The Problem of a Maritime Boundary Involving Two Horizontal Geodetic Datums

Petr Vaníč ek, Department of Surveying Engineering, University of New Brunswick, P.O. Box 4400, Fredericton, N.B., Canada E3B 5A3

Abstract

In this contribution, the following topics are discussed: the geocentric coordinate system and its different realizations used in geodetic practice; the definition of a horizontal geodetic datum (reference ellipsoid) and its positioning and orientation with respect to the geocentric coordinate system; positions on a horizontal datum and errors inherent in the process of positioning; and distortions of geodetic networks referred to a horizontal datum. The problem of determining transformation parameters between a horizontal datum and the geocentric coordinate system from known positions is then analysed.

When a maritime boundary involves two countries that use different horizontal datums to describe their shorelines, it becomes necessary to transform positions (of baseline points) from one datum to the other. These transformations are normally accomplished through the geocentric coordinate system and they include the transformation parameters of the two datums as well as the representation of the respective network distortions. Problems encountered in putting together these transformations are pointed out.

Geocentric Coordinate Systems and their Realizations

By definition, a geocentric coordinate system is a system whose origin (0, 0, 0) coincides with the centre of mass, C, of the earth, and whose axes are fixed within the earth. There exist infinitely many such systems differing from each other by the orientation of their x-, y-, and zaxes.

The geocentric system which is most often used in geodesy is the *Conventional Terrestrial* (CT) system which is oriented so that its z-axis points towards the Conventional International

Origin (CIO), the x-axis lies in the Conventional Greenwich Meridian, and the y-axis makes (with the other two axes) a right-handed Cartesian triad [Vaníč ek and Krakiwsky, 1986]. This system provides the most convenient link with metric astronomy. The other two often used geocentric coordinate systems, the Instantaneous Terrestrial, referred to the instantaneous spin axis of the earth, and the Natural Geocentric system, whose axes coincide with the earth's principal axes of inertia, provide similarly convenient links with astronomical observations and with the dynamics of the earth.

Positions in the CT-system are sometimes given in Cartesian coordinates, sometimes in curvilinear geodetic coordinates, f, l, h, i.e., in geodetic latitude, longitude, and height. The use of curvilinear geodetic coordinates, however, requires the introduction of a geocentric reference ellipsoid (co-axial with the Cartesian system) with its semi-axes denoted by "a" and "b". The geodetic height "h" is then taken as the height above that reference ellipsoid. In some literature, the biaxial reference ellipsoid is called the reference spheroid.

Curvilinear geodetic coordinates are easily transformed into Cartesian by the following formula:

$$\begin{bmatrix} x \\ N + h \right) \cos \phi \cos \lambda \\ y \\ z \\ \begin{bmatrix} (N+h) \cos \phi \sin \lambda \\ N \\ z \\ \end{bmatrix}, \qquad (1)$$

where N denotes the radius of curvature of the ellipsoid in the direction perpendicular to the meridian plane [Vaníč ek and Krakiwsky, 1986]. Simple formulae for the inverse transformation do not exist. The inverse transformation has to be solved either iteratively, or through a linearization, or using an algebraic equation of 4th degree.

The Geodetic Reference System of 1980 (GRS 80), recommended for use in geodesy by the International Association of Geodesy (IAG) in 1980 [IAG, 1980], utilizes the CT-system combined with a reference ellipsoid deemed to fit the actual shape of the earth the best. It is given by major semi-axis a = 6 378 137 m and flattening

$$f = \frac{a-b}{a} \tag{2}$$

equal approximately to 1/298.25. It is now used almost universally in geodetic works.

A practical realization of GRS 80 is the well known *World Geodetic System of 1984* (WGS 84) [Defense Mapping Agency, 1987]. Points can be positioned directly in this coordinate system either through orbits of positioning satellite systems (Transit or GPS) or by positioning relative to some already existing points. The North American Datum of 1983 (NAD 83) is another, continent-wide realization of GRS 80.

Geodetic Coordinate Systems, Their Positioning and Orientation

Before the advent of positioning satellite systems, it had not been possible to realize and use geocentric coordinate systems. *Geodetic Coordinate* (G) systems with reference ellipsoids selected to fit the shape of the earth (more rigorously, the shape of the earth's gravity equipotential surfaces) the best in a regional manner, were and still are used in most countries. More than 150 of such geodetic coordinate systems exist.

The biaxial reference ellipsoid associated with a G-system the same way as the geocentric reference ellipsoid is associated with the CT-system, is called the *Geodetic Horizontal Datum* and more than 150 such datums exist. They are non-geocentric, i.e., the centres E of these ellipsoids are displaced from the earth's centre of mass C usually by hundreds of metres. As well, the axes of horizontal datums (the axes of the G-systems), are misaligned with respect to the CT-system. The misalignments, however, are small, mostly less than 1". We note that in some literature the term "geodetic horizontal datum" is used, somewhat incoherently, for the conglomerate of the reference ellipsoid and a geodetic network (see below) referred to it. This usage should not be encouraged.

As with any coordinate system, to be of use in positioning a G-system, and thus even its Geodetic Horizontal Datum, must be positioned with respect to the earth. With geocentric positioning not available, the only way of positioning and orienting horizontal datums had been to do it with respect to the Local *Astronomical Coordinate System* (LA) of a selected point. (The LA-system is defined by local gravity vertical and the spin axis of the earth — see Vaníč ek and Krakiwsky [1986].) Six defining parameters had to be accepted at the *Initial Point*, also called the *Datum Point*. These are: geodetic latitude, geodetic longitude, geoidal height, two components of the deflection of the vertical, and the geodetic azimuth of a line originating at this point — see Vaníč ek and Krakiwsky [1986].

When satellite positioning (with respect to the CT-system) became available, it became possible to position and orient any datum with respect to the CT-system indirectly, by comparing coordinates of the same points in the G- and the CT-systems. This technique has been used to realize the CT-system as we discussed it earlier, as NAD 83. This is also the only technique we are and will be forced to use in the future when we wish to position and orient an existing horizontal datum with respect to the CT-system. A direct conversion of the 6 defining parameters into position and orientation parameters with (respect to the CT-system) is not possible; the vital link, position and orientation of the LA-system at the Initial Point with respect to the CT-system, is missing. Before we turn to the transformation between the G- and the CT-systems, let us discuss some aspects of position determination on a horizontal datum. These aspects will be of importance in the transformation.

Geodetic Horizontal Positions

Positions of various points referred to geodetic horizontal datums have been collected by geodesists and surveyors for well over a hundred years. The original use for these points was in mapping. These positions, specified by f and l on the particular datum used, are essentially horizontal positions even though heights h above the datum are known for many of the points.

Horizontal positions have been and are being obtained in a "differential" manner, also called relative positioning, by terrestrial and spatial means. Terrestrial means consist of measuring distances, angles, and azimuths. Spatial means consist of satellite- or VLBI-determined relative positions. Thus one point (and a direction), or several points, must be known by their positions to start the process of positioning. We can now see that a datum for horizontal positioning is of no practical use unless at least one position (and a direction) on it is known beforehand, to which the relative positions are tied. In the classical case, the initial point (and azimuth at that point) serves this purpose. In geodetic practice, relative horizontal positioning is applied repeatedly, resulting in *networks of points* whose horizontal positions are determined simultaneously in one adjustment process.

The height h of the network points is usually not known to the same degree of accuracy as the horizontal position because it is needed only for computing corrections needed to "reduce the observations" from the earth's surface, where they are made, to the reference ellipsoid, where they are needed for the position computation. It is thus normally determined to an accuracy of at least one order of magnitude lower than horizontal positions. It should not be mixed, therefore, with f and l to create a three-dimensional position unless some special provisions are made; precise geoid should be available and height differences should be measured following a scheme that would minimize the effect of vertical refraction.

Can satellite point position determination help with respect to heights? Not really. Point positions can be directly determined only in the CT-system. Besides, they are inherently less accurate than satellite-determined relative positions (also in the CT-system!) by at least one order of magnitude.

Horizontal positioning of the relative kind is, as with any measuring process, subject to errors, both random and systematic. Random errors are normally estimated during the position adjustment process, and these estimates should be available from responsible national agencies. These errors have the tendency to grow, as a function of distance from the initial point, somewhat more slowly than the square root of that distance.

Systematic errors in horizontal positions are more dangerous and more difficult to live with. They originate from systematic effects in measuring systems (e.g., a scale error that results from a miscalibration of a distance measuring device) and from model shortcomings (e.g., the neglect of geoidal heights in getting the heights h needed for observation reduction, or piecemeal adjustment of the network), cf. Vaníč ek and Krakiwsky [1986]. As random errors, network systematic errors generally also grow with distance from the initial point but may do it more rapidly.

Systematic and random errors combine to distort any horizontal network. These *network distortions* may easily reach tens of metres in older networks of continental dimensions. For instance, a commonly encountered 10 p.p.m. scale error alone distorts horizontal positions 1000 km from the initial point by 10 metres. It is thus very advisable to attempt to model these distortions — and we reiterate that these distortions represent horizontal shifts on the horizontal datum and should not be understood as affecting the third dimension h in any way — from comparisons with positions obtained some other way, e.g., from satellite positioning [Vaníč ek and Wells, 1974].

Transformation Between the G- and CT-Systems

Transformation between these two coordinate systems is the most simply expressed for Cartesian coordinates. Denoting the positions by the following vectors

$$\vec{r}^{G} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{G}, \quad \vec{r}^{CT} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{CT}, \quad (3)$$

we can write simply

$$\vec{r}^{CT} = \mathbf{R} \left(\omega_x, \omega_y, \omega_z \right) \vec{r}^G - \vec{t} \quad , \tag{4}$$

where **R** is the rotation matrix composed of the misalignment angles w_x , w_y , w_z and \vec{t} , is the translation vector of *E* from *C* [Vaníč ek and Krakiwsky, 1986]. The inverse transformation reads:

$$\vec{r}^{G} = \boldsymbol{R}^{T} \left(\boldsymbol{\omega}_{x}, \boldsymbol{\omega}_{y}, \boldsymbol{\omega}_{z} \right) \vec{r}^{CT} + \vec{t} \quad .$$

$$\tag{5}$$

If the positions are given in curvilinear coordinates, then these are first transformed to Cartesian form using eqn. (1).

We note that again 6 parameters, 3 misalignment angles, and 3 components of the translation vector \vec{t} , are needed for the transformations, just as 6 parameters are needed to fix a rigid body in space. No scale distortion should be considered part of the coordinate system transformations; it represents a systematic distortion of positions (coordinates) and not of a coordinate system. It is, however, often formally included in the transformations if it is not accounted for in modelling the network distortions.

Furthermore, if the horizontal datum had been positioned and oriented with respect to the LA-system of the initial point, then only one misalignment angle may be sought: that around the ellipsoidal normal at the initial point [Vaníč ek and Wells, 1974]. This can be seen when analysing the role of the 6 defining parameters (for positioning and orientation of the horizontal datum mentioned above): the prime vertical component of the deflection of the vertical, the latitude, and the azimuth selected have to satisfy a specific relation (Laplace equation) to ensure a proper alignment with the CT-system. Since this relation is generally not satisfied completely, a small rotation around the ellipsoidal normal results. It is normally very small, under 1" [Wells and Vaníč ek, 1975], for the well established datums.

This consideration restricts the rotation matrix in eqns. (4) and (5) as follows. Let us write the unit ellipsoidal normal vector at the initial point (f_0 , l_0) as:

$$\vec{n}_0 = \begin{bmatrix} \cos \phi_0 \cos \lambda_0 \\ \cos \phi_0 \sin \lambda_0 \\ \sin \phi_0 \end{bmatrix}$$
(6)

and the rotation matrix **R** as:

$$\boldsymbol{R}(\boldsymbol{\omega}_{x},\boldsymbol{\omega}_{y},\boldsymbol{\omega}_{z}) = \begin{bmatrix} 1 & \boldsymbol{\omega}_{z} & -\boldsymbol{\omega}_{y} \\ -\boldsymbol{\omega}_{z} & 1 & \boldsymbol{\omega}_{x} \\ \boldsymbol{\omega}_{y} & -\boldsymbol{\omega}_{x} & 1 \end{bmatrix},$$
(7)

valid for small misalignment angles. Imposing the constraint that w_x , w_y , w_z may be only components of the misalignment angle vector $\vec{\omega}_0$ given obviously by

$$\vec{\omega}_0 = \omega_0 \vec{n}_0 \quad , \tag{8}$$

where w_0 is the amount of misalignment along the ellipsoidal normal \vec{n}_0 , we get

$$\boldsymbol{R}(\boldsymbol{\omega}_{x},\boldsymbol{\omega}_{y},\boldsymbol{\omega}_{z}) = \boldsymbol{R}(\boldsymbol{\omega}_{0}) = \boldsymbol{\omega}_{0} \begin{vmatrix} 1 & \sin \phi_{0} & -\cos \phi_{0} \sin \lambda_{0} \\ -\sin \phi_{0} & 1 & \cos \phi_{0} \cos \lambda_{0} \\ \cos \phi_{0} \sin \lambda_{0} & -\cos \phi_{0} \cos \lambda_{0} & 1 \end{vmatrix}$$
(9)

Finally, eqn. (4) becomes

$$\vec{r}^{CT} = \mathbf{R}(\omega_0)\vec{r}^G - \vec{t} \tag{10}$$

and eqn. (5) becomes

$$\vec{r}^G = \mathbf{R}(-\omega_0)\vec{r}^{CT} + \vec{t} \quad . \tag{11}$$

Thus, only 4 transformation parameters are needed for transformations between any G-system (positioned and oriented at its initial point) and the CT-system.

Determination of Transformation Parameters

Ideally, the transformation parameters should be determined from the datum position and orientation defining parameters but, as stated earlier, this is not possible. Positions of some points, in both the G-system and the CT-system must then be known to determine the 4 transformation parameters. These are then obtained by solving a system of equations (10) or (11) for w_0 , t_x , t_y , and t_z .

There is a problem with this often used procedure, however. Based on our discussion of network point position accuracy, the heights of the network points should not be used in the transformation parameter determination, because the accuracy of so determined parameters may suffer significantly. Only horizontal positions $(f, l)_G$ should be used in the computation. But will this reduce the inherently three-dimensional problem of transformations to two dimensions? Of course not! The set of points used for the determination of (w_0, \vec{t}) will remain a three-dimensional configuration very similar to the original configuration where heights *h* were different from zero.

The construction of the set of transformation points given only by horizontal positions $(f, l)_G$ follows these steps:

$$\vec{r}^{G} = \begin{bmatrix} \phi \\ \lambda \\ h \end{bmatrix}_{G} \begin{bmatrix} \phi \\ \lambda \\ 0 \end{bmatrix}_{G} \begin{bmatrix} * \\ x \\ y \end{bmatrix} = \begin{bmatrix} N\cos\phi\cos\lambda \\ N\cos\phi\sin\lambda \\ e^{x} \end{bmatrix}^{*}$$
(12)

(for all the points in the set). We end up with a point set given by three-dimensional Cartesian coordinates of points that lie on the reference ellipsoid.

To arrive at a compatible set of positions in the CT-system, we have to map the threedimensional Cartesian coordinates in the CT-system also onto the ellipsoid of the same size and shape as the geodetic horizontal datum (a, b). To achieve this, we go through the following transformations:

$$\vec{r}^{CT} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{CT} \begin{bmatrix} \phi \\ \lambda \\ h \end{bmatrix}_{CT(a,b)} \begin{bmatrix} \phi \\ \lambda \\ \lambda \\ 0 \end{bmatrix}_{CT(a,b)} \begin{bmatrix} x \\ x \\ y \\ z \end{bmatrix}_{CT} = \overset{*}{\vec{r}}^{CT} , \qquad (13)$$

again for all the points in the set.

We now have two sets of compatible positions $(\vec{r}^{G}, \vec{r}^{CT})$ which theoretically differ only by the misalignment angle w_0 and translation vector \vec{t} of the geodetic horizontal datum with respect to the CT-system. These parameters (ω_0, \vec{t}) can now be finally evaluated, say from eqn. (11). We get immediately for each transformation point P_i :

where, with sufficient accuracy,

$$\vec{u}_{i} \neq R \begin{bmatrix} \sin \phi_{0} \cos \phi_{i} \sin \lambda_{i} - \cos \phi_{0} \sin \lambda_{0} \sin \phi_{i} \\ -\sin \phi_{0} \cos \phi_{i} \cos \lambda_{i} + \cos \phi_{0} \cos \lambda_{0} \sin \phi_{i} \\ \cos \phi_{0} \sin \lambda_{0} \cos \phi_{i} \cos \lambda_{i} - \cos \phi_{0} \cos \lambda_{0} \cos \phi_{i} \sin \lambda_{i} \end{bmatrix}$$
(15)

and *R* is the mean radius of the earth. Denoting now the 4-vector of unknown transformation parameters $[w_0, t_x, t_y, t_z]^T$ by $\mathbf{x}, \vec{r}_i^G - \vec{r}_i^C T$ by \mathbf{w}_i , and $[\vec{u}_i:\mathbf{I}]$ by \mathbf{A}_i , we can rewrite eqn. (14) as $\mathbf{A}_i \mathbf{x} = \mathbf{w}_i$. (16)

Each transformation point supplies one vectorial equation of the form of eqn. (16), i.e., a triplet of linear algebraic equations for 4 unknowns. To be able to solve these equations for \mathbf{x} , at least 4 coordinates of 2 points are needed. In practice, we will have several, say n, transformation points available and will be able to write 3n linear algebraic equations for the 4 unknowns. This system of 3n so called *observation equations* will then be solved by the least-squares technique.

There is yet another problem encountered here. Both sets of positions are burdened with errors. These should be accounted for, if at all possible, before the two sets of positions are used. If the network distortion has already been modelled, then it should be corrected for at this stage. If the distortion is not known, then at least the network point positions should be given weights inversely proportionate to some power (probably 2 would be a reasonable initial guess) of distance from the initial point. Only *f* and *l* should be considered as having errors associated with them and these errors should then be rigorously propagated into the covariance matrix of $\frac{*}{r}G$ — cf. Vaníč ek and Krakiwsky [1986].

Similarly, the positions in the CT-system, determined typically using a satellite positioning system, will have some error estimates associated with them. These should be converted into proper weights as well and used in the estimation of (ω_0, \vec{t}) .

The last issue we wish to consider under this heading is the geometrical stability of the solution (ω_0, \vec{t}) . The solution is clearly sensitive to the configuration of the transformation points:

(i) on the one hand, there exist configurations that cannot be used for estimating (ω_0, \vec{t}) because the matrix of normal equations (of the least-squares process) they produce is singular, i.e., there is a total correlation between w_0 and \vec{t} ;

(ii) on the other hand, if the area covered by the transformation points is large, the solution will be stable, i.e., there will be a minimal correlation between w_0 and \vec{t} .

To judge the quality of the least-squares estimate of (ω_0, \vec{t}) , the covariance matrix of the estimate should always be produced. This covariance matrix will allow us later to estimate the accuracy of transformed positions from the G-system into the CT-system and vice versa.

The Case Involving Two Datums

The delineation of a marine boundary between two countries on different geodetic datums has to deal with two separate transformations such as described in the previous paragraph. The transformation of positions \vec{r}^{G_1} of base points on the first datum to positions \vec{r}^{G_2} referred to the second datum has to go through the CT-system. We thus need two sets of transformation parameters, from G₁ to CT and from G₂ to CT, including their errors, and two network distortion models.

The transformation $\vec{r}^{G_1} \rightarrow \vec{r}^{G_2}$ should be carried out in the following steps:

(1) The modelled distortions of the network referred to G_1 should be subtracted from the distorted positions \vec{r}^{G_1} of base points to give undistorted positions $(f, l)_{G_1}$. (This step will not be applicable if network distortion is not known.)

(2) Horizontal undistorted positions $(f, l)_{G_1}$ are then transformed to Cartesian coordinates in the G₁-system, $\overset{*}{\vec{r}}^{G_1}$ using eqn. (12), including their covariance matrices.

(3) Cartesian coordinates \vec{r}^{G_1} in the G₁-system are transformed into the CT-system (eqn. (10)) using the first set of transformation parameters $(\omega_0, \vec{t})_1$ and their covariance matrix as well as the covariance matrix of \vec{r}^{G_1} .

(4) If the two involved geodetic datums have different shapes and sizes, $(a, b)_1 \neq (a, b)_2$, then the Cartesian coordinates \vec{r}^{CT} (and their covariance matrices) must be transformed onto the second ellipsoid $(a, b)_2$ by the following two transformations:

$$\overset{*}{\vec{r}}^{CT} \xrightarrow{\left[\begin{array}{c} \phi \end{array} \right]} \\ \rightarrow \left[\begin{array}{c} \lambda \\ \lambda \end{array} \right] \\ \left[\begin{array}{c} h \end{array} \right]_{G_2} \end{array}$$
 (17)

and

$$\begin{bmatrix} \phi \\ \vec{r} & \lambda \\ 0 \end{bmatrix}_{G_2}^{**CT} .$$

$$(18)$$

If the two datums have the same shape and size, then $\vec{r}^{CT} = \vec{r}^{CT}$ and no such transformation is required.

(5) Cartesian coordinates \vec{r}^{CT} and their covariance matrices are then transformed into the second geodetic coordinate system G₂ (eqn. (11)) using the second set of transformation parameters $(\omega_0, \vec{t})_2$, taking into account their covariance matrix.

(6) Cartesian coordinates \vec{r}^{G_2} and their covariance matrix are now transformed to $(f, l)_{G_2}$ using the inverse of eqn. (1). We note that the resulting height h_{G_2} should automatically equal to zero.

(7) Finally, the modelled distortions of the network referred to G_2 should be added to the transformed (undistorted) positions $(f, l)_{G_2}$ to give distorted positions compatible with the positions of the second country's base points referred to G_2 to begin with. (Again, this step will not be applicable if network distortion is not known.)

It must be realized that the last step requires the knowledge of the second network's distortion beyond the extent of the network itself. This is a tricky problem and requires a very careful handling. Thoughtless extension of an existing distortion model may lead to very significant errors.

On the other hand, if we neglect to model network distortions in the first, second, or both networks, we will end up with very large estimated errors for the base point positions and with very large errors for the boundary point positions.

Conclusions

As we have tried to point out, the determination of a maritime boundary based on two sets of base points referred to two different geodetic horizontal datums may lead to very large errors. This is due to:

(i) random errors in both sets of base point positions;

(ii) systematic errors in both networks to which the base points are tied;

(iii) errors in transformation parameters for both horizontal datums;

(iv) the necessity to consider position distortions beyond the extent of known positions.

Failure to implement the proper algorithm for transformation parameter estimation, as described here, may and very probably will, result in further significant deterioration of accuracy.

It should be pointed out that these problems will not be alleviated by positioning the base points of each of the countries by relative satellite methods with respect to their respective networks. Only simultaneous relative positioning of both sets of base points with respect to both networks will improve the accuracy. But this approach calls for a separate analysis which we consider beyond the scope of this contribution.

Acknowledgments

The author wishes to acknowledge that he was Humboldt Senior Distinguished Scientist at the Geodetic Institute of Stuttgart University, Germany, while preparing this paper. Many thanks go to Prof. E. Grafarend and his staff for having made the author's stay in Stuttgart a pleasant and productive one.

References

International Association of Geodesy (1980). The geodesist's handbook. Ed. I.I. Mueller, Bulletin Géodésique, Vol. 54, No. 3.

- United States Defence Mapping Agency (1987). Supplement to Department of Defence World Geodetic System 1984 technical report. Defence Mapping Agency Technical Report No. 8350.2-A, Washington, D.C..
- Vaníček, P., and E.J. Krakiwsky (1986). *Geodesy: The Concepts.* 2nd ed., North Holland, Amsterdam,.
- Vaníček, P., and D.E. Wells (1974). Positioning of horizontal geodetic datums. *Canadian Surveyor*, Vol. 28, No. 5, pp. 531-538.
- Wells, D.E., and P. Vaníč ek (1975). Alignment of geodetic and satellite coordinate systems to the average terrestrial system. *Bulletin Géodésique*, No. 117, pp. 241-257.