EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

MAP GRIDS AND DATUMS

GÁBOR TIMÁR
GÁBOR MOLNÁR

Map grids and datums

Gábor Timár
Gábor Molnár

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Map grids and datums

by Gábor Timár and Gábor Molnár
Reviewer:
György Busics
Copyright © 2013 Eötvös Lóránd University

This book is freely available for research and educational purposes. Reproduction in any form is prohibited without written permission of the owner.

Made in the project entitled "E-learning scientific content development in ELTE TTK" with number TÁMOP-4.1.2.A/1-11/1-2011-0073. Consortium leader: Eötvös Loránd University, Consortium Members: ELTE Faculties of Science Student Foundation, ITStudy Hungary Ltd.

National Development Agency
www.ujszechenyiterv.gov.hu
06 40 638 638

HUNGARY'S RENEWAL

NEW SZÉCHENYI PLAN

The project is supported by the European Union and co-financed by the European Social Fund.

ISBN 978-963-284-389-6

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Table of Contents

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 1. Introduction

Working with Geographic Information Systems (GIS), geo-reference is a methodology to

- give the coordinates of all objects of the system and
- define the coordinate system of these coordinates.

Naturally, as the coordinate systems can be of several kinds, the transformation methods between these coordinates are also a part of this field. The objects can be of vector or raster types; in the first case, the coordinates of the vertices should be given. Working with raster datasets, the coordinates of every pixel should be defined.

Fig. 1. The map of Hungary of Goetz & Probst from 1804 as a Google Earth layer: integration of completely different data technologies by the geo-reference.

The first sentence of the above paragraph is very similar to the basic exercise of the surveying. However, the GIS application supposes that the field survey has been completed, so the geo-reference is – with a very few exceptions – mostly office, computer-aided work. Besides, as it will be detailed later, the accuracy claims are often different – less – than the needs in the classical geodesy. Perhaps this is the reason, why the developing of these methods handled less important by the geodesists, albeit the methods are well known for them. However, in the GIS, the coordinate handling and conversion methods are highly needed, even if their accuracy is around one meter or even a few meters. Therefore these methods are less introduced in the literature.

The geo-reference is a crucial part of the GIS: it is the key of the uniform handling of many different input data; the key of the spatial data integration (Fig. 1). Every GIS user has already faced this problem, if his data was not in just one spatial coordinate system. I hope this book can be helpful in solving these problems correctly and exercises with the desired accuracy.

It is necessary to give here, in the introduction, the definition of the accuracy in the geo-information. It is a relative subject; in the everyday GPS practice it is mostly the one meter-few meters error, that is an acceptable level. While we work with scanned maps, it should be known that during the map making and printing process, the post-printing drying and the final scanning, the best accuracy could be around half a millimeter in the map. That's why, in this case, the aimed accuracy of the applied methods is a function of the scale of the scanned map: at 1:10000 scale, it is 5 meters while if the map has a scale of 1:50000, it is enough to apply methods with an accuracy limit of 25

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

meters. In most cases, it is not only unnecessary to apply better methods as they are less cost-effective: the input data are burdened by higher errors than our precious method is optimized for.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 2. Planar and spatial coordinate systems

2.1 Units used in geodetic coordinate systems

It is an old tradition that in our maps the angles can be read in the degree-minute-second system. The whole circle is 360 degrees, a degree can be divided into 60 minutes, a minute can be further divide into 60 seconds, so a degree consists of 3600 seconds.

Along the meridians, the physical distances connected to the angular units – supposing the Earth as a sphere – are practically equal. Along a meridian, and using the first definition of the meter, one degree distance is 40,000 km / 360 degrees = 111.111 kilometers. One second along the meridian is a 3600th part of this distance, 30.86 meters; this is the distance between two parallels, one second from each other. Along the parallels, the similar distance is also a function of the latitude and the above figures should be divided (in case of spherical Earth) by the cosine of the latitude. At the latitude of Budapest (latitude: 47.5 degrees), a longitudinal degree is 75,208 meters, a longitudinal second is 20.89 meters.

However, the degree-minute-second system is not the exclusive one. In the maps of France and the former French colonies, e.g. of Lebanon, the system of new degrees (gons or grads) is often used (Fig. 2). A full circle is 400 new degrees. One new degree consists of 100 new minutes or 10,000 new seconds.

In many cases, the GIS software packages ask some projection parameters or other coordinates in radians. Radian is also the default angular unit of the Microsoft Excel software. The full circle is, by definition, 2π radians, so one radian is approximately 57.3 degrees and one radian is 206264.806 arc seconds (this is the so called σ").

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

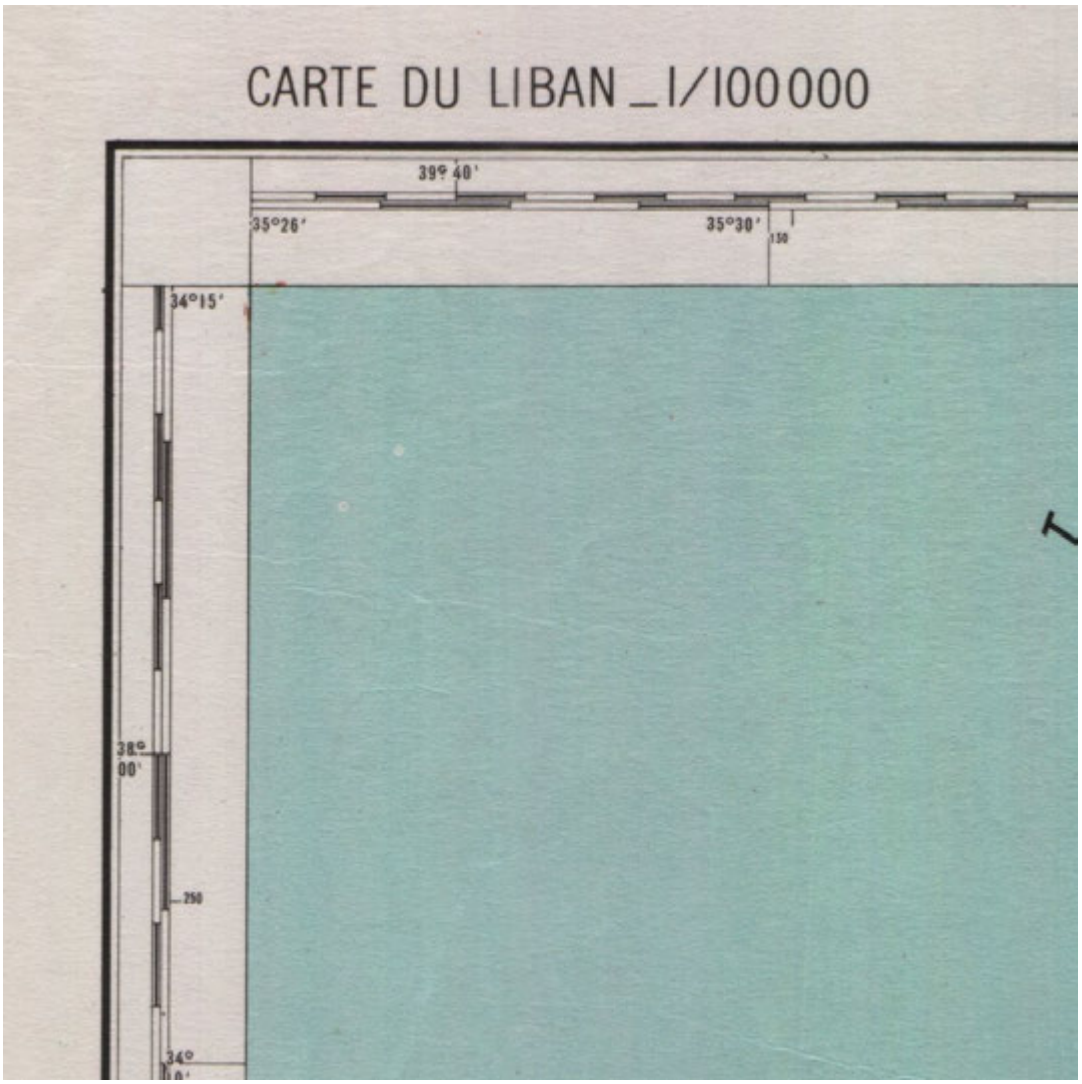


Fig. 2. In the map of Leban, a former French colony, the latitudes and longitudes are given in degrees (internal frame) and also in grads (indicated by 'G' in the external frame).

The standard international length unit is the meter. In the history, it had three different definitions. After the first one, both newer descriptions made it more accurate, keeping the former measurements practically untouched. First, the meter was introduced as the one ten millionth part of the meridian length between the pole and the equator. As this definition was far too abstract for everyday use, later a metric etalon was produced and stored in France as the physical representation of the unit. The countries have replicas of it and maintain their own national systems to calibrate all local replicas to the national ones. Nowadays, the new definition of the unit is based on quantum-physical constants that are as far from the everyday use as the first definition is. However, as it is calibrated exactly in the GPS system, it is more and more a part of our everyday life.

Using the replica system was not without side effects. During 1870s, in the newly conquered Alsace and Lorraine, the Germans connected the geodetic networks of Prussia and France. The fitting of the two systems showed an error around ten meters. Later it occurred, that French and Prussian networks was constructed using different meter replicas as scale etalons at the baselines. The length of the German metric etalon (brought also in Paris) was longer by 13.55 microns than the original French one. This makes no problem in the most cases, but in long distances, it counts: in a distance of several hundred kilometers, the error of ten meters occurs easily. The length of the German replica was later the definition of the 'legal meter', which is 1.00001355 'international' meters. There is an ellipsoid (see point 3.2), called 'Bessel-1841-Namibia, used for the German survey of southwestern Africa (Namibia); its semi-major axis is the one of the Bessel-1841 ellipsoid multiplied by this counting number between the meter and the legal meter. Thus, the legal meter is also known as 'Namibia-meter'.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

In the Anglo-Saxon cartography, different length units are also used. In the former Austro-Hungarian Monarchy, the basic unit was the ’Viennese fathom’ (Wiener Klafter). Table 1 shows the length of these units in meters.

Lenght unit	In meters
Legal meter	1.0000135965
Viennese fathom	1.89648384
Viennese mile	7585.93536
Toise	1.94906
Imperial foot	0.3047972619
US Survey foot	0.30480060966
Sazhen (Russian fathom)	2.1336
Russian Verst	1066.78

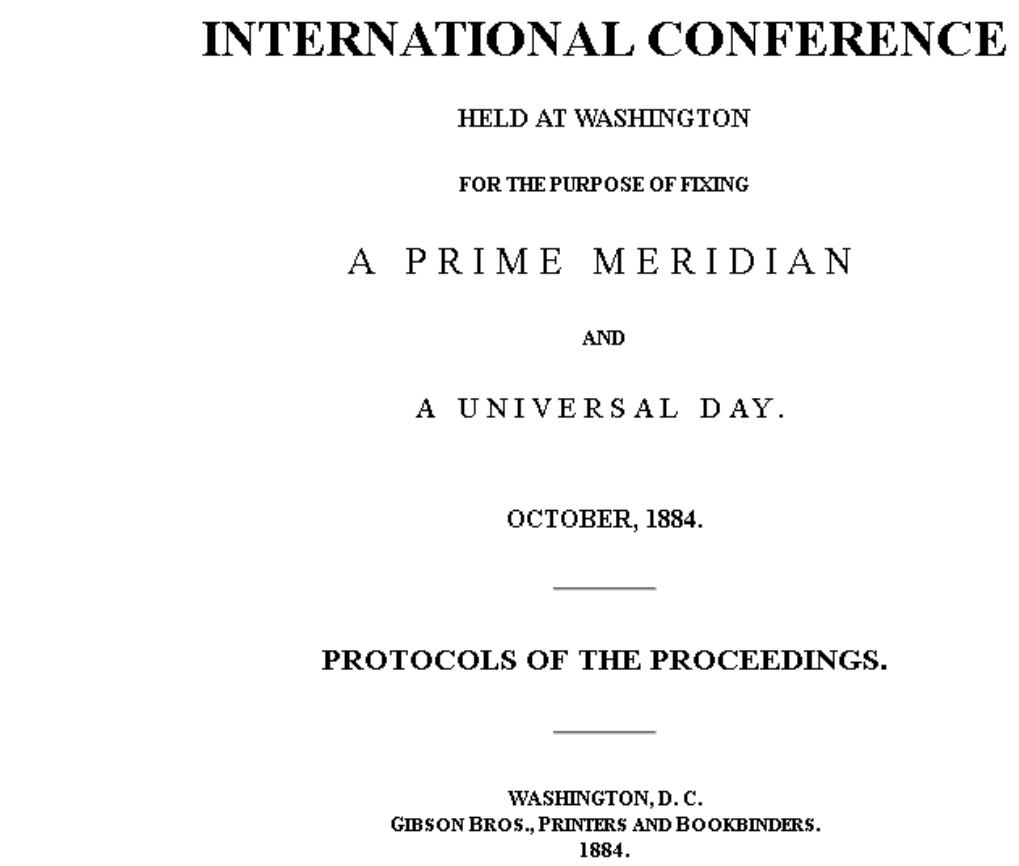
Table 1. Historical and imperial/US units in meters.

2.2 Prime meridians

There is a natural origin in the latitudes: the position of the rotation axis of the Earth provides the natural zero to start counting the latitudes from: the equator. However, in case of the longitudes, the cylindrical symmetry of the system does not offer a similar natural starting meridian therefore we have to define one.

The meridian of the fundamental point of a triangulation network (see Point 3.3) is usually selected as zero or prime meridian. Ellipsoidal longitudes of all points in the network are given according to this value. If we’d have just one system, it could work well. As we have several different networks and different prime meridians, we need to know the angular differences between them. Instead of handling the differences between all prime meridian pairs, it is worth to choose just one, and all of the others can be described by the longitude difference between it and the chosen meridian.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/



INTERNATIONAL CONFERENCE

HELD AT WASHINGTON

FOR THE PURPOSE OF FIXING

A PRIME MERIDIAN

AND

A UNIVERSAL DAY.

OCTOBER, 1884.

PROTOCOLS OF THE PROCEEDINGS.

WASHINGTON, D. C.
GIBSON BROS., PRINTERS AND BOOKBINDERS.
1884.

Fig. 3. The cover page of the protocol of the 1884 Washington conference that decided to use Greenwich as the international prime meridian.

The use the Greenwich prime meridian was proposed by the 1884 Washington Conference on the Prime Meridian and the Universal Day (Fig. 3). It was accepted by 22 votes, while Haiti (that time: Santo Domingo) voted against, France and Brazil abstained. France adapted officially the Greenwich prime meridian only in 1911, and even nowadays, in many French maps, we can find longitude references from Paris and in new degrees. It is interesting that the question of the international prime meridian was discussed in that time: the newly invented telegraph enabled to accomplish the really simultaneous astronomical observations at distant observatories. Table 2 shows the longitude difference between Greenwich and some other important meridians that were used as local or regional zero meridians.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Prime meridian	Longitude from Greenwich
Paris	2° 20’ 14,025”
Rome	12° 27’ 8,04”
Madrid	–3° 41’ 16,48”
Oslo	10° 43’ 22,5”
Pulkovo	30° 19’ 42,09”
Ferro[1]	–17° 40’
Ferro[2]	–17° 39’ 46,02”
Ferro[3]	–17° 39’ 45,975”
Vienna, Stephansdom[4]	34° 02’ 15” (from Ferro)
Vienna, Stephansdom[5]	16° 22’ 29”
Budapest, Gellérthegy[6]	36° 42’ 51,57” (from Ferro)
Budapest, Gellérthegy[7]	36° 42’ 53,5733” (from Ferro)
Budapest, Gellérthegy[8]	19° 03’ 07,5533”

Table 2. Longitude values of some prime meridians. [1]Used in Germany, Austria and Czechoslovakia. [2]The’Albrecht difference’, used in Hungary, Yugoslavia and in the Habsburg Empire. [3]According to the Bureau International de l’Heure. [4]From Ferro, in the system 1806. [5]Applying the Albrecht difference. [6]From Ferro, according to the 1821 triangulation. [7]From Ferró, according to the system1909. [8]The 1909 value, applying the Albrecht difference.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

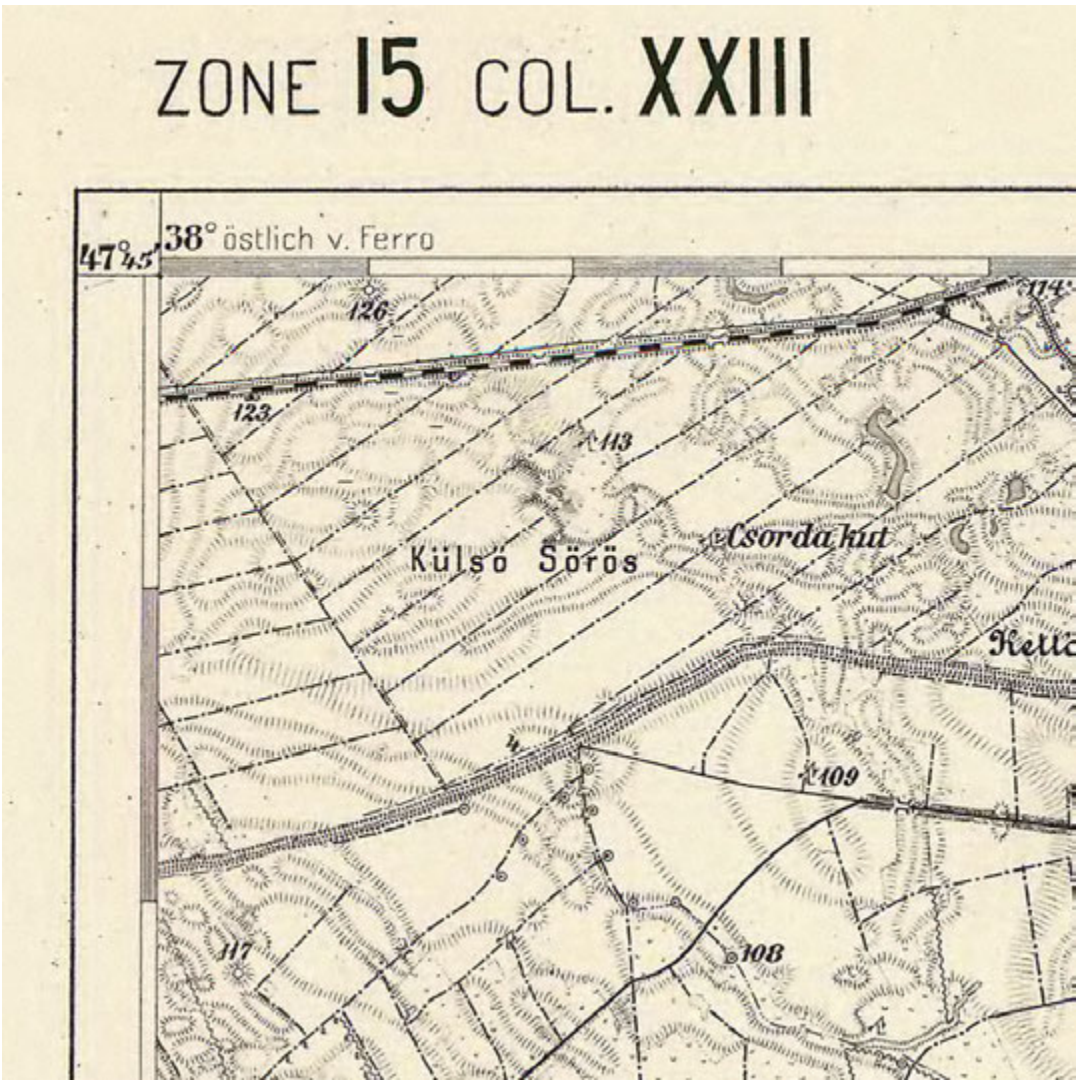


Fig. 4. „Östlich von Ferro" = East of Ferro: indication to the old Ferro prime meridian in a sheet of a Habsburg military survey.

As we see in the Table 2, some prime meridians are described by more longitude differences from Greenwich. For example, this is the situation of the Ferro meridian, which was widely, almost exclusively used in Central Europe prior to the first part of the 20th century. Ferro (Fig. 4; nowadays it is called El Hierro) is the westernmost point of the Canary Islands. The meridian 'fits to the margin of the ancient Old World' (the one without the Americas; Fig. 5). In fact, the longitude of Ferro refers to the Paris prime meridian. The longitude difference between Ferro and Paris is, according to the French *Bureau International de l'Heure* (BIH), 20 degrees, in round numbers (Fig. 6). The Ferro prime meridian itself was proposed as a commonly used one also by a – mostly forgotten – 'international conference', brokered by the French Cardinal Richelieu in the 17th century.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 5. Ferro, now El Hierro, Canary Islands, in the Google Earth. As the Ferro prime meridian is cca. 17° 40' west of Greenwich, it is quite surprising that Ferro is 'west of Ferro' indeed. This prime meridian was artificially selected and not connected to the island at all.

About the given three different values of Ferro in Table 2: the value of the BIH refers to the exact 20 degree west from Paris. The 'Albrech-difference' between Ferro and Greenwich differs from that by about one meter. Later, this difference was modified by Germany, and later by two successor states of the Monarchy. The cause was an error in the longitude observation at the old observatory tower of Berlin; this error was 13,39 arc seconds. Adding this value to the Albrecht-difference, it is 17° 39' 59.41", which can be substituted by the round number of 17° 40' with an error around 1.5 meters. So, this figure was used in Germany, Austria and Czechoslovakia, which enabled to further use the sheet system of the topographic maps.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 6. The 'Cassini meridian' of the old Paris observatory. The Ferro prime meridian was indeed defined as a meridian that is west of this line by 20 degrees in round numbers (Wikipedia).

At the Gellérthegy, the fundamental point of the old Hungarian networks, there are also several figures indicated: similarly to the latitude, the coordinates of the point are the functions of the (different) geodetic datum(s).

We can find maps, e.g. in Spain and Norway, at which the Greenwich prime meridian used, but their sheet system, remained to connected to the old, in this examples to the Madrid or Oslo meridians (Fig 7).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 7. The sheet frames of the modern 1:50,000 map of Norway follows the old Oslo meridian, however the longitudes are give from Greenwich.

Prime meridians are also applied at mapping of celestial bodies. In case of the Mars, the prime meridian is defined at the crater 'Airy-0' (named after the former director and Royal Astronomer of Greenwich). At the moon, this longitude is fixed at the Bruce Crater, in the middle of the visible part. Differently from the terrestrial coordinate system, in the sky there is a unique prime meridian, which is a good one for the celestial system. The longitude of the vernal equinox, the ascending node of the Sun's apparent orbit, is a natural possibility. The only problem is that the vernal equinox is slowly moves because of the luni-solar precession of the earth, so the celestial prime meridian should be connected to an epoch of that.

Nowadays, our terrestrial coordinate systems are not connected to the physical location of the Greenwich Observatory anymore: they are derived from the celestial system (the ICRF, the International Celestial Coordinate Frame) via the epoch of the vernal equinox and the Earth's rotation parameters. That's why in the WGS84 (see point 3.3) used by the GPS units and also by the Google Earth, the longitude of the historical Airy meridian in Greenwich is 5.31 seconds west 'from itself', indeed from the new prime meridian (Fig. 8).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 8. Surprisingly enough, the Airy meridian of the Greenwich observatory is 'west of Greenwich' by cca. 150 meters in the WGS84 datum of the Google Earth. The WGS84 is connected to the celestial reference system, not to the traditional Greenwich meridian.

2.3 Coordinate systems and coordinate frames

To locate and place any object in the plane or in the space, to define their location are enabled by coordinate systems. In the coordinate systems, or, in other words, the reference systems, the coordinates of the objects describe its location exactly. The axes to the coordinate systems are linearly independent from each other. The system types in the GIS practice:

- planar orthographic coordinate system (planar system)
- spatial orthographic coordinate system (or Cartesian system, after the Latin name of Descartes)
- spherical polar coordinate system (geocentric or spherical system)
- ellipsoidal (geodetic) coordinate system

The axes of the first two types are lines, perpendicular to each other in the plane or in the space, respectively. In the last two cases, the coordinates are one distance (from the center, or more practically, from a defined surface) and two directional angles, the longitude and the latitude. The coordinates are given in units described in Point 2.1.

Neither the coordinate systems nor the coordinates themselves are visible in the real world. That's why the coordinate systems are realized by physically discrete points and their fixed coordinates in a specific system. This physically existing, observable point set, characterized by point coordinates is called reference frame. In fact, all geodetic point networks are reference frames. Any reference frame is burdened by necessary errors, by theoretical or measure ones, based on the technology of the creation of the frame. In case of the geodetic frames, the difference between the Earth's theoretical shape, the geoid, and its ellipsoidal approximation causes theoretical errors. Besides, the limited measuring accuracy results further errors in the coordinate frame.

Longitude of a point is the same both in spherical (geocentric) and ellipsoidal (geodetic) systems. However, its latitude is different, because of the altered definition of the angle of the latitude. *In this version of the textbook, all latitudes and longitudes are interpreted in ellipsoidal (geodetic) system*.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 3. Shape of the Earth and its practical simplification

There are several approaches to define the shape of the Earth. In our study, we need one that is in a form of a function. This function should give just one value to given spheric or ellipsoidal coordinates. This value can be a length of a radius from the center to our point, or an elevation over a specific theoretical surface.

An obvious selection would be the border of the solid Earth and the hydrosphere with the atmosphere. However, this approach immediately raises some problems of definition: should the 'solid' vegetation be a part of the shape of our planet? How could we handle the buildings or the floating icebergs?

Still, if we could solve these above problems, there still is another theoretical one: this definition does not result an unambiguous function. In case of the caves or the over-bent slopes there are several altitude values connected to a specific horizontal location. The shape of the border of the phases should be somewhat smoothed.

The field of the gravity force offers exactly these kinds of smoothed surfaces. The geoid ('Earth-like') shape of the Earth can be described by a specific level surface of this force field. There are infinite numbers of level surfaces, so we choose the one that fits the best to the mean sea level. From this setup we obtain the less precise, however very imaginable definition of the geoid: the continuation of the sea level beneath the continents. Let's see, how this picture was formed in the history and how could we use it in the practical surveying.

3.1 Change of the assumed shape of the Earth in the science

The ancient Greeks were aware of the sphere-like shape of our planet. The famous experiment of Erathostenes, when in the exact time of the summer solstice (so, at the same time) the angles of the Sun elevation were measured at different geographical latitudes, to estimate the radius of the Earth, is well known. However, the accuracy of the estimation, concerning the technology of that age, is considerably good.

Although the science of the medieval Europe considered the Greeks as its ancestors, they thought that the Earth is flat. Beliefs, like 'end of the world', the answer to the question: what location we got if we go a lot to a constant direction at a flat surface, were derived from this.

The results of the 15th and 16th century navigation, especially the circumnavigation of the small fleet of Magellan (1520-21) made this view of the world obsolete. However the Church accepted this only slowly, the idea of the sphere-like Earth was again the governing one.

There were several observations that questioned the real ideal spherical shape. In the 17th century, the accuracy of the time measuring was increased by the pendulum clock. The precisely set pendulum clocks could reproduce the today's noon from the yesterday's one with an error of 1-3 seconds. If such a correctly set up clock was transferred to considerably different latitude – e.g. from Paris to the French Guyana – higher errors, sometimes more than a minute long ones, were occurred. This is because the period of the pendulum is controlled by the gravitational acceleration, that is, according to there observations, obviously varies with the latitude. Paris is closer to the mass center of our planet than the French Guyana is, thus the ideal spherical shape of the Earth must be somewhat distorted, the radius is a function of the latitude, and the real shape is like an ellipsoid of revolution.

Distorted but in which direction? Elongated or flattened? The polar or the equatorial radius is longer? Perhaps nowadays it is a bit surprising but this debate lasted several decades, fought by astronomers, geodesists, mathematicians and physicists. Finally, the angular measurements, brokered by the French Academy of Sciences, settled it. In Lapland, at high latitudes, and in Peru, at low altitudes, they measured the distances of meridian lines between points where the culmination height of a star was different by one arc degree. The answer was obvious: the Earth is flattened; the polar radius is shorter than the equatorial one.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

The flattened ellipsoid of revolution can be exactly defined by two figures, as it was shown in Chapter 2. Traditionally, one of them is the semi-major axis, the equatorial radius, gives the size of the ellipsoid. The other figure, either the semi-minor axis or the flattening or the eccentricity, gives the shape of the ellipsoid. The time of the invention of this concept, the authors usually gave the inverse flattening. This figure describes the ratio between the semi-major axis and the difference between the semi-major and semi-minor axes.

At the end of the 1700s and in the first half of the 1800s, several ellipsoids were published, as the better and better approximations of the shape of the Earth. These ellipsoids are referred to as the publishing scholar's name and the year of the publication, e.g. the Zách 1806 ellipsoid means the ellipsoid size-shape pair described be the Hungarian astronomer-geodesist Ferenc Zách in 1806.

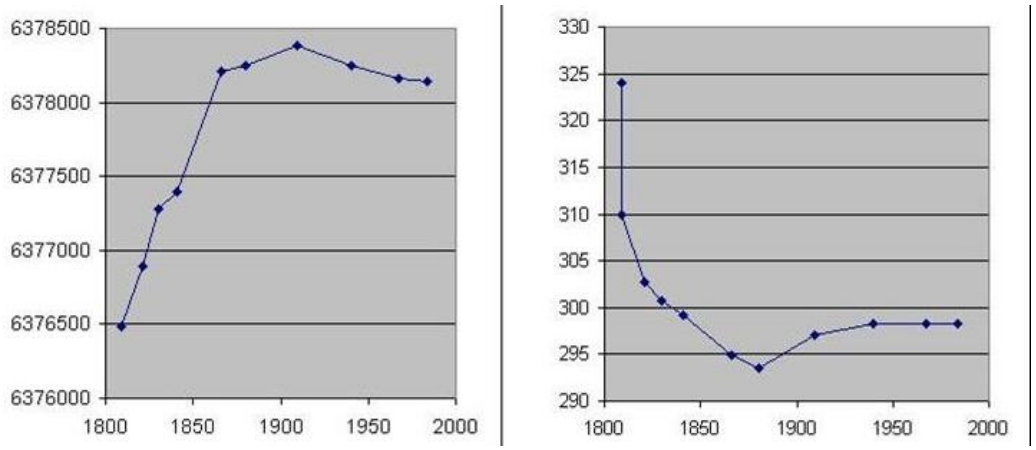


Fig. 9. The changes of the semi-major axis (left) and inverse flattening (right) of the 'most up-to-date ellipsoid' in the time. The first data indicates the geoid shape of Europe, then the colonial surveys altered these values, and finally the global values are provided.

The semi-major axis and the flattening of the estimated ellipsoids are not independent from each other. Fig. 9 shows the changes of these two figures as a function of the time, concerning the most accepted ellipsoids of that time, from 1800 to nowadays. The first part of this period was characterized by the increase of the semi-major axes and the decrease of the inverse flattening. The Earth occurred to be slightly larger and more sphere-like that it was first estimated. However, estimating the semi-major axis and the flattening is not a very complicated technical exercise. So, why were the results different, why is this whole change?

The first observers published the results based on just one arc measurement. The first ellipsoid, that was based on multiple, namely five, independent observations were set up by the Austrian scholar Walbeck in 1819. It occurred that the virtual semi-major axis and the flattening is changing from place to place. So, the whole body is not exactly an ellipsoid. It is almost that, but not completely.

This 'not completely' occurred again during the building up the triangulation networks (see point 3.3). Because of this, the shape description based on the gravity theory, mentioned in the introduction of this chapter, was defined first by Karl Friedrich Gauss in the 1820s. The name 'geoid' was proposed by Johann Benedict Listing much later, in 1872. Known the real shape of the geoid (Fig. 10) we can easily interpret the trend of the estimated ellipsoid parameters. Based on the European part of the geoid, the Earth seems to be smaller and more flattened. However, if we measure also in other continents, like in the locally different-shaped India, then we got the trend-line of the Fig. 9.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

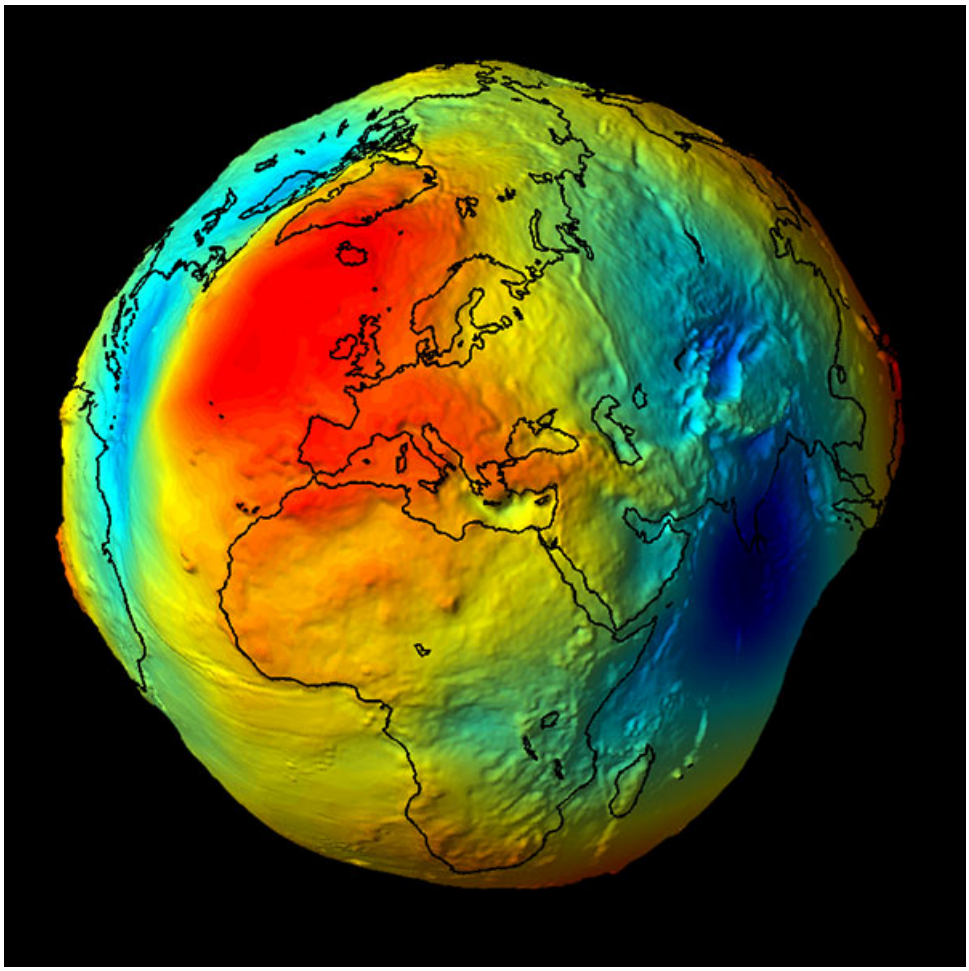


Fig. 10. The geoid, the level surface of the Earth, with massive vertical exaggeration.

The parameters of the most up-to-date ellipsoids, such as the GRS80, the WGS84, were determined by the whole geoid, with the following constraints:

- the geometric center of the ellipsoid should be at the mass center of the Earth
- the rotational axis of the ellipsoid should be at the rotational axis of the real Earth
- the volume of the ellipsoid and the geoid should be equal, and
- the altitude difference between the ellipsoid and the geoid should be on minimum, concerning the whole surface of the planet.

At a point of the surface, the geoid undulation is the distance of the chosen ellipsoid and the geoid along the plumb line. The geoid undulation from the best-fit WGS84 ellipsoid along the whole surface does not exceed the value of ±110 meters.

Summarizing: the equatorial radius of our planet is about 6378 kilometers, the difference between the equatorial and the polar radius (the error of the spherical model) is about 21 kilometers, while the maximum geoid undulation (the error of the ellipsoid model) is 110 meters.

3.2 The geoid and the ellipsoid of revolution

The mathematical description of the geoid is possible in several ways. It is possible to give the radius lengths from the geometric center to the geoid surface at crosshairs of the parallels and meridians (latitude-longitude grid). We can also give just the vertical difference of the geoid and an ellipsoid (the geoid undulations) in the same system.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

The geoid can be also described by the form of the spherical harmonics. Describing a local or regional geoid part, a grid in map projection can be also used.

Selecting any form from these possibilities, it is obvious that the geoid is a very complex surface. If we are about to make a map, we have to chose a projection. The projections, that are quite easy if we suppose the Earth as a sphere, become complicated in ellipsoidal case, while they cannot be handled at all, if the original surface is the real geoid. It was even more impossible to use in the pre-computer age, while the mathematics of the map projections was invented. So, in the geodetic and cartographic applications, the true shape of our planet, the geoid, is substituted by the ellipsoid of revolution.

The ellipsoid for this approximation is generally a well known surface with pre-set semi-major axis and flattening/eccentricity. We can note, that in case of some ellipsoids, characterized by the same name and year, it is possible to find different semi-major axis lengths (such as at the Everest ellipsoid, or the, aforementioned Bessel-Namibia, see Table 3). The cause of this is according to the original definition of these ellipsoids, the semi-major axis was not given in meters but in other units, e.g. in yards of feet. Converting to meters, it is important to give enough decimal figures in the conversion factor. Omitting the ten thousandth parts in this factor (the fourth decimal digit after the point) won't cause much difference in the everyday life, however if we have millions of feet (such in case of the Earth's radius we do) the difference is up to several hundred meters.

name	a	b	1/f	f	e
Laplace 1802	6376615	6355776.4	306.0058	0.003268	0.08078
Bohnenberger 1809	6376480	6356799.51	324	0.003086	0.07851
Zach 1809	6376480	6355910.71	310	0.003226	0.08026
Zach-Oriani 1810	6376130	6355562.26	310	0.003226	0.08026
Walbeck 1820	6376896	6355834.85	302.78	0.003303	0.08121
Everest 1830	6377276	6356075.4	300.8	0.003324	0.08147
Bessel 1841	6377397	6356078.96	299.1528	0.003343	0.08170
Struve 1860	6378298	6356657.14	294.73	0.003393	0.08231
Clarke 1866	6378206	6356583.8	294.98	0.00339	0.08227
Clarke 1880	6378249	6356514.87	293.465	0.003408	0.08248
Hayford (Int'l) 1924	6378388	6356911.95	297	0.003367	0.08199
Krassovsky 1940	6378245	6356863.02	298.3	0.003352	0.08181
GRS67	6378160	6356774.52	298.2472	0.003353	0.08182
GRS80	6378137	6356752.31	298.2572	0.003353	0.08182
WGS84	6378137	6356752.31	298.2572	0.003353	0.08182
Mars (MOLA)	3396200	3376200	169.81	0.005889	0.10837

Table 3. Data of some ellipsoids used in cartography. a: semi-major axis; b: semi-minor axis; 1/f: inverse flattening; f: flattening; e: eccentricity.

The fitting of the ellipsoid to the geoid is an important exercise of the physical geodesy. Prior to the usage of cosmic geodesy, this task could be accomplished by creating of geodetic or triangulation networks and (later) by their adjustment.

3.3 Types of the triangulation networks, their set up and adjustment

Measuring of the distance of two points is possible by making a line between them and by placing a measuring rod along it – supposed the distance is not too long between our points. As the distance becomes longer, this pro-

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

cedure starts to be complicated and expensive: distances of more than a several hundred meters are very hard to measure this way. If the terrain between our points is rough or inpassable, this method cannot be applied at all.

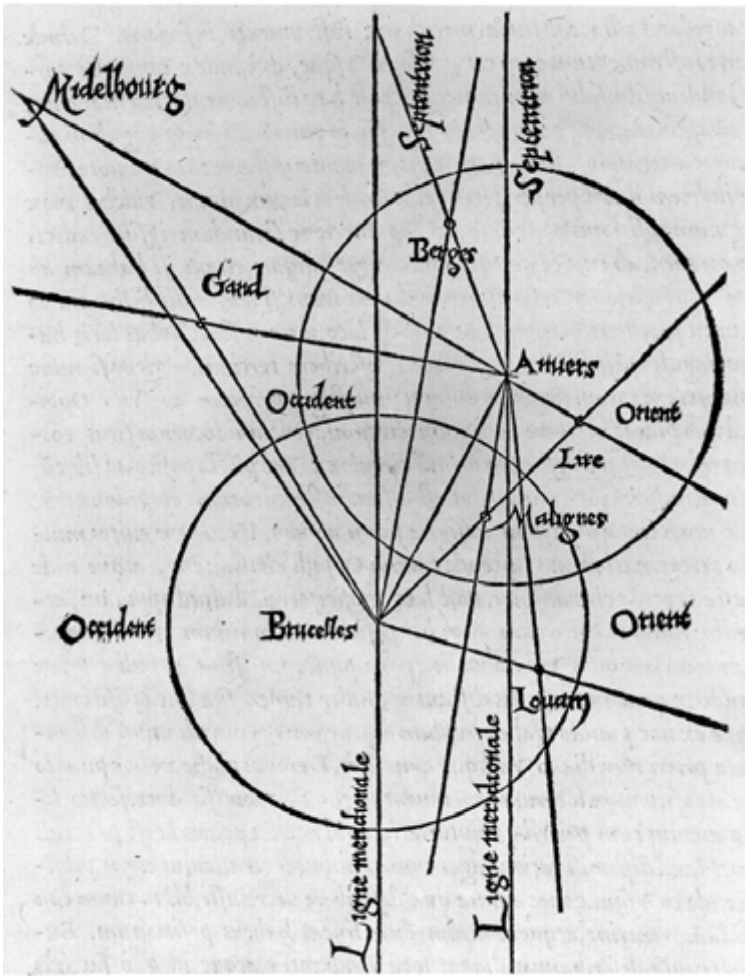


Fig. 11. The sketch of the Belgian triangulation of Gemma Frisius from the 16th century.

A new method was introduced at the end of the 16th, early 17th centuries. Measuring a longer distance can be made by measuring a shorter line and some angles. The first triangulation was proposed by Gemma Frisius (Fig. 11), then in 1615, another Dutchman, Snellius accomplished a distance measurement by triangulation between the towers of Alkmaar and Breda (true distance cca. 140 kilometers, throughout the Rhein-Maas delta swamps; Fig. 12). During the campaign, he set up triangles with church towers at the nodes and measured the angles of all triangles. Having these data, it was needed to measure only one triangle side to calculate all distances between the nodes.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 12. The distance between the towns of Alkmaar (in the north) and Breda (in the south) was determined by the 1615 triangulation of Snellius, throughout swamps, marshlands and rivers.

The Snellius-measurements provided an interesting invention: the sum of the detected inner angles of a triangle occurred to be slightly more than 180 degrees (Fig. 13). This is the consequence of the non-planar, spheroid geometry of the surface of the Earth, and this is true at spherical triangles. It was the root of a new branch of the geometry: the spherical trigonometry.

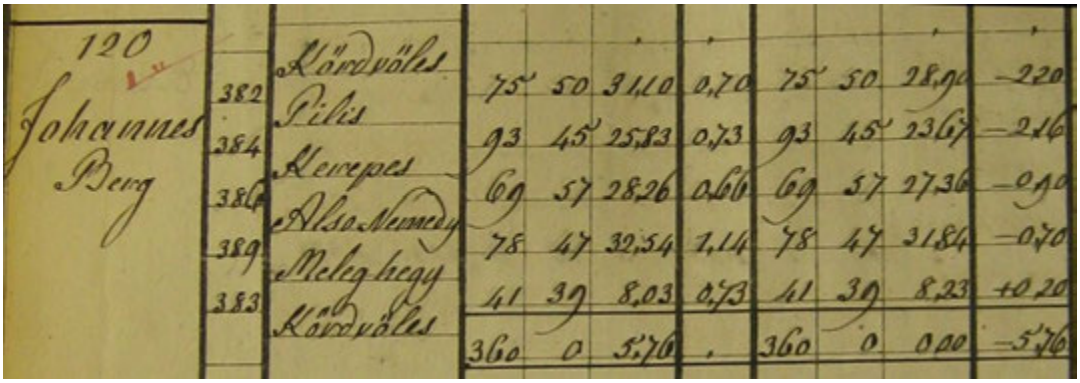


Fig. 13. Angles between far geodetic points from the point of Johannes Berg, Budapest (Habsburg survey, 1901). The sum of the angles exceed the 360 degrees, according to the spheric surface.

Using triangulation networks, not only distances but also coordinates can be determined. For this, it is first needed to measure the geographic coordinates of one point of the network. That's why there is in most cases an astronom-

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

ical observatory at the starting point of a geodetic network: these measurements can be carried out there in most simple way. Also it is necessary a baseline: a shorter distance between two network points, whose distance is measured physically, and an azimuth: the measured angle between the true north and a triangle edge to a selected network point. Of course, the angles of all triangles should be measured together with the heights of the points. Using all of these data, and assuming an ellipsoid with a pre-set semi-major axis and flattening, the coordinates of all network points can be calculated. They are called triangulated coordinates. The longitude values in these coordinates are measured from the meridian of the astronomical observatory (Fig. 14).

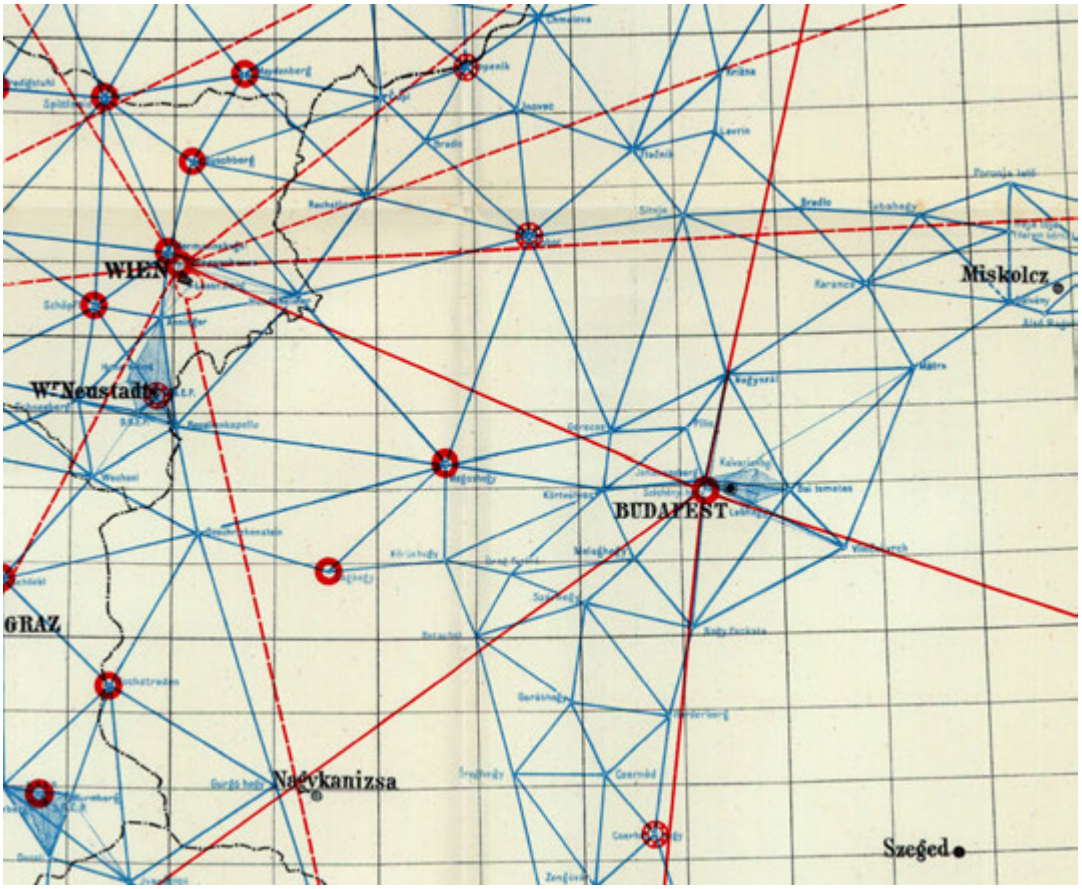


Fig. 14. Sketch of the 1901 triangulation network between Vienna and Budapest.

To check the obtained coordinates of the base points, more baselines and – which later brought a real revolution in the data processing – the astronomical coordinates were observed at several triangulation points (at the so-called Laplace-points) of the network. The *observed* positions, however, differed from the ones, *computed* by the trigonometry. The difference occurred in all cases and its magnitude was not predictable. Its cause is the geoid shape of the Earth: the astronomic observations are based on the knowledge of the local horizontal and vertical lines, which are slightly different from the tangent and normal directions of the ellipsoid. As we mentioned above, the whole body is not exactly an ellipsoid. It is almost that, but not completely.

This problem became so important in the first half of the 19th century that Gauss invented his famous method of the least squares exactly to solve that. The goal is to 'adjust' the coordinates of the base points in order to minimize the squares of the differences occurring at the Laplace-points. The method is called *geodetic network adjustment*, which is, in practical words, to homogenize the errors, mostly caused by the geoid shape, in the whole network. The result of the adjustment is a geodetic point set organized into a network, with their finalized coordinate values.

What means the adjustment from geometric point of view? What is the geometric result? An ellipsoid whose

- size and shape was pre-set during the adjustment;
- semi-minor axis fits (as much as possible) to a parallel direction of the rotation axis;

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

- surface part – the one set by the extents of the network – fits optimally to the same part of the geoid.

The geometric center of this ellipsoid is, of course, different from the mass center of the Earth (Fig. 15). This way, not only the size and shape of the ellipsoid is known but also its spatial location.

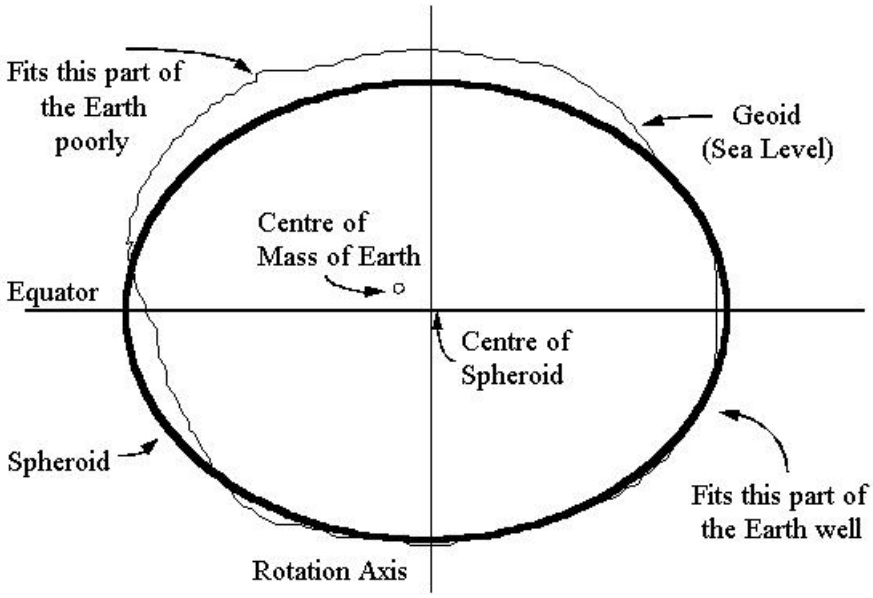


Fig. 15. Geometric result of the geodetic network adjustment: fitting the ellipsoid to the surveyed part of the geoid; the geometric centre differs from the mass center of the Earth.

From the point of view of the ellipsoid location method in space, there are three types of them:

- deliberate displacement: there is only one astronomical base-point, the network is not adjusted, the ellipsoid is fit to the geoid surface at only one point (usually the location of the astronomical observatory). This method is characteristic at the small islands in the ocean, with no continent on the horizon; the network names are often indicated by the 'ASTRO' sign. Also, this is the usual method used at the old mapping works, having geodetic basis that was build before the invention of the adjustment method.

- relative displacement: the network adjustment is accomplished, the ellipsoid is fit to a certain part of the geoid surface, practically to the extents of the survey.

- absolute displacement: the geometric center of the ellipsoid is at the mass center of the planet, the semi-minor axis lies in the rotational axis. It cannot be realized just by surface geodesy or geophysics (as the exact direction of the mass center cannot be determined from the surface by geophysical methods). For its implementation, space geodesy (Doppler measurements, GPS) is needed. Prior to the space age, before to the 1960s, there were no ellipsoids with absolute displacement. The WGS84 is a typical example of this.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 4. Geodetic datums

A geodetic datum is an ellipsoid (described by parameters of its size and shape) together with the data about its dislocation, and in some cases, its orientation and scale. It is very important to note that as the ellipsoid size, the dislocation and the orientation are different from a datum to another one, the geodetic coordinates in the different datums (according to different geodetic networks) are also different. We repeat: at a same field point the geodetic coordinates are different on different datums (Fig. 16). The GIS software packages are capable to make transformations between them, if the appropriate datum parameters are known. . This chapter shows the method of usage and estimation of these parameters.

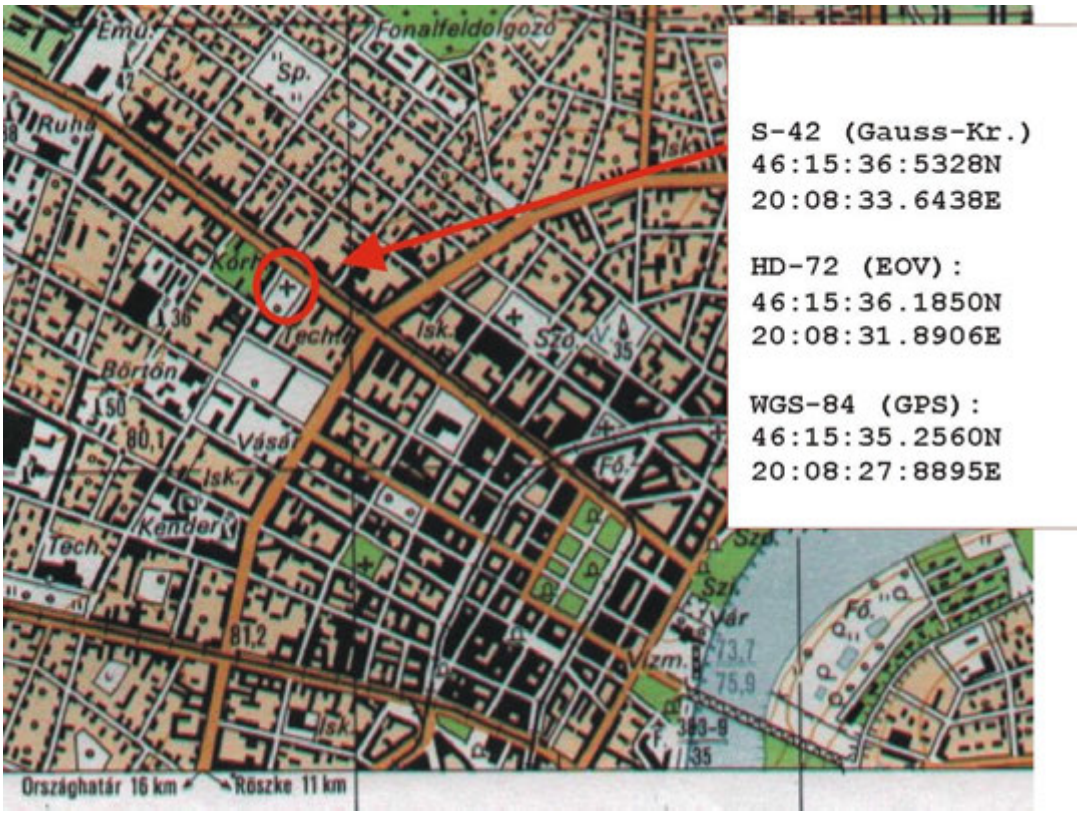


Fig. 16. The ellipsoidal coordinates of a church in the city of Szeged are different in different geodetic datum. This is the case of all terrain points.

4.1 Parameters of the triangulation networks

As it was shown in the previous chapter, the triangulation networks are characterized by its geodetic point set and the fixed geodetic coordinates of these base-points. A triangulation network is a geodetic datum. To use it in any GIS software, we have to give these data of the network in a more compressed way that is still characteristic for the whole network. We have also to know that which data is needed for a network/datum description for our very GIS software.

The most commonly used possibilities in geodetic practices to provide data at a selected point of the network (the so-called fundamental point) are as follows:

- the geodetic coordinates
- the astronomical coordinates and
- a triangulated and astronomical azimuth to a selected neighboring network point.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

As the geodetic network adjustment can be interpreted as the fit of the ellipsoid to the geoid surface, the geoid undulation at the fundamental point is usually taken as zero. If it is different by any cause, if should be given, too. For example, in the case of the Hungarian Datum 1972, the geoid undulation at the Szőlőhegy, the fundamental point, is set to 6.56 meters by the reason to fit it vertically to the unified datum of the former Warsaw Pact cartography. This value should be taken into account during our work to avoid vertical errors, if they are important.

The above set of information is considerably smaller than the one represented by the whole set of base-point coordinates in the network. It is assumed that by fitting the given ellipsoid to the fundamental point, described its own data, the coordinates of the other points can be computed. Obviously, it is not true, and the quality of a geodetic datum is given by just this accuracy of the point coordinate calculation at all points of the network. Usually, the newer the triangulation network, the better its quality is. In case of the historical Hungarian systems, the average error at the networks form the end of 19th century is 2-3 meters, 1,5-2 meters at the systems of mid-20th century, while nowadays the accuracy is as low as half a meter.

Sometimes there are other ways to giving parameters to a geodetic network: to use the three-dimensional Cartesian coordinates of the fundamental point or just giving the components of the deflection of vertical, completed by the geoid undulation.

The above parameters do not suit the GIS software needs; these programs follows a different philosophy at the datum definition. They are not using just one datum but aim to handle several ones. So, they need parameters between datums and not just for parameters of different ones. In most cases, they don't handle all possible datum pairs to convert between them but select one datum and give the transformation parameters from any other one to this. Practically, this selected datum is an absolutely displaced, globally fit WGS84, and all other (local) datums are characterized by the transformation parameters from them to the WGS84. In this method, it is needed to define the position of the geometric center of the local datum ellipsoid and – if available – the orientation difference between the local datum and the WGS84.

4.2 The 'abridging Molodensky' datum parametrization method

The easiest way to define the connection between two datums is to define the vector connecting their geometric centers (Fig. 17). This vector should be given by the components in the geocentric Cartesian coordinate system, described in the Chapter 2, expressed in meters. Obviously, if both analyzed datum are of absolute dislocations, this vector is the null vector, with the components of (0 m; 0 m; 0 m). It should be noted that the international literature often and erroneously called this method as Molodensky- or Molodensky-Badekas-type parametrization, albeit they are indeed more complex ones.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

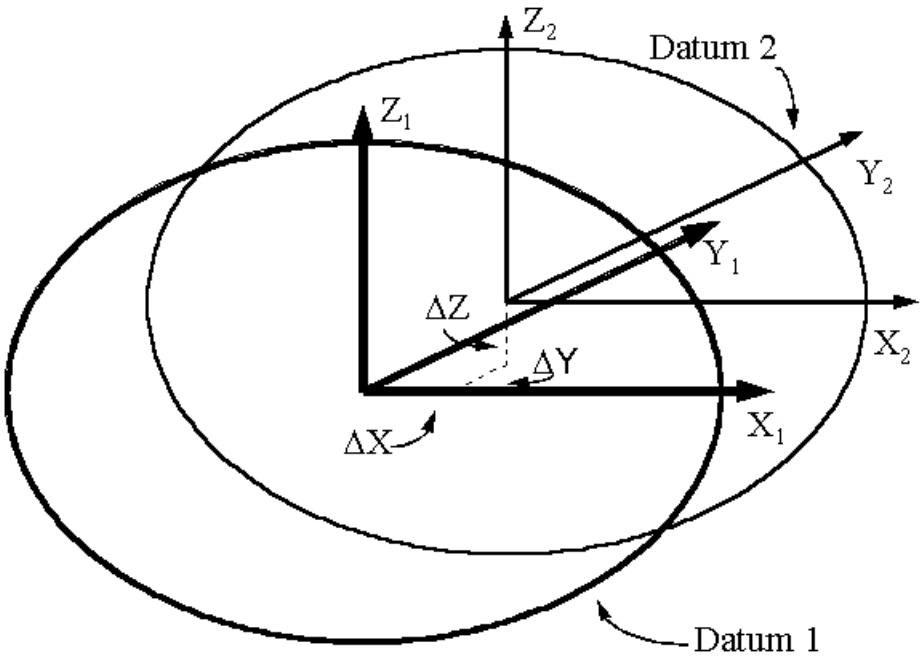


Fig. 17. The abridged Molodensky transformation is a simple shift between two datum ellipsoids, expressed by the three components of the shift vector.

So, the three parameters of the abridging Molodesky datum description are the metric distances of dX, dY and dZ, describing the spatial locations of the geometric centers of the datum locations from each other. If one of these datums is the WGS84, these dX, dY and dZ parameters give the location of the local datum with respect to the mass center of the Earth. If the coordinates of a basepoint are known on a Datum '1', the geocentric coordinates on the Datum '2' are the following:

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix} + \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} \tag{4.2.1}$$

The angular difference between the coordinates on the starting and the goal datums can be also expressed without to convert to geocentric coordinates and *vice versa*:

$$\Delta\Phi'' = \frac{-dX \sin\Phi\cos\Lambda - dY \sin\Phi\sin\Lambda + dZ\cos\Phi + (a\cdot df + f\cdot da)\sin 2\Phi)}{M \sin 1''} \tag{4.2.2}$$

$$\Delta\Lambda'' = \frac{-dX\sin\Lambda + dY\cos\Lambda}{N\cos\Phi\sin 1''} \tag{4.2.3}$$

$$\Delta h = dX\cos\Phi\cos\Lambda + dY\cos\Phi\sin\Lambda + dZ\sin\Phi + (a\cdot df + f\cdot da)\sin^2\Phi - da \tag{4.2.4}$$

where $M(\Phi) = a\frac{1-e^2}{(1-e^2\sin^2\Phi)^{3/2}}$ is the curvature in the prime meridian; $N(\Phi) = \frac{a}{\sqrt{1-e^2\sin^2\Phi}}$ is the curvature in the prime vertical; $\Delta\Phi$" and $\Delta\Lambda$" are the latitude and logitude differences between the coordinates of the two datums in arc second; Δh is the difference between the ellipsoidal heights; a and f are the semi-major axis and the flattening of the starting datum; while da and df are the differences of them between the starting and goal datums. If the ellipsoidal heights are not given, they can be estimated from leveled heights using geoid models, or we can simply omit the Equation (4.2.4) at the calculation.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

As it was mentioned, the GIS packages describe the datums by parameters between them and the WGS84 special datum, thus handling the problem that a datum cannot be this way parametrized alone, just the difference between it and another datum. If we have two different datums (not the WGS84) and we know the parameters of the transformation from them to the WGS84, the abridging Molodensky parameters between the two datums can be given because of the linearity. Let the transformation A is the one between the first datum and the WGS84 and the transformation B is the one from the second datum to it. The C shows the direct transformation from the first and second datums. The parameters of this C are (commutation):

$$\begin{bmatrix} dX_C \\ dY_C \\ dZ_C \end{bmatrix} = \begin{bmatrix} dX_A \\ dY_A \\ dZ_A \end{bmatrix} - \begin{bmatrix} dX_B \\ dY_B \\ dZ_B \end{bmatrix} \tag{4.2.5}$$

These parameters are not depending on the ellipsoids used for the different datums. For example, the datum shift parameters from the Austrian MGI datum to the WGS84 are dX=+592 m; dY=+80 m; dZ=+460 m. The same parameter set between the German DHDN77 system and the WGS84 are dX=+631 m; dY=+23 m; dZ=+451 m. Thus, the direct transformation parameters from the MGI to the DHDN77 are dX=–39 m; dY=+57 m; dZ=+9 m.

In the literature, we often find different number triplets as parameters of a transformation from a specific datum and the WGS84. Albeit it is obviously an error in spatial context, the transformation error in the horizontal coordinates (latitude and longitude) is not necessarily significant at them. Using different triplets as abridging Molodensky parameters for a datum, as it is shown below, there is always one point on the ellipsoid, where the two different parameter set result the same horizontal shift. The main question is, whether this point falls to the extents of the valid territory of the datum (the geodetic network), if possible, near to its center/fundamental point, or not. If yes, both parameter sets can be used and we can compute the vertical difference of the two datums at that point. Usually, the difference is because of the neglecting of the geoid undulation value.

Let **r1** the position vector from the center of the WGS84 to the geometric center of the Datum version 1 and **r2** is the similar one to the center of the Datum version 2. Making the difference of these position vectors in the space:

$$\mathbf{r}_{\mathbf{diff}} = \mathbf{r}_1 - \mathbf{r}_2 \tag{4.2.6}$$

Now, let's check that this vector shows to which point of the reference surface:

$$\varphi_r = \arctan\left(\frac{dZ_{diff}}{\sqrt{dX_{diff}^2 + dY_{diff}^2}}\right) \tag{4.2.7}$$

$$\lambda_r = \arctan\left(\frac{dY_{diff}}{dX_{diff}}\right) \tag{4.2.8}$$

while the length of the difference vector (the spatial difference) in meters is

$$\left|r_{diff}\right| = \sqrt{dX_{diff}^2 + dY_{diff}^2 + dZ_{diff}^2} \tag{4.2.9}$$

If the point (φ_r,λ_r) is in the area of the used triangulation network, possibly near to the fundamental point, both versions can be used. As I mentioned above, in this case, the length of $\mathbf{r}_{\mathbf{diff}}$ is usually around the geoid undulation value between the local datum and the WGS84 at the point (φ_r,λ_r). If this point falls to a distant position on the Earth's surface, one of the parameter sets is erroneous.

4.3 The Burša-Wolf type datum parameters

The Burša-Wolf type parametrization method (called after the Czech Milan Burša and the German Helmut Wolf) handles not only the difference of the positions of the geometric centers of the datum ellipsoids, but also the orientation differences and the small scale variations as one or both datum's size differs indeed from the ideal size of

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

the selected ellipsoid (Fig. 18). The transformation is expressed for the geocentric Cartesian coordinates as input and out data, as follows:

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix} + (1+k) \begin{bmatrix} 1 & \varepsilon_Z & -\varepsilon_Y \\ -\varepsilon_Z & 1 & \varepsilon_X \\ \varepsilon_Y & -\varepsilon_X & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} \tag{4.3.1}$$

This is a special case of the spatial Helmert similarity transformation for very small rotation angles (a few or a few tens of arc seconds), with the possible simplifications. In this equation, the dX, dY and dZ are the same as in case of the abridging Molodensky transformations (but, as it shown below, cannot be handle in that without analysis!), ε_x, ε_y and ε_z are the rotations along the coordinate axes and k is the scale factor. If there are no rotations and the scale difference is zero, the Equation (4.3.1) becomes the same to Equation (4.2.1).

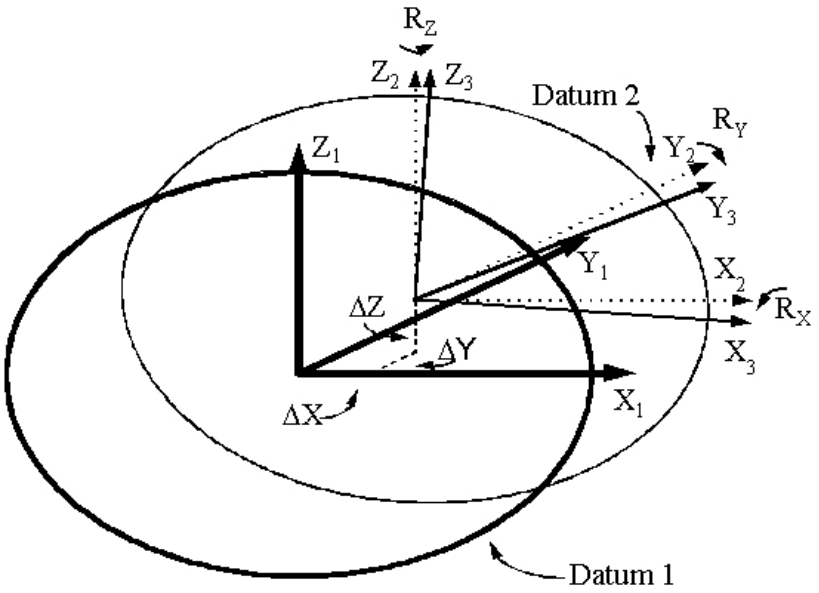


Fig. 18. The Bursa-Wolf transformation handles both the shift and the orientation differences between the two datum ellipsoids.

In fact, there are two different sign convention of the non-diagonal matrix members in Equation (4.3.1). If these signs are used like in the Equation (4.3.1), it is called *coordinate frame rotation*, which means that the coordinate axes are rotated around the fixed position vector. However, if all the signs of the non-diagonal members in the matrix of (4.3.1) are reversed, this convention is called *position vector rotation*, as this vector is rotated in the fixed coordinated frame.

Neither of the above conventions is an accepted standard. The United States, Canada and Australia use the 'coordinate frame rotation', while in Europe the 'position vector rotation' is mostly preferred. The international draft ISO19990 also proposes this latter one, however because of the U.S. refusal its international acceptation is questioned. We have to know that as most GIS software packages are developed in the U.S., Canada and Australia, the 'coordinate frame rotation' is a quasi-standard in them, while most European meta-data are published according to the 'position vector rotation' convention. If we are provided a Burša-Wolf type parameter set for a datum, first try to use it assuming the 'coordinate frame rotation', and if the results are obviously erroneous, switch all the signs of the rotation parameters.

Similarly to the abridging Molodensky transformation, the Burša-Wolf formula is commutative. It is possible to express the resultant of two transformations by simply summarize their respective parameters. This perhaps surprising statement can be easily understood mathematically:

The Equation (4.3.1) after two, successive transformation can be expressed in form

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

$$\mathbf{x'}=\mathbf{dx_2}+(1+k_2)\mathbf{A_2}[\mathbf{dx_1}+(1+k_1)\mathbf{A_1x}] \quad (4.3.2)$$

where $\mathbf{dx_1}$ and $\mathbf{dx_2}$ are the two shift vectors, k_1 and k_2 are the two scale factors, and $\mathbf{A_1}$ and $\mathbf{A_2}$ are the rotation matrices, $\mathbf{x}$ is the input position vector and $\mathbf{x'}$ is the result. Organizing this can be expressed in form

$$\mathbf{x'}=\mathbf{dx_2}+(1+k_2)\mathbf{A_2dx_1}+(1+k_2)(1+k_1)\mathbf{A_1A_2x} \quad (4.3.3)$$

where the $\mathbf{dx_r}$,k_r and $\mathbf{A_r}$ parameters of the resultant transformation are

$$\mathbf{dx_r}=\mathbf{dx_2}+(1+k_2)\mathbf{A_2dx_1} \quad (4.3.4)$$

$$k_r=k_1+k_2+k_1k_2 \approx k_1+k_2 \quad (4.3.5)$$

$$\mathbf{A_r}=\mathbf{A_1A_2} \approx \mathbf{A_1}+\mathbf{A_2} \quad (4.3.6)$$

The approximation of (4.3.5) can be immediately understood in cases when the scale factors are in order or 1-10 part per million (ppm). The approximation of (4.3.6) is a bit more difficult, we should accomplish the matrix multiplication, omitting the resulted members falling to the range of the squares of the rotation angles and the scale factor. The right side of the Equation (4.3.4) is the second transformation done to the $\mathbf{dx_1}$ shift vector. Omitting the effect of the very small scale factor, it is

$$\mathbf{dx_e}=\mathbf{dx_2}+\mathbf{A_2dx_1} \approx \mathbf{dx_1}+\mathbf{dx_2} \quad (4.3.7)$$

As the shift vector is usually much more short that the position vectors (n*100 meters compared to the Earth's radius), this approximations fits well to the practice. The three-dimensional error of this simplification is in the order of some centimeters while its horizontal component is even smaller. So, the linear commutation can be applied in the practice for the Burša-Wolf transformation, too.

4.4 Comparison of the abridging Molodensky and Burša-Wolf parametrization

The most important differences between the abridging Molodensky (AM) and the Burša-Wolf (BW) methods are shown in Table 4:

AM parametrization	BW parametrization
Easier	More complex
Usually less accurate	Usually more accurate
The parameters can be easily computed	The parameter estimation is difficult
The parameters are unambiguous	There are two conventions at the rotation parameters
Known by all GIS software packages	Known by most (but not all) GIS software packages

Table 4. Comparison of the abridged Molodensky (AM) and the Burša-Wolf (BW) datum parametrization methods.

Here we have to note that the mapping authorities of the United States follow the AM-parametrization, while the NATO adapted the BW-method.

Applying any of these methods, due to the errors of the previous geodetic network adjustments, the transformation accuracy fits to the geodetic needs (a few centimeters) only in a small area. The high-accuracy transformation exercises should be accomplished by other methods, e.g. using higher order polynomials. However the GIS software packages usually don't let the users to define polynomial transformations – however the usage of correction grid (GSB – Grid Shift Binary) files sometimes offers a good solution. However, out aim for the accuracy of a few meters (according to the map reading) is usually fulfilled by both methods. The usual errors of transformation from historical and modern Hungarian networks to the WGS84 are shown in Table 5:

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

System	Average (max.) error of AM	Average (max.) error of BW
Second survey (1821-59)	30 (200) m	Transformation not defined
Third survey (1863-1935)	5 (12) m	1,5 (4) m
DHG (1943)	2 (5) m	2 (5) m
EOV (1972)	1 m	0,2 (0,5) m
S-42 (1983)	1 m	0,2 (0,4) m

Table 5. The most frequent application errors of the two methods in Hungary. DHG (Deutsche Heeres Gitter) is the WWII German geodetic network, applied to Central and Eastern Europe.

The main source of the application errors is that usually there is no easy way to computation of the AM-parameters from the BW-type seven-parameter set. If we know the seven parameters of a BW-transformation, the three parameters of the AM-type transformation of the same datum **cannot be obtained** by just omitting the scale factor and the rotation parameters, keeping the shift ones only!

Sometimes it is tried to improve a less accurate BW-parameter set by substituting just the shift parameters from another transformation. As we see in the next chapter, it is incorrect; in most cases, the parameters of the BW transformation cannot be obtained separately.

If a parameter set (both AM or BW-type) provides incorrect results, especially if the transformation error is the double of the error without any datum transformation, try to inverse the signs of all of the parameters. If this does not correct the results, in case of the BW-method, try to change the signs of just the rotation parameters. Check whether the units we use are following the needs of the software used (arc seconds or radians). In most softwares, the scale factor should be given in ppm (part per million), while in other cases, the true value (a number close to the unity) is expected (the 'no scale difference' is expressed by zero in the first and by one in the second case). And finally; most software uses the newly set parameters only after restart.

4.5 Estimation of the transformation parameters

If we have a geodetic base-point set, containing the coordinates in two different datums, the transformation parameters between these datums can be estimated, according to both the AM and the BW methods.

The AM-parameters, the vector components between the geometric centers of the two datum ellipsoids, can be obtained easily. This calculation can be made even if the coordinates of just one common point (in most cases, the fundamental point) are known. It this case, we calculate the Cartesian coordinates of the point in both systems, using the sizes and figures of the ellipsoids and the geoid undulation values. Interpreting these two coordinate triplets as position vectors of the point in the two different systems, the desired parameters can be obtained as the components of the difference vector between them. First, the coordinates should be transformed to geocentric Cartesian ones:

$$X = (N + h)\cos\Phi\cos\Lambda$$
$$Y = (N + h)\cos\Phi\sin\Lambda \qquad (4.5.1)$$
$$Z = [N(1 - e^2) + h]\sin\Phi$$

first on the Datum 1 then on Datum 2. The first datum is usually a local one while the second is the WGS84. Then the parameters are:

$$dX = X_{WGS84} - X_{local}$$
$$dY = Y_{WGS84} - Y_{local} \qquad (4.5.2)$$
$$dZ = Z_{WGS84} - Z_{local}$$

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

In the Equation (4.5.1), *h* expresses the elevation above the ellipsoid (cf. Chapter 3). If we have the elevations above the sea level (above the geoid), we shall convert them all to ellipsoidal heights, using the geoid undulation values on local datum. If the elevations are unknown at all, they should be replaced simply by the geoid undulation values. If these values are unknown, use zero values. The geoid undulation values for the WGS84 (the second datum) can be obtained from a global geoid model, e.g. the EGM96. EGM96 data is available directly on the Internet, and free calculation programs are also available.

If we have more common points, we can repeat the above procedure for every point and the final parameters are provided by averaging.

Estimation of the BW-parameters are much more complicated. There are two approaches to do it. Usually, it is done by standard parameter estimation of the least square method. It is far beyond the goal of this handout to show the whole procedure, however it is worth to note that the method estimates the parameters simultaneously. This means that the parameters cannot be interpreted independently from each other – that's why we can't substitute the AM-type shift parameters to a BW parameter set, leaving the rotation and scale parameters untouched. In general, it is possible that the same transformation is described well by apparently very different BW parameter sets, and – contrary to the AM method – there is no easy way to show their similarity. However, there is another BW parameter estimation method, simply enough to explain, providing real, geometrically independent parameters, albeit its accuracy is a bit worse.

Let's suppose that we shall derive parameters for a transformation between the WGS84 and a local datum with known fundamental point, whose coordinates are known on both the local datum and the WGS84. In the first step, we calculate the AM type shift parameters between the two systems, using Equation (4.5.2). In the following, we choose rotation and scale parameters for them, to improve the horizontal and spatial accuracy of the transformation.

First, we shall use the fact that the effect of the scale factor to the horizontal coordinates is much less than the effect of the rotation. Moreover, we shall realize that there is a connection between the three rotation parameters and the location of the fundamental point plus the observer azimuth at it (spherical case):

$$\varphi = \arctan\left(\frac{r_Z}{\sqrt{r_X^2 + r_Y^2}}\right) \tag{4.5.3}$$

$$\lambda = \arctan\frac{r_Y}{r_X} \tag{4.5.4}$$

$$\alpha = \sqrt{r_X^2 + r_Y^2 + r_Z^2} \tag{4.5.5}$$

The inverse formulas for the ellipsoid:

$$r_X = \frac{\alpha \cos\Phi \cos\Lambda}{\sqrt{1 - e^2 \sin^2\Phi}} \tag{4.5.6}$$

$$r_Y = \frac{\alpha \cos\Phi \sin\Lambda}{\sqrt{1 - e^2 \sin^2\Phi}} \tag{4.5.7}$$

$$r_Z = \frac{\left(1 - e^2\right)\alpha \sin\Phi}{\sqrt{1 - e^2 \sin^2\Phi}} \tag{4.5.8}$$

The coordinates of the fundamental point are known. We also know that the rotation is around this point, by a single angle of α. We can estimate this angle α by calculations only if we know the azimuths from the fundamental point in both systems. However, the problem is reduced to a one-variable minimum search, even if we don't know both azimuths. We shall seek the angle α, thus rotation parameters r_x, r_y and r_z, which provides the best fit between the coordinates throughout the whole base-point set. This minimum search can be easily carried out by iteration, even in any spreadsheet software.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

The scale factor reflects to the length measurement error at the baseline, or some minor mismatch in inappropriate using of length etalons. However, if we set the shift and rotation parameters, the scale can be estimated by another iteration step.

This method is slightly less accurate than the standard simultaneous parameter estimation, as the scale and the rotation is not fully independent from each other. However, the provided parameters can be interpreted separately geometrically. The standard estimation procedure is provided in the Appendix.

4.6 The correction grid (GSB)

These datum transformation methods, discussed in the above points, however, provide enough accuracy for GIS applications, are not capable for high-precision geodetic-engineering purposes. Even the BW-method can transform between the modern triangulation based datums and the WGS84 only with a remaining error of half meter in a region like Hungary. The survey geodesy needs much higher accuracy: ten centimeters inside cities and villages and 20 centimeters outside the settlements. Therefore, the standard geodetic applications use higher order (in the Hungarian practice, e.g. fifth-order) polynomials for the calculations. Similar accuracy can be obtained using BW-transformations based on only the base points in the vicinity of our study area.

However these applications are accurate enough, they have a considerable hindrance. There is no way to define, therefore, apply them in GIS packages. These software items usually do not support these methods, we can not define them by parameter input. The BW-parameter grid (a seven-channel image, each channel containing the different BW-parameters, changing from place to place) can be used in some GIS packages, but its definition is quite difficult. There is, however, an application, whose definition is easier and is supported by many packages, including the open-source ones (e.g. the GDAL-based Quantum GIS). This is the correction grid, which is often referred to as, according to its standard file extension, Grid Shift Binary (GSB).

Similarly to the AM- and BW-methods, this does conversion between geodetic coordinates on different datums. The correction grid itself is a grid, which is equidistant along the meridians and parallels. The eastward and northward shift between the two datums should be given at their crossings, in arc seconds. We can give, if we know it, the errors of the shifts at all grid points. However, it is not compulsory, if the errors are unknown, we can simply give zeroes for these data fields. The real shift values are derived from horizontal base points, whose coordinates are known in both the source and the target datums. The eastward and northward (or, with negative sign: westward and southward) shifts are handled separately: we construct two (or, with the error grids: four) different grids. The shifts, read in the base points, are interpolated at the pre-set grid points, in both grids.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

```
NUM_OREC 11
NUM_SREC 11
NUM_FILE  1
GS_TYPE SECONDS
VERSION NTv2.0
DATUM_F HD72
DATUM_T WGS84
MAJOR_F 6378160.000
MINOR_F 6356774.516
MAJOR_T 6378137.000
MINOR_T 6356752.314
SUB_NAMEOGPSH95
PARENT  NONE
CREATED 04-10-10
UPDATED 04-10-10
S_LAT     164520.000000
N_LAT     174960.000000
E_LONG    -82800.000000
W_LONG    -57600.000000
LAT_INC      180.000000
LONG_INC     180.000000
GS_COUNT 8319
-0.00000000  -0.000000000  -0.001000  -0.001000
-0.00000000  -0.000000000  -0.001000  -0.001000
```

Fig. 19. The header of the GSB data. The first 11 rows is the general header, the next 10 rows refer to the subset extents, then follows the number of data points and the point shift and error data itself.

These grids, combined with their meta-data (e.g. resolution, extents; Fig. 19) should be converted into a binary file. The file can contain even more grids, with different resolution. Therefore we can define a correction dataset providing higher accuracy in some important regions, while we have a general transformation with unified accuracy for a larger area.

It shall be underlined again that the correction grid provides connection directly between the coordinates in the source and the target datums. Neither the AM-, nor the BW, nor any other conversions should be used; applying the grid makes all of them unnecessary. The GSB-method aims the accuracy of a few centimeters in case of transformation between modern networks. It can also provide surprising accuracy also at geo-referencing of historical maps, if the control point network is sufficiently dense and properly selected (Fig. 20).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

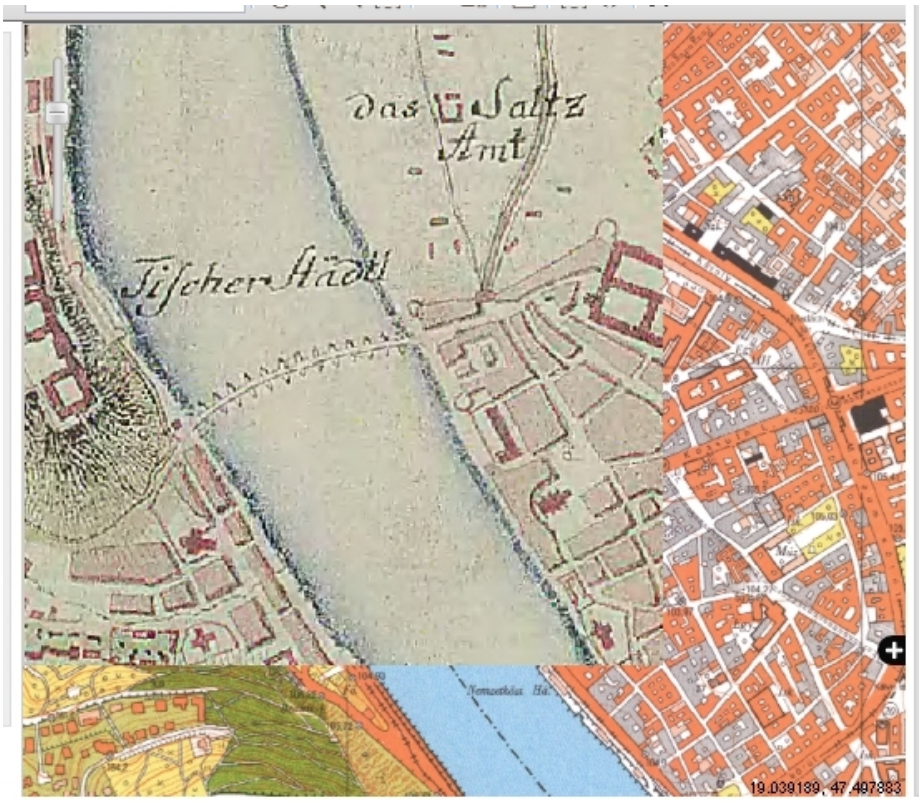


Fig. 20. The GSB technology provides excellent fit of old maps to new ones (center of Budapest in a 18th century map; note that the east bank of the river was considerably far from the other bank int hat time).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 5. Maps and projections

For the geographical information systems, maps are important data sources. In many cases, the map is represented by a scanned image, and the important data occurs as image information. Sometime we need to digitize a part of this information in vector format. To apply this information content it is needed to put this image into a pre-defined coordinate system, using the metadata and auxiliary information of the map. In this chapter we discuss the necessary meta- and auxiliary data and the methods to handle them.

Maps are planar, projected versions of the data on the base surface. Every map has a base ellipsoid, a datum (a representation of this ellipsoid), whose surface is projected to the plane of the map using some projection. The GIS packages usually know the equation information of the important map projection types. So, in this chapter we try to show the projections without to mention the projection equations.

5.1 Map projections and their parameters

For mapped representation, the surface of the Earth, the geoid, or rather its simplification, the ellipsoid should be projected to a plane. This procedure cannot be accomplished without distortion neither from the sphere, nor from the ellipsoid or from the geoid. Because of practical considerations (cf. Chapter 2), the geoid is not an input shape; the Earth is represented by sphere or ellipsoid in these computations. The procedure is called ‘projection’. The points of the surface of the sphere or the ellipsoid can be projected to a plane, to a cone or to a cylinder. The cone and the cylinder can be smoothed to the plane of the map (Fig. 21).

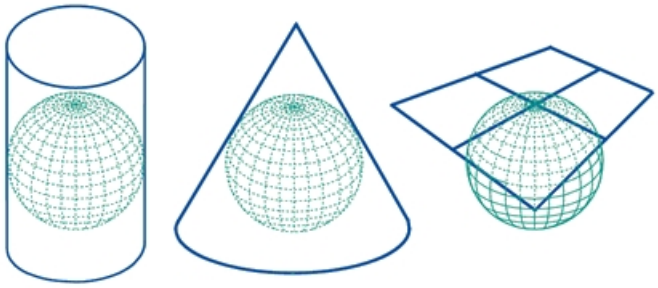


Fig. 21. To project the sphere to a plane: cylindric, conic and planar projections.

Projections are realized by projection equations. These equations create the connection between the map plane coordinates (projected coordinates) and the spherical or ellipsoidal coordinates. The most general form of the projection equations is:

$$E=f_1(\Phi,\Lambda,p_1,\ldots,p_n); \quad (5.1.1)$$

$$N=f_2(\Phi,\Lambda\, p_1,\ldots,p_n). \quad (5.1.2)$$

where E and N are the projected coordinates of a point, $p_1 \ldots p_n$ are the parameters of the projection.. Using this nomenclature (the Easting and Northing) we assume that the coordinates increase to east and to north, so the projected system has north-eastern orientation. This is true in most cases, however we discuss below the most important exceptions. The exact definition of the scale of the map is the number (usually much more less than one), which we have to multiply the resulted E and N coordinates with, to draw the map in the small piece of paper. As the Equations (5.1.1) and (5.1.2) are the direct projection equations, their inverse counterparts are

$$\Phi=g_1(E,N,p_1,\ldots,p_n); \quad (5.1.3)$$

$$\Lambda=g_2(E,N,p_1,\ldots,p_n). \quad (5.1.4)$$

The mathematical form of the functions f_1, f_2, and g_1, g_2 are based on the type of the projection. Sometimes their form is quite complicated, in some cases they are implicit functions. However, in the GIS practice it is usually not necessary to work with these equations or even to know them – in most GIS packages or GPS receivers, they are pre-programmed. All we have to know is to handle them, giving them correct parameters. The projection equations

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

can be assumed to be exact; the successive application of the direct and inverse projection equations provides the input coordinates with an error less than a millimeter.

Parameters $p_1, \ldots, p_n$ are based on the realized projection and their number n is a function of the projection type. In most cases n=5; however in some early projections e.g. the Cassini-Soldner, n=4; while in case of complicated ones, as the oblique Mercator or the oblique conic projections, n=6. These parameters should be known by the software – or, which is much more assuring, by ourselves. Let's see, what parameters are needed for the projections.

Every projection has a so called projection origin or in other words, projection center. This point is the touching point of the plane/cylinder/cone and the ellipsoid. If the touching occurs along a line in cylindrical symmetric case (the central line of the projection), a point of this line should be assigned as projection center. The ellipsoidal latitude and longitude of this point are two necessary parameters.

The projected coordinates of the projection origin are the third and fourth mandatory parameters. As a default, they are both zeroes, however for practical considerations, they are often set to different values, e.g. to obtain positive or distinguishable coordinates throughout the mapped area. Because of this shift, these parameters are called FE (False Easting) and FN (False Northing), and they are expressed usually in meters.

A further, fifth parameter is the scale factor. In some cases, the plane/cylinder/cone is not placed in toughing but in secant position, in order to enlarge the low-distortion area around the projection center or around the central line of projection (Fig. 22). The scale factor is 1 in touching cases, and it is usually less than one, showing the reduction ratio. The only exception is Ireland, where the scale factor is more than one, because of historical reasons (to have a similar scale at the center as it was provided by the British system). In case of conic projections, instead of the scale factor, the standard parallels, where the cone cuts the base surface, can be also given.

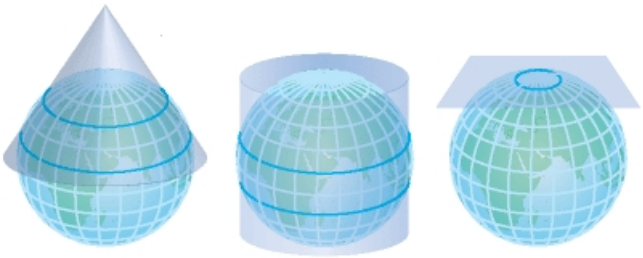


Fig. 22. If the cone, the cylinder or the plane is placed to secant position, the length distortion is zero along the secant lines.

In case of oblique cylindrical projections, the projection origin is on the central line of the projection. It can be its farthest point from the equator (called Laborde projection) or its intersection with the equator (called Hotine projection). In general case, however, any point of the central line can be a projection center; that's why a sixth parameter is needed for this projection type: the azimuth of the central line at the center (Fig. 23).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

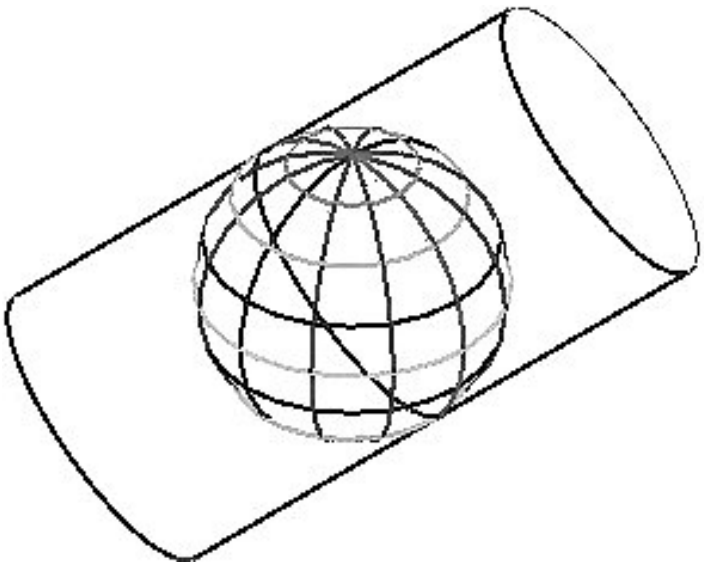


Fig 23. The tangent line is an oblique great circle in case of the oblique Mercator projection. The length distortion is minimum along this line.

It is important to mention that a projection means the type of the projection, but also its realized form, when the above mentioned parameters are fixed. On the maps, we see coordinates of realized projections.

The standards, describing the projecting procedures, often mention double projections. In these cases, the projection equations can be written in two steps: first projecting from the ellipsoid to an aposphere, then in the second projection, from the aposphere to the plane/cylinder/cone. Obviously, this was needed for the computations of the pre-computer age. In most cases, this raises no practical problems; the equations used in the GIS packages are good approximations of the double projections. If the centers of the two projections are not in the same place, the method of approximate projections (see below) can be applied.

There are a few tens of projection types that are used anywhere in the world. However, only a few of them are widespread used. Here we discuss the three most important ones; a cylindrical, a conic and a planar projection, the transversal Mercator, the Lambert conformal conic and the oblique stereographic ones, respectively. All of these discussed types are conformal projections.

At the transverse Mercator projection, the axis of the cylinder is in the plane of the equator. The origin of projection is at the equator. If the scale factor is the unity (e.g. in case of the former Warsaw Pact's Gauss-Krüger projection or the WWII German military grid), the cylinder touches the base surface along a meridian, this is the central line of the projection. In this case, the low-distortion are, where the length distortions remain under 1/10000, expands to about 180 kilometers on both sides of the central line. It can be extended by applying a scale factor less than one: e.g. in case of the UTM (Universal Transverse Mercator), where k=0.9996. The False Northing is usually (but not exclusively!) set to zero, while the False Easting is defined to avoid negative coordinates, e.g. FE=500000 m.

In case of the Lambert conformal conic projection, the axis of the cone is in the semi-minor axis of the base ellipsoid. The central line of the projection is the parallel line where the cone touches the base surface (also known as normal parallel). The projection origin is a selected point of this parallel. This projection is usually used with reduction (scale factor is less than one). This projection can be defined by the projection origin and the scale factor or by the projection origin and the two parallels where the cone cuts the ellipsoid (standard parallels).

In case of the oblique stereographic projection (also known as Roussilhe-projection) we put a plane to a selected point of the base surface, perpendicular to its normal direction at that point, which is the projection origin. If the scale is unity, the low-distortion zone is a circle with a radius of about 127 kilometers around the projection origin.

In case of every projection, there is a zone with low distortion. As we have seen, in case of the transverse Mercator, it is a stripe along the central meridian, at the conformal conic projection this stripe is along the normal parallel, while it is a circle around the origin, using stereographic projection. If the mapped area extends beyond this range

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

(in case of larger countries, or even the whole surface of the Earth), there usually are more projections defined with different origins, for different zones. The projection types are usually the same in these zones but the parameters are different, realizing different projections of course. In France, there are 4 zones of the Lambert conformal conic projection, each zone is elongated from west to east along the respective normal parallels. Using transverse Mercator projection, Austria defined 3, Germany 5 (prior to the territory losses, 7) zones, along central meridians. The zone system of Poland uses 4 stereographic and one transverse Mercator projections. These groups of projections, used as zones to map a larger area, are called projection systems.

In case of smaller countries, one zone is often enough to make low distortion maps. In the Netherlands, one stereographic projection is defined. The situation is similar in Romania, however this country is larger than the low-distortion area of the stereographic projection. The shape of the countries or regions suggests the projection type to be selected for the only zone. If the area is elongated from north to south (e.g. Chile, Portugal), the transverse Mercator projection is a good choice. For east-west elongated countries, e.g. Belgium or Estonia, the natural selection is the Lambert conformal conic projection. Switzerland and Hungary opted for oblique Mercator for the same reason, however for both countries the conic projection would have been also a good or even better possibility. The territory of Czechoslovakia between the two world wars could been mapped using one zone just by an oblique conic projection.

In the projected maps, the points of the same Easting or same Northing values are linear lines. The map grid or projection grid consists of these lines. The meridians and parallels are usually curves in the maps, just some distinct meridians and parallels can be linear. At every map points, there is an angle between the grid north and the geographic north; it is called the meridian convergence. Usually the meridian convergence is varying from place to place (an exception is the true Mercator projection where it is zero everywhere). We have to know, even seeing a low-scale map without projection grid indicated, that the invisible projection grid is there, behind the curves of parallels and meridians.

5.2 Transformation between projected coordinates

For the correct transformation from grid coordinates of a system to another one, not only the projections and their parameters should be known in both systems but also the datums of them. In ideal case, the two grids are interpreted in the same datum. However, in most cases, they are not.

The three possible ways of the coordinate transformations are shown in Fig. 24. Of course, in the datum is the same, the datum transformations are not needed. However, if the datums are different, we shall choose one of these three ways.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

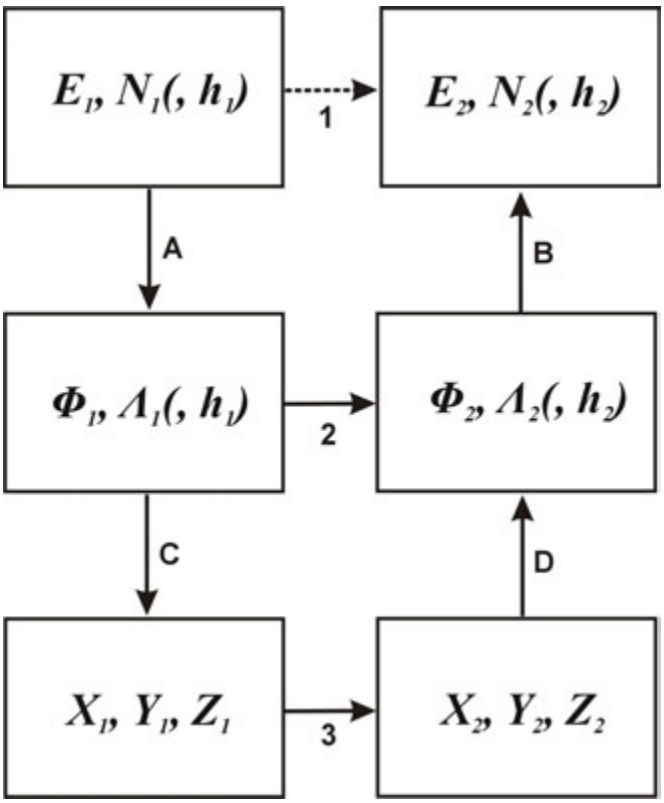


Fig. 24. The flow-chart of the coordinate transformation from the source system (indicated by the '1' index) to the target one (indicated by '2's). 1: direct polynomial transformation; 2: the abridging Molodensky transformation or the GSB shift; 3: Bursa-Wolf transformation.

The direct transformation is based on higher-order polynomials, whose several parameters are estimated from the coordinates of base points in both systems. Here this method is not discussed, as – albeit it is the most accurate method – its parameters cannot be supported in most GIS packages.

The second way starts with the usage of the inverse projection equations: we calculate the ellipsoidal coordinates in the first datum from the grid coordinates. In second step, using the abridging Molodensky formulas, we transform the ellipsoidal coordinates from the first datum to the second one. Here we shall know that this transformation can be accomplished by correction grids, too. In the final step, we transform the ellipsoidal coordinates in the second datum to the final grid coordinates, using the direct projection equations of the second grid. The errors of this method are because of the ambiguity of the datum transformation, the distortion between the two geodetic networks; the projection equations can be accepted as exact ones. This method is used by the GPS receivers with one difference: the input data is given in WGS84 ellipsoidal coordinates and the method starts with their transformation to the local datum, omitting the first step, the usage of the inverse projection equations. The method is also supported by all GIS packages.

If we know the Burša-Wolf type datum transformation parameters between the two systems, and our software supports this kind of transformation, it needs geocentric coordinates as input data. So, two more steps are applied: transformation from ellipsoidal coordinates to geocentric one and *vice versa*. The direct case is part of the trivial trigonometry, however computing the ellipsoidal coordinates from the geocentric one is surprisingly difficult in ellipsoidal case. Its closed formulas, or more precisely, its algorithm, was first given in 1989 (the Borkowski method), and its development is still an important research direction. However, the closed Bowring-formulas can be applied. The horizontal error of this approximation is below 1 centimeter; that's why this is used in GIS packages. Therefore, the error of this way is caused again by the ambiguity of the datum transformation.

In our practice, it is very rare when we shall make these computations ourselves. They are programmed in our software or GPS device, the only necessary inputs provided by us are the projection and datum parameters, if they are not pre-set in the application. However for us, the specialists, it is worth to know what is inside the 'black box'.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

5.3 Substituting projections

Sometimes we face the problem that our GIS module does not know the projection equations of one of our used systems. More frequently, we shall geo-refer a map of unknown projection. This sub-chapter discusses the procedures used in these cases.

The less-used projections are not necessarily programmed in GIS software packages. The Hungarian EOV with its double projection or the Czecho-Slovak Křovák system is often not supported in these packages. A simple user cannot make this programming work, even using the software development kits. However, it is always an option, to choose another projection type from the supported ones and to give its parameters, keeping the difference of the original and this, so-called substituting projections as low as possible. In the following, some examples are shown to use substituting projections.

A) Substituting the Hungarian EOV grid by Hotine- or Laborde-projection

The standard of the EOV grid contains a double projection: first from the IUGG67 (GRS67) ellipsoid to the aposphere, then from that surface to the cylinder. The normal parallel of the first projection is different from the latitude of the origin of the second one. In the GIS packages, the Laborde- and Hotine-projections (sometimes called also RSO; Rectified Skew Orthomorphic, or simply Oblique Mercator projection) are practically such double projections, where the origins of the two successive projections are the same. The specific case of the EOV grid is not programmed, and it is not easy to implement by ourselves.

Therefore, to use the EOV grid, we shall use substituting projection. According to analyses, this double projection is much more sensitive to changing the origin of the aposphere → cylinder step than to the position of the aposphere. So, the origin of the first projection step can be modified, to make the two origins the same. This approximation results a new, substituting projection. However, the difference between the grid coordinates provided by this method and by the original standard equations, are less than 0,2 millimeters throughout Hungary (the valid area of the projection). This makes the method applicable not only for GIS applications but also for high-accuracy geodetic use, too.

It is easy to give parameters for the Laborde-projection to use this approximation: besides the coordinates of the, now united, projection origin and the scale factor, the azimuth of the central line should be given at the origin, which is 90 degreed. The case of the Hotine-projection is slightly different: from the False Easting of the origin, we shall subtract the distance of the origin and the equator along the central line.

B) Substituting the Hungarian EOV grid by Lambert Conformal Conic projection

The Laborde- and Hotine-projections are not widespread used and in some GIS applications, they are not implemented. However, the Lambert Conformal Conic (LCC) projection is very common and known in most packages. Therefore it is worth to seek a parameter set for the LCC to approximate the EOV grid coordinates. This concept, first published by Gy. Busics, was that the central line of the LCC projection follows a parallel, which almost follows the central line of the EOV's oblique Mercator projection. The difference between these two lines is up to a few meters in Hungary. In the practice, the origin coordinates and the scale factor defined in the above point A) can be interpreted as parameters of a LCC projection. The accuracy of this approximation is a few meters in Hungary, which suits fine the aims of GIS applications.

C) Substituting the Hungarian EOV grid by Transverse Mercator projection in small area

Some GPS receivers (especially the older Garmin ones) allow the user to give the parameters of the Transverse Mercator (TM) projection, while defining a user grid. It was shown by B. Takács that – albeit the central lines of the two projections are perpendicular – it is possible to use position-specific parameters in any area with the radius not greater than 15-20 kilometers with the accuracy of GIS needs. The procedure is the following:

- To measure by GPS the longitude of a central point of the area, with known EOV grid coordinates (E_{EOV}, N_{EOV});
- To define a TM projection with the origin at the section of the above measured meridian and the equator, and with the scale factor of 0.99993;
- To read the coordinates of our selected central point (E_{TM}, N_{TM}) in this projection;

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

- The False Easting and False Northing parameters are:

$FE=E_{EOV}-E_{TM}$

$FN=N_{EOV}-N_{TM}$

D) Substituting the Budapest-centered Stereographic grid by Roussilhe-projection

The problem of the Budapest-centered Stereographic projection is of the same type as it was mentioned in point A), the double projection, with different origins. In this case, the difference is much more than at the EOV grid. However, the procedure is the same: we omit the ellipsoid → aposphere projection and giving parameters to the oblique Stereographic (Roussilhe) projection, based on just the ellipsoidal coordinates of the origin at the second projection. Because of the larger latitude difference, the accuracy is lower here, however does dot exceed 2 centimeters throughout Hungary.

E) Substituting the Czecho-Slovak Křovák grid by Lambert Conformal Conic projection

The Křovák grid is based on an oblique conformal conic projection, used exclusively in the former Czechoslovakia and its successor states. Many GIS software packages do not support this projection (or support it just because it was programmed to handle this very grid). The central line of the projection is east-west directed in the southeastern edge of Subcarpathia (Ukrainian region, formerly a part of Czechoslovakia). Going westward, it more and more leans to north. This central line cannot be defined by any other projection, so the approximation can be done only with considerable error.

It is possible to define different Lambert conformal conic grids for Slovakia and the Czech Republic, with different parameter sets. As its central line is closer to the original one in Slovakia, here the accuracy is better (average error is 6 meters, the maximum is 12 meters in Slovakia). This is acceptable for geo-referencing e.g. the 15-meter resolution Landsat ETM satellite images or topographic maps with the scale of 1:25000 or less, but not for more accurate purposes. The average approximation error in the Czech Republic is 40 meters while its extreme maximum is 82 meters. This enables the geo-reference of 1:100000 scale maps, the 90-meter resolution SRTM elevation dataset or the 250-meter resolution MODIS satellite imagery.

Finally, if we have no meta-data or reference about the projection of the map to be geo-referred, we have to choose and parametrize a projection, whose latitude-longitude grid fits well to the one of the original map.

5.4 Sheet labeling system of maps, the geo-reference provided by the labels

Map systems, covering larger regions or the whole surface of the Earth, are often consist of sheets, covering smaller parts of the target area. In this case, the sheet labels provide information about the location of the area, mapped in the respective sheet. Thus we can make a mosaic from them without following the projection or the latitude-longitude grids. Besides, there are map systems without any grid reference; in this case the geographic or the projected coordinates of the corners can be calculated from the sheet label.

The borders of the area, mapped in a sheet, are following parallels and meridians, or projection grid lines. In the first case, the shape of the sheet is an arc trapezoid, while in the second one it is a square or a rectangle. The sheet number exactly gives the coordinates of the corner points.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 25. Map sheet in Hungarian national grid without latitudes or longitudes indicated: the sheet boundary follows the projection lines.

Fig. 26. The map sheet boundaries of the Gauss-Krüger system follow the parallels and meridians.

In Hungary, sheet border of the civilian topographic maps are following the grid lines of the EOV (the national grid), without any geographic coordinate indicated (Fig. 25). The sheets of the Gauss-Krüger type military map system or the old Stereographic system are bordered by parallels and meridian arcs (Fig. 26). If we have not a topographic but a derived map, whose sheet labeling system follows the one of the topographic maps, we can use the corners as control points, even if no coordinates are given in the map (Fig. 27).

The situation is similar at the sheets of the old (second) military survey sheets of the Habsburg Empire (Fig. 28). We have no coordinates indicated (Fig. 29) in the nice and detailed 1:28800 scale sheets of the map system. However, knowing the labeling system and the physical extents of the sheet area in the field, we can easily compute the grid coordinates of the corners in its native projection. Thus, the sheet corners can be used as control points, without seeking identical terrain points in the map.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

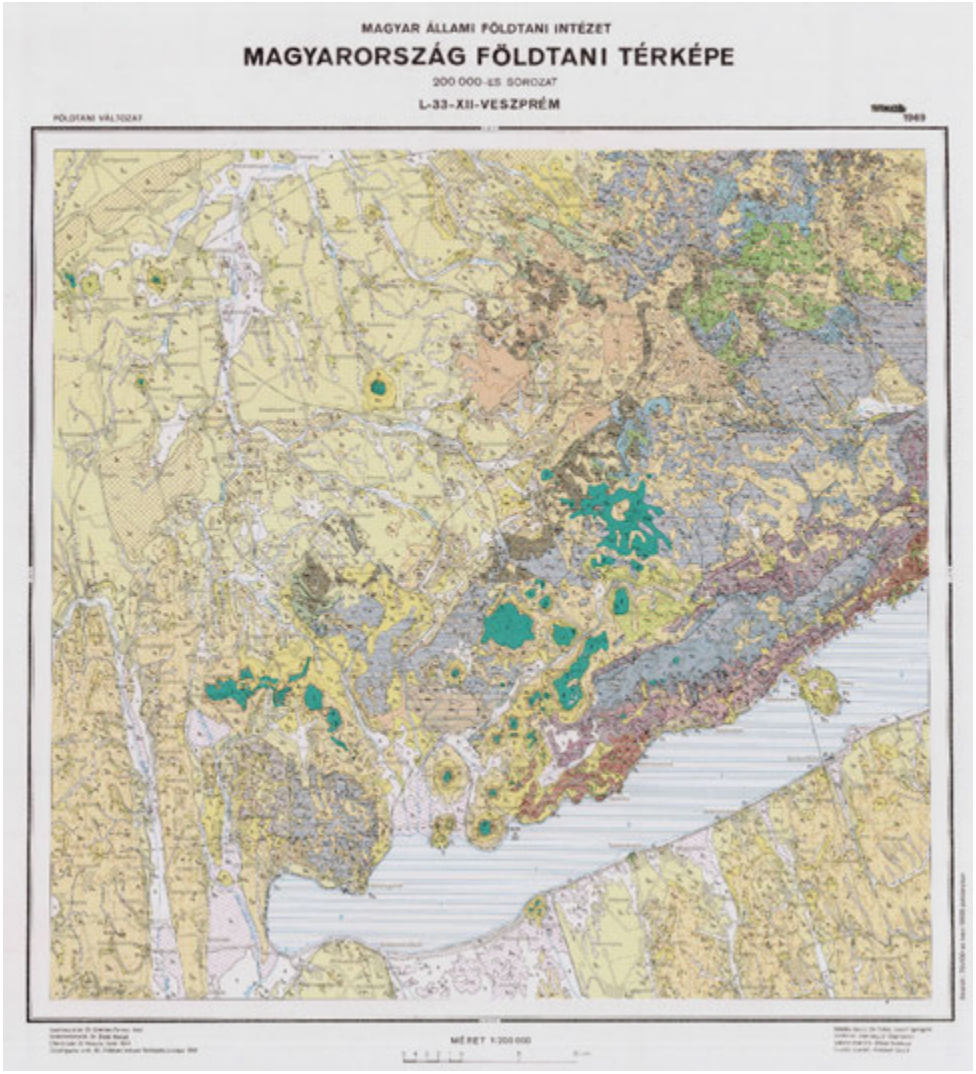


Fig. 27. If no coordinates are indicated in a map, the sheet label (here: L-33-XII) can refer to the exact location of the corners.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 28. The sheet system of the Second Military Survey of the Habsburg Empire.

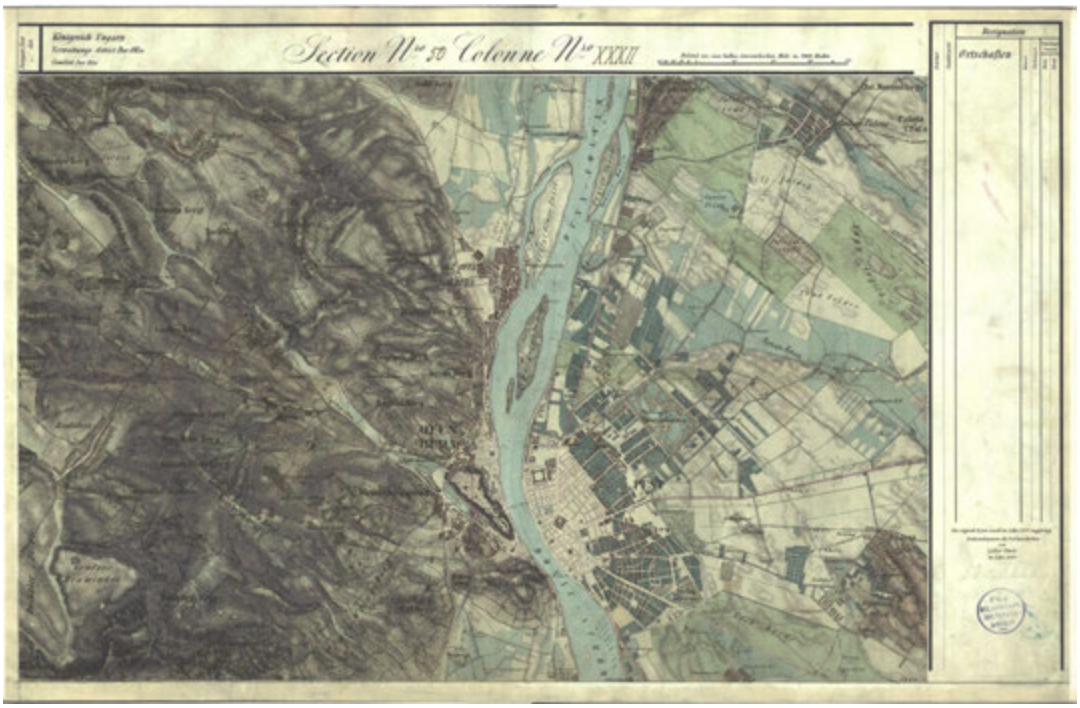


Fig. 29. There is no coordinate indication in the sheet of the Habsburg Second Military Survey; the geo-reference is given again by the sheet label (here: „Section 50 Colonne XXXII').

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 6. Geo-reference of the maps

Geo-referring of the maps and cartographic databases means that we give geo-reference to all pixels of the scanned raster data. In the beginning, the pixels of the scanned raster image have only image pixel coordinates, valid in the plane of the image. In this coordinate system, the origin is usually at the upper left corner of the image (this can be different in some GIS software packages) and every pixel means increases one unit both in horizontal and vertical directions.

During the geo-reference, we define GCPs (Ground Control Points), whose image pixel coordinates and grid coordinates are all given.

6.1 The geo-reference and the rectification

We can follow different ways while defining our GCPs. The common part of these method is that first we have to define the grid of the target coordinate system: the geodetic datum, the projection type and projection parameters. It is important that – if it is feasible – the native projection and geodetic datum of the original map should be used, and not the one that is required for the result. Also, we have to choose the method, the program should use while fitting the grid coordinate system to the raster image. The most frequently used methods are polynomial ones, with different orders, such as:

- Linear
- Quadratic
- Cubic

In case of the linear fitting, a square grid, usually somewhat rotated, is overlapping to the original image. The quadratic and cubic methods use second and third order polynomial fitting, respectively. Using these methods, better fit at the given GCPs can be achieved, however the errors occurring in between the GCPs can be considerably larger. Besides, these methods need to define more GCPs than the linear one. If possible, choose always the linear fitting method.

Besides the above discussed polynomial methods, the triangulation-based fit is also a frequent option. It provides zero error at the GCPs, while the grid in between them is fit using different linear methods in different triangles, drawn based on the GCP set. Albeit the mathematically optimum accuracy of the method has, the resulted image is usually somewhat awkward.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

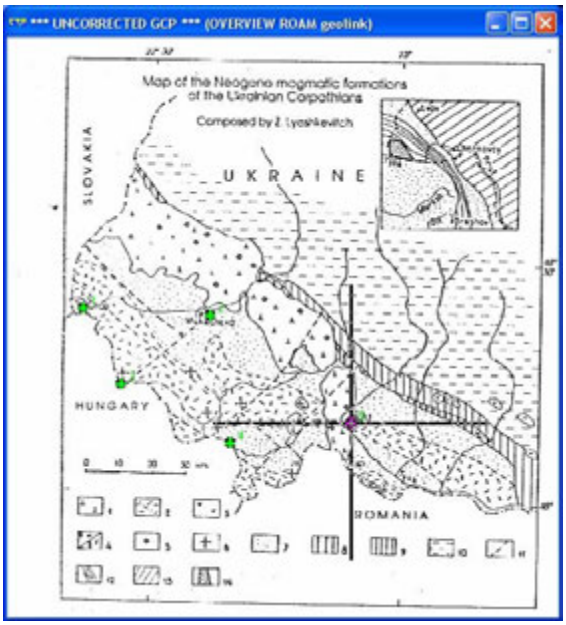


Fig. 30. Selection of ground control points (GCPs) in a map without coordinates: choose known points, here: cities, with their coordinates.

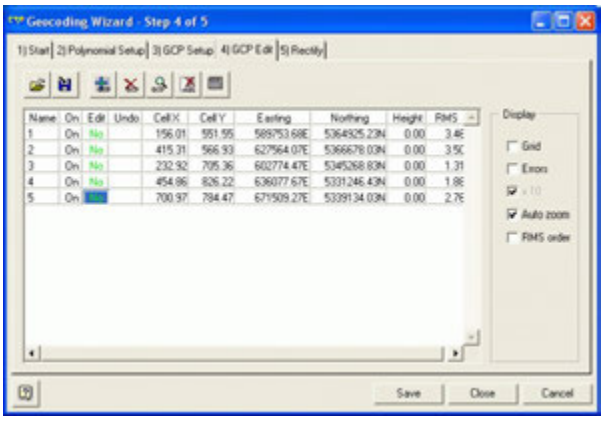


Fig. 31. The image and projected coordinates of the selected GCPs. If the magnitude of the fitting error (Column named 'RMS') is around a few pixels, the fit is acceptable.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

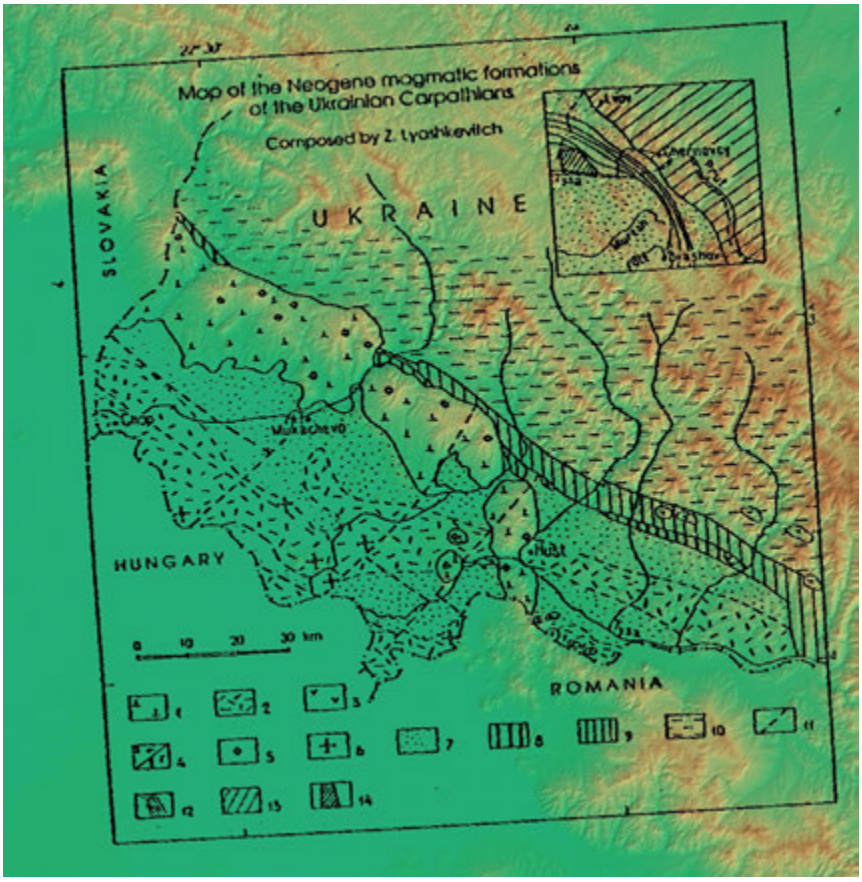


Fig. 32. Result of the rectification: a figure of a geological paper on an elevation model.

How to define the ground control points? The most widespread used, however the most uncertain method is to indentify some terrain objects in the scanned map and acquire their coordinates from auxiliary databases (Figs 30, 31 & 32). Why the ambiguity of the method? It is because the terrain points are usually less-exactly positioned in the maps than the geodetic base points, providing the maps' frame. The positions of the terrain object symbols are also affected by the map generalization; this could result a positional ambiguity of 1-2 millimeters. Moreover, this method provide a 'temptation' to omit the original projection of the map, which is a considerably error source, mainly while geo-referring medium or low scale maps.

- If there is a coordinate grid in the map, its crossing points (or the crosshairs, if they are used instead of the full grid), are the ideal control points.
- If no grid lines or crosshairs are indicated, the crossing points of the parallel and meridian lines can be also used. However, in this case we have to compute the projected coordinates of these points from the geographic ones. Most GIS packages can do these calculations (if not, this is the only application in the practice, when we do need the knowledge of the projection equations). Using only the geographic coordinates leads to unacceptably high errors!
- If even the latitude-longitude grid is not provided, we can use the corner points as GCPs if the sheet label bears the geo-reference. This is the case of many Hungarian geological or forestry maps, prior to the introduction of the EOV grid.
- Only if there is no such geo-reference, we can use identified terrain points as GCPs.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

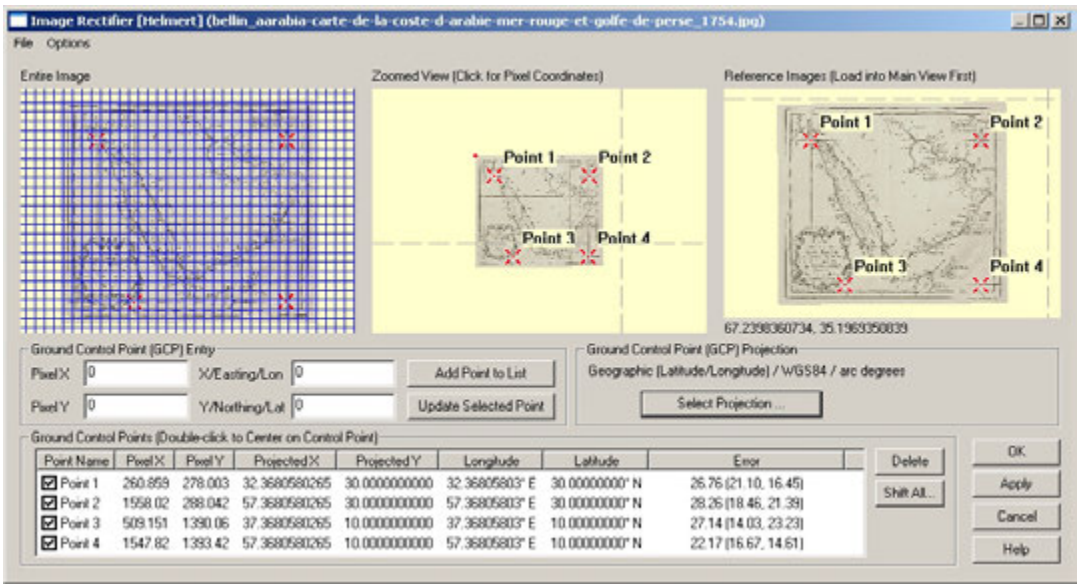


Fig. 33. As our map contains coordinate or latitude-longitude grid, its crossings provide the best GPCs. Here: because of the usage of Ferro prime meridian, the longitude values are nor round numbers. The fitting errors are too high because of the ellipsoidal coordinates given for the GCPs.

As the control points are defined (and switched on) in the rectification algorithm, the software provides their errors: the horizontal difference between their defined position and their calculated position based on the fitted grid (Figs 33, 34 & 35). The error is usually given in pixel units; errors below one – or, maximum two – pixel(s) can be accepted. In case of historical maps, the acceptable error range can be somewhat higher. However, if we have blunders at some or more GCPs, first we should check, whether we typed wrong grid coordinates or changed the easting and northing values. If the error still remains, check the identification of the used terrain objects. The dislocation of the GCPs is also important: they can't be in or near one line in the map. The aim is to have 4-6 (in case of quadratic or cubic fit, much more) GCPs, well spread in the map area and with low error.

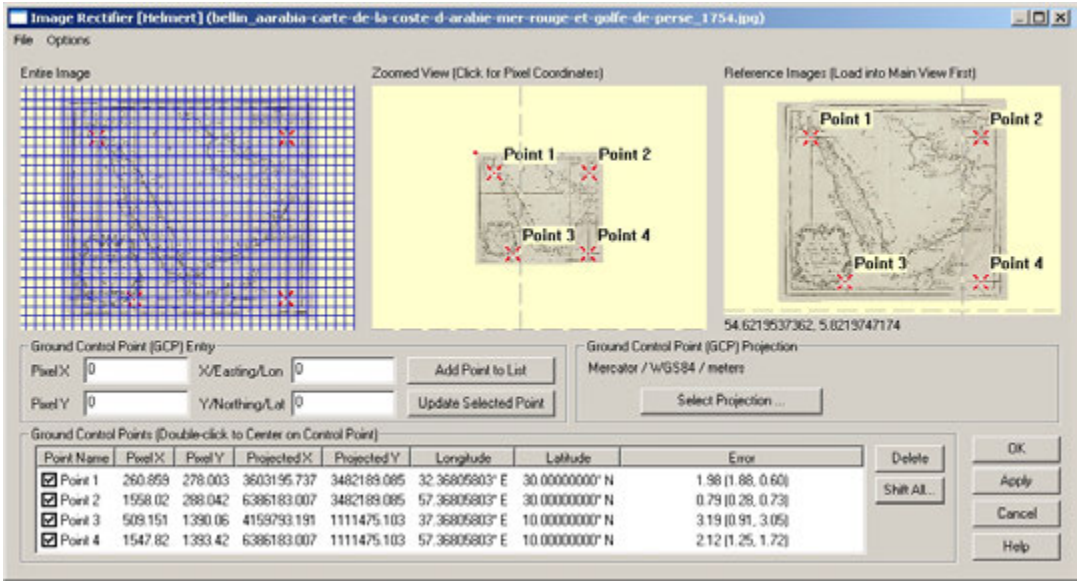


Fig. 34. The ellipsoidal coordinates are transformed to projected ones (here: in Mercator projection), and the errors are almost eliminated.

The next step is the rectification. This is a resampling method: the computer puts the grid, calculated from the image and grid coordinates of the GCPs, on the image and gives the raster values in this new image from the original one. The result is a raster image, whose rows and columns follow the east and north directions of the target grid. The resampling itself can be done by three methods:

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

- Nearest neighbor (NN)
- Bilinear
- Cubic convolution

The NN methods means that the pixels of the result image obtain their values from the original image pixel, whose center is the nearest to the target pixel center. This is the fastest procedure of all the three ones. This algorithm guarantees that there will be no other pixel values in the resulted images than they were represented in the original one. Therefore, if the pixel values refer to categories (e.g. classes in classified images) we shall choose this method for the rectification.

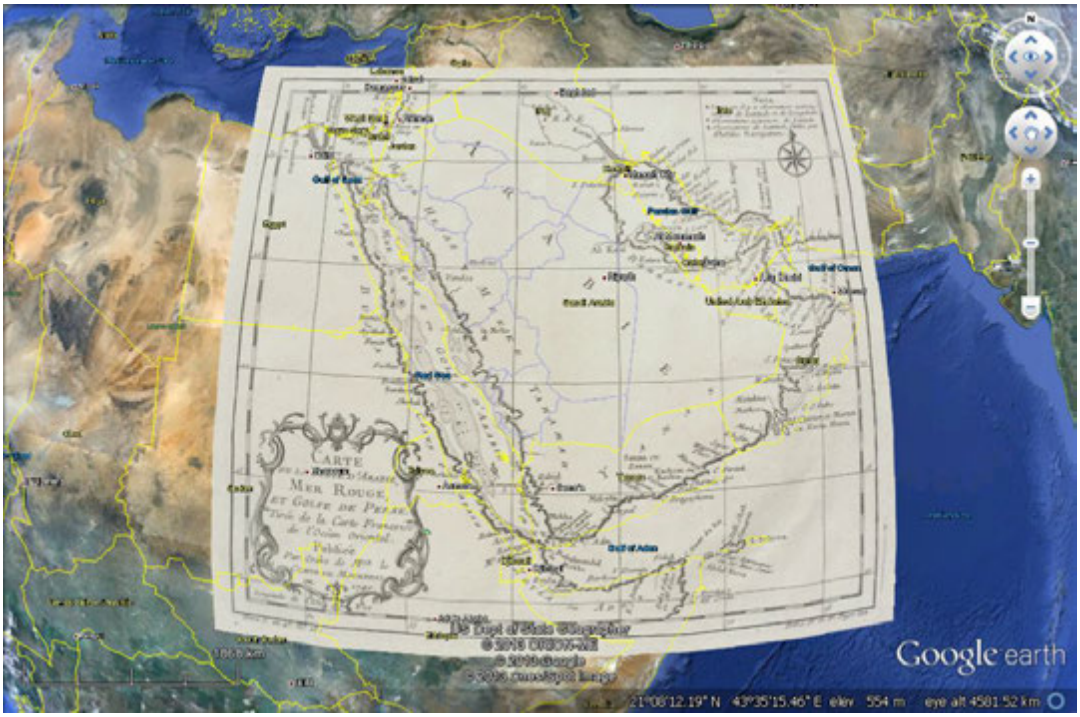


Fig. 35. Result of the rectification: the 1754 map of Arabia by Bellin (credit: David Rumsay Map Collection) on the Google Earth. Errors indicates the accuracy of the survey behind the map.

The bilinear algorithm means that the target pixel value is given by a bilinear interpolation, based on the original pixel values around it. This is the advised method, when the resolution of rectified image is considerable higher than the original one.

The convolution method provides the target pixel values based on territorial averages of the pixel fragments connected to them in the original image. In case of images with continuously varying pixel values (e.g. scanned images or satellite imagery), this method provides finer but slower solution than the NN-resampling, if the target pixel size is around or larger than the original one.

The GIS software packages stores the position of the resulted image in its coordinate system, usually in their own format. However, there is a quasi-standard description format, known by many GIS packages, this is the World File. The World File can be used to describe the position of TIFF or JPEG-type images, or even of compressed files e.g. MrSID or ECW. The very simple structure of this file is the following. It contains six values:

- The Easting increase while step one pixel to the right
- The Easting increase while step one pixel down
- The Northing increase while step one pixel to the right
- The Northing increase while step one pixel down

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

- Easting value of the upper left image corner
- Northing value of the upper left image corner

There is no absolute rule for the file extension. For a JPG image, it is a file with a same name and an extension of JPW or JGW. For TIFs, it is TFW, for SID, it is SWF. The World File does not contain any meta-data: neither the map grid and datum, nor their code are not stored. Therefore we have to know ourselves these pieces of information. Also, we have to mention that using the first four data segments, the rotation of the coordinate frame can be also handled.

As a metadata, the target coordinate grid and geodetic datum should be stored for the rectified (resampled) image. Using this, the GIS packages are able to convert to another map grid, assuming that its projection and datum parameters are known to it, according to the Chapters 4 & 5.

Example: let's assume that we have a map in the Warsaw Pact Gauss Krüger (Zone 34, Pulkovo 1942 datum) grid, and we need to convert it the the Hungarian EOV system. The following steps are to be taken:

1. set the coordinate system for the ground control points (Pulkovo 1942 datum, Gauss-Krüger Zone 34 grid);
2. define the ground control points with their grid and image coordinates;
3. rectify the scanned image to the Gauss-Krüger coordinate system, and
4. transform the result image to the EOV coordinate system.

We shall discuss again, that it would be incorrect, leading to considerable error, if we used GCPs with EOV grid coordinates. In the Gauss-Krüger system, the points of the lines, that are straight in the EOV system, form curves. If our target area is relatively small, e.g. a few kilometers, this is almost undetectable. However, at several ten or hundred kilometers distance, this deflection could be up to several ten meters and cannot be corrected or calculated. Following the above steps a)-d), this error can be avoided.

6.2 The projection analysis and the deliberate selection of projection

In many cases, we don't know the exact projection of a map or scanned cartographic database. However, when rectifying them, we shall define a coordinate system. For this we shall look up, or, if this is not feasible, we shall estimate the projection type, the projection parameters, and, if needed, the geodetic datum.

Before declaring the coordinate system of a map unknown, we shall try to look up its metadata in the literature. We can seek for references in the text of the map frame (Fig. 36). In some cases, the projection type is given, while the parameters are not. In many cases, we find a reference to a national grid, without its details; we can seek for detailed reference in textbooks or by internet search. Topographic maps are hardly made in 'unresolved' coordinates systems (however, the 1980's Hungarian hiking maps provide interesting exercise for the analyzer). The national grid and its datum of the area provide always a good starting point, even it is not referenced. If there were more standard grids of the country in the questioned time frame, all of them are worth to try. Sometimes the sheet labeling system helps to select the correct projection.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

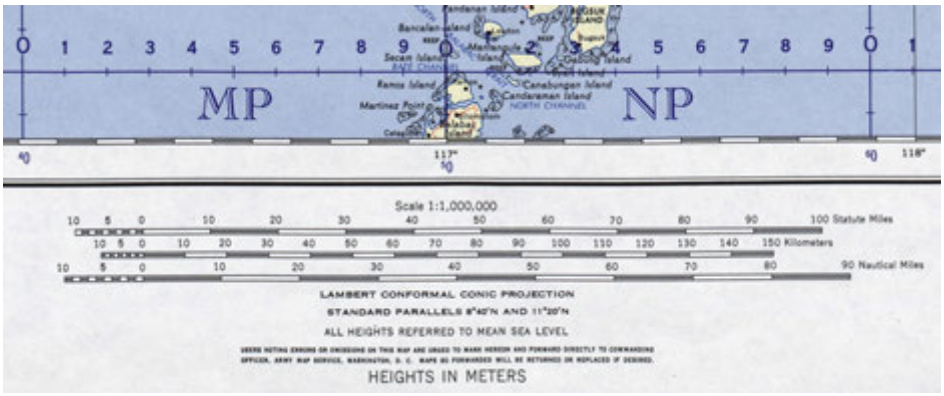


Fig. 36. In many US topographic maps not only the projection type (here: Lambert Conformal Conic) but also the projection parameters (here: the standard parallels) are given. Lucky case!

For example, the possible coordinate systems of a map or cartographic database in Hungary are: EOV, Gauss-Krüger grid, Budapest-centered Stereographic grid. The EOV has been introduced in 1975; therefore prior to this date there were no maps in this system. The Gauss-Krüger grid was used even for civilian purposes from the 1960s. However, the coordinate system was secret: in these maps there are either no coordinate reference (the geographic coordinates of the corners can be computed), or a Stereographic grid is provided. If the sheet label starts with 'L/M-33/34' (the dividers mean alternatives), the map is in Gauss-Krüger system. The sheet label of the 1:75000 scale Stereographic maps is of four digits. The label of the 1:25000 sheets of this system is completed by a hyphen and a number 1-4, the geographic longitudes are often given from the Ferro prime meridian. The above mentioned Hungarian hiking maps are in Gauss-Krüger system but they are rotated to magnetic north and their kilometer grids follow no standard system.

If the area of the map with unknown projection is small, it is practically not important, which projection is selected for the rectification. Within 10-20 kilometer distances, the deflections are not exceeding our aimed accuracy of about 5 meters. In this case, the selection of the geodetic datum is important; one base point is enough for its parametrization (see Point 4.5). The size and shape of the ellipsoid is not really important, its dislocation should be set to optimum horizontal fit.

If the scale of our map is low, and it shows a relatively large area, the situation is lucky from a point of view, that the accuracy of the map reading, the half map millimeter represents several hundred meters on the terrain. Therefore, and projection can be selected that approximates the real map projection with this, very large, error margin. Besides several hundred meters of accuracy level, the selection of the geodetic datum is either less important. For the selection of the projection, we shall analyze the latitude-longitude grid.

In mid-latitudes (e.g. in Europe), the latitude lines in the overview maps are often more or less concentric circles, while the meridians are more or less straight lines, pointing to the pole, while the angles between them are equal to each other. In this case, we can use a Lambert conformal conic projection, even if it is not the native projection of the map.

A common error is, when the rectification is made by geographic coordinates of the cross-sections of the parallels and meridians. This is an incorrect procedure, resulting large errors. The correct procedure is to analyze the latitude-longitude grid, estimate a projection type and estimate its best parameter set. Upon completing this, the coordinates of the cross-sections should be transformed to this newly defined projection for ground control point definition. When we use the real projection of the map, the rectified result is a rectangular image, without distortion at the corners (Figs. 37 & 38). In case of small-scale maps, the geodetic datum selection is not important.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

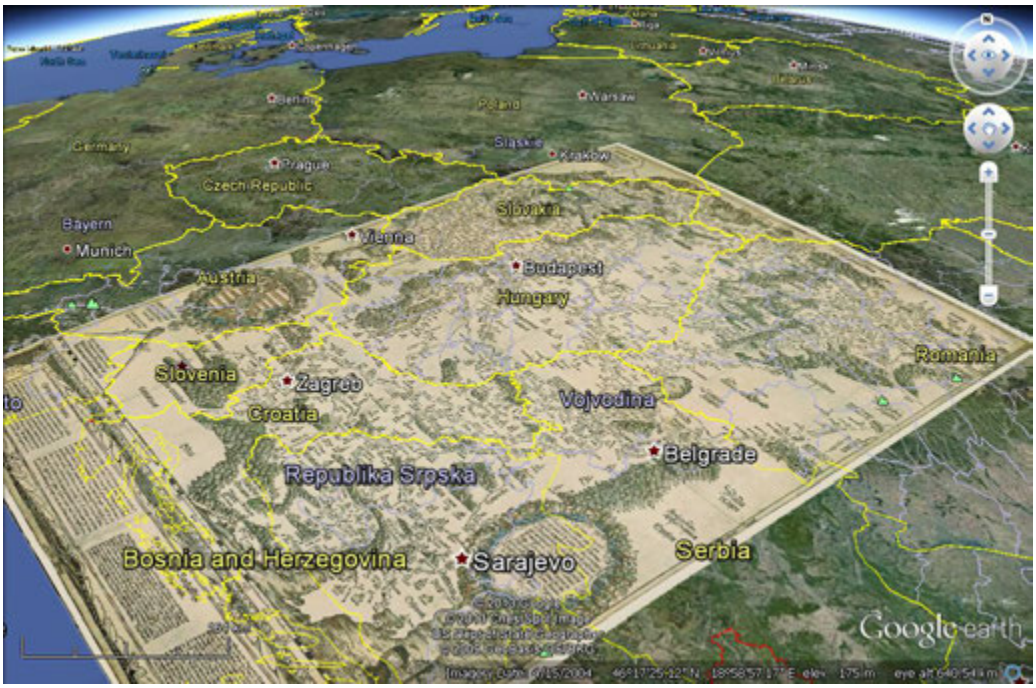


Fig. 37. The map of Hungary by Lazarus (1528), rectified in its own projection: the result is an undistorted rectangle.

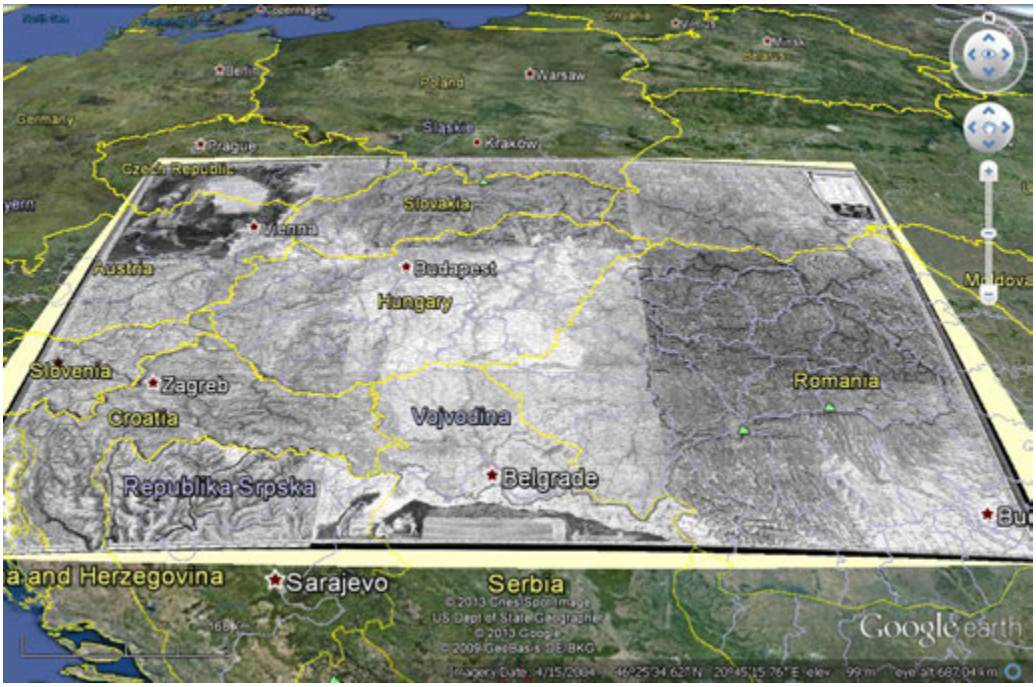


Fig. 38. The map of Lipszky (1810): the geo-referred map is an exact rectangle, the yellow zone around it is the envelopping arc trapezoid.

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

Chapter 7. Vertical geo-reference

In the above chapters we discussed only the horizontal position and geo-reference of the distinct points. In these calculations the vertical position of the points had no role. Indeed, the vertical location of the points hardly has influence on the resulted horizontal coordinates. The effect is less than a centimeter within the several hundred of kilometers of altitude, and this is much more than any elevation that occurs in the Earth's surface. Therefore, we neglected this question in the horizontal geo-reference.

There are, however, certain GIS applications, where the vertical position of the points is also important, besides their horizontal dislocation. Moreover, as it is shown in a future chapter, there is one such application – the orthorectification of the aerial photos – which definitely needs the vertical coordinates of the control point, because its image geometry. In this chapter, we summarize the knowledge, necessary for determination and interpretation of the vertical position.

7.1 Ambiguities in height definition

To define the spatial position of any point, we have to give three, linearly independent coordinates. This can be done either in a three-dimensional XYZ (Cartesian) or in polar coordinate systems. For example, the inner algorithms of the GPS system uses the first, the Earth-centered Earth fixed (ECEF) XYZ system. If the orientation of the coordinate axes is unambiguous, so is the location of the points characterized by them. However, as the Cartesian coordinates can be seldom interpreted by the average user as geographic information. Therefore, in the practice – partly because, as it was mentioned earlier, the real geo-centered position of the coordinate systems was impossible for a long time – the horizontal and vertical positions are given separated. For the GPS user, the system transforms the Cartesian coordinates to geographic (ellipsoidal) latitudes and longitudes on the WGS84 datum, and the elevation above the WGS84 datum ellipsoid.

The shape of our Earth is not an ideal sphere, so the determination of the elevation can be done in many different ways. As is was discussed in the Chapter 3.2, the real shape of the Earth is the geoid, a level surface of the summarized force fields of the gravity and the centrifuge, connected to the mean sea level. The geoid can differ from the best fitting ellipsoid vertically (geoid undulation) and the maximum of this difference is cca. 110 meters (Fig. 39). In the reality, the geoid does not fit to the sea level exactly, because of the thermo-saline differences and streams, and the typically low and high pressured meteorological zones. The vertical difference can be up to 2 meters. The elevation data shown in the topographic maps are results of precise levelings. However, because of this ambiguity it is important that which coastal point was the start of these levelings. Moreover, because of the crustal movements and the plate tectonics processes, the surface points are in constant movement, not only in horizontal but also in vertical sense. This movement causes detectable and measureable distortion in the mutual positions of the base points.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

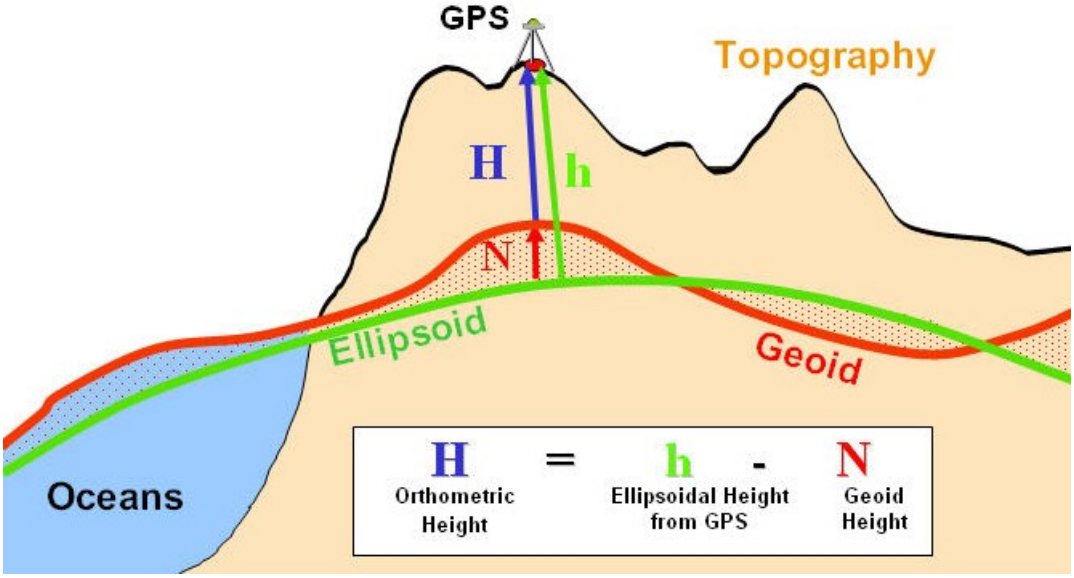


Fig. 39. The topography, the geoid and the ellipsoid.

Because of the above facts, the 'elevation above the sea level' of the point is not a fixed data. What would be unambiguous, it is the potential value of the gravity field and its difference from the pre-defined potential value of the geoid. The potential, however, cannot be directly measured, and even if we determine it, its conversion to height value can be done only approximation: the level surfaces are not parallel to each other, so the vertical distance of two level surface varies from place to place, even in very small order (Fig. 40).

These are the ambiguities and processes that mar the clarity of the determination of the height measurements 'above the sea level'.

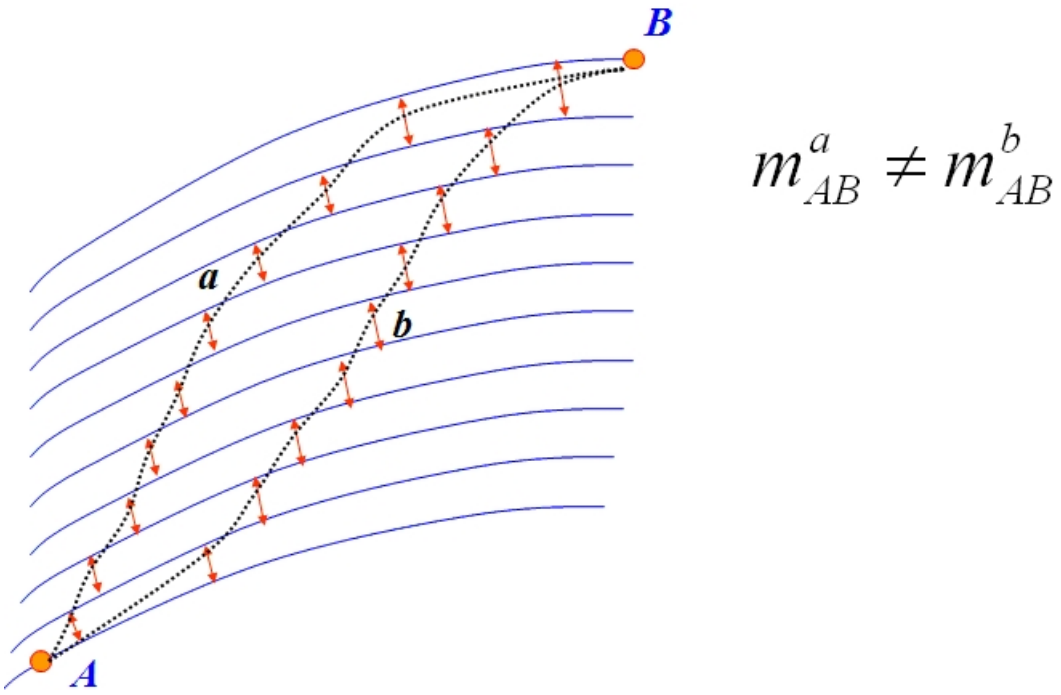


Fig. 40. Following different paths, the result of the leveling is different (Gy. Busics, 2012).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

7.2 Height definitions, elevation measurements

Height above the ellipsoid and above the sea level

Because of the ambiguities, discussed in the above point, the elevation can be defined in multiple ways. The first and most important is to discriminate the elevation above the ellipsoid and the elevation above the sea level (geoid). When we talk about the elevation above the ellipsoid, it always concerns the elevation above the WGS84 specific, Earth-fixed ellipsoid.

The difference between the above two elevations is the local geoid undulation value. This can be between +100 meter and -110 meter, in Hungary, it is between 39-46 meters. Therefore the difference is significant and the resulted error of the incorrect application is high enough to be unacceptable in any practical applications.

The difference becomes obvious, when we measure at a summit with a known height with a GPS that shows the elevation above the ellipsoid. For example, the top of the Gellért-hegy hill in Budapest is 235 meters above the sea level but a GPS (if there is no built-in geoid model) shows systematically the elevation values around 278 meters. The difference is equal to the know geoid undulation value in Budapest, which is cca. +43 meters.

Realizations of the elevation above the sea level

The standard measurement of the elevation above the sea level is made by geometric leveling along the leveling lines and by measuring the gravity acceleration connected to the points of the leveling. From these data, the difference of the potential between the endpoints of the line can be computed without any assumption:

$$K_P = \sum_0^P g_i \Delta h_i \tag{7.1}$$

In the Equation (7.1), the g_i measured gravity acceleration values are determined along the leveling line, while the Δh are the elementary height differences are determined geometrically by local leveling measurements. If an endpoint or any point of the line (or a network consisting of multiple leveling lines) is on the sea level, the geopotential value can be given with respect to this specific point. Although it is an unambiguous value, the geopotential number cannot be applied in cartography. The elevation of the point can be estimated by dividing the geopotential value by the 'characteristic' gravity acceleration value along the section below from the point to the sea level:

$$H = \frac{K_P}{\tilde{g}_P} \tag{7.2}$$

however this calculation needs assumptions. In spite of Equation (7.1), here the acceleration values are interpreted not along to leveling line but along the plumb line beneath the point. These values are not measured, so we can use theoretical models for practical use.

Determining the *orthometric height*, the 'characteristic acceleration' in the denominator of the Equation (7.2), is estimated by specific models. Interpreting the results, we shall know that the points with the same orthometric heights are usually in different level surfaces. The base level of the orthometric height is the geoid.

In case of the *normal height*, the 'characteristic acceleration' is derived from the normal formula of the Earth's gravity:

$$\gamma(\Phi) = \gamma_{eq}(1 + f^* \cdot \sin^2 \Phi + \frac{f_4}{4} \cdot \sin^2 2\Phi) \tag{7.3}$$

where the latitude of the measured point have to be taken into account (here γ_{eq} is the gravity acceleration in the equator, f^* and f_4 are constants defined in the used geodetic system. The result is further corrected using the known effect of the elevation to the gravity acceleration. The base level of the normal height is the so-called quasi-geoid or co-geoid. As the difference between the orthometric and normal heights of a point is not so high (usually a few

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

centimeters in flatlands and hills but can be up to two meters in the steep slopes of the high mountains), the height difference between the geoid and the quasi-geoid is also in this small range.

Introducing the *dynamical height* eliminates the latitude dependence of the normal height. At the calculation, instead of the normal acceleration at the latitude of the point we use the acceleration value of the 45 degrees latitude, simply substituting Φ=45º in the Equation (7.3).

The orthometric, the normal and the dynamical height are all the realizations of the elevation above the sea level. Their vertical difference are usually insignificant for any GIS application or analysis. Here we shall note again that the incorrect use of the elevation above the ellipsoid and above the sea level results a much significant (thousand or ten thousand times higher) error.

7.3 Ambiguity of the sea level: vertical datums

As it was mentioned above in the Chapter 7.1, the real sea level does not follow exactly the level surface defined as geoid. In the geodetic systems, the geoid is defined by the potential value of this level surface. The sea level can differ from this even by two meters. The difference is characteristic as a time-average from place to place, while temporal variations can be also detected.

Because of all of this, the definition of 'sea level' or 'mean sea level' is quite complex. The sea level is measured by mareographs: they record the water level at the point as a function of the time. The null points of the mareographs are *ad hoc* placed vertically. The real geopotential values are rarely determined therefore the read values at different mareographs – however their connections can be analyzed statistically – are not in direct connection. As it was above mentioned, the temporal trends of the read values are affected by the global – and in case of the inner seas: the local – sea level changes and the regional crustal movement of the area the mareograph is placed in. The first affects the real values while the second modifies the geopotential value of the local null level. The sea level is defined by the elevation data read at all mareographs, together with the horizontal position of the instruments as well as the vertical situation of their null level. The sea level can be interpreted for different time intervals (epochs), as the temporal average of all measuring points (e.g. mean sea level of 1905-1910).

In the practice, the leveling network is connected to the mean sea level at one (or more) pre-defines point. For example, the height network of the Austro-Hungarian Monarchy was set to a mareograph that was working (now abandoned) at the Molo Sartorio in Trieste (now Italy) at a given epoch (Figs. 41 & 42). The military cartography of the late Warsaw Pact used a null level connected to the Kronstadt mareograph near Leningrad (now: Sankt-Petersburg, Russia). The null level of the height system of the European Union is connected to the Amsterdam mareograph. In case of land-locked countries, such as Switzerland, the Czech Republic or Serbia, any land base point can be used as starting elevation data. In this case, of course, this value is not zero. For example, the also land-locked Hungary selected one of the eight fundamental elevation base points of the former Monarchy at Nadap. Besides the old point, a new one was set up in 1951, connected to the new elevation network (Fig. 43), the 'Hungarian zero' level is positioned beneath this point, with a distance given by tenth of millimeter accuracy (Nadap base level).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 41. The Molo Sartorio, as a yacht club base in 2007.

Fig. 42. The Molo Sartorio in Trieste in 2003, with the old position of the mareograph that was the null elevation of the Austro-Hungarian cartography (by courtesy of G. Mélykúti).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 43. The new point at Nadap (in the middle) and the old point's location at the hillslope (on the left), the fundamental point of the Hungarian elevation network (Gy. Busics, 2012).

In the topographic maps, besides the horizontal reference (the geodetic datum), also the vertical system, the so-called vertical datum should be given. Usually this refers to a null level of a mareograph and, if applicable, the epoch. In case of Hungary, the 'Adriatic system' (connected to Trieste) and the 'Baltic system' (connected to Kronstadt) are the examples, and as it was mentioned, both are realized as (different) heights from the Nadap base point. In vertical terms, the vertical difference between different vertical datums can be interpreted as constant for practical use. Because of the ocean streams, the salinity differences and the evaporation surplus of the Mediterranean Sea, the level of the Baltic sea is physically higher than the one of the Adriatic Sea. The difference is 67.47 centimeters, this value should be subtracted from the elevations given in the Adriatic height system to get the elevation in the Baltic system. When Hungary started to use the Baltic level instead of the Adriatic one, in the later (mainly at the end of 1970s) issued touristic and hiking maps, majority of the summit elevations were decreased by a meter. As the decrease is less than one meter, according to the rounding rules, the decrease does not occur in every case.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

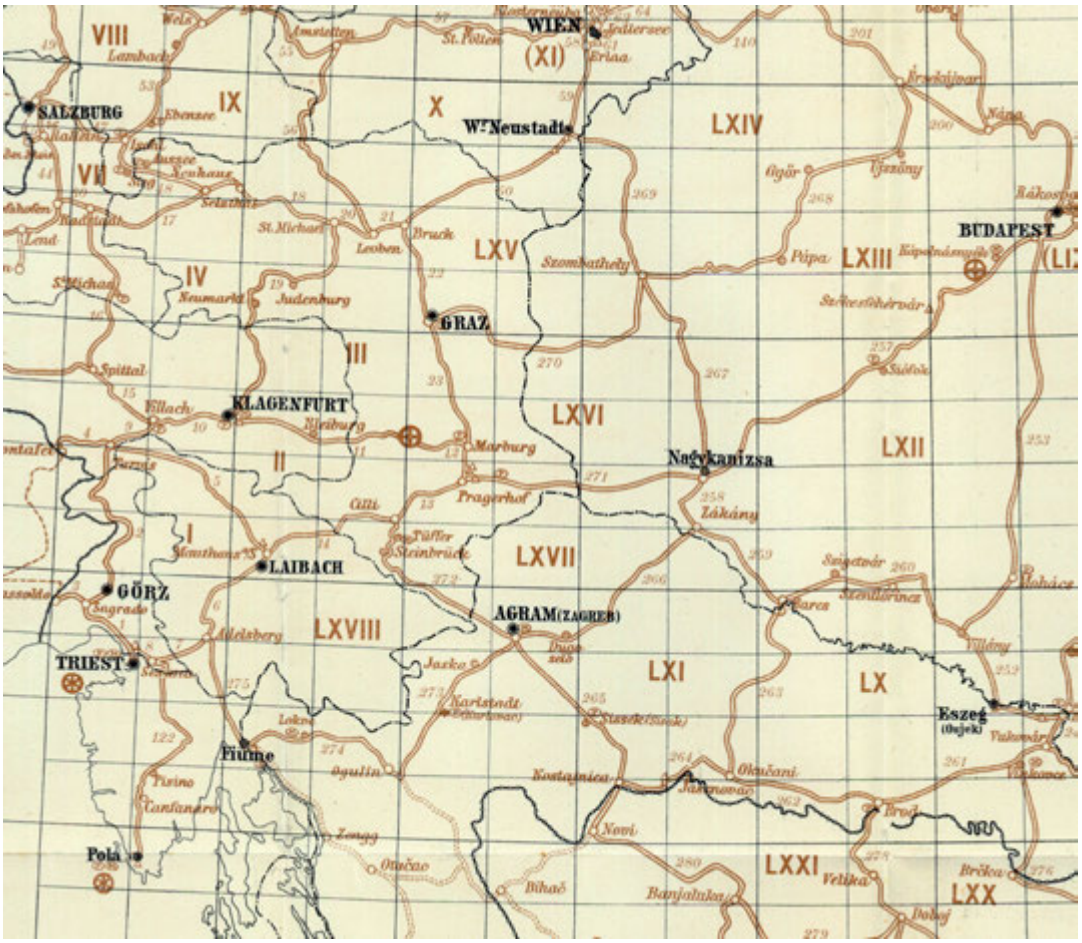


Fig. 44. The leveling lines between Trieste and Nadap and the adjoining parts of the Habsburg leveling network. The lines follow the railroad network.

Similarly to the horizontal geodetic datums, the realization of the vertical datums can be made by a system of base points. The vertical datum is characterized by the physical location of the base points, as well as their fixed elevation data. In some national systems, where the regional crustal movements are very high (e.g. in Scandinavia), the average annual uplift or subsidence values are also indicated. The elevation of the surrounding terrain points can be determined from a nearby base point by local survey technology. If the vertical point network consists of many points, it is divided to sub-networks, according to the leveling technology of their points. The accuracy of the network, however, is mostly determined by the creation method and accuracy of the first-order vertical network (Fig. 44).

The elevation correction – the simple difference making – between the different vertical datums, especially the sign (direction) of the shift, should be accomplish with special care in case of construction of such objects (bridges, tunnels), whose endpoints are in different countries using different geodetic datums. Nowadays, the European height standard is connected to the Amsterdam mareograph. The local elevation differences from this level are shown in Fig. 45.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

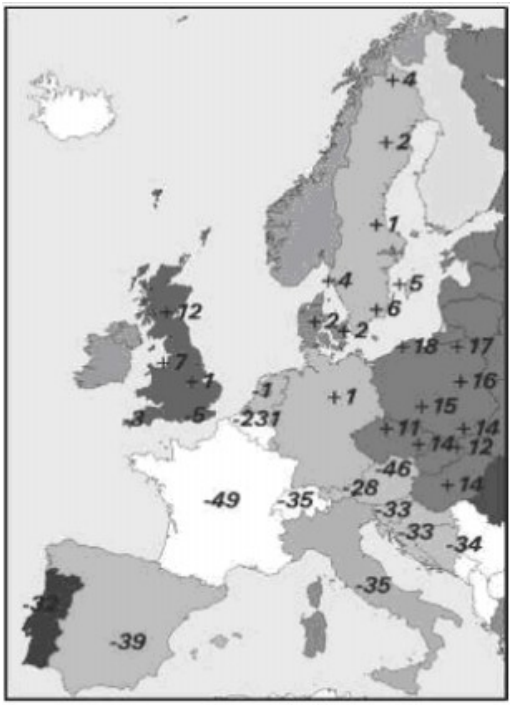


Fig. 45. Differences in centimeters between the European null level (the Amsterdam gauge) and the local elevation zeroes (Ádám et al., 2000).

Finally it should be mentioned that as the three-dimensional data collection techniques are spreading, the unification of the divided horizontal and vertical databases and networks is expected. The cause of the still-characteristic division is mainly the different methodology and accuracy in the physical-geodetic realization of the horizontal and vertical references. The geodetic use of the GPS this difference is decreasing and eliminating, the determination of the horizontal and vertical positions is unified, based on the same physical theory and geodetic practice.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 8. Terrain and elevation models

In this chapter we show the organization of the vertical data – used for geo-reference – into spatial models. We don't aim to discuss the subject in depth of the extent literature of the terrain and elevation models. However, it is necessary to introduce the definition and model types at a level that is a good overview for the reader involved in GIS and especially in the geo-reference and the ortho-rectification of aerial images (Chapter 9).

8.1 Definition and types of the terrain models

In general, we call elevation model any procedure that is able to estimate the characteristic elevation of a surface at a point, defined by its horizontal coordinates. The quality of the model depends on the accuracy of this estimation. In this definition the surface can be any three-dimensional layer, however, in the GIS technology we usually model the terrain elevation, the relief, which is also displayed by the contour lines in the topographic maps. In this case, our model is called terrain model.

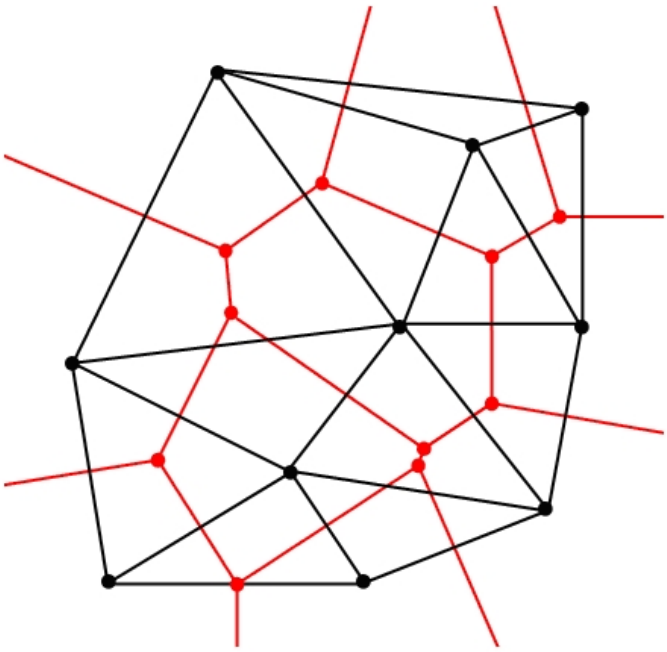


Fig. 46. The Voronoi diagram (red): connects the centers of the circumcircles of the original triangles (Wikipedia).

The terrain model can be of two kinds: vector or raster-based. The vector model expresses the irregular spatiality of the data sampling. It is based on elevation data at an irregular horizontal point set, on coordinate triplets of 3D spatial points. The elevation can be estimated between these points by some interpolation method. The easiest and most used way to do it is the application of the triangulated irregular n network (TIN). We lay an ideal triangulation net to our point set. In the practice, 'ideal' means that the sum of all triangle edges should be minimum at the whole network (Fig. 46). This way, we can arrange one and just one triangle to any point of the interpretation range on the base plane, or the point itself is an original network point. Using the planes (or any more complex but unambiguous function) fit to the different triangles, we can estimate any for any horizontal point on the interpretation range.

For our discussed practice of the geo-reference, the rectification of scanned maps and datasets requires the raster data model. This makes necessary the application of the raster variants of the elevation and terrain models. It is quite easy to estimate the elevation values at the points of any selected raster net, using the above mentioned TIN-based models. For shorter software runtime and the data-level compatibility, these raster grids are usually not realized by dynamic queries. It is easier just once to fill a raster file with data, by the TIN→GRID conversion. According to the direction of the conversion, the information content of the resulted elevation or terrain model is

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

less than the one of the original triangulated network: the original network cannot be reconstructed from the grid data. In the followings, we discuss these grid-based, raster models.

8.2 Making and characteristics of the raster-based terrain model

The raster terrain models can be constructed by using the following input data:

- Original field leveling data
- Map contours
- Stereo pairs of aerial photos
- Radar-based elevation of interferometry data
- Lased-based elevation or range (LIDAR) data.

The original field levelling data is a three-dimensional point set, which was surveyed by field measurements in some vertical datum (vertical network). The points, of course, are in irregular network in the horizontal plane. The contour lines of topographic maps were drawn using these data, prior to the widespread use of the stereo photogrammetry. In the practice, this kind of data is rarely used. After design and print of the final contour maps, their working material, such as the original field protocols and the derived point lists were often lost.

The map contours (the lines connecting the terrain points with equal elevation) were designed and drawn using the mentioned field surveys, or later by the procedures of the below discussed stereo photogrammetry. Their information content is less than the one of the original field data. The manual or automatic digitalization of these contours provides again a three-dimensional point list: for the horizontal position of the digitized points we couple the elevation value of the contour line. This point set can be interpreted as a model of the original field leveling data. A raster elevation model can be constructed similarly, by TIN→GRID conversion. However, the contour-based elevation models are distorted by three kinds of errors:

- In the sharp curves of the contours, there are one or more triangles in the irregular network, whose vertices lie on the same contour. The modeled elevations of all points of these triangles are in horizontal planes. Therefore, along the ridge lines, a ’virtual plateau’ occurs, which is not existing in the real terrain. Thus, in the histogram of the terrain model has peaks connected to every contour elevation.
- If we don’t digitize enough vertices along the long, straight sections of the contours (the point interval is less than the distance of the neighboring contour), and it is not even densified later by automatic methods, then the edges of the irregular triangulation network does intersect the contours in some places. The result is a ’fishbone pattern’ at the top or the bottom of the displayed slope (Fig. 47).
- In very flat terrains it is a frequent situation as only one contour is crisscrossing through an extent area. Even we digitize thoroughly this line, following the complex structure of oxbows and point bars of a floodplain, the result will be a single, horizontal plane. The original fine relief can be attenuated by virtual auxiliary contours, following the small ridges and valleys.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

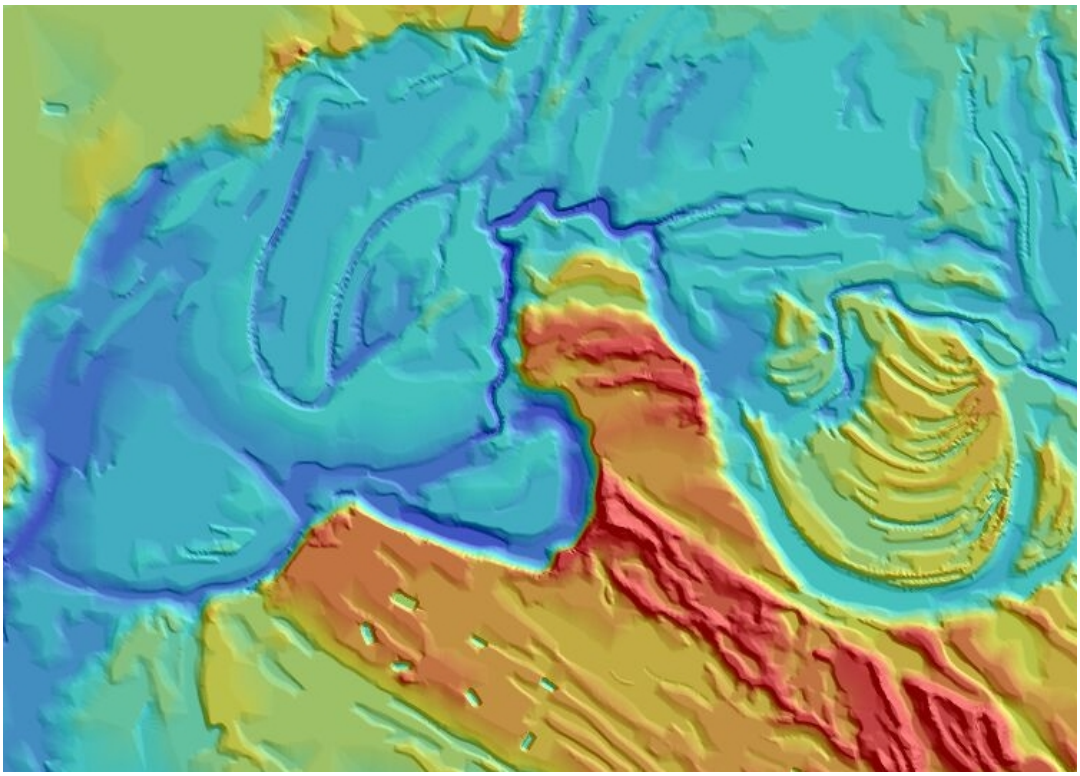


Fig. 47. 'Fishbone-like' errors along oxbow patterns in an elevation model: result of a digitized contors when the vertices are too far from each other.

The above errors can be handled by entering auxiliary data of other pieces of information into the system. At the ridges, we can define the ridge lines themselves. At the valley lines, digitizing the streamlines means such relief information, which decreases or even fully eliminates the above false effects. The most parts of the Earth's surface is formed by stream erosion. According to this, there are algorithms providing 'hydrologically correct' terrain models. These algorithms – assuming that the water runs off from all surface points – correct most of the above problems. If our assumption for the surface runoff is true, this terrain model will be the closest one to the real relief and the result in correct also in hydrological applications.

However, in the territories, where this assumption not true, or not even almost true, the resulted model can be of varying quality, sometimes even very bad. Karstic regions with gullies, dolinas, underwater creeks, or an area with many outlets (wind-formed sandy regions, or floodplains with oxbows) are killing the quality of the 'hydrologically correct' models.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 48. Stereo photo-pair: if our eyes are focusing to the infinite seeing this, the image appears in 3D.

In the case or stereo photo pairs, the elevation information of the area is represented by the different distortion of the two aerial photos, taken from different positions and angles (Fig. 48). These distortions are primarily realized by the mutual image position of the base points of the horizontal geodetic and/or elevation networks (whose coordinates are well-known both in horizontal and vertical sense). Other terrain objects with unknown geodetic position can also be identified in both images, which gives auxiliary information for the above pieces of information. The result of the aerial triangulation was mostly a contour map – and we can derive the grid model according to the above mentioned procedure. With the advance of the computer technologies, however, it is sometimes possible to make a raster terrain model directly from the photo pairs and the detected points on them. For this procedure, not only aerial photos but also satellite images (better than mid-resolution, e.g. the ASTER data) can be processed. It should be mentioned that the identified and paired image elements correspond not necessarily to the terrain but they can be in elevated positions (vegetation, buildings). These ones can be omitted, or if we use them all, we shall recognize the resulted dataset not as terrain but elevation model, containing these elements, too.

The radar technology is connected to the terrain modeling in two different manners. The here only mentioned but not discussed radar-interferometry primarily detects the vertical movements of the surface. However, the radar-based altitude measurement is capable to determine directly the distance of the terrain or terrain objects from radar source and detector, which should be in known position with respect to the Earth. This is the technology, which revolutionized the availability of the digital elevation models in the early 2000s, giving a huge push to any research that needs these data sources. The radar was invented to follow the position of aircrafts from the surface by electromagnetic rays propagating through the relatively dense atmosphere the Earth. However, the reverse way is also possible to detect the surface from board of the air- and spacecrafts with localized radar beams. This ability was clearly shown by Zoltán Bay in 1946, who measured the distance of the Moon by radar experiments. In 2003, using a source and detector pair placed on board of the Space Shuttle *Endeavour*, the majority of the Earth's surface was surveyed, resulted in the *Shuttle Radar Topography Mission* (SRTM) dataset. This model became the most used elevation dataset worldwide, primarily because of its free availability and globally unified characteristics. According to the used technology, the partial effect of the built environment and the vegetation is in the data. Nowadays, the cca. 100 meter spatial resolution is considered to be quite low, however at the time of the publication of the dataset, it brought a real breakthrough for a wide spectra of the sciences.

To improve the resolution, the newest of the discussed technologies, the LIDAR should be applied. The laser range measurement became a part of the toolbox of geodesy in the last decades, after the invention of the portable lasers. The most up-to-date application, the laser scanning is based on the scanner that is capable to alter the direction of

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

the laser beam in a pre-set range and to record the backscattered signal. First because of the huge amount of this data, this application is widespread used only in the very last decade. The laser scanner can be applied in the field and can be also mounted onboard of aircrafts. Its satellite application is limited because of the atmospheric scattering. The laser signals are reflected back from the buildings. Because of the high density of the measurement points (up to several points per square meter), there are soil-reflections even while surveying of vegetated area. The high resolution of the method is ideal for surveying the micro-topography of near-flat terrains.

The quality characteristics of the raster-based terrain and elevation models are:

- The horizontal resolution (pixel size);
- The numerical representation of the elevations;
- The vertical accuracy.

If the source is a contour map, we shall also give:

- The scale of the original map
- The regular contour interval of the original map, as well as the smallest contour interval (halving and/or auxiliary contours).

It should be underlined that the numerical representation and the accuracy of the elevations are not the same. The representation (e.g. „integer") shows the smallest elevation difference (e.g. one meter) that can be represented in the model. This is not the same to the accuracy of the elevation estimation (e.g. 3 or 5 meters) that is based on the whole technology chain led to the elevation model. Of course, the representation should be finer than the accuracy, otherwise the representation itself mars the accuracy. The raster-based elevation models are images, whose pixel lines and columns are parallel to the axes of some geodetic or projected coordinate system. This coordinate system, and the place of our image in this system are also very important pieces of meta-data of the terrain or elevation model.

8.3 Availability of the terrain models

There was a long period, when the national geodetic/geoinformation data providers offered the elevation models, developed on the base of contours their own topographic maps. Nowadays, the availability of models, resulted from laser scanning is more and more frequent. These models show the elevation in the horizontal coordinate system of the given country, and similarly, the elevations are represented in the local vertical datum. The quality characteristics are also determined by the technology level of the providing country and by the scale and quality of the available topographic maps. In most cases, the accuracy and the resolution of the laser scanned models are better than the ones of the contour-based models. However the national data providers are constantly working on actualization and quality improvement of their data, there are no such data available for the huge majority of the Earth's surface.

The situation is different, and sometimes surprisingly better in case of the medium-resolution terrain models. Different international groups were formed in the 1990s to compile global models using the local ones. At the end of the last millennium, such datasets (e.g. the GTOPO30) were issued and widespread used. However, according to their edited/mosaicked being, the data quality of these models are heavily varies from place to place. The situation was fundamentally improved by the Shuttle Radar Topography Mission (SRTM) dataset, published in 2003.

This program was started in 1996 by the American NASA (*National Aeronautic and Space Administration*), aiming to mapping the relief of cca. 80% of the Earth's surface, using a radar system, onboard of the Space Shuttle (Fig. 49). After some delays, the space shuttle Endeavour has been launched in 11 February, 2000, onboard with all necessary instruments for the measurement. The whole survey campaign lasted 11 days. The space measurements were completed and supported by extent surface GPS-measurements as well as placing many (around 70 thousands) artificial radar reflectors at pre-set positions, to provide geo-reference. The data processing took 18 months, led by the NIMA (*National Imagery and Mapping Agency*) of the US Ministry of Defense. According to the agreement between the NASA and the NIMA, with the permission of the NASA, the dataset is archived and published by the USGS (*United States Geological Survey*).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

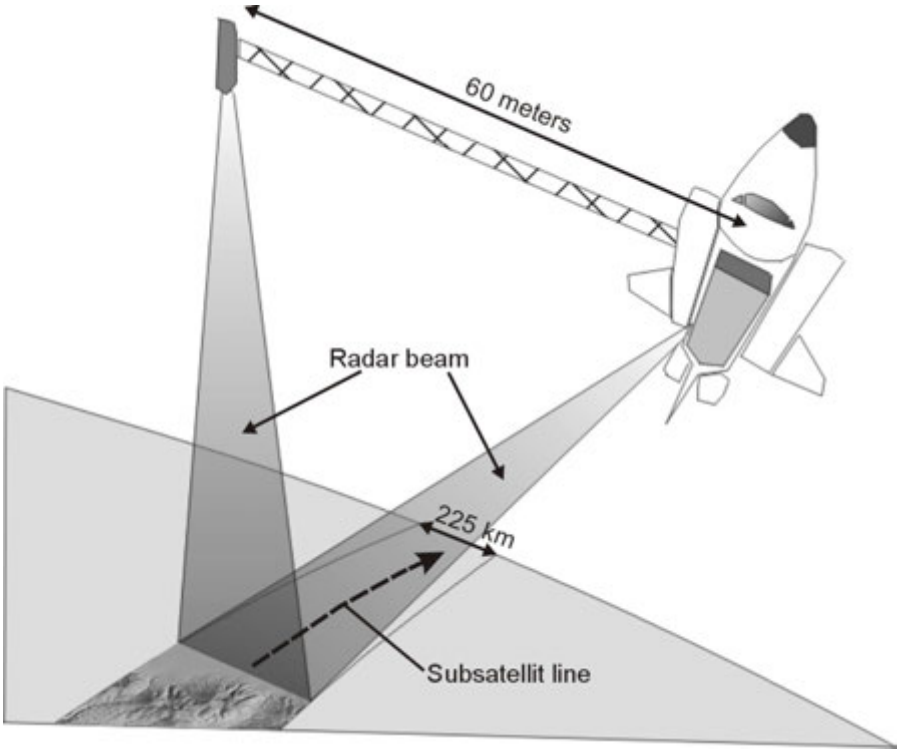


Fig. 49. The settings of the SRTM measurement onboard of the Space Shuttle.

In the frame of the project, the digital elevation model of the mapped area was completed in two different resolutions: the pixel size of the finer version is one arc second (available publicly only for the territory of the United States) while the general version has the pixel size of 3 arc seconds (cca. 90-100 meters is mid-latitudes). Thus, such a public database was created, whose existence and use should be known for any specialists, working with geo-information technology (Fig. 50).

For the measurement, onboard radar equipment was used. As the orbit inclination of the space shuttle in the experiment was 57 degrees, it didn't fly over the polar regions. In the frame of the SRTM program, therefore, was between the 60^{th} degrees of northern and the 57^{th} degrees of southern latitudes. For example, the database is not covering Finland; its topography is not available in it. The resulted 3-arcdegree resolution data is available for everybody on the Internet. The latitude-longitude grid follows the parallels and meridians, the horizontal datum is the WGS84. The elevations are interpreted above the level of the EGM96 global geoid model.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

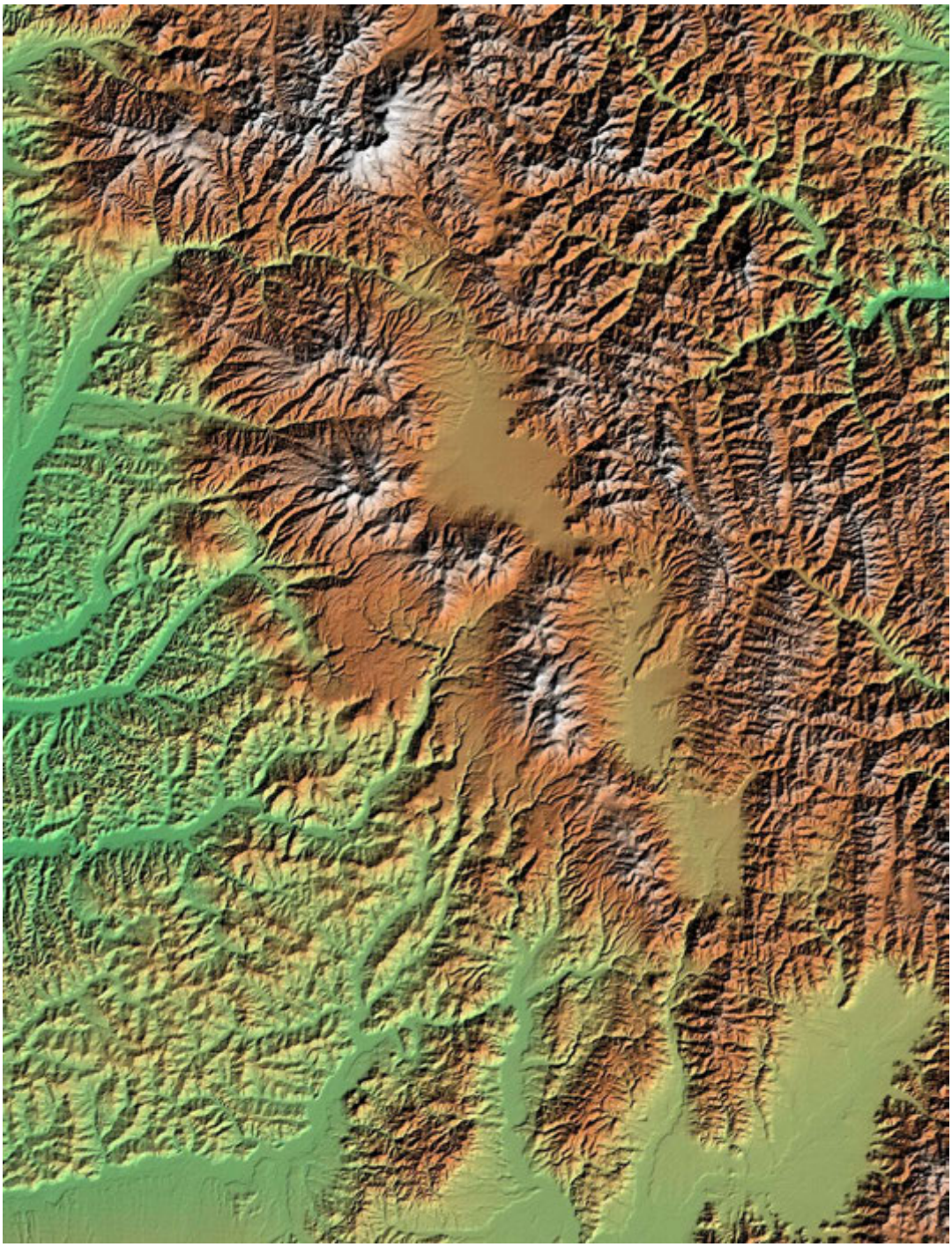


Fig. 50. The elevation model of the Székely Land (eastern Transylvania) in the SRTM dataset.

While using the dataset, we shall keep in mind that it was constructed with radar technology. We have uncertain signals from water surfaces (because of the unavoidable waves), so at the seas, lakes and rivers, we have false data. Majority of them was filtered out during the data processing, and these pixels have NULL cell values. Similar NULL value have been arranged for many mountainous pixels, mainly in deep valleys, which were in radar shadow, according to the survey geometry, and we don't have radar backscatter signal from. This kind of data absence is more frequent in the high mountains. If necessary, the missing data can be completed from other, lower resolution models. The 5.6 centimeter wavelength radio signals are not penetrating the dense or even the medium foliage and, of course, scattered back from the solid roofs or walls of the buildings. Thus, the elevation values of the model represent the geoid height of the layer that is the reflector for the 5.6 centimeter wavelength electromagnetic signal. In the regions of cities or forests, the effect of the buildings and the trees is in our data.

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

On the Mars, the thin atmosphere enables to survey the surface elevation by laser altimetry. The resulting MOLA (Mars Orbiter Laser Altimetry) project provides an SRTM-like elevation model with a horizontal resolution around half kilometer, of course without any artificial of vegetation ‘noise’ (Fig. 51). In the last decade, the altimetry of the Mars has been significantly improved.

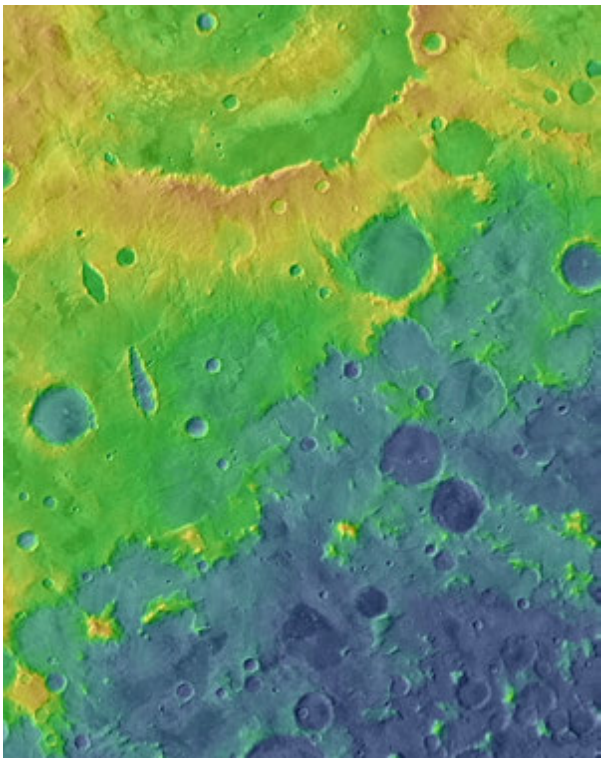


Fig. 51. The southern foreland of the Huygens crater in the Mars, shown by the MOLA elevation dataset.

8.4 The effect of the built environment and the vegetation: elevation models

As it was earlier discussed, some technologies of the elevation model creation cannot discriminate – or only with serious post-processing – the height of the soil, the vegetation and the buildings. And, as it is shown in the next chapter, for the ortho-rectification of the aerial photographs, these pieces of information are also needed to handle the effects of the partially oblique-photographed buildings. Therefore, besides the terrain models, showing the elevation of the terrain itself, elevation models that represent the real photographed surfaces, are also needed. Their construction can be made in two ways:

- The terrain model can be over-written by the elevation of the estimated, modeled height of the vegetation and the buildings. The built objects can be modeled by some three-dimensional prism or a combination of prisms. The vegetation effect is represented by an added constant elevation, characteristic for the plant species (forest trees, agriculture crops). This method is somewhat similar to the ’railroad model’ toys: we add the extra elevation of the objects to the already existing terrain model.

- The elevation model can be directly computed from laser scanned (lidar) data. The active reflecting surface can be any solid object (building roof or walls, foliage of forests). Using post-processing algorithms, the elevation model can be provided from the original three-dimensional point set that is the result of the laser scanning (Fig. 52).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

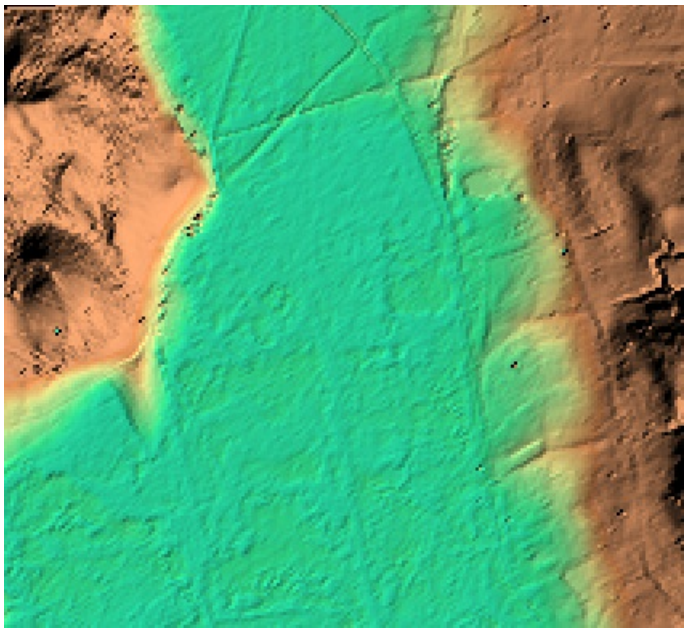


Fig. 52. Artificial objects (roads, railroads, dykes) in a Hungarian flatland, shown in a lidar-based elevation model (Zlinszky et al., 2012).

It should be mentioned again here, that the above discussed SRTM elevation model contains height elements referring to the vegetation and the built environment. However, in this dataset, the systematic difference of the model height and the terrain height refers only to the extents of the towns and forests, and this vertical difference is far from the real surplus. Thus, the SRTM cannot be used as a certified elevation model.

In the practice of the geo-reference, the elevation models are raster-based datasets. This always causes some model errors, whose order of magnitude is depending on the horizontal resolution. The raster model cannot correctly describe the vertical walls and forest-boundaries in three dimensions. However, this ambiguity causes only subpixel registration errors at geo-reference of aerial photos and ultrahigh resolution satellite images. This small error is much more insignificant than the one occurs when no elevation model is used.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 9. Ortho-rectification of aerial photos

The basic geometry and distortion are considerably different from the ones of the maps and map-based raster datasets. Maps are made to show the downscaled version of the landscape in a plane, projected all map objects to this, not depending on their vertical position. The distortions of the photographs are completely different. Here projection is central, the perspective distortion is characteristic, because of the optical realization (Fig. 53).

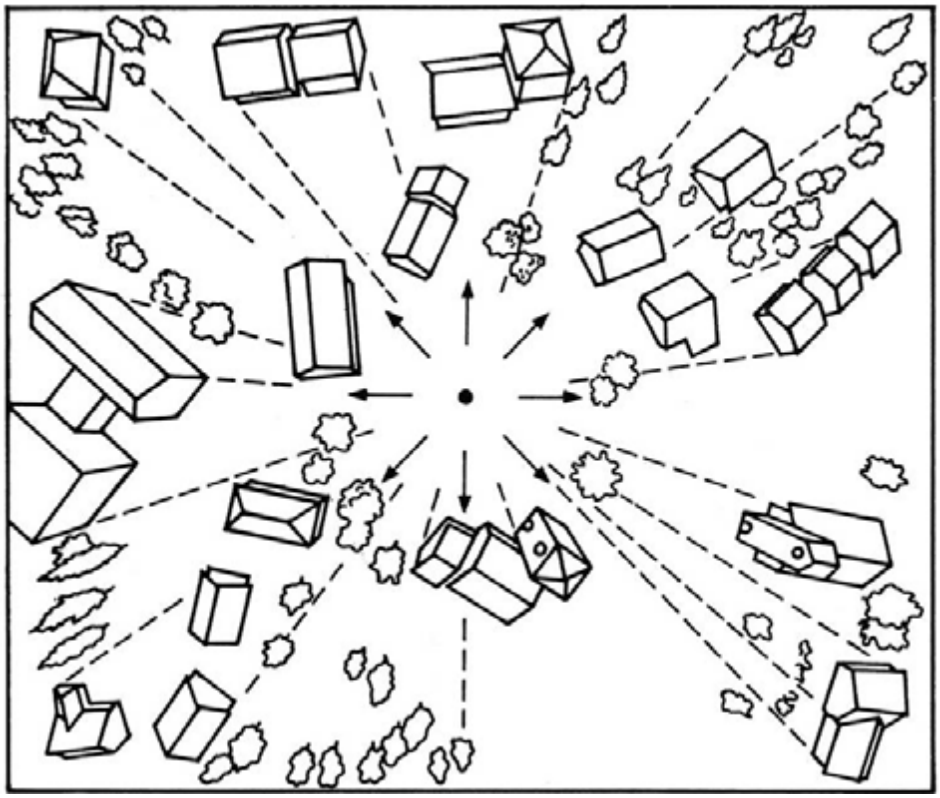


Fig. 53. Characteristic distortions in an aerial photo.

Though geo-reference can be assigned to pixels of photos with any orientation, and the geo-scientific value of surface photos and landscapes is also significant, in this chapter we discuss the aerial photos, mostly taken from aboard of aircrafts. These images approximate the map-like representation of the target, and their fit to standard coordinate systems is of great value in the geoinformatic analyses.

9.1 The goal of the ortho-rectification

The goal of the ortho-rectification is to resample the pixels of an aerial photo to a coordinate system that is interpreted in a selected horizontal surface (practically in a level surface of the region of the airphoto). This coordinate system should be defined in the geographic information system, according to the above chapters (e.g. a map projection plane).

Geo-referring the aerial photos, two different distortion effects should be corrected:

- The perspective distortion, which is the result of the geometry of the photograph taking.
- The distortion effect of the relief and/or the surface.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Up to now, just because of the planar model characteristics of the maps, the vertical geo-reference was neglected in the rectification, here this simplification is no more possible. And this is not even impossible; it is very important, which terrain or elevation model is used. In the aerial photos, the soil, the real physical terrain surface is invisible in many place, it is covered by the vegetation or the artificially built objects. It is our decision, based on the available data and the terrain, how to take into account the elevation of the terrain itself, the different, vertically extent objects and vegetation foliage.

There are some auxiliary information, needed to ortho-rectify the image:

- The camera model and the internal (or in other term: interior) orientation data,
- The external orientation data, and
- A terrain or elevation model, covering the area of the photograph.

9.2 The camera model and the internal orientation

The camera model summarizes the optic geometry from the optic center of the photo geometry (from the center of the object lens of the camera) to the image. Its parts are:

- The focal length, and
- The geometrical position of the fiducial points.

In case of the professional aerial photocameras, mainly of the older ones, the focal length is a constant at a certain camera. In the image plane, some pre-fixed points, the so-called fiducial points are placed. These are positioned near to the image corners and/or the halving points of the sides, their position is constant with respect to the image center (the principal point of the image). Their positions are expressed in a local coordinate system in the plane of the image, the origo is the principal point, the axes are parallel to the image sides. The positions are described in millimeters or centimeters (Fig. 54).

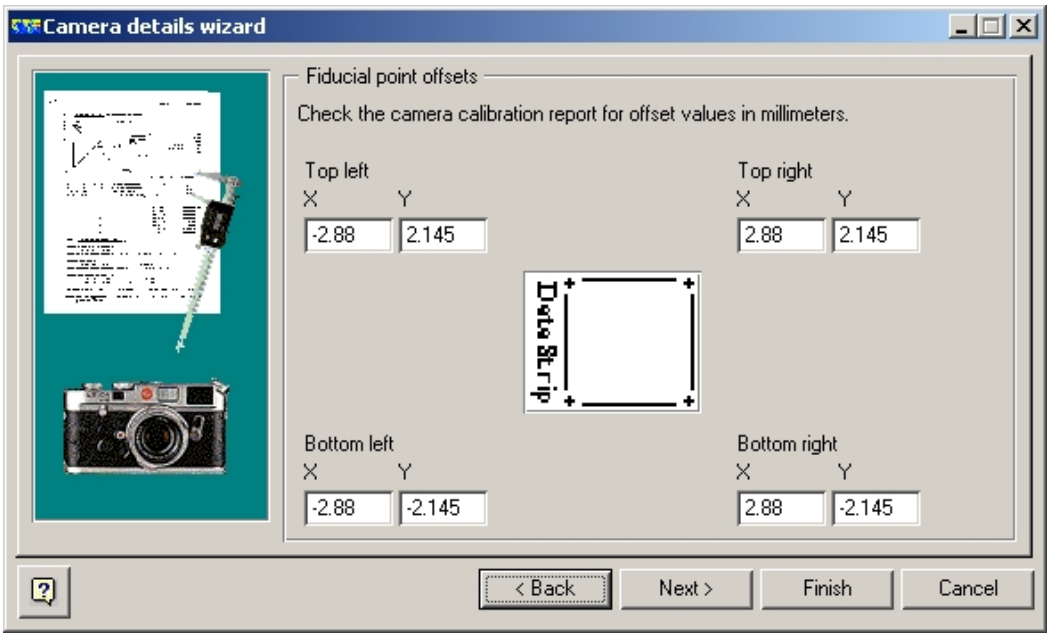


Fig. 54. Defining of the positions of the CCD corners as fiducial points in a camera model of a 1/2.5" CCD (cf. Table 6) in a GIS software. The camera model also needs the focal length.

These meta-data are strictly needed for the exact geo-reference. For rectifying an archive aerial photo, the original camera type, and thus its camera model parameters are obligatory subjects of our investigation. For the ortho-rec-

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

tification, the GIS softwares ask for these camera models. The necessary data of some 'standard', widespread used cameras are often built-in. Also, we can define new camera models for our own instruments, knowing the necessary data.

Beyond the camera model, a further element of the internal orientation is the image coordinates of the frame points in our digital format image. In GIS software environment, it is practically given by moving the cursor to the frame points, and record them (e.g. by mouse click) in correct order.

9.3 The external orientation

For the ortho-rectification, the part of the photo geometry between the optical center and the object is also need to know. The most important elements of this are the three-dimensional position of the optical center and the camera orientation. We have to know that the photo was taken from where and to which direction.

The location of the optical center is best to know in the wanted coordinate system of the later ortho-rectified image. The elevation of this point is also has to be known and given, practically from the sea level (the geoid). If we want to measure them during the flight, onboard an GPS instrument should be used, however its data can be applied only with some corrections. The exact position of the camera, valid at the time of the photo taking, should be interpolated from the continuous position string of the GPS. Besides, it has to be taken into correction (and this correction is never fully correct) that the GPS antenna and the optical center are not at the same place. Their position is fixed only in the coordinate system fixed to the aircraft, but its heading, roll and pitch affect the difference vector element in any external (ground-fixed) coordinate system.

The direction vector of the optical axis of the camera is also important to describe the image geometry, and their onboard recording is also can be attempted. For this, an inertial navigation system (INS) can be used. This contains gyroscopes (rotation sensors) and accelerometers (motion sensors) and records the actual angle difference in three dimensions from a reference direction, which is pre-set prior to the flight. In this case, it is again an exercise to get the orientation angle data exactly at the time of photo taking. The six elements of the external position are the three locations and the three orientation data; they can be input directly into the GIS system used for ortho-rectification.

In the practice of the geo-reference, however, these data are seldom presented, even with preliminary accuracy. Fortunately, the elements of the external orientation can be completed in indirect way. They can be estimated using ground control points, moreover, this method often provides better accuracy than the built-in navigation system. Most of the widely used GIS software packages offer the possibility of the estimation of the six external orientation parameters instead of their direct input. To perform this estimation, the ground control points should be given in the target coordinate system and the image positions of these points should be given also in the aerial photo. The target coordinate system, however, should be a projected one, latitude-longitude based geographic systems should be avoided here. The elevation of these point should be defined as precisely as possible; this affects the accuracy, and sometimes the possibility of the parameter estimation. The elevations can be obtained from elevation models or can be read from topographic maps. We have to be prepared to do a meticulous work with many clarifications, difficult point identifications and switching the already recorded points on and off, during the process (Figs. 55 & 56).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

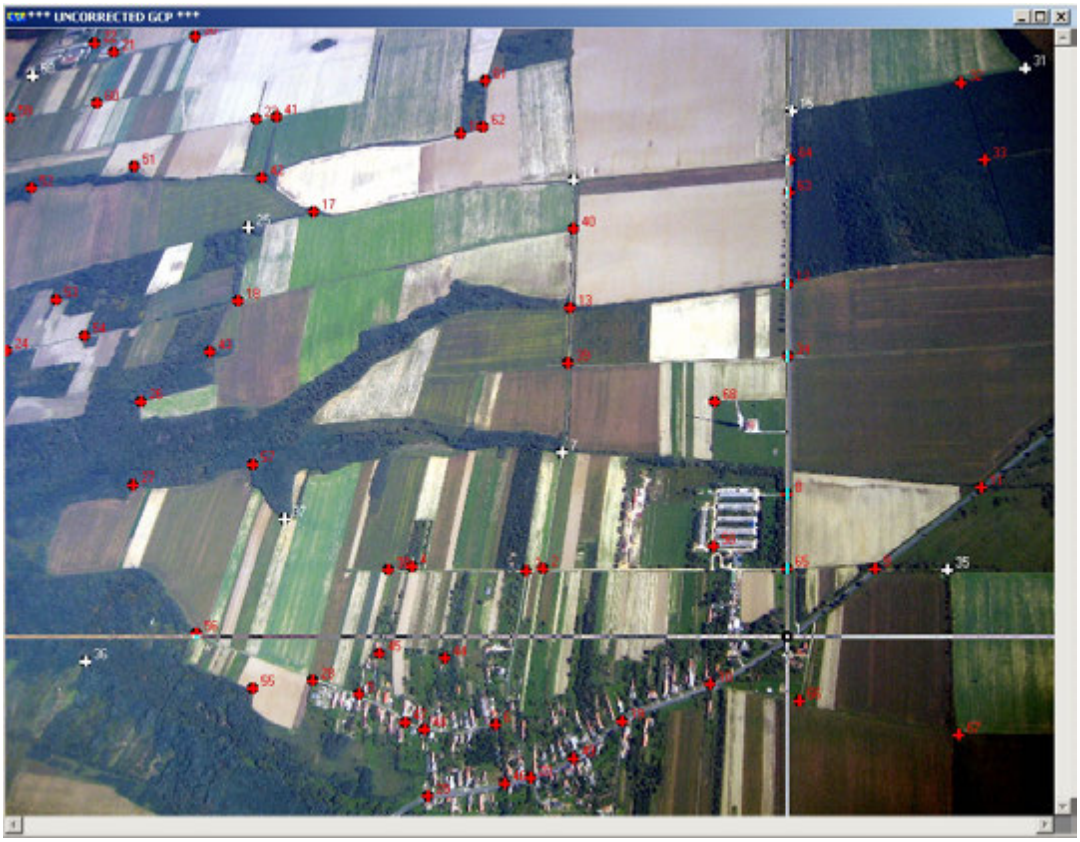


Fig. 55. GCPs in ortho-rectification: it is quite a work to find the best ones.

According to the recorded point data – three positional data and two image coordinates for each point – and the already defined interior orientation the software estimates and gives the six parameters of the external orientation. However, the parameter estimation is often burdened by significant error. Therefore, a dense control point system, covering the whole image, should be created. The closer the optical axis to the vertical, the better is the quality of the parameter estimation. If the image contains the horizon, it is almost impossible to estimate the elements of the external orientation. However, the image should be used as a whole – no part of it can be cropped out – to save the internal orientation data!

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

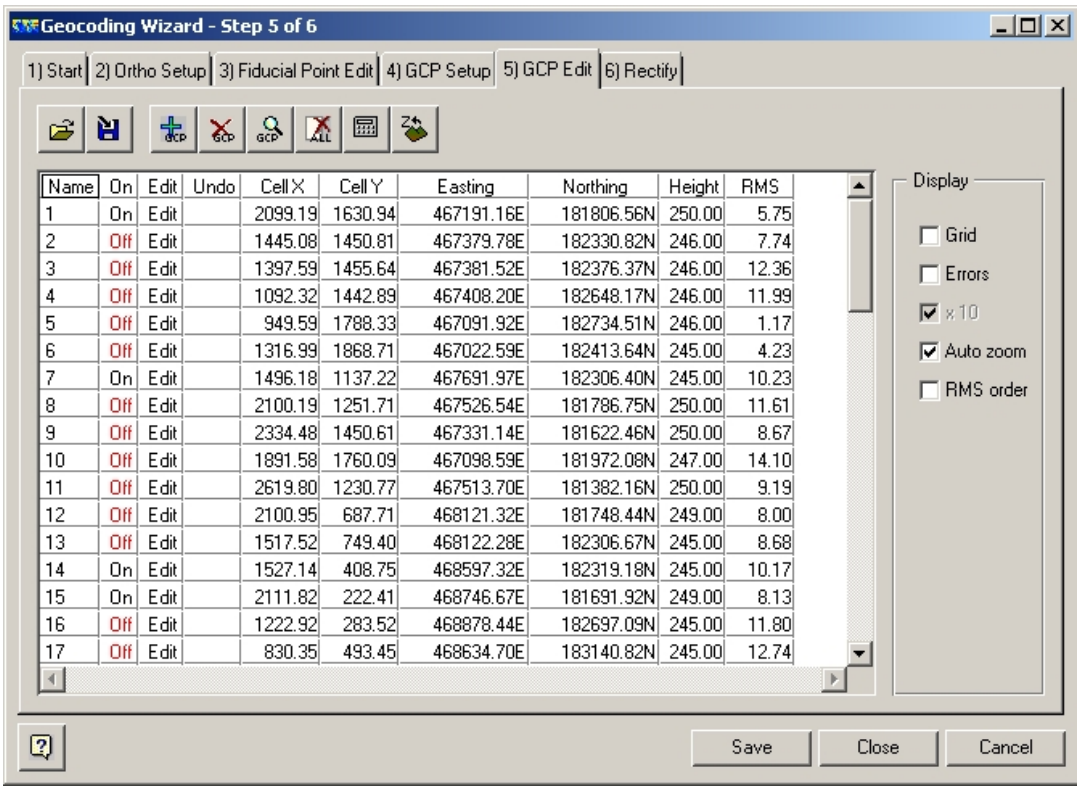


Fig. 56. The elevations should be also given for the GCPs at the ortho-rectification.

9.4 Camera model of compact digital photo-cameras

The above data are mostly used while using professional aerial photographing instruments. They work with focal length and photo-negative size of several decimeters, with constant geometrical settings. However, we can ortho-rectify the images taken from aircrafts by compact hobby cameras (Figs. 57 & 58). Of course, the focal length of the camera can be altered by the zoom function and can be different from image to image. The actual focal length is stored in the meta-data of the digital image (practically in the EXIF tag of the image; Fig. 59). The focal length can be altered step by step, therefore we should define several camera model for a single camera, one model to one focal length value.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 57. Aerial photo taken by a compact digital camera (by the courtesy of Z. Barcza).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Fig. 58. Rectified version of the above photograph. The optical axis is far from the nadir direction, that's why the strange shape – however the fitting is good even in the far corner.

There are no fiducial in the hobby cameras, so they should be substituted by other positions. Practically, the corner points of the images can be used as fiducial points. This solution can be quite inaccurate at traditional negatives or dia-positives but provides surprisingly good results with digital cameras. The problem with the traditional film is the not exact planar position of the film material in the camera, there are small undulation remained. Therefore the frames are not exactly in the same position, with respect to the camera mechanics. A further error source is the film development: usually not the original frame is processed, which means the lost of the internal orientation. These problems do not occur at digital cameras. The film frame is represented by the CCD sensor. Its size is a characteristic constant for the camera. Therefore, the position of the sensor corners can be defined as frame points. The exact internal orientation can be obtained by defining the image coordinates of the four corners of the images (Table 6).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Sensor type	Width (mm)	Height (mm)
1/10”	1.28	0.96
1/8”	1.6	1.2
1/6”	2.4	1.8
1/4”	3.2	2.4
1/3.6”	4	3
1/3.2”	4.54	3.42
1/3”	4.8	3.6
1/2.7”	5.37	4.04
1/2.5”	5.76	4.29
1/2.3”	6.17	4.55
1/2"	6.4	4.8
1/1.8”	7.18	5.32
1/1.7”	7.6	5.7
1/1.6”	8.08	6.01
2/3”	8.8	6.6
1”	12.8	9.6
1.5”	18.7	14

Table 6. Physical size of different CCD sensors of digital cameras, for defining the camera models.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

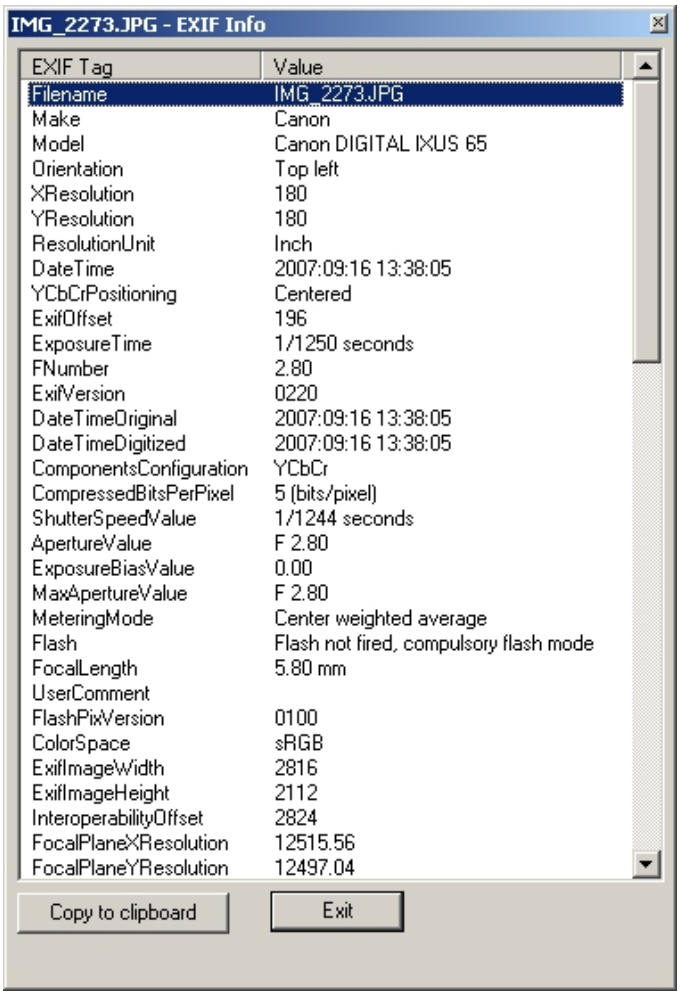


Fig. 59. The EXIF tag of the picture shown in Fig. 57. The make and the type of the camera makes the CCD-size (Table 6) searchable. The focal length is also needed for the camera model.

9.5 The ortho-rectification process

When we have all of the above mentioned parameters of both the internal and the external orientations, we can start the main part of the procedure. During this, the algorithm computes the real spatial position of all image pixels. Then the image is resampled, using these positions, into a target coordinate system, which was pre-selected by the user. To accomplish this step, we shall know also the elevation of the image points; that's why a terrain or elevation model is asked for by the algorithm. The accuracy of the elevations can be lower than it was needed for the control point definition at the estimation of the external orientation elements – while the accuracy of one point there affects the whole image, now it controls only its near vicinity.

The result should be always verified, e.g. by a topographic map (practically the one used for control point definition). The horizontal fit is usually the best near to the corner that is the closest one to the vertical axis from the camera. The fit could be unacceptably poor around the far corner, which is caused by the errors of the external element estimation. We can do a feedback to making a new estimation or just retain the good fitting parts of the resulted image (Fig. 60).

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

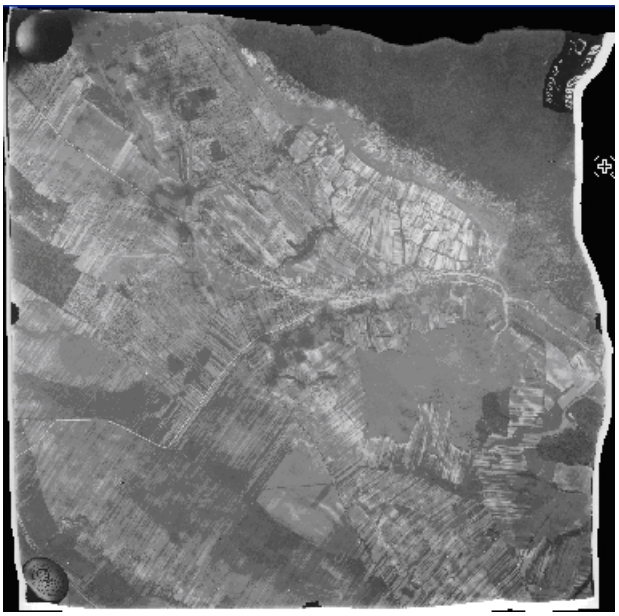


Fig. 60. The edge of the rectified airphoto is an irregular line because of the relief.

9.6 The effect of the applied elevation model

In most cases, we have a terrain model for the surveyed region, which defines the terrain elevation with more or less accuracy. However, as it was discussed, the aerial photographs often show not the soil/terrain itself, but the top of the covering vegetation (field crops) or roofs of the buildings. If we omit this fact, e.g. because of missing data of the building heights, the fit of the image will be good at the terrain level. The top of the buildings will be shifted by several meters from the vertical axis from the camera (Fig. 61).

Fig. 61. If the elevation model does not contain the building heights, the fit is valid at the terrain level only.

In case of accurate models, showing also the height settings of the buildings, all points of the resulted images will be in correct horizontal position. We will have data absences at the occultation pixels (e.g. the ones covered by buildings, higher towers). This is not an error but a consequence of the survey geometry: indeed, we don't have any information about the covered terrain parts in the photo.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

9.7 Making of digital anaglif images

There is an application, in which we don't eliminate the distortion effect of the relief but, on the contrary, we use it for our purposes. The so-called anaglif image can be constructed for a section area of the aerial photographs, taken from different positions. The black-and-white versions of the two images are turned to different colors and a unified color image is compiled from them. If this image is observed through eyeglasses with the same colors used for the anaglif, it appears as a three-dimensional image in our brains (Fig. 62).

There is nothing more to do than processing both images. However, at the ortho-rectification step, we shall use the same horizontal planes as an elevation model for both images. It should be repeated: at this step only; for the estimation of the external element parameters, the vertical positions of the control points should be known. While displaying the anaglif, it is important that the different colors in the print and the eyeglass should be in same order (e.g. the red on the right, the green on the left both in the images and at the eyeglass). Otherwise, the image shown three-dimensional character only after a rotation by 180 degrees.

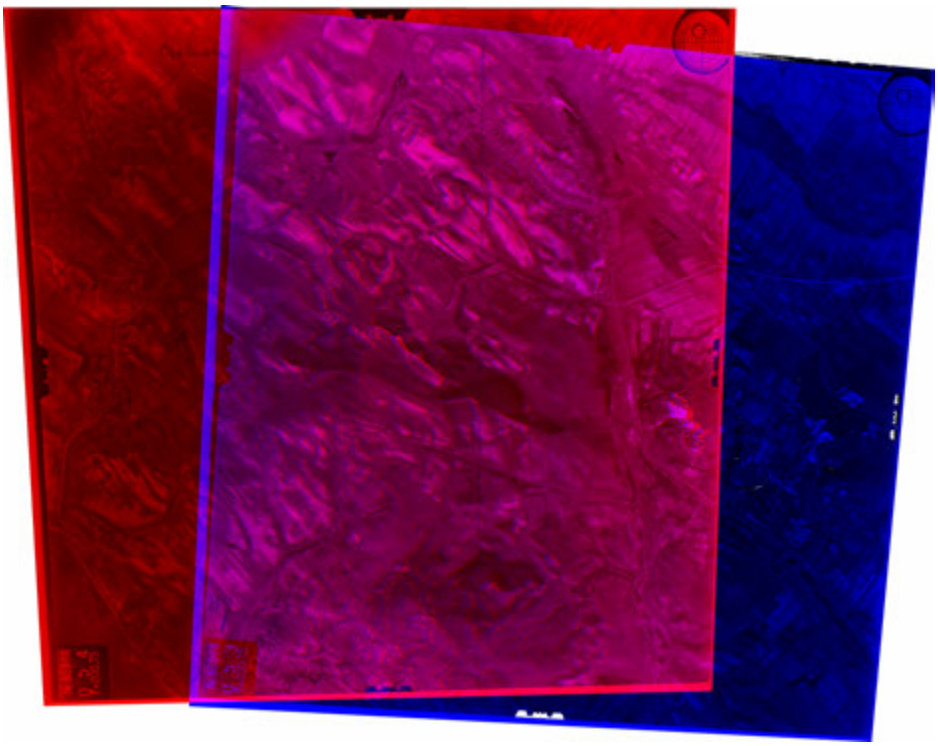


Fig. 62. Anaglif image: throughout anaglif glasses, the terrain is viewable in 3D.

9.8 Rectification of the photographed documents and maps

The above discussed method can be used not only to fit photographs taken onboard of aircrafts into map coordinate systems. When we take a picture about a document or a map sheet holding the camera in hand, the target is depicted by the same perspectic distortion. If we aim to reconstruct the original geometry of the planar target, or to rectify the photographed map in its own projection, we shall apply the method of this very chapter.

In case of text documents, it is – because of the difficulty of control point selection – not always easy. If needed, we can make slightly, by pencil, some small signs at pre-measured points of the documents for control points. After taking the photo, these signs should be removed without any damage of the original document. In case of photographed maps, this problem does not occur. The control points should be selected the same way we discussed in Chapter 6. The difference is that the rectification is to be done by the procedure discussed in this chapter. We usually don't have any information about the vertical position of the photographed material. Thus the elevation of

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

the control points are set to zero, as well as the elevation model pixel values. Applying this method, we can reconstruct the geometry of the photographed map and we can fit it to map coordinate system in the same algorithm.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Chapter 10. References – Recommended literature

Ádám, J. (1982): On the determination of similarity coordinate transformation parameters. *Bollettino di Geodesia e Scienze Affini* **41**: 283-290.

Ádám J. (2000): Magyarországon alkalmazott geodéziai vonatkoztatási rendszerek vizsgálata. *Geodézia és Kartográfia* **52/12**:9-15.

Ádám J. (2009): Geodéziai alapponthálózataink és vonatkoztatási rendszereink. *Geodézia és Kartográfia jubileumi különszám,* **61:** 6–20.

Ádám J., Bányai L., Borza T., Busics Gy., Kenyeres A., Krauter A., Takács B. (2004): Műholdas helymeghatározás. Műegyetemi Kiadó, Budapest, 458 p.

Badekas, J. (1969): Investigations related to the establishment of a world geodetic system. *Report 124*, Department of Geodetic Science, Ohio State University, Columbus.

Bíró P. (1985): Felsőgeodézia. Tankönyvkiadó, Budapest, 196 p.

Borkowski, K. M. (1989): Accurate algorithms to transform geocentric to geodetic coordinates. *Bulletin Géodésique* vol. **63**: 50-56.

Bowring, B. (1976): Transformation from spatial to geographical coordinates. *Survey Review* **XXIII**:323-327.

Burša, M. (1962): The theory for the determination of the non-parallelism of the minor axis of the reference ellipsoid and the inertial polar axis of the Earth, and the planes of the initial astronomic and geodetic meridians from the observation of artificial Earth satellites. *Studia Geophysica et Geodetica***6**:209-214.

Busics Gy. (1996): Közelítő transzformációk a GPS és az EOV koordináta-rendszerei között. Geodézia és Kartográfia 48(6): 20-26.

Defense Mapping Agency (1986): Department of Defense World Geodetic System 1984 – Its Definition and Relationships With Local Geodetic Systems. Technical Report 8350.2. St. Louis, Missouri, USA.

Defense Mapping Agency (1990): Datums, Ellipsoids, Grids and Grid Reference Systems. DMA Technical Manual 8358.1. Fairfax, Virginia, USA

Homoródi L. (1953): Régi háromszögelési hálózataink elhelyezése és tájékozása. *Földméréstani Közlemények* **5**: 1-18.

Hotine, M. (1947): The orthometric projection of the spheroid. *Empire Survey Review* **9**: 25-166.

Hőnyi E. (1967): Két földi ellipszoid relatív helyzetének meghatározása a háromszögelési hálózat alapján. *Geodézia és Kartográfia* **19**:263-268.

ICW (without author indication, 1884): International Conference Held at Washington for the Purpose of Fixing a Prime Meridian and a Universal Day. October 1884, Protocols of the Proceedings, Gibson Bros., Printers and Bookbinders, 212 p. Elérhetőség: The Project Gutenberg EBook of ~, e-book #17759

Kis K. (2002): Általános geofizikai alapismeretek. ELTE Eötvös Kiadó, 384 p.

Mélykúti G., Alabér L. (2010): Magyarországi térképezések története. Nemzeti Digitális Tankönyvtár, Nyugat-Magyarországi Egyetem.

Mihály Sz. (1995): A magyarországi geodéziai vonatkozási és vetületi rendszerek leíró katalógusa, 4. kiadás, FÖMI, Budapest.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Mihály Sz. (1996): Description Directory of the Hungarian Geodetic References. *GIS***4**:30-34.

Molnár G., Timár G. (2002): (2002): Az EOV-koordináták nagypontosságú közelítése Hotine-féle ferdetengelyű Mercator-vetülettel. *Geodézia és Kartográfia* **54**(3): 18-22.

Molnár G., Timár G. (2005): Determination of the parameters of the abridging Molodensky formulae providing the best horizontal fit. *Geophysical Research Abstracts* **7**: 01018.

Molnár, G., Timár, G. (2009): Mosaicking of the 1:75,000 sheets of the Third Military Survey of the Habsburg Empire. *Acta Geodaetica et Geophysica Hungarica* **44**(1): 115-120.

Molodensky M.S., Eremeev, V.F., Yurkina, M.I. (1960): Metody izucheniya vnesnego gravitatsionnogo polya i figuri Zemli. *Tr. CNIIGAiK*, vyp. **131**., Moszkva.

Papp, E., Szűcs, L., Varga, J. (1997): GPS network transformation into different datums and projection systems. *Reports on Geodesy*, Warsaw University of Technology, No. 4 (27).

Papp E., Szűcs L., Varga J. (2002): Hungarian GPS network transformation into different datums and projection systems. *Periodica Polytechnica Ser. Civ. Eng.* **46**(2): 199-204.

Snyder, John P. (1987): Map projections – a working manual. *USGS Prof. Paper* **1395**: 1-262.

Takács B. (2001): EOV koordináták beállítása GARMIN vevőkön. Elektronikus jegyzet, http://www.agt.bme.hu/staff_h/bence/eov_gar.html - utolsó elérés: 2013. január 2.

Timár, G. (2004): GIS integration of the second military survey sections – a solution valid on the territory of Slovakia and Hungary. *Kartografické listy* **12**: 119-126.

Timár G. (2007): A ferrói kezdőmeridián. *Geodézia és Kartográfia* **59**(12): 3-7.

Timár G., Molnár G. (2002): Az HD72→ETRS89 transzformáció szabványosítási problémái. *Geodézia és Kartográfia* **54**(12): 28-30.

Timár, G., Danišík, M. (2003): Aproximácia Křovákovho zobrazenia Lambertovým konformným kužeľovým zobrazením na území Slovenska pre potreby GIS a GPS. *Kartografické listy* **11**: 100-102.

Timár G., Molnár G., Pásztor Sz. (2002): A WGS84 és HD72 alapfelületek közötti transzformáció Molodensky-Badekas-féle (3 paraméteres) meghatározása a gyakorlat számára. *Geodézia és Kartográfia***54**(1): 11-16.

Timár G., Varga J., Székely B. (2003): Ismeretlen paraméterezésű valódi kúpvetületen készült térkép térinformatikai rendszerbe integrálása. *Geodézia és Kartográfia* **55**(2): 8-11.

Varga J. (1982): Átszámítás az egységes országos vetületi rendszer (EOV) és a korábbi vetületi rendszereink között. *Geodézia és Kartográfia* **34**(2):

Varga J. (2000): Vetülettan. Műegyetemi Kiadó, Bp., 296 p.

Völgyesi, L., Tóth, Gy., Varga, J. (1996): Conversion between Hungarian Map Projection Systems. *Periodica Polytechnica Civ. Eng.* **40**(1): 73-83.

Wolf, H. (1963): Geometric connection and re-orientation of three-dimensional triangulation nets. *Bulletin Géodésique* **68**:165-169.

Zlinszky, A., Mücke, W., Lehner, H., Briese, Ch., Pfeifer, N. (2012): Categorizing Wetland Vegetation by Airborne Laser Scanning on Lake Balaton and Kis-Balaton, Hungary. Remote Sensing 4(6): 1617-1650.

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

Appendix A. Appendix: procedures to estimate the datum transformation parameters

The abridging Molodensky formulae describe the datum transformation simply by the components of the position vector pointing from the geometric center of the target datum ellipsoid to the one of the source datum ellipsoid. It does not take the orientation and scale differences into account. It is also referred to as *three-parameter datum transformation*. The neglected orientation and scale parameters are applied in the Burša-Wolf method; besides the three position shift parameters, uses three orientation and one scale parameters, too. Therefore it is also called as seven-parameter datum transformation method. The parameters of both transformations (as well as the ones of other procedures) are derived from coordinates of geodetic base points, whose coordinates are known in both the source and the target datums.

In this Appendix we show the estimation methods of

- The abridging Molodensky parameters, providing the best horizontal fit, and
- The Burša-Wolf parameters, providing the best spatial fit.

Estimation of the abridging Molodensky-parameters, providing the best horizontal fit

The procedure – verifying its name – provides direct connection between the geodetic coordinates and ellipsoidal heights in the source and the target datums. To estimate the shift parameters, we need base points, whose ellipsoidal coordinates are known on both datums. In the practice, usually the low-order geodetic base points are used as common points, whose coordinates are given in well-defined projection systems. The inverse projection parameters should be used to obtain the ellipsoidal coordinates.

The abridging Molodensky formulae are given in Equations (4.2.2), (4.2.3) and (4.2.4). Using the base point set with the coordinates both in the source and the target system, the differences between the observed and the calculated coordinates should be minimized, as follows:

$$\sum_{i=1}^{N}\left(\Phi_T^{(i)} - \Phi_S^{(i)} + \Delta\Phi''^{(i)}\left(\Phi_S,\Lambda_S\right)\right)^2 + \sum_{i=1}^{N}\left(\cos\Phi_S^{(i)}\cdot\left(\Lambda_T^{(i)} - \Lambda_S^{(i)} + \Delta\Lambda''^{(i)}\left(\Phi_S,\Lambda_S\right)\right)\right)^2 = \min \tag{10.1}$$

where the 'S' lower index indicates the source coordinates and the 'T' indicates the target ones. These values can be calculated using the Molodensky formulae as a function of the geodetic coordinates. To get the optimum in a planar system instead of the geodetic system, the longitude difference is scaled by $\cos(\Phi)$ The condition of the minimum is that the partial derivative of the square sums of the differences in the first two equations, by the parameters, should be all zeroes.

Doing the partial derivations and using the value C=a·df+f·da, the Equation (10.1) can be expressed in the following matrix form:

$$\mathbf{Ax} = \mathbf{b} \tag{10.2}$$

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

$$A_{11} = \sum_{i=1}^{N}\left[\left(\frac{\sin\Phi^{(i)}\cos\Lambda^{(i)}}{M^{(i)}\sin 1''}\right)^2 + \left(\frac{\sin\Lambda^{(i)}}{N^{(i)}\cos\Phi^{(i)}\sin 1''}\right)^2\right]$$

$$A_{12} = \sum_{i=1}^{N}\left[\frac{\sin^2\Phi^{(i)}\sin\Lambda^{(i)}\cos\Lambda^{(i)}}{\left(M^{(i)}\sin 1''\right)^2} - \frac{\sin\Lambda^{(i)}\cos\Lambda^{(i)}}{\left(N^{(i)}\cos\Phi^{(i)}\sin 1''\right)^2}\right]$$

$$A_{13} = \sum_{i=1}^{N}\left[\frac{-\sin\Phi^{(i)}\cos\Phi^{(i)}\cos\Lambda^{(i)}}{\left(M^{(i)}\sin 1''\right)^2}\right]$$

$$A_{22} = \sum_{i=1}^{N}\left[\left(\frac{\sin\Phi^{(i)}\sin\Lambda^{(i)}}{M^{(i)}\sin 1''}\right)^2 + \left(\frac{\cos\Lambda^{(i)}}{N^{(i)}\cos\Phi^{(i)}\sin 1''}\right)^2\right]$$

$$A_{23} = \sum_{i=1}^{N}\left[\frac{-\sin\Phi^{(i)}\cos\Phi^{(i)}\sin\Lambda^{(i)}}{\left(M^{(i)}\sin 1''\right)^2}\right]$$

$$A_{33} = \sum_{i=1}^{N}\left[\left(\frac{\cos\Phi^{(i)}}{M^{(i)}\sin 1''}\right)^2\right]$$

where the elements of the (symmetric) matrix **A** and the vector **b** are:

$$b_1 = \sum_{i=1}^{N}\left[\frac{-\Delta\Phi^{(i)}\sin\Phi^{(i)}\cos\Lambda^{(i)}}{M^{(i)}\sin 1''} + \frac{C\sin 2\Phi^{(i)}\sin\Phi^{(i)}\cos\Lambda^{(i)}}{\left(M^{(i)}\sin 1''\right)^2} - \frac{\Delta\Lambda^{(i)}\sin\Lambda^{(i)}}{N^{(i)}\cos\Phi^{(i)}\sin 1''}\right]$$

$$b_2 = \sum_{i=1}^{N}\left[\frac{-\Delta\Phi^{(i)}\sin\Phi^{(i)}\sin\Lambda^{(i)}}{M^{(i)}\sin 1''} + \frac{C\sin 2\Phi^{(i)}\sin\Phi^{(i)}\sin\Lambda^{(i)}}{\left(M^{(i)}\sin 1''\right)^2} + \frac{\Delta\Lambda^{(i)}\cos\Lambda^{(i)}}{N^{(i)}\cos\Phi^{(i)}\sin 1''}\right]$$

$$b_3 = \sum_{i=1}^{N}\left[\frac{\Delta\Phi^{(i)}\cos\Phi^{(i)}}{M^{(i)}\sin 1''} - \frac{C\sin 2\Phi^{(i)}\cos\Phi^{(i)}}{\left(M^{(i)}\sin 1''\right)^2}\right] \qquad (10.3)$$

In the Equation (10.3) all the coordinates, the M and N values are interpreted in the source system. This is an inhomogeneous linear equation system, whose solution is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \qquad (10.4)$$

where $\mathbf{A}^{-1}$ is the inverse of the matrix **A.** The solution vector **x** contains the *dX*, *dY* and *dZ* parameters. In the practice, the parameters can be easily determined by the Cramer rule.

Estimation of the Burša-Wolf parameters

Here the goal is to provide such parameters of the BW-transformation (Equation 4.3.1) that provide minimum difference between the measured (given) coordinates of a point set and the calculated coordinates of the same set, deriving from the coordinates in the other datum by the BW-method. It is formulated as follows:

$$\sum_{j=1}^{M=2,3}\sum_{i=1}^{N}\left(\tilde{X}_{(2),j}^{(i)} - X_{(2),j}^{(i)}\left(\tilde{X}_{(1),1}^{(i)}, \tilde{X}_{(1),2}^{(i)}, \ldots\right)\right)^2 = \min \qquad (10.5)$$

In Equation (10.5) the ~ sign refers to the measured data, while the lower index indicates the source (1) and the target (2) systems. The index *i* runs for the point set, the number of the point in this set is *N*, while the index *j* refers to the dimension – in case of planar coordinates, it is 2 while at spatial coordinates, *j*=3.

If we substitute the following variables in the Helmert transformation:

$$A = (1+\kappa)$$
$$B = -(1+\kappa)\cdot\gamma$$
$$C = (1+\kappa)\cdot\beta$$
$$D = -(1+\kappa)\cdot\alpha \qquad (10.6)$$

:

XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/

it appears as follows

$$\begin{aligned}
\mathbf{X}_{(2)} &= \mathbf{dX} + A\cdot\mathbf{X}_{(1)} + B\cdot\mathbf{Y}_{(1)} + C\cdot\mathbf{Z}_{(1)} \\
\mathbf{Y}_{(2)} &= \mathbf{dY} - B\cdot\mathbf{X}_{(1)} + A\cdot\mathbf{Y}_{(1)} + D\cdot\mathbf{Z}_{(1)} \\
\mathbf{Z}_{(2)} &= \mathbf{dZ} - C\cdot\mathbf{X}_{(1)} - D\cdot\mathbf{Y}_{(1)} + A\cdot\mathbf{Z}_{(1)}
\end{aligned} \tag{10.7}$$

These equations are linear ones to the parameters to be estimated. Again, the least squares method can be applied for the estimation of the parameters. The minimum condition is:

$$\begin{aligned}
&\sum_{i=1}^{N}\left[\mathbf{X}_{(2)}^{(i)} - \left(\mathbf{dX} + A\cdot\mathbf{X}_{(1)}^{(i)} + B\cdot\mathbf{Y}_{(1)}^{(i)} + C\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]^2 + \\
&\sum_{i=1}^{N}\left[\mathbf{Y}_{(2)}^{(i)} - \left(\mathbf{dY} - B\cdot\mathbf{X}_{(1)}^{(i)} + A\cdot\mathbf{Y}_{(1)}^{(i)} + D\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]^2 + \\
&\sum_{i=1}^{N}\left[\mathbf{Z}_{(2)}^{(i)} - \left(\mathbf{dZ} - C\cdot\mathbf{X}_{(1)}^{(i)} - D\cdot\mathbf{Y}_{(1)}^{(i)} + A\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]^2 = \min
\end{aligned} \tag{10.8}$$

The equation system (10.8) contains the measured coordinates of the identical points, **X, Y, Z** are their coordinates in the source and the target systems. The minimum condition is set for the square-sums of the minimums between the measured and the calculated coordinates. In other words, it is set for the squares of the metric distances. The condition is similar to minimization of the absolute value of the distance difference.

The condition of the minimum is that the partial derivatives according to the parameters (the **dX, dY, dZ,** *A, B, C* and *D* values) should be zeroes. These partial derivatives are:

$$\begin{aligned}
&\frac{\partial}{\partial\mathbf{dX}}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{X}_{(2)}^{(i)} - \left(\mathbf{dX} + A\cdot\mathbf{X}_{(1)}^{(i)} + B\cdot\mathbf{Y}_{(1)}^{(i)} + C\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right] = 0 \\
&\frac{\partial}{\partial\mathbf{dY}}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{Y}_{(2)}^{(i)} - \left(\mathbf{dY} - B\cdot\mathbf{X}_{(1)}^{(i)} + A\cdot\mathbf{Y}_{(1)}^{(i)} + D\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right] = 0 \\
&\frac{\partial}{\partial\mathbf{dZ}}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{Z}_{(2)}^{(i)} - \left(\mathbf{dZ} - C\cdot\mathbf{X}_{(1)}^{(i)} - D\cdot\mathbf{Y}_{(1)}^{(i)} + A\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right] = 0 \\
&\frac{\partial}{\partial A}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{X}_{(2)}^{(i)} - \left(\mathbf{dX} + A\cdot\mathbf{X}_{(1)}^{(i)} + B\cdot\mathbf{Y}_{(1)}^{(i)} + C\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{X}_{(1)}^{(i)} + \\
&\quad ,\sum_{i=1}^{N} -2\cdot\left[\mathbf{Y}_{(2)}^{(i)} - \left(\mathbf{dY} - B\cdot\mathbf{X}_{(1)}^{(i)} + A\cdot\mathbf{Y}_{(1)}^{(i)} + D\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Y}_{(1)}^{(i)} + \\
&\quad ,\sum_{i=1}^{N} -2\cdot\left[\mathbf{Z}_{(2)}^{(i)} - \left(\mathbf{dZ} - C\cdot\mathbf{X}_{(1)}^{(i)} - D\cdot\mathbf{Y}_{(1)}^{(i)} + A\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Z}_{(1)}^{(i)} = 0 \\
&\frac{\partial}{\partial B}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{X}_{(2)}^{(i)} - \left(\mathbf{dX} + A\cdot\mathbf{X}_{(1)}^{(i)} + B\cdot\mathbf{Y}_{(1)}^{(i)} + C\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Y}_{(1)}^{(i)} + \\
&\quad ,\sum_{i=1}^{N} 2\cdot\left[\mathbf{Y}_{(2)}^{(i)} - \left(\mathbf{dY} - B\cdot\mathbf{X}_{(1)}^{(i)} + A\cdot\mathbf{Y}_{(1)}^{(i)} + D\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{X}_{(1)}^{(i)} = 0 \\
&\frac{\partial}{\partial C}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{X}_{(2)}^{(i)} - \left(\mathbf{dX} + A\cdot\mathbf{X}_{(1)}^{(i)} + B\cdot\mathbf{Y}_{(1)}^{(i)} + C\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Z}_{(1)}^{(i)} + \\
&\quad ,\sum_{i=1}^{N} 2\cdot\left[\mathbf{Z}_{(2)}^{(i)} - \left(\mathbf{dZ} - C\cdot\mathbf{X}_{(1)}^{(i)} - D\cdot\mathbf{Y}_{(1)}^{(i)} + A\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{X}_{(1)}^{(i)} = 0 \\
&\frac{\partial}{\partial D}:\ \sum_{i=1}^{N} -2\cdot\left[\mathbf{Y}_{(2)}^{(i)} - \left(\mathbf{dY} - B\cdot\mathbf{X}_{(1)}^{(i)} + A\cdot\mathbf{Y}_{(1)}^{(i)} + D\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Z}_{(1)}^{(i)} + \\
&\quad ,\sum_{i=1}^{N} 2\cdot\left[\mathbf{Z}_{(2)}^{(i)} - \left(\mathbf{dZ} - C\cdot\mathbf{X}_{(1)}^{(i)} - D\cdot\mathbf{Y}_{(1)}^{(i)} + A\cdot\mathbf{Z}_{(1)}^{(i)}\right)\right]\cdot\mathbf{Y}_{(1)}^{(i)} = 0
\end{aligned} \tag{10.9}$$

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

In the Equation (10.9), the seven parameters can be moved before the summation in each row. After rearrangement, this can be written as an inhomogeneous linear equation system, similar to the form of Equation (10.2). Making the derivations,

$$\begin{bmatrix} -1 & 0 & 0 & \mathbf{X}_{(1)} & \mathbf{Y}_{(1)} & \mathbf{Z}_{(1)} & 0 \\ 0 & -1 & 0 & \mathbf{Y}_{(1)} & -\mathbf{X}_{(1)} & 0 & \mathbf{Z}_{(1)} \\ 0 & 0 & -1 & \mathbf{Z}_{(1)} & 0 & -\mathbf{Y}_{(1)} & -\mathbf{X}_{(1)} \\ \mathbf{X}_{(1)} & \mathbf{Y}_{(1)} & \mathbf{Z}_{(1)} & \mathbf{X}_{(1)}^2+\mathbf{Y}_{(1)}^2+\mathbf{Z}_{(1)}^2 & 0 & 0 & 0 \\ \mathbf{Y}_{(1)} & -\mathbf{X}_{(1)} & 0 & 0 & \mathbf{X}_{(1)}^2+\mathbf{Y}_{(1)}^2 & \mathbf{Y}_{(1)}\mathbf{Z}_{(1)} & -\mathbf{X}_{(1)}\mathbf{Z}_{(1)} \\ \mathbf{Z}_{(1)} & 0 & -\mathbf{Y}_{(1)} & 0 & \mathbf{Y}_{(1)}\mathbf{Z}_{(1)} & \mathbf{X}_{(1)}^2+\mathbf{Z}_{(1)}^2 & \mathbf{X}_{(1)}\mathbf{Y}_{(1)} \\ 0 & \mathbf{Z}_{(1)} & -\mathbf{X}_{(1)} & 0 & -\mathbf{X}_{(1)}\mathbf{Z}_{(1)} & \mathbf{X}_{(1)}\mathbf{Y}_{(1)} & \mathbf{Y}_{(1)}^2+\mathbf{Z}_{(1)}^2 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{dX} \\ \mathbf{dY} \\ \mathbf{dZ} \\ A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(2)} \\ \mathbf{Y}_{(2)} \\ \mathbf{Z}_{(2)} \\ \mathbf{X}_{(1)}\mathbf{X}_{(2)}+\mathbf{Y}_{(1)}\mathbf{Y}_{(2)}+\mathbf{Z}_{(1)}\mathbf{Z}_{(2)} \\ \mathbf{X}_{(2)}\mathbf{Y}_{(1)}-\mathbf{Y}_{(2)}\mathbf{X}_{(1)} \\ \mathbf{X}_{(2)}\mathbf{Z}_{(1)}-\mathbf{Z}_{(2)}\mathbf{X}_{(1)} \\ \mathbf{Y}_{(2)}\mathbf{Z}_{(1)}-\mathbf{Z}_{(2)}\mathbf{Y}_{(1)} \end{bmatrix}$$

occurs. The elements of the matrix **A** and the vector **b** – similarly to the solution of the abridging Molodensky method – show the sums of the values of all base points of the set. We shall use this simplification to avoid an equation image too complex and make it ready to print. Where the squares or the mixed products of the coordinates occur among the matrix or vector elements, the summarization should be made for them. Together with the omission of the sum sign, we also omit the index *i*.

The vector **x**, containing the estimated parameters, can be provided by the inverse of the matrix A, similarly to the Equation (10.4). Afterwards, according to the Equation (10.6), the κ scale factor and the α, β and γ angle values can be computed from the *A, B, C* and *D* values. For this procedure, we have to have at least three base points with their **X, Y, Z** coordinates both in the source and target datums.

View publication stats
XML to PDF by RenderX XEP XSL-FO F ormatter, visit us at http://www.renderx.com/