Transformation of coordinates between

two horizontal geodetic datums

by

Petr Vaní□ek

Department of Geodesy and Geomatics Engineering, University of New Brunswick,

PO Box 4400, Fredericton, NB, Canada, E3B 5A3

and

Robin R. Steeves

Microsearch Corporation

PO Box 59, Juniper, NB, Canada, E0J 1P0

Abstract. The following topics are discussed in this paper: 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.

It is often found necessary to transform positions from one horizontal datum to another. 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 these transformations together 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 by convention. There are infinitely many such coordinate systems differing from each other by the orientation of their axes.

The most common geocentric system used in geodesy is the *Conventional Terrestrial* (CT) system which is oriented so that the 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 [Vaní□ek and Krakiwsky, 1986]. This system provides the most convenient link with geodetic astronomy.

Two other 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 respectively.

Positions in the CT-system are sometimes given in Cartesian coordinates (x,y,z), and sometimes in curvilinear geodetic coordinates (ϕ, λ, 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 major semi-axis *a* and minor semi-axis *b*. The geodetic height *h*, sometimes called the ellipsoidal height, is the height of a point above the surface of that reference ellipsoid. In some literature, the biaxial reference ellipsoid is called the reference spheroid.

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

, (1)

where *N* denotes the radius of curvature of the ellipsoid in the direction perpendicular to the meridian plane [Vaní□ek and Krakiwsky, 1986]. The inverse transformation has to be solved either iteratively, through a linearization, using an algebraic equation of fourth degree [Vaní□ek and Krakiwsky, 1986], or some other techniques.

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 closely fit the actual shape of the Earth. It is given by major-semi axis *a* = 6 378 137 m and flattening

(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* (G) coordinate systems [Vaní□ek and Krakiwsky, 1986], 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 G-systems exist.

A biaxial reference ellipsoid associated with a G-system (in the same way that the geocentric reference ellipsoid is associated with the CT-system) is called a *Geodetic Horizontal Datum.* A geodetic horizontal datum is non-geocentric, i.e., the centre *E* of the ellipsoid is displaced from the Earth's centre of mass *C,* usually by hundreds of metres. As well, the axes of a geodetic horizontal datum (the axes of the G-system), are misaligned with respect to the CT-system. The misalignments, however, are small, mostly less than one second.

We note that the term "geodetic horizontal datum" is sometimes used for the conglomerate of the reference ellipsoid and a geodetic network (see below) referred to it. This usage can be misleading and should not be encouraged.

To be of practical use in positioning, a G-system, and thus even its geodetic horizontal datum, must be positioned with respect to the Earth. Prior to the advent of satellite positioning, when geocentric positioning was not possible, the only way of positioning and orientating horizontal datums had been to do it with respect to the *Local Astronomical* (LA) coordinate system of a selected point. (The LA-system is defined by the local gravity vertical and the spin axis of the Earth [Vaní□ek and Krakiwsky, 1986].) Six defining parameters had to be chosen at the *Initial Point (*also called the *Datum Point)*: 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 [Vaní□ek and Krakiwsky, 1986].

In the case of some prior information being available, the values of the geodetic latitude and longitude of the Initial Point were chosen so as to achieve the best fit between astronomic and geodetic latitudes and longitudes at a number of deflection points. Such was the case with NAD 27 and ED 1950. This approach was designed to ensure that the reference ellipsoid would fit the geoid in the region of interest. As the adjustment was carried out in two dimensions, the result was equivalent to the case described above except that the defining parameters would implicitely refer to another point, that point being the centroid of the deflection points used.

With the advent of satellite positioning (with respect to the CT-system), it became possible to indirectly position and orient a datum with respect to the CT-system by comparing coordinates of the same points in the G-system and the CT-system. This technique has been used, for example, to realize the CT-system as NAD 83, and is, in fact, the only technique available to position and orientate an existing horizontal datum with respect to the CT-system. A direct conversion of the six defining parameters into position and orientation parameters (with respect to the CT-system) is not possible; the vital link, the position and orientation of the LA-system at the Initial Point with respect to the CT-system, is missing. We know only the directions of the LA-axes in the G-system (through Φ, Λ, and $A-\alpha$) while the geocentric position of the Initial Point is not known at all.

Before we turn to the transformation between the G-system and the CT-system, 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 (coordinates) of various points referred to a geodetic horizontal datum have been determined by geodesists and surveyors for over several hundred years. The original use of these points was in mapping, and their positions, specified by latitude ϕ and longitude λ on the particular datum used, are essentially horizontal positions even though heights h above the datum might have been known for some of the points. Let us discuss this statement in some detail.

Horizontal positions have been, and are being, determined by relative positioning (determining coordinates of new points by making measurements from points with known coordinates). Relative positioning by terrestrial means consists of measuring distances, angles, and azimuths. Relative positioning by spatial means consists of making satellite or VLBI measurements. Thus the coordinates of one point and the direction from one point to another, or the coordinates of several points, must be known to start the process of positioning. A horizontal datum is therefore of no practical use unless at least one position and a direction is known beforehand, to which the relative positions may be tied. In the classical case, the initial point and the azimuth from that point to another point serve this purpose. In practice, relative horizontal positioning is applied repeatedly, resulting in *networks of points* whose horizontal positions are determined simultaneously in one adjustment computation.

The height *h* of a "horizontal" network point is normally determined (e.g. by measuring vertical angles) to an accuracy suitable for computing corrections needed to reduce terrestrial observations from the Earth's surface, where they are made, to the reference ellipsoid, where they are needed for position computations. Because levelling is not normally used to determine orthometric heights H of network points, and geoidal heights N were known to a relatively poor accuracy, the geodetic heights were usually determined to an accuracy of about ten times lower than horizontal positions, and should therefore not be mixed with latitude and longitude to create a three-dimensional position. Such "heights" h are not intended to be used as coordinates, and if they are combined with accurately determined latitudes and longitudes to form three-dimensional positions, unwanted distortions will be introduced into the network.

Note that satellite point positioning cannot help us improve the accuracy of heights that were determined for the establishment of horizontal networks. Satellite positioning can only produce heights that are relative to the CT-system and, without knowing how to transform CT-system coordinates to G-system coordinates we cannot use them in the G-system. Of course, once we know how to transform between the CT-system and the G-system, the situation changes. But as this determination of the parameters needed for the transformation is our objective here, we must consider the transformation to be unknown at this stage. Even if they can be used, satellite-determined heights are inherently at most half as accurate as satellite-determined relative horizontal positions (also in the CT-system).

Relative horizontal positioning is, as with any measuring process, subject to errors, both random and systematic. Random errors are normally estimated during the adjustment computations, and these estimates should be available from responsible national agencies. Random 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. If a set of deflection points has been used instead of the initial point, the random errors are more evenly distributed in the network.

Systematic errors in horizontal positions are more dangerous and more difficult to handle. 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 obtaining heights needed for observation reduction, or piecemeal adjustment of the network) [Vaní□ek and Krakiwsky, 1986]. As with random errors, network systematic errors generally grow with distance from the initial point but usually do it more rapidly.

Systematic and random errors combine to distort all horizontal networks. These *network distortions* may easily reach tens of metres in older networks of continental dimensions. See, for example, [Ehrnsperger, 1991]. For instance, in older networks, a commonly encountered 10 p.p.m. scale error alone distorts horizontal positions that are 1000 km from the initial point by 10 metres. It is thus very advisable to attempt to model these distortions. We emphasize that these distortions represent horizontal shifts on the horizontal datum and should not be understood as affecting the third dimension *h* in any way.

Transformation Between a G-System and the CT-System

In theory, transformation between these two coordinate systems is most simply expressed in Cartesian coordinates. Denoting the positions by the following vectors

, and , (3)

we can write simply

, (4)

where **R** is the rotation matrix composed of the misalignment angles ωx, ωy, ωz and

is the translation vector of *E* from *C* [Vaní□ek and Krakiwsky, 1986]. The inverse transformation reads:

,

or, with sufficient precision in the case of small rotations,

. (5)

If the positions are given in curvilinear coordinates, they are first transformed to Cartesian form using equation (1).

We note again that six parameters (in this case three misalignment angles and three components of the translation vector **t**) are needed for the transformations, just as six parameters are needed to fix a rigid body in three dimensional space. No scale distortion should be considered part of the coordinate system transformations because a scale difference 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 network distortions.

Furthermore, if the horizontal datum had been positioned and oriented with respect to the LA-system of the initial point (or the centroid), then only one misalignment angle may be sought: that around the ellipsoidal normal at the initial point (centroid) [Vaní□ek and Wells, 1974]. This can be seen when analysing the role of the six 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 (the Laplace equation) to ensure a proper alignment with the CT-system. Because this relation is generally not satisfied completely, a small rotation around the ellipsoidal normal remains (see Figure 1). It is normally very small, under one second for well established datums [Wells and Vaní□ek, 1975].

We note that additional Laplace points supply additional geodetic azimuths in the network adjustment. These contribute individually to local changes in the shape of the network and collectively to the network orientation at the initial point or centroid. The additional Laplace points thus do not affect our argument here.

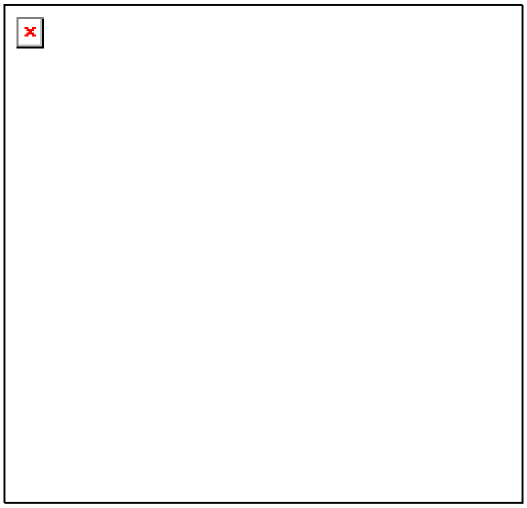


Fig. 1. Misalignment of a horizontal datum.

This consideration restricts the rotation matrix in equations (4) and (5) as follows. Let us write the unit ellipsoidal normal vector at the initial point ($\phi 0$, $\lambda 0$) as:

(6)

and the rotation matrix **R** as:

, (7)

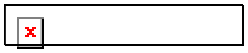
which is valid for small misalignment angles. Imposing the constraint that ωx, ωy, ωz may only be components of the datum misalignment angle vector given by

, (8)

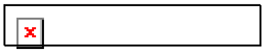
where is the amount of misalignment around the ellipsoidal normal , we get [Wells and Vaní□ek, 1975]

. (9)

Finally, equation (4) becomes

 (10)

and equation (5) becomes

. (11)

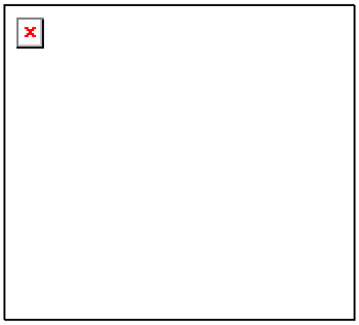
Thus, only four transformation parameters, rather than the currently used 6 parameters, are needed for transformations between any G-system and the CT-system if the G-system had been positioned and oriented at its initial point.

Determination of Transformation Parameters

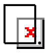
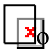
Ideally, the transformation parameters should be determined from the datum position and orientation parameters, but, as stated earlier, this is not possible. Positions of some points, in both the G-system and the CT-system must therefore be known to determine the four transformation parameters ($\omega 0$, tx, ty, and tz) by solving a system of equations based on either of equations (10) or (11).

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 would suffer significantly. Only positions $(\phi, \lambda, 0)_G$ on the surface of the reference ellipsoid should be used in the computation. The inherently three-dimensional character of our problem is thereby retained, without being adversely affected by inaccurate heights.

To construct our system of equations based on either equation (10) or (11), we first select a set of network points for which we have both geodetic coordinates $(\phi, \lambda)_G$ and CT coordinates $(x,y,z)_{CT}$. For each point, we first transform the geodetic datum coordinates as follows:

 (12)

which provides us with a set of three-dimensional Cartesian coordinates for points that lie on (the surface of) the reference ellipsoid. Note that the ellipsoid in this case is our non-geocentric horizontal datum.

We must now transform our CT coordinates for each point so that they are compatible with the coordinates . To be compatible, the transformed CT coordinates must be on the surface of an ellipsoid with the same size and shape as the geodetic horizontal datum (i.e., with semi-axes a and b). In other words, we want the transformed CT coordinates to differ from  only by the four transformation parameters that we wish to determine. We transform the CT coordinates as follows:

, (13)

again for all the points in the set. While the 3D positions and refer to the same point, the positions and generally do not. Associating with , after the appropriate transformation parameters are applied, will result in diffenences of horizontal positions on the reference ellipsoid in the order of at most , i.e., 1.5 cm for h = 1000 m and |t| = 100 m. If either h or |t| are large, these differences may be significant and the evaluation of would have to be iterated.

We now have coordinates and for each network point . Substituting these in equation (11) results in:

.

Rearranging and substituting previous results, we can write

. (14)

Now, substituting for an expression parallel to (12) and replacing N by R and b/a by 1, we get with sufficient accuracy,

, (15)

R being the mean radius of the Earth. Equation (14) may be written as

, (16)

where

, ,

and .

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

We note that in the least squares adjustment outlined above, the vector is used as a triplet of observations. These observations have errors associated with them that must be modelled. Note that equation (16) does not model any systematic errors that these observations may contain, and therefore such systematic errors (distortions) must be removed before using them to determine in our least squares adjustment. This can be done by means of a number of techniques such as described by [Andersson and Poder, 1981; Junkins, 1991]. If the distortions are not known, then the network coordinates should, at least, be assigned appropriate weights that are inversely proportional to some function of the distance from the initial point. Random errors in can be modelled with an appropriate covariance matrix that is derived from the covariance matrices of the $(\phi, \lambda)_G$ and $(x,y,z)_{CT}$ coordinates.

The stability of the least squares adjustment outlined above is clearly a function of the geometrical configuration of the transformation points. If the area covered by the transformation points is large, the solution will generally be stable. The covariance matrix of the adjusted parameters should always be computed so that the quality of the determined parameters may be analysed. Furthermore, the observation residuals resulting from the adjustment should always be examined for outliers. Large residuals would indicate that our observations contain distortions that have not been appropriately modelled.

The Case Involving Two Datums

Sometimes, geodetic work is done in an area referred to two geodetic datums. The transformation of positions referred to the first datum to positions referred to the second datum has to go through the CT-system. We thus need two sets of transformation parameters: one set to transform from G_1 to CT, and another to transform from G_2 to CT, including their covariance matrices, and the two network distortion models. The transformation should be carried out in the following steps:

1. The modelled systematic distortions of the network referred to G_1 should be subtracted from the distorted positions of base points to give undistorted positions (ϕ, λ)G1. This step will not be applicable if the network distortions are not known.

1. Horizontal undistorted (corrected for distortions) positions (ϕ, λ)G1 are then transformed to Cartesian coordinates (for h = 0) in the G_1-system, using equation (12), including their covariance matrices.

1. Cartesian coordinates in the G1-system are transformed into the CT-system using equation (10) with the first set of transformation parameters and their covariance matrix as well as the covariance matrix of .

1. If the two involved geodetic datums have different shapes and sizes, $(a, b)1$ $(a, b)2$, then the Cartesian coordinates (and their covariance matrices) must be transformed onto the second ellipsoid $(a, b)2$ by the following transformations:

1. (17)

1. If the two datums have the same size and shape, then and no such transformation is required.

1. Cartesian coordinates and their covariance matrices are then transformed into the second geodetic coordinate system G2 using equation (11) with the second set of transformation parameters , taking into account their covariance matrix.

1. Cartesian coordinates and their covariance matrix are now transformed to (ϕ, λ)G2 using the inverse of equation (1). We note that the resulting height should automatically equal to zero.

1. Finally, the modelled distortions of the network referred to G2 should be added to the transformed (undistorted) positions (ϕ, λ)G2 to give distorted positions compatible with the positions of the points referred to G2 to begin with. Again, this step will not be applicable if network distortion is not known.

We note that 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 point positions.

Conclusions

Procedures for transforming coordinates between a horizontal geodetic datum and the CT-system, and for transforming coordinates from one geodetic datum to another were given. The importance of clearly separating these transformations (between coordinate systems) from the treatment of systematic and random errors in network coordinates was pointed out.

Acknowledgments

The first 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.

The authors wish to thank Dr. Richard Langley, Prof. C. C. Tscherning, and two anonymous reviewers for their critical review of this paper.

References

Andersson, O. and K. Poder (1981). Koordinattransformationer ved Geodaetisk Institut. Landinspektoeren, Vol. 30, pp. 552-571.

Ehrnsperger, W. (1991). The ED 87 Adjustment. *Bulletin Géodésique*, Vol. 65, pp. 28-43.

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

Junkins, D. (1991). The National Transformation for Converting Between NAD27 and NAD83 in Canada, in *Moving to NAD83* (ed. D. C. Barnes), CISM, Ottawa, pp. 16-40.

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..

Vaní□ek, P., and E.J. Krakiwsky (1986). *Geodesy: The Concepts.* 2nd ed., North Holland, Amsterdam,.

Vaní□ek, 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. Vaní□ek (1975). Alignment of geodetic and satellite coordinate systems to the average terrestrial system. *Bulletin Géodésique*, No. 117, pp. 241-257.