 0 1 (3A)

Note that the first column of the matrix has become the first column of a unit matrix. Now we operate on the second column.

B(i) Reduce the second element of the second row to 1.0 by dividing the whole row by 1.25:

0 1 0.4 -0.4 0.8 0 (2B)

B(ii) Reduce the second element of the first row to zero by subtracting 0.5 x (2B):

1 0 0.8 1.2 -0.4 0 (1B)

B(iii) Reduce the second element of the third row to zero by subtracting 0.5 x (2B):

0 0 0.3 -0.8 -0.4 1 (3B)

Pause for a re-write and to note that the first 2 columns are those of a unit matrix. One column to go.

1 0 0.8 1.2 -0.4 0 (1B)
0 1 0.4 -0.4 0.8 0 (2B)
0 0 0.3 -0.8 -0.4 1 (3B)

C(i) Reduce the third element of the third row to unity by dividing by 0.3:

0 0 1 -8/3 -4/3 10/3 (3C)

C(ii) Reduce the third element of the first row to zero by subtracting 0.8 x (3C):

1 0 0 10/3 2/3 -8/3 (1C)

C(iii) Reduce the third element of the second row to zero by adding 0.4 x (3C):

0 1 0 2/3 4/3 -4/3 (2C)

Re-writing in full, we see that we have got there:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 10/3 & 2/3 & -8/3 \\ 2/3 & 4/3 & -4/3 \\ -8/3 & -4/3 & 10/3 \end{bmatrix} \begin{bmatrix} -5 \\ 6 \\ -1 \end{bmatrix} \quad \begin{matrix} (1C) \\ (2C) \\ (3C) \end{matrix}$$

The unknowns are obtained by multiplying out:

$$\begin{aligned} x &= (10/3)\ (-5) + (2/3)\ (6) + (-8/3)\ (-1) = -10 \\ y &= (2/3)\ (-5) + (4/3)\ (6) + (-4/3)\ (-1) = 6 \\ z &= (-8/3)\ (-5) + (-4/3)(6) + (10/3)\ (-1) = 2 \end{aligned}$$

3 Matrix Inversion with Pivotal Searching

For an exact solution, such as that given above, there are no problems with rounding errors. However, they are usually important when inverting large matrices on a computer. It is important to reduce rounding errors as far as possible by using pivotal searching. The method is essentially the same as that given above, except that, instead of working through the rows in sequence reducing the diagonal element to unit, one selects the row with the largest element. It can be shown that, for a normal equations matrix, this *pivotal element* is invariably on the leading diagonal. In compressed form, the procedure is as follows:

1 1/2 1 1 0 0 (1)
1/2 3/2 1 0 1 0 (2)
1 1 3/2 0 0 1 (3)

The largest element is the middle one, so we reduce this to 1 and the ones above and below it to zero:

(1)-1/2(2A):	5/6	0	2/3	1	-1/3	0	(1A)
(2)÷3/2:	1/3	1	2/3	0	2/3	0	(2A)
(3)-(2A):	2/3	0	5/6	0	-2/3	1	(3A)

The largest element outside row 2 (which has already been done) is the first. Here we go again:

(1A)÷5/6:	1	0	4/5	6/5	-2/5	0	(1B)
(2A)-1/3(1B):	0	1	2/5	-2/5	4/5	0	(2B)
(3A)-2/3(1B):	0	0	3/10	-4/5	-2/5	1	(3B)

The largest diagonal element outside rows 1 and 2 is the last:

(1B)-4/5(3C):	1	0	0	10/3	2/3	-8/3	(1C)
(2B)-2/5(3C):	0	1	0	2/3	4/3	-4/3	(2C)
(3B)÷3/10:	0	0	1	-8/3	-4/3	10/3	(3C)

Which is the same as we had before. (Phew!)

4 Matrix Inversion with Pivotal Searching and Compact Storage

The next stage is to save some storage space. At all stages of the above calculation, half the columns are those of a unit matrix and need not be stored. In the compact form of the solution these are omitted, and the inverse matrix overwrites the original:

Open						Compact		
1	1/2	1	1	0	0	1	1/2	1
1/2	3/2	1	0	1	0	1/2	3/2	1
1	1	3/2	0	0	1	1	1	3/2
5/6	0	2/3	1	-1/3	0	5/6	-1/3	2/3
1/3	1	2/3	0	2/3	0	1/3	2/3	2/3
2/3	0	5/6	0	-2/3	1	2/3	-2/3	5/6
1	0	4/5	6/5	-2/5	0	6/5	-2/5	4/5
0	1	2/5	-2/5	4/5	0	-2/5	4/5	2/5
0	0	3/10	-4/5	-2/5	1	-4/5	-2/5	3/10
1	0	0	10/3	2/3	-8/3	10/3	2/3	-8/3
0	1	0	2/3	4/3	-4/3	2/3	4/3	-4/3
0	0	1	-8/3	-4/3	10/3	-8/3	-4/3	10/3

A further saving of space can be achieved by noting that both the normal equations and its inverse are symmetrical, so the elements below the diagonal are redundant. This is not true of the intermediate stages, but the changes of sign follow a systematic pattern that can be deduced. The numbers (including those from the right hand side of the equations) are stored in a linear array, whose contents change as shown below:

Row 1			RHS 1	Row 2		RHS 2	Row 3	RHS 3
1	1/2	1	-5	3/2	1	6	3/2	-1
5/6	-1/3	2/3	-7	2/3	2/3	4	5/6	5
6/5	-2/5	4/5	-42/5	4/5	2/5	34/5	3/10	3/5
10/3	2/3	-8/3	-10	4/3	-4/3	6	10/3	2
			x			y		z

This is the method of Malin, Barraclough & Hodder, called ZAPP

5 QR Decomposition

We require a least squares solution of the system

$$C\,\vec{x} = \vec{d}$$

where C is the matrix of coefficients of the equations of condition, $\vec{x}$ is the column vector of the unknowns (to be determined) and $\vec{d}$ is the column vector of the right hand sides of the equations.

C has dimensions m x n where $m \geqslant n$; if m = n the solution is unique, provided C is non-singular.

Suppose the least squares solution for the unknowns is $\vec{x}'$; then $C\,\vec{x}' - \vec{d} = \vec{r}$ where $\vec{r}$ is the column vector of residuals.

We may write C = QR where $Q^TQ = I$, and R is upper triangular on top of zeros.

$$\text{Then}\quad |\vec{r}|^2 = ||C\,\vec{x}' - \vec{d}||^2$$

$$= ||QR\,\vec{x}' - \vec{d}||^2 \qquad\qquad R = \begin{bmatrix} 0\cdots\triangledown \\ 0 \end{bmatrix}\begin{matrix} \updownarrow n \\ \updownarrow m-n \end{matrix}$$

Multiplying by Q^T does not change the length (since Q = $\sqrt{T}$, effectively)

$$|\vec{r}|^2 = ||R\vec{x}' - Q^T\vec{d}||^2$$

Put $Q^T\vec{d} = \begin{bmatrix} d_1 \\ \cdots \\ d_2 \end{bmatrix}$ where d_1 is n x 1, d_2 is (m − n) x 1, $|\vec{r}|^2 = r^2$, the sum of squares of the residuals.

$$r^2 = \left\|\begin{bmatrix} 0\cdots\triangledown \\ 0 \end{bmatrix}\vec{x}' - \begin{bmatrix} d_1 \\ \vdots \\ d_2 \end{bmatrix}\right\|^2$$

$$= \left\|\begin{matrix} R\vec{x}' - [d_1] \\ \cdots\cdots \\ [d_2] \end{matrix}\right\|^2$$

$$= \left\| R\vec{x}' - [d_1] \right\|^2 + \left\|[d_2]\right\|^2$$

The second term is a constant. Condition for least squares solution is

$$R\vec{x}' = \vec{d}_1$$

Since R is upper triangular, we get the unknowns by back substitution.

The problem has now been reduced to that of finding R and $\vec{d}_1$, $\vec{d}_2$

Denote the columns of C by $\underline{c}_i$, i = 1, n, where $\underline{c}_i = \begin{bmatrix} c_{1i} \\ c_{2i} \\ c_{3i} \\ \vdots \\ c_{mi} \end{bmatrix}$

The first stage of the transformation is equivalent to a 'reflection' in m-dimensional space so that the transformed $\underline{c}_1$ has all elements after the first zero. Striking out the first row and column of **C** leaves an m − 1 x n − 1 matrix which can similarly be transformed so that its first column has zeros below the leading diagonal. This is repeated until all the sub-diagonal elements are zero and the resulting matrix is R above. If the right-hand sides are transformed in the same way at each stage, we obtain $\begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$. Back substitution gives the least squares solution for $\vec{x}$ and $||(d_2)||^2$ is the sum of squares of the residuals.

Let $\underline{c}_1$ denote the position vector of P $(c_{11}, c_{21}, c_{31}, \ldots c_{m1})$, and $\underline{c}^*$ the vector $(\epsilon_{11}\,|\underline{c}_1|, 0, 0, \ldots, 0)$. The required axis of reflection is $\vec{OA}$ where O is the origin and A has position vector $\underline{a} = \underline{c}_1 + \underline{c}^*$, if $c_{11} > 0$, $\underline{a} = \underline{c}_1 - \underline{c}^*$ if $c_{11} < 0$.

The reflection of P in this axis is Q, position vector $\underline{c}^*$ (or $-\underline{c}^*$). The reflection of R $(c_{1i}, c_{2i}, c_{3i}, \ldots, c_{mi})$ in the axis is **S** where

$$\vec{OS} = \frac{2\,(\underline{c}_i.\underline{a})\,\underline{a}}{|\underline{a}|^2} - c_i$$

$$|\underline{a}|^2 = 2l\,(l + c_{11}) \qquad\qquad \text{where } l = |\underline{c}_1|$$

$$\vec{OS} = \frac{(\underline{c}_i \cdot \underline{a})}{l(l+c_{11})}\,\underline{a} - \underline{c}_i$$

In the programme, the 'reflection' of P is set as $(-\varepsilon_{11}|c_1|, 0, 0, 0,.., 0)$ and the components of the other transformed vectors are calculated as

$$\underline{c}_i - \frac{(\underline{c}_i \cdot \underline{a})}{l(l+c_{11}}\,\underline{a}$$

This merely has the effect of multiplying the equations through by -1.