Survey Review, **42**, 315 pp. 20-26 (January 2010)

AN INTEGRAL FOR GEODESIC LENGTH

AFTER DERIVATIONS BY P. D. THOMAS

Craig M. Rollins
National Geospatial-Intelligence Agency

ABSTRACT

Using a change of variable suggested by P. D. Thomas (1952), the arclength of a segment of a geodesic curve on an ellipsoid becomes an integral having the same form as arclength on an ellipse, a simpler problem. This leads to a succinct theoretical solution to the Direct and Indirect Problems of geodesics. With modern mathematical software, it is also a practical solution.

KEYWORDS: Geodesic curve. Arc length. Ellipsoid.

On a smooth surface, a geodesic is a curve which is as straight as possible while constrained to lie on the surface. On an ellipsoid of revolution about its minor axis, a geodesic is infinite in two directions, in general, as discussed in [**2**] page 78. A geodesic segment is the portion of a geodesic between two endpoints. With qualifications, it has the property that no shorter path exists that connects its two endpoints. This property may be the main interest about geodesic segments, but the theory begins with a pair of integral formulas which stand apart from the shortest-path qualifications, provide a basis for calculations, and allow geodesic segments to be arbitrarily long. The integral formulas are of first importance — unwanted long solutions can be discarded — and a new formula for geodesic segment length is presented here.

REDUCED LATITUDE SOLUTION

A succinct theoretical statement of the solutions to the Direct and Indirect Problems of geodesics, covering all cases, is a desirable item to have in hand before developing these problems' numerical and software solutions. This can be done in more than one way. Using β, the reduced latitude, as the variable of integration, [**4**], [**6**] and others have:

$$\lambda_2 - \lambda_1 = c\int_{\beta_1}^{\beta_2} \frac{\sqrt{1-\varepsilon^2 \cos^2\beta}}{(\cos\beta)\sqrt{\cos^2\beta - c^2}}\,\mathrm{d}\beta \tag{1}$$

$$s_2 - s_1 = a\int_{\beta_1}^{\beta_2} \frac{(\cos\beta)\sqrt{1-\varepsilon^2 \cos^2\beta}}{\sqrt{\cos^2\beta - c^2}}\,\mathrm{d}\beta \tag{2}$$

where c is the non-metric Clairaut constant of the geodesic defined by:

$$c = \frac{\sin\alpha\ \cos\phi}{\sqrt{1-\varepsilon^2 \sin^2\phi}} = \sin\alpha\ \cos\beta \tag{3}$$

and every point on the geodesic curve has a geodetic longitude λ, a geodetic latitude

Contact: Craig M Rollins e-mail: Craig.M.Rollins@nga.mil

DOI 10.1179/003962609X451663

ϕ, a reduced latitude $\beta = \arctan(\sqrt{1-\varepsilon^2}\ \tan\phi)$, an azimuth α of the forward direction measured clockwise from North, and a distance s ("mile-marker") along the geodesic from some reference point on it. These notations are also subscripted "1" and "2" consistently for the two endpoints of the segment of the geodesic. The eccentricity of the ellipsoid is ε and its semi-major axis is a. All angles are in radians.

Eq. (1) and (2) are correct for the case that $\beta_1 < \beta_2$ and the geodesic segment does not pass through a "vertex" — defined as a point of extreme latitude on the geodesic, i.e. a point where $\beta = \pm\beta_{\max}$ with $\beta_{\max} = \arccos(|c|)$. Other situations are handled by dividing the geodesic segment into subsegments that start or end at a vertex, and applying symmetry considerations. The various cases are illustrated by the following representative examples, where (λ_1, β_1) is the first city, (λ_2, β_2) is the second city, and λ_1, λ_2 both lie in the interval $(-\pi, \pi]$ except where noted. For the geodesic from New York to Paris, the integrations would be $\int_{\beta_1}^{\beta_{\max}} + \int_{\beta_2}^{\beta_{\max}}$ in place of $\int_{\beta_1}^{\beta_2}$. The same is true for the geodesic from New York to Tokyo, where also it is noted that $\lambda_2 - \lambda_1 < 0$ because $c < 0$ and both cities need to have an East (positive) longitude or both need to have a West (negative) longitude. This is required because Eq. (1) does not recognize the equivalence of 180°E with 180°W. In this article, longitude sits on the real line, not on the unit circle. Capetown to Sydney will utilize $\int_{-\beta_{\max}}^{\beta_1} + \int_{-\beta_{\max}}^{\beta_2}$. Reykjavik south to Montevideo, the short way, is $\int_{\beta_2}^{\beta_1}$. Reykjavik to Montevideo, the long way over both polar regions is $\int_{\beta_1}^{\beta_{\max}} + \int_{-\beta_{\max}}^{\beta_{\max}} + \int_{-\beta_{\max}}^{\beta_2}$ provided Montevideo is given an East longitude. It is essential that the integrations be applied to Eq. (1) and (2) consistently. Let geodesics with $c > 0$ be called prograde, and geodesics with $c < 0$ be called retrograde.

Since $c = \pm\cos(\beta_{\max})$, the integrals in Eq. (1) and (2) converge but are improper when β_1 or β_2 equals $\pm\beta_{\max}$, or the integration is split to include $\pm\beta_{\max}$ as discussed above. This causes numerical difficulties, especially when the integral is embedded in iterative procedures like those discussed below for the Direct and Indirect Problems. An alternative approach discussed next avoids this difficulty.

LONGITUDE DIFFERENCE INTEGRAL IN P. D. THOMAS

On pages 64 - 65 of [**5**], P. D. Thomas uses the change of variable, $\sin\phi = k\sin\theta$ with $\theta = 0$ at some Equator crossing $\phi = 0$, and with the constant k defined by:

$$k = \sqrt{\frac{1-c^2}{1-c^2\varepsilon^2}} \tag{4}$$

(The Clairaut constant in [**5**] is $a \cdot c$ in this article.) By this transformation, a longitude difference integral like Eq. (1) is converted to an Elliptic Integral of the 3rd Kind:

$$\lambda_2 - \lambda_1 = \frac{c\,(1-\varepsilon^2)}{\sqrt{1-c^2\varepsilon^2}} \int_{\theta_1}^{\theta_2} \frac{1}{(1-k^2\sin^2\theta)\sqrt{1-k^2\varepsilon^2\sin^2\theta}}\,\mathrm{d}\theta \tag{5}$$

where c and k are given by Eq. (3) and (4) and $\theta_i = \arcsin((\sin\phi_i)/k)$ for $i = 1, 2$. The same caveats as above apply, namely that $\theta_1 < \theta_2$, and the geodesic segment does not pass through a vertex, now identified as a point where $\theta = \pm\pi/2$. Other situations are handled exactly like the formulations that used β as the variable of integration, now with $\pi/2$ taking the place of $\beta_{\max}$. For example, New York to Paris would employ $\int_{\theta_1}^{\pi/2} + \int_{\theta_2}^{\pi/2}$ in place of $\int_{\theta_1}^{\theta_2}$. Note that for $0 < k < 1$, the integral is never improper.

COMPANION INTEGRAL FOR GEODESIC ARCLENGTH

Just as Eq. (2) is the companion to Eq. (1), so also is the following equation a companion to Eq. (5). It is new here or it deserves to be more widely known:

$$s_2 - s_1 = \frac{a\,(1-\varepsilon^2)}{\sqrt{1-c^2\varepsilon^2}} \int_{\theta_1}^{\theta_2} \frac{1}{\left(1-k^2\varepsilon^2\sin^2\theta\right)^{3/2}}\,\mathrm{d}\theta \tag{6}$$

It employs the same caveats and extensions to other situations as adopted for Eq. (5). Note that for $0 < k < 1$, the integral is never improper.

To prove Eq. (6), the integrand in Eq. (2) must be converted from expressions in β to expressions in θ. There are three pieces. Let the first piece be $\cos\beta\,\mathrm{d}\beta$. From Eq. (3) and the definition of β, we have $\cos\beta = (\cos\phi)/w$, and $\mathrm{d}\beta = \sqrt{1-\varepsilon^2}\,\mathrm{d}\phi / w^2$ where $w = \sqrt{1-\varepsilon^2\sin^2\phi} = \sqrt{1-k^2\varepsilon^2\sin^2\theta}$ is converted as required. After $\sin\phi = k\sin\theta$ is differentiated to give $\cos\phi\,\mathrm{d}\phi = k\cos\theta\,\mathrm{d}\theta$, the above can be combined to give $w^{-3}k\sqrt{1-\varepsilon^2}\cos\theta\,\mathrm{d}\theta$ for piece 1. The second piece is $\sqrt{1-\varepsilon^2\cos^2\beta}$ and converts to $w^{-1}\sqrt{w^2-\varepsilon^2\cos^2\phi} = w^{-1}\sqrt{1-\varepsilon^2}$. The third piece is $\sqrt{\cos^2\beta - c^2}$ which can be similarly manipulated to produce $w^{-1}\sqrt{\cos^2\phi - c^2w^2}$ and then $w^{-1}\sqrt{(1-c^2)-k^2(1-\varepsilon^2c^2)\sin^2\theta}$. Replacing k^2 according to Eq. (4) yields $w^{-1}\sqrt{(1-c^2)(1-\sin^2\theta)}$ or $w^{-1}\sqrt{1-c^2}\cos\theta$ for piece 3. Combining pieces 1, 2, and 3 and the factor a gives $\mathrm{d}s = a(1-\varepsilon^2)\,w^{-3}k/\sqrt{1-c^2}\,\mathrm{d}\theta$. This is the differential form of Eq. (6) after simplifying $k/\sqrt{1-c^2}$ and replacing w^{-3}. (In like manner, Eq. (5) can be derived from Eq. (1).)

Remarkably, Eq. (6) is also the formula for the arclength of a segment of a meridian on an ellipsoid of semi-major axis $a_1 = a\sqrt{1-c^2\varepsilon^2}$ and eccentricity $\varepsilon_1 = k\,\varepsilon$ between geodetic latitude L_1 taken to be the same number as θ_1 and geodetic latitude L_2 taken to be θ_2. For this second ellipsoid, the meridional arclength is

$a_1(1-\varepsilon_1{}^2)\int_{L_1}^{L_2}(1-\varepsilon_1{}^2\sin^2 L)^{-3/2}\mathrm{d}L$. Substituting for a_1, ε_1, L_1, and L_2 and simplifying yields Eq. (6). More than a curiosity, this implies that the literature and algorithms for meridional arclength, an easier problem, apply here with the above substitutions. The integral in Eq. (6) is a special case of the Elliptic Integral of the 3rd Kind, treated extensively in the literature.

WHY THETA IS PREFERRED OVER BETA

The main purpose of this paper is to (re)introduce Eq. (6) and explain its advantages over Eq. (2). Some advantages have been mentioned — the integral is proper and the formula coincides with the oft-studied meridional arclength computation. Another advantage, which Eq. (5) likewise has over Eq. (1), is that the variable θ can be extended in meaning and values to allow the single integral form $\int_{\theta_1}^{\theta_2}$ to handle all geodesic segments of any length. To introduce this idea, reconsider the integrations with respect to θ for the geodesic from New York to Paris. Then, $\int_{\theta_1}^{\pi/2}+\int_{\theta_2}^{\pi/2}=\int_{\theta_1}^{\pi-\theta_2}$ because the integrands of Eq. (5) and (6) depend on θ only as part of $\sin\theta$, and $\sin\theta$ is symmetric with respect to $\theta=\pi/2$. This suggests that the value of θ for Paris should be $\pi-\theta_2>\pi/2$ instead of $\theta_2<\pi/2$, i.e. that θ_2 should be so redefined.

Because the variable k occurs solely as k^2 in Eq. (5) and (6), there is room to amend Eq. (4) and give k a sign. Let $k>0$ if α_1 belongs to the NE or NW quadrants, and let $k<0$ if α_1 belongs to the SE or SW quadrants. Then, the equation by which θ was introduced, namely $\sin\phi=k\sin\theta$, is retained, but now the relationship between θ and ϕ is many-to-one. At the initial point, $|\theta_1|\le\pi/2$ is the rule, and at the terminal point, $\sin\theta_2=(\sin\phi_2)/k$ such that the interval $\theta_1\le\theta\le\theta_2$ includes all the vertices lying on the geodesic segment. The integrations are now always $\int_{\theta_1}^{\theta_2}$ with $\theta_1<\theta_2$. The previous examples are reworked as follows, where (λ_1,θ_1) is the first city, (λ_2,θ_2) is the second city, and the previous stipulations about city longitudes λ_1 and λ_2 still hold. For the geodesic segment from New York to Paris, α_1 belongs to the NE quadrant, $k>0$, $0<\theta_1<\pi/2$, and $\pi/2<\theta_2<\pi$. New York to Tokyo has α_1 in the NW quadrant, and again $k>0$, $0<\theta_1<\pi/2$, and $\pi/2<\theta_2<\pi$. Capetown to Sydney will have α_1 in the SE quadrant, $k<0$, $\phi_1<0$ yielding $0<\theta_1<\pi/2$, and $\phi_2<0$ yielding $\pi/2<\theta_2<\pi$. Reykjavik to Montevideo, the short way, will have α_1 in the SW quadrant, $k<0$, $-\pi/2<\theta_1<0$, and $0<\theta_2<\pi/2$. Reykjavik to Montevideo, the long way, will have α_1 in the NE quadrant, $k>0$, $0<\theta_1<\pi/2$, and $3\pi/2<\theta_2<2\pi$.

The set of allowed values of (θ_1, θ_2) is independent of the constant c, whereas the set of allowed values of (β_1, β_2) is not. In particular, the extreme latitudes always correspond to $\theta = \pm\pi/2$. This allows an assignment of geometric properties to parameters such that c determines the infinite geodesic up to a rotation of the ellipsoid, θ_1 and θ_2 identify the finite segment of interest, and λ_1 rotates it into place. In a variation of the above theme, let $|\theta_1| \le \pi/2$ be dropped and $k > 0$ be adopted as requirements. Then, the members of the quadruple $(c, \theta_1, \theta_2, \lambda_1)$ can be chosen independently subject only to $0 \le |c| < 1$ and $\theta_1 < \theta_2$ and will exhaust (with repetitions) all the possibilities of constructing non-Equatorial geodesic segments. Evaluating Eq. (5) and (6) in this way is a methodology for generating test vectors, $(\lambda_1, \phi_1, \alpha_1, s_2 - s_1, \lambda_2, \phi_2, \alpha_2)$, for the Direct and Indirect Problems without first developing and debugging the case logic and the iteration logic that these problems' solutions require.

DIRECT PROBLEM

The next purpose of this article is to make the case that Eq. (3) through (6) with iteration logic (as needed) can solve the Direct and Indirect Problems of geodesics.

In the Direct Problem, $(\lambda_1, \phi_1, \alpha_1, s_1, s_2,)$ is given, and $(\lambda_2, \phi_2, \alpha_2)$ is to be found. As an example, suppose an oceanographic research drone is launched from $\lambda_1 = -72°$, $\phi_1 = -34°$ (near Santiago, Chile) on a geodesic path starting at $\alpha_1 = -100°$ and $s_1 = 0$. Let the GRS80 ellipsoid be adopted for this exercise, i.e. $a = 6378137\,\text{m}$, $\varepsilon^2 = f(2-f)$, and $f = 1/298.257222101$. Where will it be after $s_2 = 10\,000\,\text{km}$ of travel?

Eq. (3) gives $c = -0.8172985069$, Eq. (4) with sign convention gives $k = -0.5775071640$, and the definition of θ gives $\theta_1 = 75.53201382°$. (Angles are computed in radians, but reported here in degrees). The quantity θ_2 is the only unknown quantity in Eq. (6) and is found by iteration of Eq. (6). Trial $\theta_2 = 90°$ yields $s_2 = 1609$ km, well short of the required 10 000 km. Therefore, the geodesic will pass the point of extreme southern latitude and head northward (at some angle). Trial $\theta_2 = 180°$ produces $s_2 = 11599$ km, slightly too long. Linear interpolation produces the next trial, and one Newton-Raphson iteration later, we get $\theta_2 = 165.5689530°$, accurate to 10^{-7} degrees. Eq. (5) produces $\lambda_2 - \lambda_1 = -95.42822834°$ from which $\lambda_2 = -167.4282283°$. Then, $\phi_2 = -8.274933372°$ is obtained from θ_2, and $\sin\alpha_2 = -0.8258397811$ is obtained from Eq. (3). Approaching the terminal endpoint, α lies in the NW quadrant, so $\alpha_2 = -55.67372482°$. The quantity α_2 is a forward azimuth. The back azimuth $\alpha_2 + 180°$ is often reported instead. This finishes the Direct Problem.

INDIRECT PROBLEM

In the Indirect Problem, $(\lambda_1, \phi_1, \lambda_2, \phi_2)$ is given, and $(s_2 - s_1, \alpha_1, \alpha_2)$ is to be found. As an example, let the initial point be Mumbai (72° E, 19° N for this example) and let the terminal point be Los Angeles (119° W, 34° N). Using the GRS80 ellipsoid, what is the length of the shortest path between these points, and what are the forward directions of travel at the endpoints of this path?

To begin, λ_2 must be changed to $-119° + 360° = 241°$ since the shortest route is the prograde (generally eastward) geodesic, and longitude must be continuous. Although α_1 is yet to be found, it should lie in the NE quadrant, making $k > 0$. With the maximum latitude occurring enroute, $\theta > 90°$ will be the case approaching Los Angeles. Therefore at the endpoints, the extraneous solutions for θ are eliminated in favor of $\theta_1 = \arcsin((\sin\phi_1)/k)$ and $\theta_2 = \pi - \arcsin((\sin\phi_2)/k)$ where the values of $\arcsin x$ lie in the interval $[-\pi/2, \pi/2]$.

In Eq. (5), k is a function of c, and θ_1 and θ_2 are the above functions of k, hence functions of c. In this view, c is the only unknown in Eq. (5) from which it can be found by iteration. Trial $c = 0.1$ produces 174.10693373°, larger than the target $\lambda_2 - \lambda_1 = 169°$, and trial $c = 0.3$ produces 161.47438509°, too small. Seven iterations of a crude secant method, or use of *Mathematica*'s `FindRoot` routine produces $c = 0.184155171333$. The game is to find c somehow and the methods of [4] are an alternative, with Eq. (5) as the check. Then, Eq. (4) gives $k = 0.9830087745$, $\theta_1 = 19.34135798°$, and $\theta_2 = 145.3293358°$. Eq. (6) is now evaluated to get $s_2 - s_1 = 14024777.42$ m for the length of the geodesic segment, and Eq. (3) gives $\alpha_1 = 11.22703573°$ and $\alpha_2 = 167.1794253°$. This finishes the typical Indirect Problem.

Special Cases

Lastly, this article aims to show that Eq. (3) through (6) with an addendum for the Equator cover the special cases.

In the Indirect Problem, if $\phi_1 = \phi_2 = 0$ and $\pi\sqrt{1-\varepsilon^2} < |\lambda_2 - \lambda_1| < \pi$, the starting value for $|c|$ is determined by linear interpolation between $|c| = 1$ at $|\lambda_2 - \lambda_1| = \pi\sqrt{1-\varepsilon^2}$ (the "lift-off longitude" given in [3] page 42) and $|c| = 0$ at $|\lambda_2 - \lambda_1| = \pi$. The starting value for c takes its sign from $\lambda_2 - \lambda_1$.

For the meridional case, Eq. (6) is correct as given, with substitutions $c = 0$, $k = 1$, and $\theta = \phi$ or substitutions $c = 0$, $k = -1$, and $\theta = -\phi$. For fixed θ_1 and θ_2, Eq. (5) holds in the limit as $c \to 0^+$ (respectively, $c \to 0^-$) and simplifies to $\lambda_2 - \lambda_1 = n\pi$, for some integer $n \geq 0$ (respectively, some integer $n \leq 0$). Neither Eq. (5) nor Eq. (6) handles the Equatorial case where $c = \pm 1$, $k = 0$, and θ is undefined, but this case is covered by $s_2 - s_1 = a \cdot c(\lambda_2 - \lambda_1)$ instead. The quadrant ambiguity of α in Eq. (3) is usually resolved by inspection, but also provides the Indirect Problem's two solutions where that is the case. This deserves further comment:

A solution to the Indirect Problem is a path, as short as possible, that connects the two given points. The number of solutions has to respect the symmetries of the ellipsoid and for $\varepsilon > 0$ is either 1, 2, or ∞. If it is infinite, the two given points are the North and South Poles. If it is two, $\phi_2 = -\phi_1$, and the two paths, portrayed on a ϕ versus λ plot, will be symmetric to each other through the point given by $\lambda = (\lambda_1 + \lambda_2)/2$ and $\phi = 0$. Note that the case of nearly antipodal points does not receive separate treatment in the methods presented here.

Some geodesics are closed curves, but most are not. For a given geodesic, the azimuth of every Equator crossing from south to north (respectively, from north to south) is a constant, namely $\arcsin c$ (respectively, $\pi - \arcsin c$). Therefore, a geodesic will retrace itself if it crosses the Equator from south to north a second time at the same point. If $c = 0$ or $c = \pm 1$ or c has the property that $\lambda_2 - \lambda_1 = 2\pi q$ for some rational number q when Eq. (5) is evaluated using $\theta_1 = 0$ and $\theta_2 = 2\pi$, then the geodesic is a closed curve. For other values of c, the geodesic winds around the ellipsoid forever, never crossing the Equator twice at the exact same point.

IMPLEMENTATION, TESTING, COLLABORATION

The Direct and Indirect Problem formulations above were implemented in the *Mathematica* programming environment where Elliptic Integrals, 20-digit arithmetic, and root-finding routines were available and intelligent enough to overpower any numerical difficulties. Tests were run against calculators available from the National Geospatial-Intelligence Agency (NGA) [**8**], and the National Geodetic Survey [**9**]. The results were in full agreement (to the precision of the other programs) except where the other programs were stressed beyond their claims of validity such as in cases of nearly antipodal points. The *Mathematica* implementation revealed the shortcomings of the other calculators.

In another vein, this work was part of NGA's participation in the development of [**1**]. In that document, "closed form" solutions for all coordinate conversions and spatial operations are preferred, if such can be obtained. The equations presented here serve that end. Helpful criticisms of drafts of this article were provided by Ralph M. Toms of SRI, International and Paul D. Berner of SEDRIS and some checking with *Mathematica* was provided by Thomas H. Meyer of the University of Connecticut.

References

1. ISO/IEC 18026:2006(E), Information Technology -- Spatial Reference Model (SRM).
2. Rapp, R. H., 1991. *Geometric Geodesy* Part I, The Ohio State University.
3. Rapp, R. H., 1993. *Geometric Geodesy* Part II, The Ohio State University.
4. Sjoberg, Lars E., 2007. Precise determination of the Clairaut constant in ellipsoidal geometry, *Survey Review*, **39**(303), pages 81-86.
5. Thomas, Paul D., 1952. Conformal Projections in Geodesy and Cartography, Special Publication No. 251, U.S. Department of Commerce, Coast and Geodetic Survey, Washington, D.C. Reprinted 1978.
6. Thomas, Paul D., 1970. *Spheroidal Geodesics, Reference Systems, & Local Geometry*, U.S. Naval Oceanographic Office.
7. Wolfram, Stephen, 2003. The *Mathematica* Book, 5th Edition, Wolfram Media.
8. http://geoengine.nima.mil/geospatial/SW_TOOLS/NIMAMUSE/webinter/rangbear.html
9. http://www.ngs.noaa.gov/TOOLS/Inv_Fwd/Inv_Fwd.html