Journal of Geodetic Science

• 3(2) • 2013 • 87-94 DOI: 10.2478/jogs-2013-0012 •

Various parameterizations of "latitude" equation – Cartesian to geodetic coordinates transformation

Research Article

M. Ligas*

AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering, Department of Geomatics, al. A. Mickiewicza 30, 30-059 Krakow, Poland

Abstract:

The paper presents a solution to one of the basic problems of computational geodesy – conversion between Cartesian and geodetic coordinates on a biaxial ellipsoid. The solution is based on what is known in the literature as "latitude equation". The equation is presented in three different parameterizations commonly used in geodesy – geodetic, parametric (reduced) and geocentric latitudes. Although the resulting equations may be derived in many ways, here, we present a very elegant one based on vectors orthogonality. As the "original latitude equations" are trigonometric ones, their representation has been changed into an irrational form after Fukushima (1999, 2006). Furthermore, in order to avoid division operations we have followed Fukushima's strategy again and rewritten the equations in a fractional form (a pair of iterative formulas). The resulting formulas involving parametric latitude are essentially the same as those introduced by Fukushima (2006) (considered the most efficient today). All the resulting variants are solved with Newton's second-order and Halley's third-order formulas. It turns out that all parameterizations of the "latitude equation" show a comparable level of performance.

Keywords:

biaxial ellipsoid • Cartesian and geodetic coordinates • latitude equation

Received 24-05-2012; accepted 11-04-2013

1. Introduction

There is a lot of space devoted to the problem of conversion between Cartesian and geodetic coordinates in the geodetic literature. This paper is yet another example, except that, it presents a very elegant geometric concept leading to the solution that arose outside the geodetic community (Nurnberg 2006). The paper presents not only the geometric concept which was intended to find the distance from a point to an ellipse/ellipsoid but also its considerable numerical improvements and adjustment to geodetic applications. This enhancement leads to various "latitude equations" depending on the parameterization of the ellipsoid involved. The resulting equations are solved effectively with Newton and Halley's methods after Fukushima (2006). In fact, this work extends Fukushima's algorithms (Fukushima 2006) to geodetic and geocentric parameterizations of the ellipsoid.

The relations tying Cartesian and geodetic coordinates may be expressed as follows (e.g. Heiskanen and Moritz 1967):

$$x = (N + h)\cos\phi\cos\lambda \tag{1a}$$

$$y = (N + h)\cos\phi\sin\lambda \tag{1b}$$

$$z = \left[N\left(1 - e^2\right) + h\right]\sin\phi \tag{1c}$$

where: x, y, z – Cartesian coordinates of a point P ϕ, λ, h – geodetic coordinates of the point P, latitude, longitude and ellipsoidal height, respectively $N = \frac{a}{\sqrt{1-e^2\sin^2\phi}}$ – radius of curvature in the prime vertical e^2 – the first eccentricity squared $e^2 = 1 - \left(\frac{b}{a}\right)^2$ a, b – semi – major and semi – minor axes of the ellipsoid; respectively

[*] E-mail: marcin.ligas@agh.edu.pl

As it may be seen the transformation $(\phi, \lambda, h) \longrightarrow (x, y, z)$ is a straightforward task to do. On the contrary, the transformation from (x, y, z) to (ϕ, λ, h) carries some load of difficulty. The exception to the latter is to find the longitude on the basis of (x, y) (Vermeille 2004) and the geodetic latitude for $h = 0$. These may be expressed as: *longitude (formulas for any h)*

$$\lambda = \arctan \frac{y}{x} \tag{2}$$

$$\begin{cases} \lambda = \frac{\pi}{2} - 2 \arctan \frac{x}{\sqrt{x^2+y^2}+y}, \text{ for } y \geqslant 0 \\ \lambda = -\frac{\pi}{2} + 2 \arctan \frac{x}{\sqrt{x^2+y^2}-y}, \text{ for } y < 0 \end{cases} \tag{3}$$

Geodetic latitude (h = 0)

$$\phi = \arctan \left[\frac{z}{(1-e^2)\sqrt{x^2+y^2}} \right] \tag{4}$$

Equation 4 is used in those methods of converting Cartesian to geodetic coordinates where an intermediate step of finding the projection of a point of interest P onto the surface of the ellipsoid along the normal is involved (e.g. Lin and Wang 1995, Zhang et al. 2005). In general, if $h \neq 0$ then there is no simple transformation from (x, y, z) to (ϕ, λ, h).

The geodetic literature is full of solutions to the problem which are either approximate (different iterative formulas) see e.g: Heiskanen and Moritz (1967), Bowring (1976), Fukushima (1999, 2006), Lin and Wang (1995), Feltens (2008) or exact (solution to a quartic equation) e.g: Borkowski (1987), Hedgley (1976), Vermeille (2002). Despite the plethora of solutions to the problem we would like to broaden this already vast part of computational geodesy with some algorithms concerning the problem.

2. Basic Formulas

An ellipse centered at the origin has the parameterization $\mathbf{r}_0(\xi) = [x(\xi), y(\xi)]$ (at this stage we do not differentiate between various parameterizations: geodetic, geocentric and parametric latitude). Any point P is represented by the vector $\mathbf{r} = [x, y]$. The shortest distance between the point P and the ellipse will be obtained when the scalar product between the vector $\mathbf{r}$-$\mathbf{r}_0(\xi)$ and the tangent vector $\mathbf{r}'_0\,(\xi)$ will be equal to zero (see Fig. 1), this may be expressed as:

$$[\mathbf{r} - \mathbf{r}_0\,(\xi)] \circ \mathbf{r'}_0\,(\xi) = 0 \tag{5}$$

Equation 5 leads to a nonlinear equation with respect to the unknown parameter ξ. After solving this equation with respect to ξ Nurnberg (2006) uses $|\mathbf{r}$-$\mathbf{r}_0(\xi)|$to find the shortest distance between a point P and the ellipse.

In order to adjust this general idea to geodetic applications we use parametric (reduced) latitude ψ, geodetic latitude ϕ, and geocentric latitude u as the parameter ξ in Eq. 5. Due to the symmetry of a

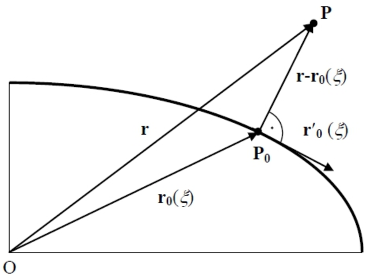

Figure 1. Geometric illustration of the solution.

rotational ellipsoid the problem of conversion is solved on a meridian section $(p = \sqrt{x^2+y^2}, z)$. To find a suitable expression for geodetic height h we use various modifications of Eq. 6 (see e.g. Borkowski 1989):

$$h = (p - a\cos\psi)\cos\phi + (z - b\sin\psi)\sin\phi \tag{6}$$

Since, it is well known that the use of trigonometric functions slows down algorithms considerably; we limited their use by applying the same substitution as Fukushima did (Fukushima 1999, 2006); namely:

$$t_\xi = \tan\xi, \quad \cos\xi = \frac{1}{\sqrt{1+t_\xi^2}}, \quad \sin\xi = \frac{t_\xi}{\sqrt{1+t_\xi^2}} \tag{7}$$

Hence, trigonometric equations resulting directly from Eq. 5 are treated as a basis for deriving more efficient form of equations (irrational form) to be solved, and although they are presented in the text they will not take part in the algorithms' efficiency test. Also, due to reasons of comparison to the most efficient algorithm known nowadays developed by Fukushima (2006) we split the irrational form of the equations (with respect to tangent of latitude) into a pair of iterative formulas with respect to S and C (sine and cosine of latitude). Due to the fact that one of the resulting algorithms (parametric latitude involved) is nothing but Fukushima's one, the purpose of the paper is to check whether it is the most efficient parameterization or perhaps remaining ones are equally efficient and in case of geodetic latitude the most natural.

3. Numerical methods involved

In order to solve the nonlinear equations resulting from various parameterizations of the latitude equation we shall apply the second – order Newton and the third – order Halley's methods (after Fukushima 2006). The iterative process for the two methods may briefly be summarized as (Householder 1970, Kincaid and Cheney 1991):

Newton's method

$$\xi_{n+1} = \xi_n - \frac{f\,(\xi_n)}{f'\,(\xi_n)} \tag{8}$$

Halley's method

$$\xi_{n+1} = \xi_n - \frac{2f(\xi_n) f'(\xi_n)}{2f'^2(\xi_n) - f(\xi_n) f''(\xi_n)} \tag{9}$$

where: n – iteration number, $f(\xi)$, $f'(\xi)$, $f''(\xi)$ – function resulting from a nonlinear equation $f(\xi) = 0$ and its first and second derivatives; respectively, ξ – iterated solution to the equation $f(\xi) = 0$.

4. Summary of the algorithms

Algorithms to transform from Cartesian (x, y, z) to geodetic (ϕ, h) coordinates are presented in the following order (computations of the longitude are not included due to their simplicity). First, three algorithms concerning three different parameterizations derived directly from the basic Eq. 5 are listed. This includes the basic vectors $\mathbf{r}$, $\mathbf{r}_0(\xi)$, $\mathbf{r}'_0(\xi)$, resulting equation $f(\xi) = 0$ and the first $f'(\xi)$ and second $f''(\xi)$ derivatives, starting value for the iterative process ξ_0 and the final expressions for ϕ and h. These algorithms have been named as A1 (parametric latitude), A2 (geodetic latitude) and A3 (geocentric latitude). Algorithms A1, A2, A3 are presented for the sake of consistency of the entire algorithms' derivation flow line and will not be the subject of comparisons in the following section. This is mainly due to an overload with trigonometric functions. Next, more appealing algorithms will be presented, to begin with the abovementioned algorithms rewritten in an irrational form by the substitution defined by Eqs. 7. These are presented twofold, in a direct form derived from the basic algorithms A1, A2, A3 and in a nondimensionalized (unitless) form in order to avoid potential under/over flow when executing algorithms (Hedgley 1976, Fukushima 2006). Unitless variants of the algorithms are named as A1n, A2n, A3n (n – normalized) and will be the subject of mutual comparisons as to the speed and accuracy. A1n, A2n, A3n will be solved both with Newton (N) and Halley's (H) methods. Next in order are the algorithms derived from A1n, A2n, A3n (for N and H) by replacing t_ξ with S_ξ/C_ξ (Fukushima 2006). This leads to two independent iterative formulas solving the conversion problem. These algorithms are called A1nSC, A2nSC, A3nSC (N and H) and will be compared with one another and with A1n, A2n, A3n (N and H).

Algorithm A1 – "latitude" equation in terms of parametric (reduced) latitude $(\xi = \psi)$

$$\mathbf{r} = \begin{bmatrix} p \\ z \end{bmatrix}, \tag{10a}$$

$$\mathbf{r}_0(\psi) = \begin{bmatrix} p_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} a\cos\psi \\ b\sin\psi \end{bmatrix}, \tag{10b}$$

$$\mathbf{r}'_0(\psi) = \begin{bmatrix} p'_0 \\ z'_0 \end{bmatrix} = \begin{bmatrix} -a\sin\psi \\ b\cos\psi \end{bmatrix} \tag{10c}$$

$$f(\psi) = \left(a^2 - b^2\right)\sin\psi\cos\psi - pa\sin\psi + zb\cos\psi = 0 \tag{11}$$

$$f'(\psi) = \left(a^2 - b^2\right)\left(\cos^2\psi - \sin^2\psi\right) - pa\cos\psi - zb\sin\psi \tag{12a}$$

$$f''(\psi) = -2\left(a^2 - b^2\right)\sin 2\psi + pa\sin\psi - zb\cos\psi \tag{12b}$$

$$\psi^0 = \arctan\left(\frac{z}{\sqrt{1-e^2}\,p}\right) \tag{13}$$

$$\begin{cases} \phi = \arctan\left(\frac{\tan\psi}{\sqrt{1-e^2}}\right) \\ h = \frac{p\sqrt{1-e^2}\cos\psi + z\sin\psi - b}{\sqrt{1-e^2\cos^2\psi}} \end{cases} \tag{14}$$

Algorithm A2 – "latitude" equation in terms of geodetic latitude $(\xi = \phi)$)

$$\mathbf{r} = \begin{bmatrix} p \\ z \end{bmatrix}, \tag{15a}$$

$$\mathbf{r}_0(\phi) = \begin{bmatrix} p_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} N\cos\phi \\ N\left(1-e^2\right)\sin\phi \end{bmatrix}, \tag{15b}$$

$$\mathbf{r}'_0(\phi) = \begin{bmatrix} p'_0 \\ z'_0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{1-e^2\sin^2\phi}^3}\,a\left(1-e^2\right)\sin\phi \\ \frac{1}{\sqrt{1-e^2\sin^2\phi}^3}\,a\left(1-e^2\right)\cos\phi \end{bmatrix} \tag{15c}$$

$$f(\phi) = e^2 N\sin\phi\cos\phi - p\sin\phi + z\cos\phi = 0 \tag{16}$$

$$f'(\phi) = e^2 N\left(\frac{\cos^2\phi}{1-e^2\sin^2\phi} - \sin^2\phi\right) - p\cos\phi - z\sin\phi \tag{17a}$$

$$f''(\phi) = N_1 + p\sin\phi - z\cos\phi \tag{17b}$$

where:

$$\begin{aligned} N_1 = {} & e^2 N\sin 2\phi\left[\frac{e^2-1}{\left(1-e^2\sin^2\phi\right)^2} - 1\right] \\ & + \frac{1}{2}e^4 N\sin 2\phi\frac{\cos 2\phi + e^2\sin^4\phi}{\left(1-e^2\sin^2\phi\right)^2} \end{aligned}$$

$$\phi^0 = \arctan\left(\frac{z}{\left(1-e^2\right)p}\right) \tag{18}$$

ϕ is directly obtained through the iteration process

$$h = p\cos\phi - a\sqrt{1-e^2\sin^2\phi} + z\sin\phi \tag{19}$$

Algorithm A3 – "latitude" equation in terms of geocentric latitude $(\xi = u)$

$$\mathbf{r} = \begin{bmatrix} p \\ z \end{bmatrix}, \tag{20a}$$

$$\mathbf{r}_0(u) = \begin{bmatrix} p_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} \rho \cos u \\ \rho \sin u \end{bmatrix}, \tag{20b}$$

$$\mathbf{r}'_0(u) = \begin{bmatrix} p'_0 \\ z'_0 \end{bmatrix} = \begin{bmatrix} -\dfrac{1}{\sqrt{1+e'^2\sin^2 u}^3}\, a\left(1+e'^2\right)\sin u \\ \dfrac{1}{\sqrt{1+e'^2\sin^2 u}^3}\, a\left(1+e'^2\right)\cos u \end{bmatrix} \tag{20c}$$

where:

$$\rho = \frac{a}{\sqrt{1+e'^2\sin^2 u}}, e'^2 = \frac{a^2-b^2}{b^2}$$

$$f(u) = e'^2\rho\sin u\cos u - \left(1+e'^2\right)p\sin u + z\cos u = 0 \tag{21}$$

$$f'(u) = e'^2\rho\left(\frac{\cos^2 u}{1+e'^2\sin^2 u} - \sin^2 u\right) - \left(1+e'^2\right)p\cos u - z\sin u \tag{22a}$$

$$f''(u) = N_2 + \left(1+e'^2\right)p\sin u - z\cos u \tag{22b}$$

where:

$$N_2 = -\frac{1}{2}e'^4\rho\sin 2u\frac{\cos 2u - e'^2\sin^4 u}{\left(1+e'^2\sin^2 u\right)^2} - e'^2\rho\sin 2u\left[1+\frac{1+e'^2}{\left(1+e'^2\sin^2 u\right)^2}\right]$$

$$u^0 = \arctan\left(\frac{z}{p}\right) \tag{23}$$

$$\begin{cases} \phi = \arctan\left(\dfrac{\tan u}{1-e^2}\right) \\ h = \dfrac{p\left(1-e^2\right)\cos u + z\sin u - b\sqrt{1-e^2\sin^2 u}}{\sqrt{1-e^2\cos^2 u\,(2-e^2)}} \end{cases} \tag{24}$$

In order to avoid the use of trigonometric functions the algorithms are rewritten in terms of substitution defined by Eqs. 7 and in this way resulting in irrational equations solved with respect to t_ξ. Their mutual similarity made it possible to present them in a general form:

$$f\left(t_\xi\right) = \frac{E_\xi t_\xi}{\sqrt{F_\xi t_\xi^2+1}} - G_\xi t_\xi + H_\xi = 0 \tag{25}$$

$$f'\left(t_\xi\right) = \frac{E_\xi}{\sqrt{F_\xi t_\xi^2+1}^3} - G_\xi \tag{26}$$

$$f''\left(t_\xi\right) = \frac{-3E_\xi F_\xi t_\xi}{\sqrt{F_\xi t_\xi^2+1}^5} \tag{27}$$

$$t_\xi^0 = \frac{H_\xi}{G_\xi\left(1-e^2\right)} \tag{28}$$

where: *for the parametric latitude:* (dimensionalized):

$$t_\xi = t_\psi, \tag{29a}$$

$$E_\psi = a^2 - b^2, \tag{29b}$$

$$F_\psi = 1, \tag{29c}$$

$$G_\psi = ap, \tag{29d}$$

$$H_\psi = bz \tag{29e}$$

Algorithms A1Nn, A1Hn (non – dimensionalized, divided by a^2):

$$t_\xi = t_\psi, \tag{30a}$$

$$E_\psi = e^2, \tag{30b}$$

$$F_\psi = 1, \tag{30c}$$

$$G_\psi = \frac{p}{a}, \tag{30d}$$

$$H_\psi = \frac{bz}{a^2} \tag{30e}$$

$$\begin{cases} \phi = \arctan\left(\dfrac{t_\psi}{\sqrt{1-e^2}}\right) \\ h = \dfrac{p\sqrt{1-e^2} + zt_\psi - b\sqrt{1+t_\psi^2}}{\sqrt{1-e^2+t_\psi^2}} \end{cases} \tag{31}$$

for the geodetic latitude: (dimensionalized):

$$t_\xi = t_\phi, \tag{32a}$$

$$E_\phi = ae^2, \tag{32b}$$

$$F_\phi = 1 - e^2, \tag{32c}$$

$$G_\phi = p, \tag{32d}$$

$$H_\phi = z \tag{32e}$$

Algorithms A2Nn, A2Hn (non – dimensionalized, divided by a):

$$t_\xi = t_\phi, \tag{33a}$$

$$E_\phi = e^2, \tag{33b}$$

$$F_\phi = 1 - e^2, \tag{33c}$$

$$G_\phi = \frac{p}{a}, \tag{33d}$$

$$H_\phi = \frac{z}{a} \tag{33e}$$

$$\begin{cases} \phi = \arctan\left(t_\phi\right) \\ h = \dfrac{p - a\sqrt{1 + \left(1 - e^2\right) t_\phi^2} + z t_\phi}{\sqrt{1 + t_\phi^2}} \end{cases} \tag{34}$$

for the geocentric latitude: (dimensionalized):

$$t_\xi = t_u, \tag{35a}$$

$$E_u = a e'^2, \tag{35b}$$

$$F_u = 1 + e'^2, \tag{35c}$$

$$G_u = \left(1 + e'^2\right) p, \tag{35d}$$

$$H_u = z \tag{35e}$$

Algorithms A3Nn, A3Hn (non – dimensionalized, divided by a):

$$t_\xi = t_u, \tag{36a}$$

$$E_u = e'^2, \tag{36b}$$

$$F_u = 1 + e'^2, \tag{36c}$$

$$G_u = \frac{\left(1 + e'^2\right) p}{a}, \tag{36d}$$

$$H_u = \frac{z}{a} \tag{36e}$$

$$\begin{cases} \phi = \arctan\left(\dfrac{t_u}{1 - e^2}\right) \\ h = \dfrac{p\left(1 - e^2\right) + z t_u - b\sqrt{1 - e^2 + t_u^2}}{\sqrt{\left(1 - e^2\right)^2 + t_u^2}} \end{cases} \tag{37}$$

Iterative process for the above – listed algorithms will be executed with Newton's method Eq. 8 (**A1Nn**, **A2Nn**, **A3Nn**) and Halley's one Eq. 9 (**A1Hn**, **A2Hn**, **A3Hn**). Following Fukushima's strategy we also reduce the number of division operations by rewriting the above equations (with respect to t_ξ) in a fractional form replacing t_ξ with S_ξ/C_ξ (sine and cosine). This approach results in two independent iterative formulas. This may also be presented in a general form applicable to the three parameterizations with the set of suitable substitutions for $E_\xi, F_\xi, G_\xi, H_\xi$. Thus, an initial guess may be expressed as:

$$t_\xi^0 = \frac{H_\xi}{G_\xi\left(1 - e^2\right)} = \frac{S_\xi^0}{C_\xi^0} \Rightarrow S_\xi^0 = H_\xi \wedge C_\xi^0 = G_\xi\left(1 - e^2\right) \tag{38}$$

General formula for the Newton's method in a fractional form is then given by:

Algorithms A1NnSC, A2NnSC, A3NnSC

$$S_\xi^{n+1} = H_\xi \sqrt{F_\xi\left(S_\xi^n\right)^2 + \left(C_\xi^n\right)^2}^3 + E_\xi F_\xi\left(S_\xi^n\right)^3 = \alpha_\xi^n \tag{39}$$

$$C_\xi^{n+1} = G_\xi \sqrt{F_\xi\left(S_\xi^n\right)^2 + \left(C_\xi^n\right)^2}^3 - E_\xi\left(C_\xi^n\right)^3 = \beta_\xi^n \tag{40}$$

General formula for the Halley's method in a fractional form is given by:

Algorithms A1HnSC, A2HnSC, A3HnSC

$$S_\xi^{n+1} = \alpha_\xi^n \beta_\xi^n - \gamma_\xi^n S_\xi^n \tag{41}$$

$$C_\xi^{n+1} = \left(\beta_\xi^n\right)^2 - \gamma_\xi^n C_\xi^n \tag{42}$$

where:

$$\gamma_\xi^n = 1.5 E_\xi F_\xi S_\xi^n \left(C_\xi^n\right)^2 \left[\left(G_\xi S_\xi^n - H_\xi C_\xi^n\right)\sqrt{F_\xi\left(S_\xi^n\right)^2 + \left(C_\xi^n\right)^2} - E_\xi S_\xi^n C_\xi^n\right]$$

$$= 1.5E_{\xi}^{2}F_{\xi}\left(S_{\xi}^{n}\right)^{2}\left(C_{\xi}^{n}\right)^{2}G_{\xi}\left[\frac{e^{2}\sqrt{F_{\xi}\left(S_{\xi}^{n}\right)^{2}+\left(C_{\xi}^{n}\right)^{2}}-\left(1-e^{2}\right)E_{\xi}}{E_{\xi}}\right] \tag{43}$$

which simplifies to: *for the parametric latitude*:

$$\gamma_{\psi}^{n} = 1.5E_{\psi}^{2}\left(S_{\psi}^{n}\right)^{2}\left(C_{\psi}^{n}\right)^{2}G_{\psi}\left[\sqrt{\left(S_{\psi}^{n}\right)^{2}+\left(C_{\psi}^{n}\right)^{2}}-\left(1-e^{2}\right)\right] \tag{44}$$

for the geodetic latitude:

$$\gamma_{\phi}^{n} = 1.5E_{\phi}^{2}F_{\phi}\left(S_{\phi}^{n}\right)^{2}\left(C_{\phi}^{n}\right)^{2}G_{\phi}\left[\sqrt{F_{\phi}\left(S_{\phi}^{n}\right)^{2}+\left(C_{\phi}^{n}\right)^{2}}-\left(1-e^{2}\right)\right] \tag{45}$$

for the geocentric latitude:

$$\gamma_{u}^{n} = 1.5e^{2}E_{u}F_{u}\left(S_{u}^{n}\right)^{2}\left(C_{u}^{n}\right)^{2}G_{u}\left[\sqrt{F_{u}\left(S_{u}^{n}\right)^{2}+\left(C_{u}^{n}\right)^{2}}-1\right] \tag{46}$$

The final values of (φ, h) for the Newton and Halley's methods in the fractional form are obtained by (where S and C are sufficiently correct values from the iteration process): *for the parametric latitude*:

$$\begin{cases} \phi = \arctan\left(\dfrac{S_{\psi}}{C_{\psi}\sqrt{1-e^{2}}}\right) \\ h = \dfrac{C_{\psi}p\sqrt{1-e^{2}}+zS_{\psi}-b\sqrt{S_{\psi}^{2}+C_{\psi}^{2}}}{\sqrt{C_{\psi}^{2}\left(1-e^{2}\right)+S_{\psi}^{2}}} \end{cases} \tag{47}$$

for the geodetic latitude:

$$\begin{cases} \phi = \arctan\left(\dfrac{S_{\phi}}{C_{\phi}}\right) \\ h = \dfrac{C_{\phi}p+zS_{\phi}-a\sqrt{\left(1-e^{2}\right)S_{\phi}^{2}+C_{\phi}^{2}}}{\sqrt{S_{\phi}^{2}+C_{\phi}^{2}}} \end{cases} \tag{48}$$

for the geocentric latitude:

$$\begin{cases} \phi = \arctan\left(\dfrac{S_{u}}{C_{u}\left(1-e^{2}\right)}\right) \\ h = \dfrac{C_{u}p\left(1-e^{2}\right)+zS_{u}-b\sqrt{\left(1-e^{2}\right)C_{u}^{2}+S_{u}^{2}}}{\sqrt{\left(1-e^{2}\right)^{2}C_{u}^{2}+S_{u}^{2}}} \end{cases} \tag{49}$$

5. Numerical tests

Numerical tests relied on mutual comparisons of twelve presented algorithms (A1Nn, A1Hn, A2Nn, A2Hn, A3Nn, A3Hn, A1NnSC, A1HnSC, A2NnSC, A2HnSC, A3NnSC, A3HnSC) resulting from the three different parameterizations of the "latitude equation", representations (irrational or fractional form) and numerical methods applied (Newton or Halley). Since, Fukushima's methods (Fukushima 2006) have been considered to be the most effective so far and presented algorithms with respect to parametric (reduced) latitude are essentially nothing but Fukushima's algorithms it is justifiable to make comparisons only among different parameterizations without involving any other algorithms known from the literature. This will give the answer whether anyone of the adopted parameterizations behaves better than others.

The adopted reference ellipsoid was GRS80. All algorithms were tested on two height ranges. The first one from -10 to 10 km with the step of 12.5 m (Case A) and the second one from 0 to 36 000 km with the step of 12.5 km (Case B). Latitudes varied from 0° to 90° with the step of 0.05° (3′). The test procedure was divided into two stages: forward step and backward step. In the forward step, for the two height ranges and the mentioned range of latitudes Cartesian coordinates were produced. In the backward step geodetic coordinates were recovered form Cartesian ones. Along with the retransformed coordinates the time of execution and angular and distance accuracy were stored. Both angular and distance accuracy were measured as $\log_{10}$ from the maximum absolute deviate from the theoretical values from the first step of the procedure i.e. $\log_{10}$(abs(max(ϕ_t-ϕ_c)), $\log_{10}$(abs(max(h_t-h_c)) where ϕ_t, h_t denote theoretical values of geodetic coordinates from the first step and ϕ_c, h_c denote recalculated values on the basis of Cartesian coordinates. The algorithms were executed with the limit of only one iteration.

Table 1. Scaled CPU times and accuracy characteristics for the algorithms.

	Case A			Case B		
Algorithm	Scaled CPU Time	Max error in ϕ	Max error in h	Scaled CPU Time	Max error in ϕ	Max error in h
A1Nn	1.09	-11.09	-14.67	1.09	-6.32	-12.07
A2Nn	1.06	-11.09	-14.67	1.06	-6.32	-12.07
A3Nn	1.10	-11.09	-14.68	1.11	-6.32	-12.07
A1Hn	1.13	-14.44	-14.69	1.13	-8.82	-13.85
A2Hn	1.15	-14.44	-14.71	1.15	-8.82	-13.85
A3Hn	1.16	-14.44	-14.68	1.16	-8.82	-13.85
A1NnSC	1.00	-11.09	-14.67	1.00	-6.32	-12.07
A2NnSC	0.96	-11.09	-14.72	0.96	-6.32	-12.07
A3NnSC	1.01	-11.09	-14.61	1.01	-6.32	-12.07
A1HnSC	1.06	-14.44	-14.67	1.06	-8.82	-13.85
A2HnSC	1.08	-14.44	-14.65	1.08	-8.82	-13.85
A3HnSC	1.09	-14.44	-14.61	1.10	-8.82	-13.85

All the methods were coded in Borland Delphi 7 with double precision floating point arithmetic (extended type 10B in size, 19 – 20 significant digits) and run under Windows XP Professional operating system, on HP Pavilion notebook with AMD Athlon 64X2 Dual – Core Processor TK – 55, 1.80 GHz, 960 MB RAM. Constant values used in the algorithms were declared only once at the beginning of a driver program and had no impact on measuring the time of execution of any particular procedure. Results of comparison are presented in Table 1. CPU times are scaled to A1NnSC algorithm (A1 in fractional form solved with Newton's method).

Table 1 reveals that for Case A all algorithms assure more than satisfactory level of accuracy (0.00000003″). For Case B a considerable decrease in accuracy especially for the geodetic latitude is visible and for the algorithms solved with Newton's method maximum error reaches the value of 0.002″. It is also visible that algorithms in the fractional form are superior to algorithms in the irrational form with respect to timing what is the same result as Fukushima's one. Within the same algorithm but with different parameterizations the timing and the accuracy are maintained roughly on the same level what makes all parameterizations equal. Taking into consideration the overall performance of the algorithms the superiority certainly goes towards the ones in the fractional form solved with Halley's third order formula (A1HnSC, A2HnSC, A3HnSC) which are mutations of the recommended variant (f) in Fukushima (2006). Whilst interpreting the timing results one must remember that there are always other processes running on computers in the background and this is the reason that timings of the same numerical procedure may vary from one run to another. Thus, a small differences between the timings should be treated as negligible.

As far as the stability of the presented algorithms is concerned the explanation provided by Fukushima (2006) is fully valid here due to the fact of similarly looking equations that make up the new algorithms. Also the numerical tests did not reveal any singularities for the points tested (millions of points).

6. Conclusions

We have presented new algorithms, in the sense of various parameterizations, to transform Cartesian to geodetic coordinates. Although the use of iterative numerical methods to solve various "latitude" equations it turns out that they can be used in non – iterative way because only one iteration is needed in order to obtain a satisfactory level of accuracy. Algorithms based on the parametric latitude are essentially the same as those introduced by Fukushima (2006) considered the most effective today. Algorithms based on remaining parameterizations (geodetic and geocentric) are equally efficient both in terms of accuracy and timing. The only advantage of the new algorithms (from academic viewpoint rather than from practical) over Fukushima's ones is the use of geodetic latitude which is the most natural choice since after all this latitude is of interest.

Acknowledgements

The paper is the result of research on geospatial methods carried out within statutory research no. 11.11.150.006 at the Department of Geomatics, AGH University of Science and Technology, Krakow.

References

Borkowski K.M., 1987, Transformation of geocentric to geodetic coordinates without approximations, Astrophy. Space Sci., 139, 1–4

Borkowski K.M., 1989, Accurate algorithm to transform geocentric to geodetic coordinates, Bull. Geod., 63, 50–56

Bowring B.R., 1976, Transformation from spatial to geographical coordinates, Surv. Rev., 23, 323–327

Feltens J., 2008, Vector methods to compute azimuth, elevation, ellipsoidal normal, and Cartesian (X,Y,Z) to geodetic (ϕ,λ,h) transformation, J. Geod., 82, 493–504

Fukushima T., 1999, Fast transform from geocentric to geodetic coordinates, J. Geod., 73, 603–610.

Fukushima T., 2006, Transformation from Cartesian to geodetic coordinates accelerated by Halley's method, J. Geod., 79, 689–693

Hedgley D.R., 1976, An exact transformation from geocentric to geodetic coordinates for nonzero altitudes, NASA TR R – 458, Washington

Heiskanen W. A. and Moritz H., 1967, Physical Geodesy, W. H. Freeman and Company, San Francisco

Householder A. S., 1970, The numerical treatment of a single nonlinear equation, McGraw – Hill, New York

Kincaid D. and Cheney W., 1991, Numerical Analysis, Brooks/Cole Publishing Company, Pacific Grove, California

Lin K.C. and Wang J., 1995, Transformation from geocentric to geodetic coordinates using Newton's iteration, Bull. Geod., 69, 300–303.

Nurnberg R., 2006, Distance from a point to an ellipse/ellipsoid, http://www2.imperial.ac.uk/~rn/distance2ellipse.pdf

Vermeille H., 2002, Direct transformation from geocentric coordinates to geodetic coordinates. J. Geod., 76, 451–454

Vermeille H., 2004,Computing geodetic coordinates from geocentric coordinates. J. Geod., 78, 94–95

Zhang C. D., Hsu H. T., Wu X. P., Li S. S., Wang Q. B., Chai A. Z. and Du L., 2005, An alternative algebraic algorithm to transform Cartesian to geodetic coordinates, J. Geod., 79, 413–420

Journal of Geodetic Science

• 3(2) • 2013 • 95-102 DOI: 10.2478/jogs-2013-0015 •

The effect of tunnelling on repeated precise levelling measurements for vertical deformation control of the Metro4 project

Research Article

Cs. Égető[1], L. Földváry[1]*, T. Huszák[2]

1 *Department of Geodesy and Surveying, Budapest University of Technology and Economics, Műegyetem rkp 3, H-1111 Budapest, Hungary, phone: (36)1463-3092, fax: (36)1463-3192*

2 *Department of Geotechnics, Budapest University of Technology and Economics*

Abstract:

The effect of the construction of the 4th subway line of Budapest (Metro4) on the potential surfaces of the gravity field has been simulated, using the prism modelling technique. In the study mass loss due to the excavation of the two tunnels and of the stations has been considered. Mass variations deform the level surfaces; as so, the vertical reference surface of the levelling measurements is changing by time. In this study, the effect of the mass loss on levelling measurements was determined at a level 1 m above the ground, roughly simulating common instrument heights. Subsequently, the effect of the actual deformations of the surface on the levelling measurements has also been determined. According to the results, under certain arrangements of the levelling line with respect to the position of the excavations, the error due to the change of the vertical reference is in the 1 µm order of magnitude, thus negligible.

Keywords:

Budapest Metro4 project • levelling • prism modelling • temporal variations of gravity

Received 11-04-2013; accepted 29-05-2013

1. Introduction

In case of a huge industrial project, certain geodetic aspects should be considered for surveying engineering applications. The excavation of various layers of soil during the construction of a subway line affects the local gravity field. As the gravity field serves as the local reference frame of the simultaneously performed vertical control measurements, it affects the results of vertical measurements. In the study, gravity field variations due to the excavation of the new subway line of Budapest, Metro4 is analysed from the aspect of the accompanying surveying tasks, particularly the vertical control measurements.

2. Theoretical background

In case of excavating tunnels, deformations on the surface are expected to be accompanied. Furthermore, the removal of a notable amount of bed rock and soil changes the local mass distribution, hence it changes the structure of the gravity field too. Therefore, the observed height variation of repeated levelling measurements, δH is influenced not only by the actual deformation, but also an apparent deformation due to the change of the gravity field by time. The observed height variation is thus formulated as (Biró 1983):

$$\delta h = \delta H + \delta N, \qquad (1)$$

where δh is the actual surface deformation, δH is the change in orthometric height, and $\delta N$is the change of the geoid undulation due to the mass variation. (Temporal variation between two

*E-mail: fl@sci.fgt.bme.hu

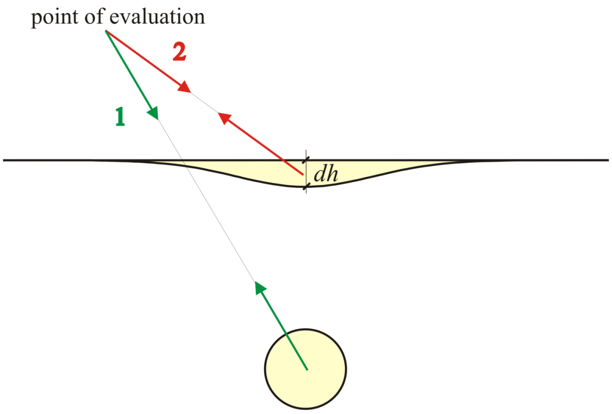

Figure 1. The deformation of the equipotential surfaces occurs due to (1) the tunnelling and (2) vertical deformations. The arrows visualize those gravitational effects, which are cancelled due to the mass removal.

epochs, t_0 and t_1 is henceforth noted by the symbol δ throughout this study.) The change of the geoid undulation does affect the height measurements, but means no actual deformation, so it has no effect on the neighbouring structures, buildings. Still, its appearance affects the geometrical levelling.

The change of the geoid undulation by time, δN in Eq. (1) can be derived as (Strang van Hees 1977, Heck 1982, Sjöberg 1982, Biró 1983, Sjöberg 1987, Biró et al. 1986):

$$\delta N = \frac{R}{4\pi\gamma}\iint \delta g\, S(\psi) d\sigma + \frac{R}{4\pi\gamma}\iint \frac{2g}{R}\delta H\, S(\psi) d\sigma. \tag{2}$$

In Eq. (2) R is a the mean Earth radius, γ is the normal gravity, δg is the change of gravity anomaly in time, $S(\psi)$ is the Stokes function with ψ being the spherical distance between the source point and the point of evaluation, $d\sigma = \cos\varphi d\varphi d\lambda$ is an infinitely small surface element on the unit sphere, φ and λ refers to the latitude and the longitude. The derivation of Eq. (2) is presented in Appendix A based on Biro et al. (1986) and Biro (1983).

The terms on the right hand side of Eq. (2) consist of two independent sources of mass variations: 1) deformation of the equipotential surface due to the mass loss corresponding to the tunnelling, 2) deformation of the equipotential surface due to the actual vertical displacement of the physical surface (c.f. Fig. 1).

Considering a mass anomaly, δm occurring between times t_0 and t_1, the major component of δN, i.e. the first term of the right hand side of Eq. (2), $\delta N_{main} = \frac{R}{4\pi\gamma}\iint \delta g\, S(\psi) d\sigma$, can directly be obtained by applying the Bruns' equation on the Newtonian potential:

$$\delta N_{main} = \frac{k}{\gamma}\int \frac{\delta m}{r}. \tag{3}$$

In Eq. (3) r refers to the distance between mass element and the point of evaluation, and k is the gravitational constant.

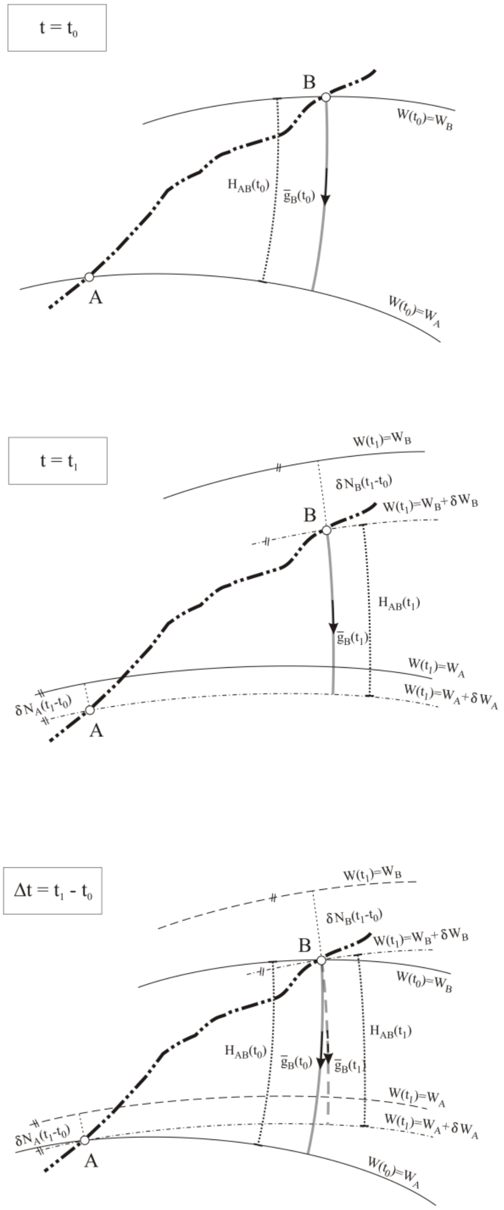

Figure 2. The ground, potential surfaces, plumb lines and height differences before the construction (top), after the construction (middle) and comparison of the two (bottom).

In the followings a single height difference between stations A and B with repeated levelling measurements performed at times t_0 and t_1 is considered (c.f. Fig. 2). The temporal variation of the geoid undulation between the two stations can be defined as the difference of the temporal changes in each point:

$$\delta N_{AB} = \delta N_B - \delta N_A. \tag{4}$$

The measured vertical displacement is interpreted as the temporal change of the observed height difference:

$$\delta H_{AB} = H_{AB}\left(t_1\right) - H_{AB}\left(t_0\right). \tag{5}$$

Then the relationship between change of geoid undulation and observed vertical displacement can be derived from the equations above as follows (c.f. Biró 1983, Eq. (212.5) on page 31):

$$\delta H_{AB} = -\delta N_{AB} - \frac{\tilde{g}_B\left(t_1\right) - \tilde{g}_B\left(t_0\right)}{\tilde{g}_B\left(t_1\right)} H_{AB}\left(t_0\right). \tag{6}$$

In this equation $\tilde{g}_B$ refers to the mean gravity between levels of A and B along the plumb line through point B. This is approximately equal to the gravity at the half of the H_{AB} height difference. Derivation of Eq. (6) can be found in Appendix B.

In the present study the effect of the excavation on the level surfaces was estimated based on Eqs. (3)-(6) for the newly constructed subway line of Budapest, the Metro 4 line. The integral of Eq. (3) was approximated by a summation over mass elements. A preceding study has already been presented by Égető and Földváry (2011), where the effects of vertical displacements due to the tunnelling were neglected. In the present study, actual vertical displacements based on the vertical monitoring measurements were included to the analysis.

3. Data

The calculations have been performed by actual data of the Metro 4 construction. The used data are the following:

1. longitudinal sections and cross sections of the tunnels (provided by the DBR Metro Project Directorate)

2. soil density along the tunnels (provided by the Geovil Ltd)

3. gravity anomaly data in the vicinity of the tunnels (provided by the Hungarian Geological and Geophysical Institute)

4. vertical displacements in the control points (provided by the Consortium of the Soldata Ltd and Hungeod Ltd)

The top view of the tunnels and the location of the surrounding gravimetric data are displayed on Fig. 3 with a sketch of the streets of Budapest in the background. The used coordinates are local stereographic coordinates in Budapest Municipal Projection system.

Figure 4 provides a closer view of the south-west part of the subway line. At this scale the two tunnels already appears separately. The blue lines show the levelling lines, where vertical control data is available.

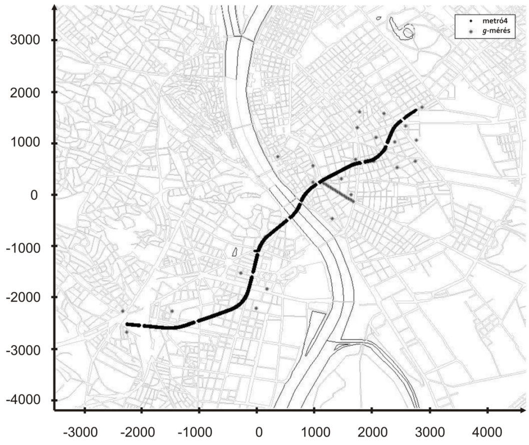

Figure 3. The tunnel (black line interrupted at the stations) and the gravimetric points (grey stars) with main structures of Budapest in the background. The axes are in meter, referring to coordinates in the Budapest Municipal Projection system.

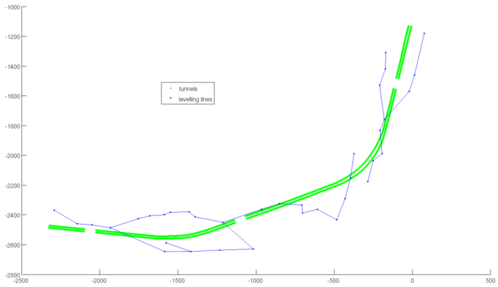

Figure 4. The location of the vertical control lines with respect to the tunnels (note that the interruptions of the tunnels referring to the presence of a stations). The axes are in meter, referring to coordinates in the Budapest Municipal Projection system.

4. Calculations for Metro 4

First a 3D geometry model of the tunnels and of the stations was determined based on the longitudinal sections. The corresponding mass model of the excavated soil has been derived by using the soil density data. The resolution of the mass model was 30 cm, i.e. the tunnel was decomposed into cubes with 30 cm long edges. Using this mass model, its effect on the equipotential surfaces can be determined by Eq. (3). For that, points of evaluation are defined. The points of evaluation were defined to be 1 m above the benchmarks of the vertical control levelling lines (c.f. Fig. 4). The vertical control network in the test area incorporates 80 levelling lines, containing 223 benchmarks. As the study intends to analyse the effect of the deformation of the equipotential surfaces on re-

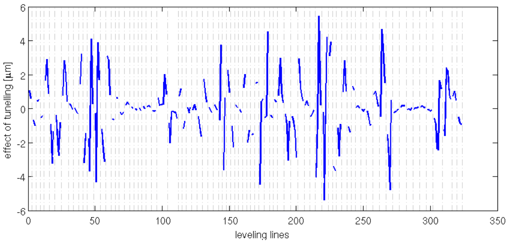

Figure 5. Effect of the level surface deformations on the levelling measurements due to the tunnelling. The horizontal axis counts the number of benchmarks, grouped according to the 80 levelling lines, separated by grey vertical dashed lines.

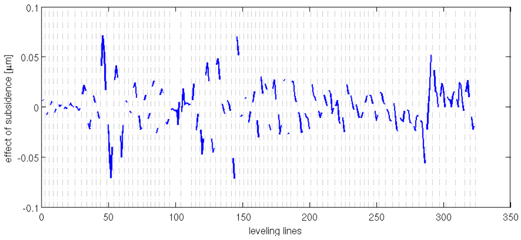

Figure 6. Effect of the level surface deformations on the levelling measurements due to the subsidence. The horizontal axis counts the number of benchmarks, grouped according to the 80 levelling lines, separated by grey vertical dashed lines.

peated levelling measurements, the value of the 1 m was added to the heights of the benchmarks in order to roughly simulate a typical instrument height.

Then the strict formulas of rectangular prisms (c.f., Holstein 2003, Nagy et al. 2000) were simplified to simple point masses, and Eq. (3) was used. The accuracy loss due to this simplification was verified by Égető and Földváry (2013), and it was found to cause negligible difference. Along the subway line there are altogether 10 subway-stations (the term "subway-station" is used instead of "station" to avoid confusion with gravimetric stations or surveying markings). In Fig. 3 at the subway-stations there is a gap, since it was modelled separately from the tunnel. In case of the subway-stations, the exact rectangular prism formula was used. The mass removals due to excavation along the tunnels and in the subway-stations were summed, resulting in an estimate of the effect of the excavation only. Thus no vertical displacements were assumed to occur at this stage (i.e. the right hand side of Eq. (2) is simplified to its first term).

Subsequently, the gravity value was interpolated from the neighbouring gravity data to each point of evaluation, and for every single height differences of the levelling line the effect of the tunnelling on the measurements was determined by Eq. (6). The obtained effect is presented in Fig. 5. As the effect was calculated along many levelling lines, they are separated by grey vertical dashed lines in the figure.

In the next step, the effect of the actual vertical movements on the gravity field has been determined, i.e. the second term on the right hand side of Eq. (2). A rectangular prism model was developed from the observed subsidence. The horizontal resolution of the model is 10 m, and subsidence values were interpolated for each grid cells. The grid was covered by Delaunay triangulation, and interpolation was performed assuming linear variation of the subsidence along every edge of the triangles.

Then, similarly to Fig. 5, for every single height difference of the levelling line the effect of the subsidence on the measurements was determined by Eq. (6). The change of the level surfaces due to the subsidence is shown in Fig. 6.

The effect of the deformation of the levelling surfaces due to the tunnelling on the levelling (c.f. Fig. 5) is in the μm order. The observed subsidence has an even smaller effect on the level surface 1 m above, in the order of 0.01 μm (c.f. Fig. 6).

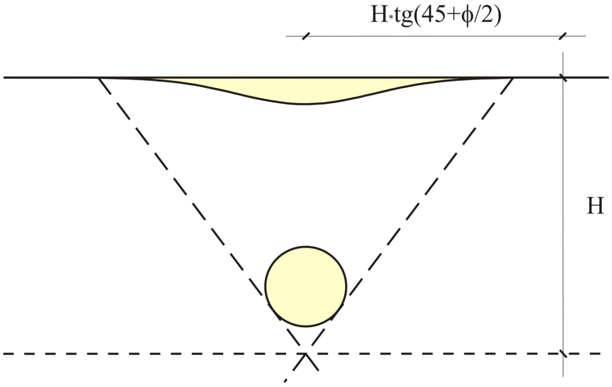

Figure 7. The relation between the depth of tunnelling and the extension of the subsided area.

In the present case, the observed vertical deformations are small, showing some millimetres subsidence and uplift only. Among the observed deformations at the 223 benchmarks, in 166 cases the deformation is less than ±5 mm. Only in 3 cases the deformation was found to be more than 20 mm (−50 mm, −36 mm and −27 mm), the rest of the observed vertical deformations is within ±16 mm, and shows a normal distribution. As this amount of vertical motion is quite small, and was found not to be relevant subsidence, further investigations are performed for reliable simulated cases in the next section.

5. Simulation of the effect of the subsidence on equipotential surfaces

From Fig. 6 it is obvious that the effect of the subsidence on level surfaces is negligible in the case of Metro4. In order to generalize the conclusion on negligibility of this effect, different reliable scenarios are tested.

The subsidence due to the tunnelling takes place over a large area, depending on the depth of the tunnelling and of the internal friction angle of the soil (c.f. Fig. 7).

The internal friction angle, φ varies between the theoretical 0° and about 35°, ranging values between 0° and 10° for clay and silt, between 15° and 25° for sand and between 20° and 35° for gravel (Look 2007). Considering such internal friction angles, the tunnelling at a depth of H causes vertical surface deformation over a radius between H and $2H$. According to that, the modelled subsiding area was defined by a maximal radius of 100 m corresponding to more than 50 m in depth in gravel.

Theoretically, the shape of the subsidence can be approximated by a bell curve (Széchy 1966). However, in real cases the subsidence is governed by local effects, so the *de facto* subsidence rarely shapes a bell. In the lack of any generalizable realistic models, in the present study the bell curve shape is used.

The depth of the subsidence depends on several parameters, including the size of tunnelling, the soil, the building technology, etc. It is so variable that instead of analytical choice of depth values, reliable orders of magnitude of subsidence are taken. The analysed subsidence values are 1 mm, 10 mm and 100 mm.

The model has considered two cases: subsidence due to a point mass change and subsidence due to mass change along a line. These arrangements are visualized in Fig. 8. (Note, that this figure is for visualization purposes only. As different subsidence values are tested in the study, no scale for the vertical axis is presented. The horizontal axes are also not scaled, since that would be misleading: the actual resolution of the model is much finer, than the presented resolution, which is just defined for the visualization.) In the first case the effect was modelled by a circular subsidence with maximal value at its centre exactly above the mass anomaly. Different estimates of the effect were determined by varying the maximal subsidence and the extension radius of the subsided area. With these two parameters a bell curve was defined that at the radius of the extension it reaches the 3σ probability. The derived subsidence model was then approximated by a fine set of orthogonal columns, and the effect of each prism was calculated by the exact rectangular prism formula (Holstein 2003).

In the second case, the linear excavation was considered to occur along one of the horizontal axes. The bell curve is the characteristic of the profile of the subsidence, perpendicularly to the line of tunnelling. The profile was defined similarly to the point mass case, by free parameters of the maximal subsidence and of the extension radius (as in this case the subsidence is not circular, but a stripe, 'extension half-distance' would be a more exact term for this case). By having the subsidence profile in 2D, it is extended along the direction of tunnelling by setting the size of the prisms to 'large' values along the subsidence axis. In order to be rigorous, the value of convergence with increasing length was determined by Richardson-extrapolation.

In both cases, the calculation is evaluated 1 m above the surface, similarly to the former assumption, i.e. the levelling instrument height is modelled with this altitude. Since the maximal displacement of the level surfaces occurs above the mass anomaly, the calculation is presented in this singular point.

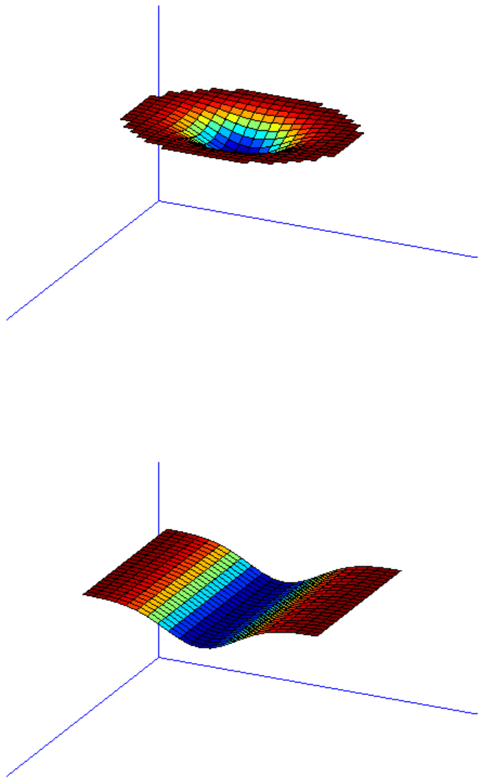

Figure 8. Prism models of circular (top) and linear (bottom) subsidence.

The results are shown in Fig. 9 for the point mass case and Fig. 10 for the linear tunnelling case. The maximal effect can be detected in the extreme case of having an infinitely long linear subsidence with maximal value of 10 cm, vanishing gradually along a bell curve at a distance of 100 m. Even in this case, the level surfaces have maximal displacement in the μm range.

All in all, the effect is found to be negligible for most cases. According to Eq. (6), this effect is differentiated, so it is even smaller in case of a height difference.

Note, that the reality can be extremely different from the simulated cases. Much larger subsidence can be assumed, due to some additional, local circumstances. When obvious holes appear in the test region, then the effect can be relevant. However, in such cases the emphasis is not on the precise engineering surveying tasks, but on surveying tasks of prevention of landslides or building damages.

6. Discussion

Classical surveying techniques are tied to the horizontal and vertical directions. As so, the local features of the gravity field implic-

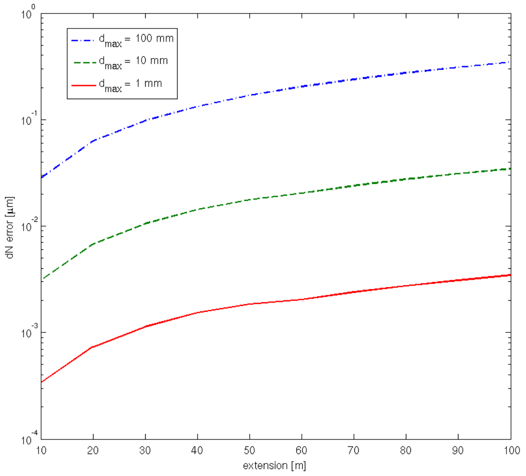

Figure 9. The effect of circular subsidence on the gravity field as function of the radius of the subsiding area.

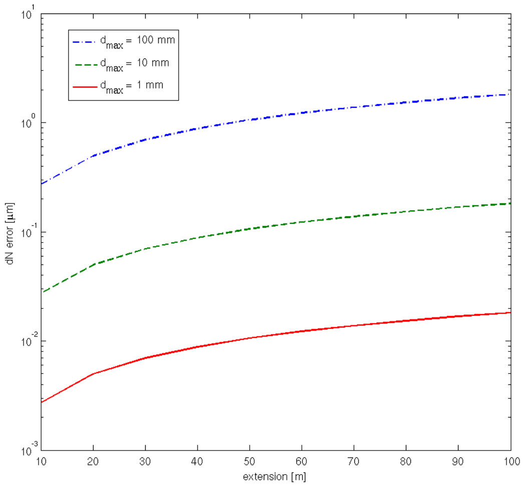

Figure 10. The effect of linear subsidence on the gravity field as function of the length of the subsiding area.

itly have an essential impact on the surveying tasks. However, the gravity field is changing by time. As gravity is partially composed of the gravitation of the surrounding masses, mass variations in the vicinity generate changes in the gravity field, causing changes in the local horizontal and vertical direction. Several constructions involve notable motion of soil mass. Large earthworks are often requires horizontal and vertical monitoring to prevent damages. And just such cases can be influenced by the change of the reference frame, i.e. the local horizontal and vertical directions.

In the study, the effect of the excavation of a subway line was investigated. Based on exact information on the new subway line of Budapest, the effect of the gravity field variations on the levelling measurements was found to be in the range of some µm, which is generally negligible.

As the actual vertical deformation in the case of Metro4 was found to be very small, theoretical scenarios of vertical deformation were also analysed and found to be negligible in all cases.

As long the earthwork does not exceed the excavation of the presented subway line, the variations of the gravity field have only theoretical effect on the accompanying surveying measurements.

Acknowledgement

The used data have been provided by the DBR Metro Project Directorate, the Geovil Ltd, the Hungarian Geological and Geophysical Institute and the Consortium of the Soldata Ltd and Hungeod Ltd, which is highly appreciated.

References

Biró P., 1983, Time variation of height and gravity, Wichmann Verlag, Karlsruhe, Akadémiai Kiadó, Budapest.

Biró P., Thong N. C. and Weisz E., 1986, Modelling of secular variations in gravity and in geoidal undulations, Periodica Polytechnica Separatum, Civil Engineering, 30/1-2, 23-36.

Égető Cs. and Földváry L., 2011, The effect of the Metro 4 tunnel system on the gravity field (in Hungarian), Geomatika, XIV, 1, 17-26.

Égető Cs. and Földváry L., 2013, Numerical accuracy analysis of modeling excavation induced gravity field variations, Proceedings of GV-CONF 2013 (in press).

Heck B., 1982, Combination of leveling and gravity data for detecting real crustal movements, Deutsche Geodätische Kommission, Reihe B, Heft Nr. 258/VII., 20-30.

Holstein H., 2003, Gravimagnetic anomaly formulas for polyhedra of spatially linear media, Geophysics, 68, 157–167.

Look B. G., 2007, Handbook of Geotechnical Investigation and Design Tables, Taylor & Francis Group, London.

Nagy D., Papp G. and Benedek J., 2000, The gravitational potential and its derivatives for the prism, J. Geod., 74, 7-8, 552-560.

Sjöberg L.E., 1982, Studies on the land uplift and its implications on the geoid in Fennoscandia. Depth. Geod. Rep. No. 14, Uppsala University, Uppsala, Sweden

Sjöberg L.E., 1987, Combination of temporal changes of gravity, height and potential coefficients for the determination of secular changes of the geoid, ZfV 112: 167-172

Strang van Hees G.L., 1977, Zur zeitlichen Änderung von Schwere und Höhe, ZfV 102, 444-450.

Széchy K., 1966, The art of tunneling, Akadémiai Kiadó, Budapest.

APPENDIX A

Derivation the formula for the true vertical displacement

Since most of the literature, which has developed the formula for the true vertical displacement is not widely available or not written in English (e.g. Strang van Hees 1977, Heck 1982, Sjöberg 1982, Biró 1983, Biró et al. 1986, Sjöberg 1987), and since the corresponding formulae, namely Eq. (2) and Eq. (6) are not trivial, it is worth to deduce them. The formulation is introduced according to Biró (1983) and Biró et al. (1986). Note that this way of derivation the formulae is not the only option; other authors derive the same result in more or less different ways (Strang van Hees 1977, Heck 1982). Figure 11 shows an arbitrary point P in its original position at the Earth's surface before the displacement at time $t = t_0$. The potential surface, $W = W_P$ passes through the point. There is a displacement assumed to occur between t_0 and a subsequent time t_1.

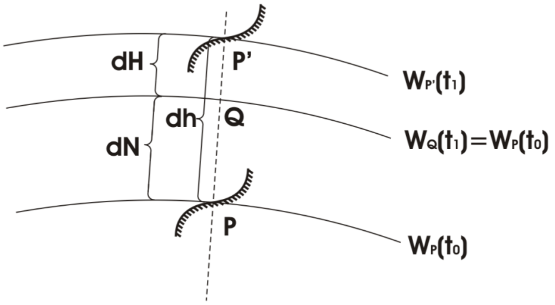

Figure 11. The effect of the time variation of the gravity field at the Earth's surface.

The displacement, δh physically shifts by the point P to P'. The displacement of the ground means mass redistribution, which causes changes in the gravity field as well. The change of the gravity field during this time span is noted by $\delta W = W(t_1) - W(t_0)$, which is in the original point P results in a new potential value, $W_P(t_1) = W_P(t_0) + \delta W$. There is a point Q along the plumbline passing through P, where the value of the potential after the displacement coincidently becomes equal to the potential value in P before the displacement, i.e. $W_Q(t_1) = W_P(t_0)$. The distance between P and Q shows the shift of the level surface by time, i.e. δN. The shift of the level surface from P to Q means that the vertical reference has been changed, and subsequent levelling measurements can detect δH instead of the actual δh. Figure 11 also gives a visual explanation for Eq. (1).

The potential value at t_1 in P is

$$W_P(t_1) = W_P(t_0) + \delta W. \qquad \text{(A1)}$$

It can also be defined through a linear approximation based on the potential in point Q:

$$W_P(t_1) = W_Q(t_1) - \left(\frac{\partial W}{\partial h}\right)_Q \delta N. \qquad \text{(A2)}$$

Relating Eq. (A1) and Eq. (A2) to each other, making use of the equality of $W_Q(t_1) = W_P(t_0)$, and defining the vertical gradient of the potential by the negative gravity vector, $\left(\frac{\partial W}{\partial h}\right)_Q = -g_Q$, the following relationship between δW and δN can be found:

$$\delta N = \frac{\delta W}{g} \qquad \text{(A3)}$$

Since the investigation deals with generally slight variations, $g_Q \doteq g_P$ is a good approximation both in t_0 and t_1.

The change of gravity in Q is

$$\delta g_Q = g_Q(t_1) - g_Q(t_0) = -\left(\frac{\partial W_Q(t_1)}{\partial h} - \frac{\partial W_Q(t_0)}{\partial h}\right)$$
$$= -\left(\frac{\partial}{\partial h}\delta W\right)_Q. \qquad \text{(A4)}$$

Assuming the gravity field to be homogeneous between Q and P (before the displacement) and between Q and P' (after the displacement), linear approximation of $g_Q(t_1)$ and $g_Q(t_0)$ can be obtained by expansion into series:

$$g_Q(t_1) = g_{P'}(t_1) - \left(\frac{\partial g}{\partial h}\right)_{P'} \delta H. \qquad \text{(A5)}$$

$$g_Q(t_0) = g_P(t_0) + \left(\frac{\partial g}{\partial h}\right)_P \delta N. \qquad \text{(A6)}$$

Supposing that $\left(\frac{\partial g}{\partial h}\right)_P \doteq \left(\frac{\partial g}{\partial h}\right)_{P'}$, and inserting Eqs. (A5) and (A6) to Eq. (A4) the following formula can be derived:

$$\left(\frac{\partial}{\partial h}\delta W\right)_Q - \left(\frac{\partial g}{\partial h}\right)_P \delta N$$
$$= -\left[g_{P'}(t_1) - g_P(t_0) - \left(\frac{\partial g}{\partial h}\right)_P \delta H\right]. \qquad \text{(A7)}$$

Considering the left hand side of Eq. (A3) and the right hand side of Eq. (A7) the following differential equation is gained:

$$\left(\frac{\partial}{\partial h}\delta W\right)_Q - \left(\frac{1}{g}\frac{\partial g}{\partial h}\right)_P \delta W_Q = -\left[g_{P'}\left(t_1\right) - g_P\left(t_0\right) - \left(\frac{\partial g}{\partial h}\right)_P \delta H\right]. \tag{A8}$$

In Eq. (A8) quantities at the right hand side can be observed, since $\delta g = g_{P'}\left(t_1\right) - g_P\left(t_0\right)$ is the time variation of the gravity in P, and δH is the levelled change in height. The vertical gravity gradient, $\frac{\partial g}{\partial h}$ can be approximated by neglecting the Earth's flattening, for the gravity field of a sphere with radius R and spherically symmetric mass distribution is the following form:

$$\left(\frac{\partial g}{\partial h}\right)_P \doteq -\frac{2g}{R}. \tag{A9}$$

Inserting Eq. (A9) and using $\delta g = g_{P'}\left(t_1\right) - g_P\left(t_0\right)$ for the change in gravity, Eq. (A8) becomes:

$$\left(\frac{\partial}{\partial h}\delta W\right)_Q + \frac{2}{R}\delta W_Q = -\left[\delta g + \frac{2g}{R}\delta H\right]. \tag{A10}$$

As the right hand side is considered to be known from measurements, this linear differential equation is suitable for to serve as a boundary condition written for the displaced level surface, for the determination of δW_Q. As a zero order solution of the third boundary value problem, Stokes' integral is given for the surface of a simplified Earth (sphere with radius R):

$$\delta W_Q = \frac{R}{4\pi}\iint\left(\delta g + \frac{2g}{R}\delta H\right) S(\psi) d\sigma. \tag{A11}$$

By applying the Bruns formula on Eq. (A11), and separating by the two terms of the integrand, Eq. (2) is obviously derived.

APPENDIX B

Derivation the formula for observed height differences due to true vertical displacement

The derivation of Eq. (6) is presented here, which is performed according to Biró (1983). Let us consider levelling observations before and after deformation between stations A and B according to Fig. 2. Temporal variation of geoid undulation difference at points A and B is δN_{AB}, and the observed vertical displacement is δH_{AB}, which are defined by Eqs. (4) and (5), respectively. The height difference, H_{AB} at these two epochs can be written as

$$H_{AB}\left(t_0\right) = -\frac{1}{\tilde{g}_B\left(t_0\right)}\left(W_B\left(t_0\right) - W_A\left(t_0\right)\right) \tag{A12}$$

$$\begin{aligned} H_{AB}\left(t_1\right) = & -\frac{1}{\tilde{g}_B\left(t_1\right)}\left(W_B\left(t_1\right) - W_A\left(t_1\right)\right) = \\ & -\frac{1}{\tilde{g}_B\left(t_1\right)}\left(\left(W_B\left(t_0\right) + \delta W_B\right) - \left(W_A\left(t_0\right) + \delta W_A\right)\right) = \\ & -\left(\frac{W_B\left(t_0\right) - W_A\left(t_0\right)}{\tilde{g}_B\left(t_1\right)} + \frac{\delta W_B - \delta W_A}{\tilde{g}_B\left(t_1\right)}\right). \end{aligned} \tag{A13}$$

In Eqs. (A12) and (A13) $W_i\left(t_j\right)$ refers to the potential at point i at the time t_j, the symbol δ refers to temporal variation of a quantity from t_0 to t_1, and $\tilde{g}_B$ is the mean gravity along the plumb line through point B between level surfaces of A and B.

The change in the vertical displacement using Eq. (A12) and Eq. (A13) can be explained as:

$$\begin{aligned} \delta H_{AB} = & H_{AB}\left(t_1\right) - H_{AB}\left(t_0\right) = -\frac{1}{\tilde{g}_B\left(t_1\right)}\left(\delta W_B - \delta W_A\right) \\ & - \frac{W_B\left(t_0\right) - W_A\left(t_0\right)}{\tilde{g}_B\left(t_1\right)} + \frac{W_B\left(t_0\right) - W_A\left(t_0\right)}{\tilde{g}_B\left(t_0\right)}. \end{aligned} \tag{A14}$$

The last two terms on the right hand side of Eq. (A14) may be simplified considering

$$\begin{aligned} & -\frac{W_B\left(t_0\right) - W_A\left(t_0\right)}{\tilde{g}_B\left(t_1\right)} + \frac{W_B\left(t_0\right) - W_A\left(t_0\right)}{\tilde{g}_B\left(t_0\right)} = \\ & \frac{\left(W_B\left(t_0\right) - W_A\left(t_0\right)\right)\left(\tilde{g}_B\left(t_1\right) - \tilde{g}_B\left(t_0\right)\right)}{\tilde{g}_B\left(t_1\right)\tilde{g}_B\left(t_0\right)} = \\ & -H_{AB}\left(t_0\right)\frac{\tilde{g}_B\left(t_1\right) - \tilde{g}_B\left(t_0\right)}{\tilde{g}_B\left(t_1\right)}. \end{aligned} \tag{A15}$$

Substituting Eq. (A15) to Eq. (A14), it simplifies to

$$\delta H_{AB} = -\frac{1}{\tilde{g}_B\left(t_1\right)}\left(\delta W_B - \delta W_A\right) - H_{AB}\left(t_0\right)\frac{\tilde{g}_B\left(t_1\right) - \tilde{g}_B\left(t_0\right)}{\tilde{g}_B\left(t_1\right)}. \tag{A16}$$

By substitution of (A3) and (3) to the first term of the right hand side of Eq. (A16), Eq. (6) is derived.

Journal of Geodetic Science

• 3(2) • 2013 • 103-109 DOI: 10.2478/jogs-2013-0013 •

The geoid or quasigeoid – which reference surface should be preferred for a national height system?

Research Article

L. E. Sjöberg

Royal Institute of Technology (KTH), Stockholm, Sweden

Abstract:

Most European states use M. S. Molodensky's concept of normal heights for their height systems with a quasigeoid model as the reference surface, while the rest of the world rely on orthometric heights with the geoid as the zero-level. Considering the advances in data caption and theory for geoid and quasigeoid determinations, the question is which system is the best choice for the future. It is reasonable to assume that the latter concept, in contrast to the former, will always suffer from some uncertainty in the topographic density distribution, while Molodensky's approach to quasigeoid determination has a convergence problem. On the contrary, geoid and quasigeoid models computed by analytical continuation (e.g., rcr technique or KTH method) have no integration problem, and the quasigeoid can always be determined at least as accurate as the geoid. As the numerical instability of the analytical continuation is better controlled in the KTH method vs. the rcr method, we propose that any future height system be based on normal heights with a quasigeoid model computed similar to or directly based on the KTH method (Least squares modification of Stokes formula with additive corrections).

Keywords:

analytical continuation • geoid • normal height • orthometric height • quasigeoid

Received 01-02-2013; accepted 31-05-2013

1. Introduction

Today the Earth's surface and its geometric height above the selected reference ellipsoid (the geodetic height) can be regarded as known quantities as determined by satellite positioning, and, in particular from Global Navigation Satellite System (GNSS) data. However, the geoid, which for most countries is the primary vertical reference surface, cannot be provided from such data. For those countries the orthometric height (primarily determined by levelling) defines the height system as the height above the defined geoid model. For many countries the national height networks are rather sparse, and densifying heights by GNSS-levelling is a versatile technique. This technique needs the adopted geoid model, and along with that GNSS positioning improves, the quality of the geoid model has usually become the limiting factor for such height determination. The world-wide task of the geodetic community today to determine "the 1-cm geoid" is not easy, in particular in mountainous regions, as there is a major problem stemming from the geoid dependence on the only partly known topographic mass distribution, and this problem occurs also in determining orthometric heights.

In 1945 M. S. Molodensky (Molodensky et al. 1962) suggested substituting the geoid and orthometric height by the concepts of the quasigeoid and normal height, a brilliant idea to avoid the above problem with the topographic mass distribution. Since then former Soviet Union states and most of additional European states have changed their height systems to normal heights related with a quasigeoid surface.

In many countries today the qualities of the geoid and quasi-geoid models limit the use of GNSS-levelling, i.e. the technique to do levelling by GNSS and a gravimetric geoid model. Looking into the future, we may expect large improvements in GNSS technology for accurate positioning, which will increase the need for even more accurate geoid models. In parallel, the global and regional gravity field data will be known with much higher accuracy and density then today, which will be of great benefit for accurate geoid and quasi-geoid modelling. This is particularly the case if such development goes hand in hand with the necessary improvements in geoid computational techniques with thorough corrections for all types of errors and optimal combination of various data types (cf. Ågren and Sjöberg 2012).

Let us now imagine that all such errors are reduced to nearly zero, but the uncertainty in the topographic density distribution remains. Is the normal height concept attached with the quasi-geoid then the right way to go? Or are there also other limitations in the two approaches that make this choice less evident? According to Vanicek et al. (2012) it is so, as Molodensky's method is based on a geometric technique of integrating gravity over the Earth's surface (or, more precisely a smoothed version of it; the telluroid), which is too rough a surface in many regions to be a practical tool for accurate quasigeoid determination. On the other hand, there could also be other, more attractive, alternatives to compute the quasi-geoid.

Below we will discuss the two future alternatives for a precise vertical reference surface: the geoid or the quasi-geoid? The answer to the question is definitely very important to most countries, as the longer they wait to change system (if this is the choice), the more complicated and costly it will certainly be. And, in any case, each country deserves being confident in its choice of height system for the future.

2. The geoid

The geoid is an important tool for geophysical interpretation, etc., but this aspect is not at stake in this article. Here we will only consider it as a technical reference surface for height systems.

Gravimetric geoid modelling is based on Stokes' formula (Stokes 1849), which integrates the gravity anomaly on a sphere, approximating the Earth's surface, to provide the geoid height, i.e. the height of the geoid above the reference ellipsoid. In modern techniques, Stokes' formula is modified in various ways, primarily for taking into account an Earth Gravitational Model (EGM) for representing the long wavelengths of the geoid, while, more or less, only short wavelengths are determined by surface gravity (e.g., Sjöberg 2003b).

The application of Stokes' formula requires a number of corrections to the primarily determined surface gravity anomaly data. As the formula is performed on a sphere, and it does not allow masses external to the sphere, the data must be a) corrected for the topographic effect on gravity anomaly (direct topographic effect on gravity anomaly), b) corrected for the atmospheric effect (direct atmospheric effect) and c) reduced from the Earth's surface to the sphere (the downward continuation, dwc, effect). After these reductions to the gravity anomaly, Stokes' formula yields *the co-geoid*, which needs to be corrected for the removed topographic and atmospheric effects (indirect topographic and atmospheric effects) and for the approximation of sea level by a sphere rather than an ellipsoid (ellipsoidal effect) to provide the geoid height. It is important to remember that also in the modified Stokes formula, the EGM reduction to the sphere needs corrections for direct and indirect topographic and atmospheric effects.

The above technique for removing the effects of the external masses was originally called *regularization* of the geoid, resulting in the *co-geoid* or *regularized geoid* (Heiskanen and Moritz 1967, p. 289). Due to the mass shifts, the level surfaces have changed, which calls for two corrections (the primary and secondary indirect effects) to the co-geoid to finally reach the geoid. Today, this type of geoid determination is included in the more general term *remove-compute-restore (rcr) technique* for geoid determination. Although there are different versions of the rcr geoid estimator, it can generally be presented as follows:

$$\begin{aligned}\tilde{N}^{L,M} = & \frac{R}{4\pi\gamma_0}\iint\limits_{\sigma_0} S^L(\psi)\left[\Delta g^M + \delta\Delta g^T_{dir} + \delta\Delta g^a_{dir}\right.\\ & \left. +\delta\Delta g^{dwc} + \delta\Delta g^e_{dir}\right] d\sigma + c\sum_{n=0}^{M}\left(\frac{2}{n-1}\right)\Delta g_n^{EGM}\\ & + \delta N^T_{dir,M} + \delta N^T_I + \delta N^a_I + \delta N^e_I \qquad (1a)\end{aligned}$$

where R is the mean Earth radius, $\gamma_0$is normal gravity on the reference ellipsoid, $S^L(\psi)$ is the modified Stokes formula with geocentric angle ψ as argument, σ_0 is the cap of integration around the computational point,

$$\Delta g^M = \Delta g - \sum_{n=2}^{M}\Delta g_n^{EGM} \qquad (1b)$$

and the remaining correction terms in the bracket [] are in order the direct gravity anomaly effects of the topography, atmosphere and ellipsoid-to-sphere, $\delta N^T_{dir,M}$ is the direct topographic effect on the geoid of the EGM, and the last three terms are the indirect effects on the geoid of the topography, atmosphere and sphere-to-ellipsoid correction. In practice, some of these terms are usually neglected. [Note that in Eq. (1) the direct topographic effect refers to the gravity anomaly and not to the gravity attraction, which is more common in the rcr models. This leads to that the commonly added secondary indirect effect does not apply.] The indices $M$and L are the maximum degrees of the EGM and the modification of Stokes formula, respectively.

To keep the direct and indirect topographic effects small, they usually do not act on the complete attraction of the topographic masses, but on its difference to a compensation mass model located on or below sea level. By this trick one achieves both the original need for the DTE in removing the effect of the forbidden topographic masses, and a reduction of its magnitude. This reduction is dependent on the choice of compensation model, e.g., Helmert's 2nd method of condensation. To avoid systematic errors in the geoid process, the same compensation model must be used also for the indirect topographic effects.

Alternatively, the modified Stokes formula can be computed preliminary without the gravity anomaly corrections above, and each correction is added afterward as a combined effect of direct and indirect effects. This is the case in the KTH approach called *Least Squares Modification of Stokes formula with Additive corrections (LSMSA;* Sjöberg 2003a and 2003b):

$$\tilde{N}^{L,M} = \frac{R}{4\pi\gamma_0}\iint_{\sigma_0} S^L(\psi)\Delta g d\sigma + c\sum_{n=0}^{M}\left(Q_n^L + s_n\right)\Delta g_n^{EGM}$$
$$+ \delta N^T_{comb} + \delta N_{dwc} + \delta N^a_{tot} + \delta N^e_{tot}, \qquad (2a)$$

where s_n and $Q_n^L$are so-called modification parameters (chosen to minimize the error of the geoid estimator) and the Molodensky truncation coefficients, respectively. Moreover,

$$\delta N^T_{comb} = \frac{R}{4\pi\gamma_0}\iint_{\sigma_0} S^L(\psi)\,\delta\Delta g^T_{dir} d\sigma + \delta N^T_I \qquad (2b)$$

$$\delta N_{dwc} = \frac{R}{4\pi\gamma_0}\iint_{\sigma_0} S^L(\psi)\,\delta\Delta g_{dwc} d\sigma \qquad (2c)$$

$$\delta N^a_{tot} = \frac{R}{4\pi\gamma_0}\iint_{\sigma_0} S^L(\psi)\,\delta\Delta g^a_{dir} d\sigma + \delta N^a_I \qquad (2d)$$

$$\delta N^e_{tot} = \frac{R}{4\pi\gamma_0}\iint_{\sigma_0} S^L(\psi)\,\delta\Delta g^e_{dir} d\sigma + \delta N^e_I \qquad (2e)$$

are the additive corrections. Hence, the first row of Eq. (2a) is the modified Stokes formula, which uses the original EGM and gravity anomaly data. The second row of the equation consists of the additive corrections. For more details, see, e.g., Sjöberg (2003b).

Some advantages of the LSMSA method are that

- it uses least squares to minimize the effects of errors in the data and truncation.
- the method becomes more flexible, as the additive corrections can be added whenever the data for them becomes available. (It means that the repetition of the main computational steps can be avoided.)
- some additive corrections are much easier to compute than the corresponding direct and indirect effects.
- The direct and indirect effects are consistently combined into an effect on the geoid height. (This means also that the topographic compensation is meaningless.)

Assuming a constant topographic density ρ (e.g., 2.67 g/cm^3) the combined topographic effect on the geoid height (the direct plus the indirect topographic effects), called the topographic bias by Sjöberg (2007a), can be determined by the simple formula

$$dN^T_{comb} = -\frac{2\pi G\rho}{\gamma_0}\left(H^2 + \frac{2}{3}\frac{H^3}{R}\right), \qquad (3)$$

For Mount Everest this effect is of the order of 9.8 m.

Sjöberg (2004) used Eq. (3) to estimate the error in the geoid determination caused by the uncertainty in the topographic density. As the formula is directly proportional to the density, an uncertainty in this parameter of 10% may range to about 1 m for the highest mountain. At the elevations of 1 and 3 km the uncertainties become 1 and 11 cm, respectively. Although the assumption of a 10% error in density variation from the normal mean value may be pessimistic, these figures still show that this error source could be significant already for the 1-cm geoid model in mountainous regions. These error estimates are larger than those reported by Martinec (1998) and Vanicek et al. (2012), and the reason for this discrepancy is that our estimates are based on the total topographic effect, while the cited estimates only refer to the uncertainty in either the DTE or the (primary) indirect effect, which are only part of the error and they vary with the chosen method for reduction of gravity.

Theoretically, the computational results in the rcr and LSMA techniques should agree, but numerically significant differences may occur. For a comparative discussion, see Sjöberg (2005).The most important difference for this study is that the additive correction for the dwc effect to the geoid height by the LSMSA method is numerically much more stable than the dwc effect on the gravity anomaly used in the rcr technique, as discussed in Sect. 4.

Interestingly, the LSMSA technique initially uses a Stokes integral with surface gravity anomaly, and this integral is the same as the zero-order solution in the quasigeoid determination by Molodensky's technique (Molodensky et al. 1962).

3. The determination of the quasi-geoid

3.1. Molodensky's approach

Molodensky's original approach to quasigeoid height determination (Molodensky et al. 1962) is a geometric-physical method that deals with solving a Fredholm integral equation of the second kind over the known surface of the Earth by successive approximations, an iterative procedure whose convergence is doubtful unless the Earth's surface is sufficiently smoothed. For the real topography the solution is definitely divergent. Moritz (1980, Sect. 47) concludes that the Molodensky series can be regarded as an asymptotic series, implying that the iteration improves up to some high order, say $n_{\max}$, beyond which the series diverges. If this is the case, the truncation error below order $n_{\max}$ could well be negligible, suggesting that Molodensky's approach is practical. On the other hand, the Earth's surface is partly too rough to allow a sufficiently smooth integration, making the series divergent also at low orders of iteration. This is in agreement with the arguments used by Vanicek et al. (2012).

3.2. Remove-compute-restore technique

The quasi-geoid can also be determined by the rcr technique, which goes back to the method of analytical dwc of the gravity anomaly to sea level (Bjerhammar 1962 and 1963). The main difference to the geoid determination is that Stokes formula is replaced by the extended Stokes formula $S(r_P, \psi)$ (e.g., Heiskanen and Moritz 1967, p. 320). The resulting formula for the height anomaly (ζ) becomes:

$$\tilde{\zeta}^{L,M} = \frac{R}{4\pi\gamma} \iint_{\sigma_0} S^L(r_P,\psi)\left[\Delta g^M + \delta\Delta g^T_{dir} + \delta\Delta g^a_{dir} + \delta\Delta g^{dwc} + \delta\Delta g^e_{dir}\right] d\sigma + c\sum_{n=0}^{M}\left(\frac{2}{n-1}\right)\left(\frac{R}{r_P}\right)^{n+2}\Delta g_n^{EGM} + \delta\zeta^{T,M}_{dir} + \delta\zeta^T_I + \delta\zeta^a_I + \delta\zeta^e_I \tag{4}$$

where $S^L(r_P,\psi)$ is the extended, modified Stokes formula and γ is normal gravity at normal height.

As the integration now is carried out over a sphere (in contrast to Molodensky's approach above), there is no convergence problem of the integral. The only substantial problem is the same as with the geoid determination, namely to compute the dwc effect of the gravity anomaly from the surface to the internal (Bjerhammar) sphere. Thus we conclude that from a practical point of view this computational method is at least as simple as that for geoid determination by the rcr technique.

3.3. LSMSA technique

The LSMSA technique can also be used for computing the quasi-geoid height by the formula

$$\tilde{\zeta}^{L,M} = \frac{R}{4\pi\gamma} \iint_{\sigma_0} S^L(r_P,\psi)\Delta g\, d\sigma + c\sum_{n=0}^{M}\left(Q_n^L + s_n\right)\left(\frac{R}{r_P}\right)^{n+2}\Delta g_n^{EGM} + \delta\zeta_{dwc} + \delta\zeta^a_{tot} + \delta\zeta^e_{tot} \quad . \tag{5}$$

Note that there is no topographic effect in this equation, because the direct and indirect topographic effects cancel each other. Alternatively, the formula is written as a Stokes integral at point level (Ågren 2004, Sjöberg 2007b, Ågren et al. 2009):

$$\tilde{\zeta}^{L,M} = \frac{R}{4\pi\gamma} \iint_{\sigma_0} S^L(\psi)\Delta g\, d\sigma + c\sum_{n=0}^{M}\left(Q_n^L + s_n\right)\left(\frac{R}{r_P}\right)^{n+2}\Delta g_n^{EGM} + \delta\tilde{\zeta}_{dwc} + \delta\zeta^a_{tot} + \delta\zeta^e_{tot}, \tag{6}$$

which uses the original (modified) Stokes formula and a slightly different dwc effect, denoted $\delta\tilde{\zeta}_{dwc}$. For further details, see the next section, where the dwc problem is paid a special attention.

4. The dwc problem

Here we discuss the downward continuation effects by the rcr and LSMSA methods.

4.1. The dwc effect in the rcr technique

For the rcr technique the dwc problem is exactly the same for geoid and quasigeoid determination, as the problem is that of analytically continuing the surface gravity anomaly to sea level approximated by the computational sphere used in Stokes' formula.

The downward continuation of the gravity anomaly to the sphere can be approximated by a Taylor series, but this approach has limited use as the surface radial derivatives can hardly be determined sufficiently accurate. (One way to achieve the derivatives is by Proposition 1 of the Appendix.)

A mathematical rigorous model for the dwc problem is given by Poisson's integral formula, as first proposed by Bjerhammar (1962) and (1963). As the downward continued gravity anomaly on the sphere is the unknown under the integral, the formula is a Fredholm integral equation of the first kind and the problem is improperly posed (e.g., Payne 1975 and Hansen 1998). In the present case this implies that a detailed solution will be sensitive to errors in the data, and the propagated uncertainty will increase with topographic elevation and the resolution requested for the solution. Although the rcr technique uses topographically reduced gravity anomaly data as input (while Bjerhammar used original surface gravity anomaly data), the numerical ill-conditioning of the solution is a severe problem for a high resolution discretized grid step size, e.g., of 30"x 60" as demonstrated by Martinec (1998, Sect. 8.6.4) using data from the Canadian Rocky Mountains. However, as suggested by Novak et al. (2001) and demonstrated by Kingdon and Vanicek (2011), the ill-conditioning can be mitigated through regularization, and even without regularization Huang (2002) claimed that the geoid error only reaches a few centimetres in regions with elevations over 3 kilometres. Goli and Najafi (2011) showed that planar approximation of the Poisson equation can considerably speed up the laborious computational process at the prize of about 1 cm additional uncertainty in the geoid height.

4.2. The dwc problem by the LSMSA method

In the LSMSA approach the downward continuation effect on the gravity anomaly is directly estimated on the geoid or height anomaly. This approach is advantageous from the point of view that it avoids the ill-conditioning in the dwc of the gravity anomaly by Poisson's integral. Here we limit the discussion to that on the height anomaly when considering the original approach with a global set of gravity anomalies and no contribution from an EGM.

The preliminary estimator of the height anomaly is Eq. (5) with the dwc error (here considered for the limiting case with $\sigma_0 = \sigma$; for a more detailed study, see Sjöberg 2007b)

$$d\zeta_{dwc} = \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(r_P,\psi)\left[\Delta g^* - \Delta g\right]d\sigma, \tag{7}$$

which can be split into "the spherical shell effect" ($d\zeta^1_{dwc}$) and "the terrain effect" ($d\zeta^2_{dwc}$):

$$d\zeta_{dwc} = d\zeta^1_{dwc} + d\zeta^2_{dwc}, \tag{8a}$$

where

$$d\zeta^1_{dwc} = \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(r_P,\psi)\left[\Delta g^* - \Delta g\left(r_P,Q\right)\right]d\sigma_Q \tag{8b}$$

and

$$d\zeta^2_{dwc} = \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(r_P,\psi)\left[\Delta g\left(r_P,Q\right) - \Delta g\right]d\sigma_Q. \tag{8c}$$

Here Δg^* is the downward continued surface gravity anomaly. By expanding $S(r_P,\psi)$, Δg^* and $\Delta g\left(r_P,Q\right)$ as Laplace harmonic series and substituting into Eq. (7) and considering the following relation between the height anomaly harmonic (ζ_n) and gravity anomaly harmonic (Δg^*_n):

$$\zeta_n = \frac{R\Delta g^*_n}{\gamma\left(n-1\right)}\left(\frac{R}{r_P}\right)^{n+1}, \tag{9}$$

the spherical shell effect becomes

$$d\zeta^1_{dwc} = \sum_{n=2}^{\infty}\zeta_n\left[1-\left(\frac{R}{r_P}\right)^{n+2}\right]. \tag{10}$$

This formula can be applied numerically by utilizing a high degree EGM such as EGM2008 for estimating ζ_n. Alternatively one may expand the term $(R/r_P)^{n+2}$ a la Taylor to end up with the (truncated) series (Sjöberg 2007b)

$$d\zeta^1_{dwc} \approx \frac{H_P\Delta g_P}{\gamma} + 3\zeta_P\frac{H_P}{r_P} + \frac{H_P^2}{2\gamma}\left(\frac{\partial\Delta g}{\partial H}\right)_P - \frac{H_P^2\Delta g_P}{r_P\gamma} - 3\zeta_P\left(\frac{H_P}{r_P}\right)^2, \tag{11}$$

where the second and last terms are not significant at the 1 cm and one mm levels, respectively.

The terrain effect can also be estimated by a Taylor series (here limited to first order):

$$d\zeta^2_{dwc} \approx \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(r_P,\psi)\left(H_P - H_Q\right)\left(\frac{\partial\Delta g}{\partial H}\right)_Q d\sigma_Q \tag{12}$$

Eqs. (10)-(12) can be applied numerically without problems with respect to the roughness of the terrain. However, what could be required in rough terrain is that the Taylor series in Eq. (12) is extended to second or higher order (which should not imply a problem of convergence). An integral formula for obtaining higher derivatives of the gravity anomaly is given in Appendix.

We now consider the dwc effect in the slightly different approach for height anomaly by Eq. (6). Again we limit the study here to the limiting case with a global Stokes integration. Then the dwc effect can be written (Sjöberg 2007b)

$$d\tilde{\zeta}_{dwc} = \frac{R}{4\pi\gamma}\iint\limits_{\sigma}\left[S(r_P,\psi)\Delta g^* - S(\psi)\Delta g\right]d\sigma, \tag{13}$$

which can be split into the two components

$$\begin{aligned} d\tilde{\zeta}^1_{dwc} &= \frac{R}{4\pi\gamma}\iint\limits_{\sigma}\left[S(r_P,\psi)\Delta g^* - S\left(\psi\right)\Delta g\left(r_P,Q\right)\right]d\sigma_Q \\ &= \frac{H_P}{r_P}\zeta_P, \end{aligned} \tag{14a}$$

and

$$\begin{aligned} d\tilde{\zeta}^2_{dwc} &= \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(\psi)\left[\Delta g\left(r_P,Q\right) - \Delta g\right]d\sigma_Q \\ &\approx \frac{R}{4\pi\gamma}\iint\limits_{\sigma} S(\psi)\left(H_P - H_Q\right)\left(\frac{\partial\Delta g}{\partial H}\right)_Q d\sigma_Q. \end{aligned} \tag{14b}$$

Slightly different forms of these equations were derived by Ågren (2004), who substituted the factor γ^{-1} by the approximation $\gamma_0^{-1}\left(1+2H_P/r_P\right)$ in Eq. (6) and the subsequent equations. See Ågren et al. (2011).

5. Conclusions and final remarks

It is obvious that Molodensky's geometric/physical approach to determine the height anomaly cannot be applied with high accuracy in rough terrain due to convergence problems in the Molodensky series.

The rcr technique has a problem in the analytical downward continuation of the gravity anomaly in rough terrain. Hence, this is a problem both in geoid and quasigeoid determination.

In contrast to the rcr technique, the LSMSA method computes directly the dwc effect on the geoid or quasigeoid height. This procedure is much more stable than computing the dwc effect on the gravity anomaly. In Sect. 4 we have shown that the dwc effect on the height anomaly can be practically handled without any convergence problem related with the terrain. Considering also that the determination of the geoid (in contrast to the quasigeoid) requires information on the topographic density distribution (which is frequently not well known), we conclude that the height anomaly can be determined more accurately than the geoid. As there is a similar problem for the uncertainty in the topographic density distribution in determining orthometric heights (but not for normal heights), we conclude that a normal height system is the best choice for a future height system. Once the normal gravity field is defined, the normal heights and the quasigeoid can be determined without any error stemming from the topographic mass distribution, and the quasigeoid can be estimated more precisely than the geoid as the reference surface.

One aspect on orthometric heights vs. normal heights is also that the latter are smoother (as they are geopotential numbers divided my mean gravity and mean normal gravity, respectively, and normal gravity is the smoother of the two quantities), which makes sense, e.g., when interpolating between data points. On the other hand, the geoid, being an equipotential surface of the Earth's gravity field, is smoother than the quasigeoid surface.

References

Ågren J., 2004, Regional geoid determination methods for the era of satellite geodesy. PhD thesis in geodesy, Royal Institute of Technology, Stockholm.

Ågren J. and Sjöberg L E., 2012, Investigations of the requirements for a future 5 mm quasigeoid model over Sweden, GGHS 2012, Venice, Italy. (Poster available)

Ågren J., Sjöberg L. E. and Kiamehr R., 2009,The new gravimetric geoid model KTH08 over Sweden, J. Appl. Geod. 3, 143-153.

Bjerhammar A., 1962, Gravity reduction to a spherical surface, Royal Institute of Technology, Division of Geodesy, Stockholm, Sweden.

Bjerhammar A., 1963, A new theory of gravimetric geodesy, Royal Institute of Technology, Division of Geodesy, Stockholm, Sweden.

Goli M. and Najafi-Alamdari M., 2011, Planar, spherical and ellipsoidal approximations of Poisson's integral in near zone. J. Geod. Sci. 1, 1, 17-27.

Hansen P. C., 1998, Rank-deficient and discrete ill-posed problems. SIAM monographs on mathematical modelling and computation, Philadelphia, Pennsylvania.

Heiskanen W. A. and Moritz H., 1967, Physical geodesy, W H Freeman and Co., San Francisco and London.

Huang J., 2002, Computational methods for the discrete downward continuation of the Earth gravity and effects of lateral topographical mass density variation on the gravity and the geoid, Ph.D. dissertation, Dept. Geod. and Geom. Eng. Techn. Rep. No. 216, UNB, Fredricton, New Brunswick, Canada.

Kingdon R. and Vanicek P., 2011, Poisson downward continuation solution by the Jacobi method, J. Geod. Sci. 1, 1, 74-81.

Martinec Z., 1998, Boundary-value problems for gravimetric determination of a precise geoid, Lecture Notes in Earth Sciences, Springer.

Molodensky M.S., Eremeev V.F. and Yurkina M.I., 1962, Methods for study of the external gravitational field and figure of the earth, Transl. from Russian (1960), Israel program for Scientific Translations, Jerusalem, Israel.

Moritz H., 1980, Advanced physical geodesy. Herbert Wichmann Verlag, Karlsruhe.

Novak P., Kern M., Schwarz K., 2001, "Numerical studies on the harmonic downward continuation of band-limited airborne gravity." Stud. Geophys. Geod. 45, 327-345.

Payne L. E., 1975, Improperly posed problems in partly differential equations, SIAM, Philadelphia, Pennsylvania.

Sjöberg L. E., 2003a, A computational scheme to model the geoid by the modified Stokes's formula without gravity reductions, J. Geod. 77, 423-432.

Sjöberg L. E., 2003b, A general model of modifying Stokes' formula and its least-squares solution, J. Geod. 77, 459-464.

Sjöberg L. E., 2004, The effect on the geoid of lateral topographic density variations, J. Geod. 78, 34-39.

Sjöberg L. E., 2005, Discussion on the approximations made in the practical implementation of the remove-compute-restore technique in regional geoid modelling, J. Geod. 78, 645-653.

Sjöberg L. E., 2007a, The topographic bias in geoid determination, J. Geod. 81, 345-350.

Sjöberg L. E., 2007b,The downward continuation effects for geoid and quasigeoid heights in Stokes' formulas, Research report from research visit to NGS, Washington, D.C., March 2007.

Stokes G. G., 1849, On the variation of gravity on the surface of the earth, Trans Cambridge Phil. Soc. 8, 672-695.

Vanicek P., Kingdon R. and Santos M., 2012, Geoid versus quasigeoid: a case of physics versus geometry, Contr. Geophys. Geod. 42, 1, 101-117.

APPENDIX

From Heiskanen and Moritz (1967, p. 38) one obtains the radial derivative of the disturbing potential T (or any harmonic function) on the sphere of radius $R$by the following surface integral:

$$\left(\frac{\partial T}{\partial r}\right)_P = -\frac{T_P}{R} + \frac{R^2}{2\pi}\iint_\sigma \frac{T - T_P}{l_0^3} d\sigma, \qquad \text{(A1)}$$

where

$$l_0 = 2R\sin\left(\psi/2\right). \qquad \text{(A2)}$$

Notation: $g^{(k)} = \partial^k \Delta g / \partial r^k$.

The following proposition was derived by Sjöberg (2007b).

Proposition .1.
Under spherical approximation it holds that

$$g_P^{(k+1)} = -\frac{k+1}{r_P} g_P^k + \frac{1}{16\pi r_P}\iint_\sigma \frac{g^{(k)} - g_P^{(k)}}{\sin^3\left(\psi/2\right)} d\sigma. \qquad \text{(A3)}$$

Proof. Under spherical approximation it holds that

$$\Delta g = -\frac{\partial T}{\partial r} - 2\frac{T}{r}, \qquad \text{(A4)}$$

from which follows that $U = r^k g^{(k)}$ is a harmonic function. Hence by putting $U = T$ and $R = r_P$ in Eq. (A1) the proposition follows. □

Journal of Geodetic Science

• 3(2) • 2013 • 110-110 DOI: 10.2478/jogs-2013-0014 •

Errata to article Sjöberg, L.E. (2012), Journal of Geodetic Science 2: 162-171 entitled Solutions to the ellipsoidal Clairaut constant and the inverse geodetic problem by numerical integration

Erratum

Lars E. Sjöberg,

Royal Institute of Technology (KTH)

Received 27-05-2013; accepted 31-05-2013

Location	reads	should read
p 162,col 1; p 166, col 2; ref. list	Schmidt (2006a) (2006b)	Schmidt (2000) (2006)
p 163, col 2	Sjöberg and Shirazian (2011)	Sjöberg and Shirazian (2012)
Eqs. (14a) and (14b)	4	2
Eq. (15)	$and1+$	and 1+
p 165, below Eq. (18)	(4b)	(4c)
p 167, col 2	(??)	(A16a)

Missing in reference list:

Sjöberg L.E. and Shirazian M., 2012, Solving the direct and inverse geodetic problems on the ellipsoid by numerical integration. J. Surv. Eng. 138: 9-16

Corollary 4.2 should read:

If $\beta_2 = -\beta_1$, then

$$c = \frac{\sin\left(D\lambda/2\right)}{\sqrt{\sin^2\left(D\lambda/2\right) + t_1^2}}. \tag{16}$$

The proof follows directly from Eq. (14b).