Journal of Geodetic Science

• 1(1) • 2011 • 82-87 DOI: 10.2478/v10156-010-0010-7 •

A contemporary perspective of geoid structure

Letter to the Editor

H. Moritz*

Institute of Navigation and Satellite Geodesy Graz University of Technology

Abstract:

The present paper reviews the contemporary state of definition and theory of the geoid. Key features are: quasigeoid, external gravitational field from satellites and its analytical downward continuation to the Earth's interior, data combination by least-squares collocation, and a new view of gravity reduction. This is done under the modern systematic perspective provided by the possibility of a purely geometric satellite determination of the Earth' surface by GPS combined with satellite altimetry.

Keywords:

Analytical continuation • geoid • least squares collocation • physical geodesy • quasi geoid

Received 27 November 2010; accepted 21 December 2010

1. Introduction

Modern geoid definition and determination have developed remarkably since about 1950, for the following main reasons:

1. Molodensky's theory from 1945 of doing physical geodesy without using the geoid, working directly with the topographical Earth surface;
2. The launch of the first Sputnik and Explorer satellites (1957/58);
3. The first practically useful gravimetric geoid determinations at about the same time;
4. The use of statistical theories of gravity at about the same time;
5. The first satellite-determined and combined gravitational field solutions, a few years later;
6. The theory of least-squares collocation for an exact combination of heterogeneous data (1968); and
7. The determination of points on the Earths' surface by GPS since about 1980.

Most work in physical geodesy is concerned with the determination of the geoid, but there is no unique workable — and uniquely accepted — definition of the geoid. We have three main definitions of the geoid:

– The classical definition,

– The quasigeoid,

– The harmonic geoid.

All three definitions have their merits and their drawbacks, which can befound in the textbooks of physical geodesy. The aim of the present article is to describe the basic ideas of this rather difficult problem from a different perspective, focusing on principal features rather than computational methods and details, which can be found in the literature. For details, formulas and references we shall, for simplicity, frequently refer to the references in the book (Hofmann-Wellenhof and Moritz, 2005), denoted briefly by HWM. However, we shall frequently discuss the matters from a "fresh" perspective. Extensive references are given in this book; see also the Journal of Geodesy and the Internet.

*E-mail: helmut.moritz@tugraz.at

VERSITA

We shall disregard small temporal variations such as waves, tides and other small effects, which are the subject of special investigations. The Earth model, used here as usual, is a rigid Earth rotating uniformly about a rigid axis.

A basic progress, achieved since a few decades, has been the possibility to determine the topographic Earth surface S geometrically. This has fundamentally changed the character of physical geodesy, which has been relieved its classical task of determining S. It makes the basic problem of physical geodesy an overdetermined boundary-value problem, which is a definite advantage over Stokes and Molodensky. The present paper takes this feature into account systematically, which may be new in this consistent form.

2. The classical definition of the geoid

The geoid has been first proposed by C.F. Gauss as the "mathematical Earth surface". On the oceans, it is the sea surface; on the continents it is the surface above which the "heights above sea level" or "orthometric" heights are counted. Introducing the gravity potential or geopotential W, the geoid is a surface of constant geopotential $W = W_o = \text{const}$.

Gravity is the resultant of gravitational attraction and centrifugal force of Earth rotation. The latter can be computed by a simple formula and can be taken for granted. The difficulty rests with the gravitational attraction of the "topographic masses" above sea level. The big problem is the principal impossibility of the determination of the density of the topographic masses below the Earth surface. We shall call it the density problem.

(The second problem is practical: the gravity field has to be given at every point of the Earth surface; it is always measured at discrete problems only and interpolated in between. We shall call this the interpolation problem. It is not a theoretical problem since, in principle, we can make the net of gravity stations as dense as required). The geoid's companion is the reference ellipsoid, above which the ellipsoidal height or "GPS height". It is directly measured by GPS and poses no problem in the sense of this paper. It is purely geometric and thus perfectly simple and unique.

Figure 1 provides a simple view of the situation. The geoid height N is the difference between the orthometric height H and the GPS height h. They are counted along the same normal to the ellipsoid, which, for the present purpose, is considered to be normal to the geoid as well. This is a permissible simplification, so that the basic equation

$$h = H + N \text{ or } H = h - N \tag{1}$$

holds: if we measure h by GPS and know the geoid height N, we can determine the orthometric height H. Or, again knowing the geoid height N, we get the ellipsoidal height h by levelling, which is appropriate for purposes of engineering surveying, for instance in a tunnel, where GPS does not work.

(Inertial surveying (INS) does work in tunnels but depends on the gravity field. Among other problems, engineering surveying keeps alive the consideration of the gravity field inside the earth in spite of the density problem mentioned above.)

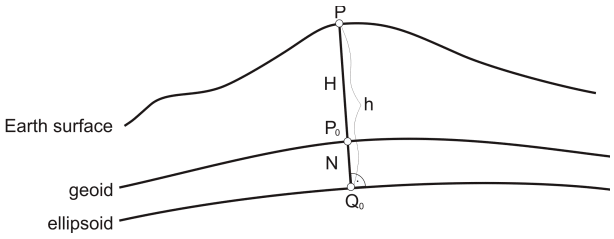

Figure 1. The basic geometry.

If there were no masses above the geoid, then gravity g could be measured at sea level, since in this case the Earth surface coincides with the geoid.

Let us first consider this case. Considering the reference ellipsoid E as an approximation to the geoid linearizes the problem. A suitable reference gravity field of normal gravity γ and normal potential U is defined. The gravity potential, or geopotential, has introduced above, and the ellipsoidal normal potential U is also called spheropotential (spheroid is an obsolete name of ellipsoid). The corresponding linear quantities are the disturbing potential $T = W - U$ and the gravity disturbance $\delta g = g - \gamma$. These two linear quantities are all referred to the same geoidal point P_0. (This is in contrast to the old concept of gravity anomaly Δg, where gravity referred to the geoid but normal gravity to the ellipsoid. This is now obsolete if the heights of the gravity stations have been measured by GPS. Therefore, the gravity anomaly Δg will no longer be used in this paper.)

The geoid height N is related to the gravity disturbance δg by the integral formula by Neumann, Hotine and Koch, briefly called Koch's formula, which is the modern equivalent of Stokes' formula. We have

$$T = \frac{R}{4\pi} \iint_{\sigma} K(\psi)\, \delta g \, d\sigma \tag{2}$$

with

$$K(\psi) = \frac{1}{\sin(\psi/2)} - \ln\left(1 + \frac{1}{\sin(\psi/2)}\right) \tag{3}$$

and get by Bruns' formula

$$N = T/\gamma. \tag{4}$$

All this is quite analogous to the familiar Stokes method; see HWM pp. 299-303 for details.

3. The quasigeoid

The density problem was first clearly recognized by the Russian geodesist and geophysicist M.S. Molodensky in 1945. He solved it

by the following idea, which is certainly one of the greatest ideas of all time in geodesy.

Molodensky proposed to give up the geoid and the internal gravity field altogether and to work at the Earth surface only. Stokes' formula involves an integral extended over the sea surface represented, for this purpose, by an ellipsoid (which is first approximated by a sphere but the extension to the reference ellipsoid poses no problem as Molodensky himself recognized). Molodensky (1945) found an integral equation extended over the Earth solved by a series whose first term is Stokes' formula and the higher terms represent the topography (HWM p. 304, 308).

In the place of Figure 1 we shall now consider Figure 2. Instead of the geoid height N at sea level we now have the height anomaly ζ at the Earth surface:

$$N = Q_0P_0 \tag{5}$$

and

$$\zeta = QP. \tag{6}$$

Let us repeat from Sec. 2: The geoid is

$$W = W_0 = \text{constant}, \tag{7}$$

and the normal potential U is defined such that the reference ellipsoid is a surface of constant spheropotential

$$U = U_0 = \text{const.} = U(Q_0). \tag{8}$$

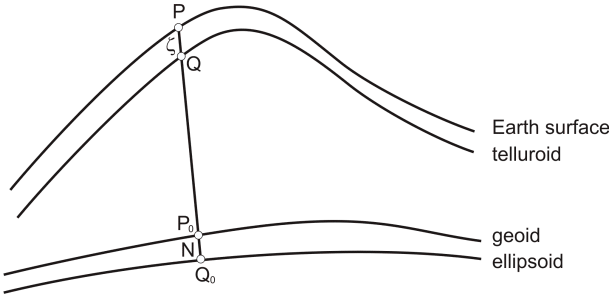

Figure 2. The classical and the modern view.

The constant U_0 is chosen to be equal to W_o so that also

$$U(Q_0) = W(P_0). \tag{9}$$

This is for the geoid at sea level, and is quite simple indeed (Figure 1). Around 1945, Molodensky had the brilliant idea to transfer this relation to the Earth surface, defining

$$U(Q) = W(P). \tag{10}$$

To each point P on the Earth surface, there corresponds a point Q along the vertical. All points Q define a surface, which Hirvonen in 1962 called the telluroid.

Of course, neither the Earth surface nor the telluroid are level surfaces, and the "height anomaly" ζ is a function of the variable point Q. It is the result of Molodensky's solution, but it is just the opposite of N: ζ is "above" at the Earth surface, whereas N is "below" at sea level. However, their values are pretty similar. Also Bruns' Equation (4) holds and gives for the height anomaly

$$\zeta = T/\gamma. \tag{11}$$

Of course, T and γ now refer to the Earth surface S rather than to the geoid; cf. Eqs. (5) and (6).

Now we work on S, which is not an equipotential surface. Thus the height h must be taken into account. So for T, the disturbing potential at the Earth surface, we get Molodensky-Koch' series

$$T = \frac{R}{4\pi} \iint_\sigma K(\psi)\, \delta g \; d\sigma + \kappa_1 + \kappa_2 + \kappa_3 \ldots, \tag{12}$$

where the "Molodensky corrections" κ_i are decreasing functions of height h; cf. HWM Eq.8-87.

The height anomalies ζ are similar in magnitude to the geoid height N, but have quite a different geometric interpretation. Still, surveyors like to have "heights above sea level", above some geoid-like surface. To satisfy them, Molodensky introduced an artificial "quasigeoid" by plotting the height anomaly ζ above the ellipsoid (if it is positive and below if it is negative). The quasigeoid is pretty close to the geoid but has no direct physical interpretation whatsoever. On the hand, it can be computed theoretically rigorously without needing any hypotheses about the density of the topographic masses. Thus we have the "geoid dilemma":

- geoid: direct physical interpretation but theoretically not computable;
- quasigeoid: directly computable but no physical interpretation.

4. The satellite geoid and analytical continuation

The geoid has the basic definition as a surface of constant geopotential (Eq. (7) above). The geopotential W at satellite heights can be developed into a convergent infinite series of spherical (or ellipsoidal) harmonics. (As usual, we disregard the centrifugal force, which is elementary.)

If we wish to use this series for the definition (7) of the geoid, we have to continue the spherical harmonic series down the sea level, more precisely to the reference ellipsoid. However, if we regard the series as infinite, it will in general not converge anymore and, thus, cannot be used to describe the geoid $W = \text{constant}$.

A trivial matter is the centrifugal force: it can easily be added wherever appropriate and we shall thus disregard it. The real difficulty, and it is tremendous, is the following fact. Outside the Earth surface, the geopotential is harmonic: it satisfies Laplace's equation

$$\Delta W = 0 \tag{13}$$

where

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{14}$$

is the Laplace operator (we take the centrifugal potential for granted).

Inside the Earth, however, we have Poisson's equation (setting the gravitational constant = 1)

$$\Delta W = -4\pi\rho, \tag{15}$$

where ρ denotes the density of the "topographic masses" above the reference ellipsoid (ad generally of all masses inside the Earth surface S). Outside S we have $\rho = 0$, inside the Earth we naturally have $\rho > 0$. Thus, by Eqs. (14) and (15), second derivatives have a jump discontinuity across S, which shows that W outside S and W inside S are really two different analytical functions. We can, however, analytically continue W from the outside into the Earth's interior, so that $\Delta W = 0$ holds throughout space. Inside the Earth, of course, we must replace W by W_c, which is the analytical continuation of the geopotential into the Earth' interior.

In HWM p. 313, we have tried to explain analytical continuation by a simple example from daily life. Imagine you drive a road that at first is completely straight. (The straight line is the simplest analytical curve.) At some point the road goes into a circular curve, which also is an analytical curve, but a different one. Thus the straight line is analytical, but the road "straight line + curve" as a whole is not analytical. In the curve, the straight line becomes a circle, and the driver is well advised to turn the steering wheel! As a good mathematician but bad driver, he would continue in the original straight direction and get off the road. He would have mistaken the straight line, which is the analytical continuaton of the road, for the curved road, causing an accident.

Returning to geodesy, $\Delta W = 0$ outside the Earth, but inside the Earth we have $\Delta W_c = 0$ only for the analytical continuation W_c.

Any function satisfying $\Delta f = 0$ is called harmonic, so instead of analytical continuation we also speak of harmonic continuation, so W_c is the harmonic continuation of W into the Earth's interior.

Unfortunately, most external potentials cannot be continued analytically into the Earth's interior. This is directly related with the fact that most spherical-harmonic series of the geopotential are divergent (as far as they are infinite). Thus, the harmonic geoid, that is the surface of constant harmonic downward continuation of the geopotential,

$$W_c = W_0, \tag{16}$$

does not in general exist ! (Please carefully distinguish the quasi-geoid, as determined by analytical continuation to point level, from the harmonic geoid, which corresponds to analytical continuation to sea level, cf. HWM p.304.)

This bitter truth is somewhat sweetened by the fact that empirically from satellites determined spherical (or ellipsoidal) harmonic series must anyway be truncated and then it is continuable. Also in the terrestrial case, any not-continuable potential W can, outside the Earth, be approximated by a continuable potential as closely as we wish. This Runge theorem, called so by Krarup in 1969, is theoretically very important. It is an existence theorem but does not give us a means to compute it for a given accuracy.

The problem is theoretically very difficult. We can, and in fact we must, disregard it but we must know that a Runge theorem exists. The most recent review of this topic is H. Moritz' article "Classical Physical Geodesy" in Freeden et al. (2010 pp. 127-158). (Other references are in HWM as we have already remarked.)

5. The new geodetic boundary value problem

A very general formulation of Molodensky's problem is in terms of a boundary-value problem.

If the Earth surface S and the potential W on S are given, then W in space outside S can be determined by a solution of the Dirichlet boundary problem for harmonic functions (W is harmonic outside S if the centrifugal force is disregarded; see Sec. 4).

The gravity vector $\mathbf{g}$ is the gradient of the geopotential W:

$$\mathbf{g} = gradW = (W_x, W_y, W_z) = \left(\frac{\partial W}{\partial x}, \frac{\partial W}{\partial y}, \frac{\partial W}{\partial z}\right). \tag{17}$$

Thus, on S, the following symbolic relation holds (g is the norm of $\mathbf{g}$):

$$\mathbf{g} = f_1(S, W). \tag{18}$$

Note that (18) is not an ordinary equation, but a functional equation in the sense of nonlinear functional analysis. This is a difficult mathematical subject, but gives a simple symbolic expression. Its difficulty is also seen from the fact that the boundary problem in the modern sense is overdetermined as we have remarked in Sec. 1 and will better understand in Sec. 6.

If we solve symbolically Eq. (18) for W, then we get

$$W = f_2(g, S). \tag{19}$$

What does this mean? It is really a very difficult functional equation. We try to solve it by linearization. Actually, the Earth surface S is known. The linearized version of W is the disturbing potential T, and for the linearization of g we take the gravity disturbance δg; see Sec. 2.

Then Eq. (19) may be written

$$T = \text{Koch}\ (\delta) + \text{Molodensky corrections}\ \kappa_i, \tag{20}$$

which is nothing else than the Molodensky-Koch series (3) !

6. Solution by least-squares collocation

So far, we have assumed, that the measured function δg is known continuously at every point of the Earth surface. This, of course, is only an idealization because gravity can be measured at a set of discrete points only. This is shown in Figure 3, which may picturesquely be called the "geodetic porcupine" (in German, "der geodätische Igel"). It shows the body of the porcupine, which is the reference ellipsoid, and a finite, though very great, number of "spines" or "quills", orthogonal to the ellipsoid, whose ends are the points P_i of measurements on the Earth surface situated at height h_i above the ellipsoid.

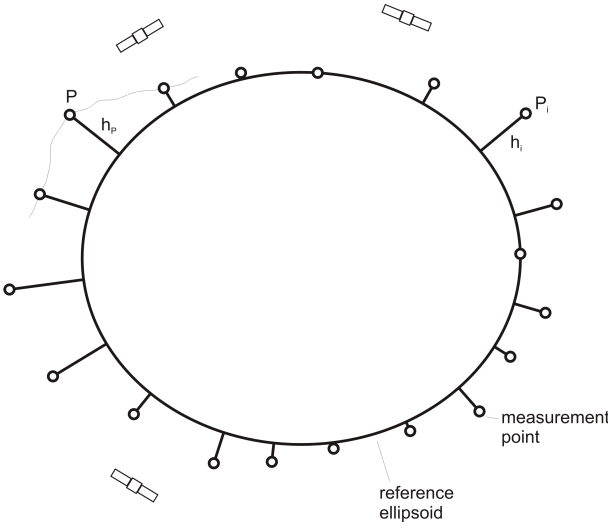

Figure 3. The basic structure of collocation.

The model consists only of the surface of the ellipsoid and its spines; initially there is nothing else, especially no Earth surface. The latter will be computed later.

Since the data are given at a finite set of points only, any exact linearized solution must be a matrix rather than an integral formula. The method is called "least-squares collocation". Since this has become a geodetic household word, we shall not derive it here (see again HWM chapter 10), but immediately give the solution:

$$f(P) = [C_{P1}\ C_{P2}\ \ldots C_{Pq}] \begin{bmatrix} C_{11}\ C_{12}\ldots C_{1q} \\ C_{21}\ C_{22}\ldots C_{2q} \\ \ldots\ldots\ldots\ldots\ldots \\ C_{q1}\ C_{q2}\ldots C_{qq} \end{bmatrix}^{-1} \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ldots \\ \ell_q \end{bmatrix} . \tag{21}$$

Here ℓ_i denotes the measurements at points P_i, which may be gravity disturbances δg or any other "linear functionals", that is, linearized measurements such as deflections of the vertical, or even linearized satellite data such as spherical harmonic coefficients or the data obtained by satellite gradiometry.

The analytical and geometric structure of the (linearized) gravity field enters through the "kernel function" $K(P, Q)$. This is a given spatial function of points P and Q. It is harmonic at any point outside the ellipsoid, that is, it satisfies Laplace's equation both as a function of P and of Q. It is a positive definite function. (Positive-definiteness of functions is defined exactly like with matrices.)

The matrix coefficients C_{Pi} and C_{ik} are computed exactly from the structure of the model as linear functions of the kernel function $K(P, Q)$ independent of the data but embodying their mathematical structure.

The computed function $f(P)$ may be the disturbing potential T at any point P on, outside or inside the Earth surface. Because of the harmonic character of the kernel function $K(P, Q)$, it will represent the disturbing potential outside and on S. Inside S it will represent the harmonic downward continuation of T, which is now a regular analytic function by the definition of the kernel function $K(P, Q)$. Through collocation we automatically get a regular harmonic geoid in the sense of Eq. (16)! In other terms, collocation allows and implies analytical continuation in the sense of Runge (Sec. 5).

The computed function may, however, also be T/γ, γ being normal gravity. Then, by Eqs (5) and (6), the same formula (21) gives the height anomaly ζ if we set the height parameter equal to h, and it gives the harmonic geoid height N_c if we put the height parameter equal to zero. In the first case, we get, point by point, the Earth surface S.

Now we also understand that our problem is overdetermined, as we have remarked in Sections 1 and 5. We can determine differences of the geopotential between two arbitrary points of the Earth surface in two ways:

1. By applying the basic formula (21) to determine W at the two prescribed points and form the difference; or
2. To determine the potential difference ("geopotential number" by classical spirit levelling.

We shall here use these two methods, and the small differences they give, only as a check. We shall not continue here; it will probably not need a Gauss to turn this idea into an adjustment procedure; this may, however, be the subject of a nice research paper.

Another simple application is the case that in Eq. (21) both the function f and the measurements ℓ_i are gravity disturbances δg at the same level. This reduces collocation to least-squares prediction of gravity with which all this began in 1963 (cf. HWM Sec. 9.4) before Krarup in 1969 turned it into full-fledged least-squares collocation for arbitrary geodetic data.

The term "least-squares" indicates some possible relation to statistics. The kernel function $K(P, Q)$ is an analytic function and as such independent of statistics. It has, however, been shown that frequently is advantageous to interpret the kernel function statistically, identifying it with a statistical covariance function of the linearized gravity potential. The "covariances" C in Eq. (21) must

be exactly and carefully derived from the kernel function because they carry the burden of the precise analytical and geometrical structure.

The collocation model may be extended to include random errors and systematic parameters in order to obtain a general synthesis of least-squares collocation-adjustment. The matrix structure is very suitable for the application of computers; depending in the number of observations the matrices to be inverted may be very large.

Least-squares collocation is theoretically able to extract all information from the data in an optimal way. It as also realistic in the sense that it takes into account the fact that the data given are finite number and not as continuous functions on the Earth surface as presupposed in the integral formulas discussed above. (Interpolation between discrete data is automatically included in collocation.)

7. Interpolation and gravity reduction

The implementation the classical geoid definition is theoretically impossible because the density of the topographic masses, the masses above sea level, cannot be empirically determined, so it we have a principal error, the density error (Section 2).

The classical method was to remove a hypothetical model of the topographic masses by gravity reduction. The intention was to computationally remove the topographic masses either completely (Bouguer reduction) or to move it into the interior of the ellipsoid according some model of isostasy (topographic-isostatic reduction). This procedure is theoretically inadequate because of the density error, but it works practically to, say, decimeter accuracy.

The primary effect of classical gravity reduction is that (hopefully) there are no more masses outside the reference ellipsoid, so that the Koch integral and similar formulas for harmonic gravity can be applied.

Topographic-isostatically gravity even has an advantage: it is much smoother than original gravity and therefore can be interpolated much more easily and accurately. This advantage may even to a certain extent compensate the basic density uncertainty.

In Molodensky's theory and in collocation, gravity reduction may (and in high mountains must) be applied also to get smoother data that may be better interpolated. This is the remove-restore process: the topographic model (together with its isostatic compensation) is removed, an integral formula or collocation is performed, and the topographic model is restored again at the same point.

In contrast to classical geoid computation (where the ''remove point'' is different from the ''restore'' point), errors in the topographic-isostatic density model do not matter here if the same density model is used for removal and restoration: removing a wrong density model is no worse than not removing it at all. So, the density error will not play a role here.

The remove-restore model is not limited to gravity reduction or deflections of the vertical: we may also remove (and restore) a global satellite model to get a smaller and smoother field for interpolation.

The present paper only outlines the basic structures, which should be taken into account for actual computations. Also, least-squares collocation is theoretically optimal, but in some practical computations, integral formulas may be preferable in some cases, e.g. with gravity measurements only.

Acknowledgment

The author expresses his thanks to his colleagues in the Institute of Navigation and Satellite Geodesy at TU Graz for constant support and help, especially to Norbert Kuehtreiber for critically reading the paper and to Bernadette Wasle for drawing the pictures.

References

Freeden W., Nashed M.Z., Sonar T. (Eds.), (2010), Handbook of Geomathematics, Springer.

Hofmann-Wellenhof B., Moritz M., (2005), Physical Geodesy. Springer, Vienna and New York.