Mapping the sphere

It is not possible to map a portion of the sphere into the plane without introducing some distortion. There is a lot of evidence for this. For one thing you can do a simple experiment. Cut a grapefruit in half and eat one of the halves. Now try to flatten the remaining peel without the peel tearing. If that is not convincing enough, there are mathematical proofs. One of the nicest uses the formulas for the sum of the angles of a triangle on the sphere and in the plane. The fact that these are different shows that it is not possible to find a map from the sphere to the plane which sends great circles to lines and preserves the angles between them. The question then arises as to what is possible. That is the subject of these pages.

We will present a variety of maps and discuss the advantages and disadvantages of each. The easiest such maps are the central projections. Two are presented, the gnomonic projection and the stereographic projection. Although many people think so, the most important map in navigation, the Mercator projection, is not a central projection, and it will be discussed next. Finally we will talk briefly about a map from the sphere to the plane which preserves area, a fact which was observed already by Archimedes and used by him to discover the area of a sphere. All of these maps are currently used in mapping the earth. The reader should consult an atlas, such as those published by Rand Macnally, or the London Times. On each of the charts in such an atlas the name of the projection used will be indicated. The variety of projections used may be surprising.

There is a quantitative way of measuring distortion, and how it changes from place to place on the sphere. The distortion ellipse provides a way of graphically displaying this information. We will compute and display the distortion ellipse for each of the maps we discuss.

Properties of the sphere

We will specify the location of a point on the sphere in terms of latitude and longitude. These must be defined with respect to some reference point R on a fixed reference great circle E called the ***equator.*** Corresponding to E there are two poles which we will label with N and S and refer to as the ***north and south poles.*** The equator splits the sphere into two hemispheres, called the ***northern and southern hemispheres.***

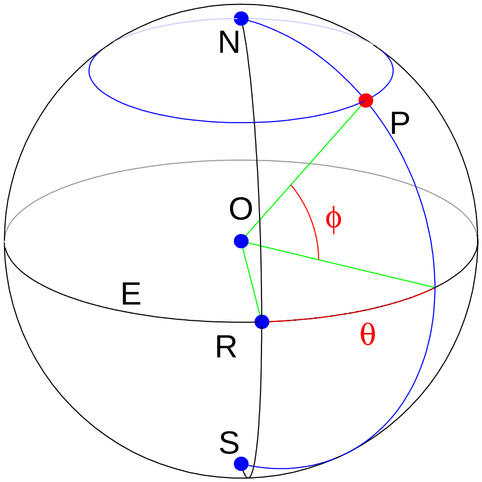

Figure 1. Latitude and longitude.

Given any point P on the sphere, which is different from the poles, there is a unique great circle through P and the two poles. The half of that great circle containing P and terminated by the poles is called the ***meridian*** of P. The angle measured at the center of the sphere along the meridian from P to the intersection of the meridian and the equator is called the ***latitude*** of P. We will consider latitude to be positive (or north) in the northern hemisphere and negative (south) in the southern hemisphere. Look at Figure 1, where the latitude is the angle ϕ. The meridian through our reference point R is called the ***prime meridian.***

The locus of points having a constant latitude is called a ***parallel of latitude.*** It is actually a (small) spherical circle with center at one of the poles. The only parallel which is a great circle is the equator itself. Figure 1 shows the parallel of latitude and the meridian for the point P.

The counter-clockwise direction along the equator E from R when viewed from the north pole is called the positive (or eastern) direction. For our point P, the angle along the equator as measured from the center of the sphere from R to the intersection of the meridan of P is called the ***longitude*** of P. In Figure 1 the longitude of P is the angle θ. Longitude is postive (east) or negative (west) depending on the direction the angle is measured from R.

On the earth, by international agreements going back to 1894, the prime meridian is the meridian which passes through the center of the transit at the observatory in Greenwich, England. The reference point R is the intersection of the equator with the prime meridian. Longitude and latitude are measured in degrees. Longitude is denoted as east or west depending on whether the location in question is east or west of the prime meridian. Similarly latitude is designated to be either north or south. With these conventions Salt Lake City is located at $40^\circ\ 46'$N, and $111^\circ\ 53'$W. The meridian located at 180°E is the same as that at 180°W. This meridian runs almost entirely through the Pacific Ocean, and coincides for the most part with the international date line.

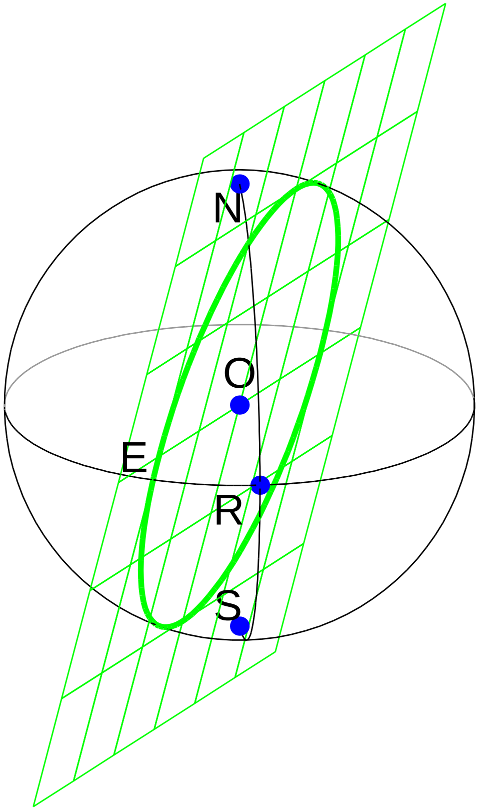

Figure 2. A great circle.

There are two classes of special curves on the sphere. The first is the class of ***geodesics***, i.e., the curves of shortest length connecting two points. It turns out that a geodesic on the sphere is a segment of a ***great circle***, i.e., the intersection of a plane through the center of the sphere with the sphere itself (see Figure 3). The importance of such curves for navigation is therefore clear. To get from one point to another in the shortest time we should follow a great circle. This is what airplanes do when traveling long distances such as from America to europe.

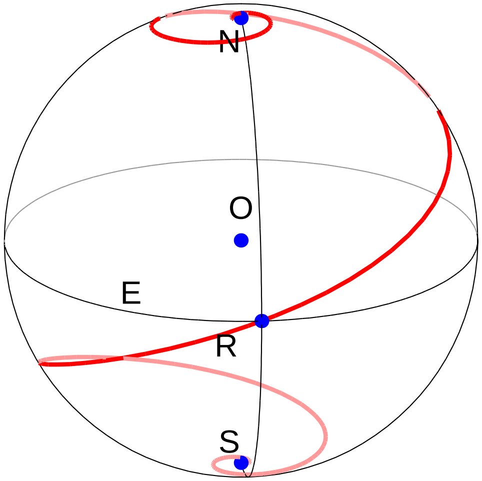

Figure 3. A rhumb line.

The second important type of curve is the ***rhumb line*** or ***loxodrome.*** A rhumb line is a curve which intersects all of the meridians of longitude at the same angle. For example, if two points have the same latitude, then the rhumb line connecting them is the parallel of that latitude. This example makes it clear that rhumb lines are not the same as great circles. A more typical example of a rhumb line is shown in Figure 3. For ships equipped with compasses, the easiest course to steer is one with a constant compass direction. Such courses are precisely the rhumb lines. On the other hand, steering a course along a great circle requires constant course changes, unless the great circle happens to be a rhumb line. Consequently rhumb lines are also very important to navigators. When steering a ship across an ocean, a navigator will plot a great circle to minimize distance, but he will then approximate the great circle with rhumb line segments to make it easy on the helmsman.

Exercise: Describe all great circles which are rhumblines.

The planar gnomonic projection

The maps of the sphere which are easiest to understand are the central projections. For these we choose a point called the center of the projection and an image plane, which is usually tangent to the sphere at some point. Then to find the image of a point we simply take the line through the center and the point and find where it intersects the plane. This process can be likened to taking a photograph of the sphere.

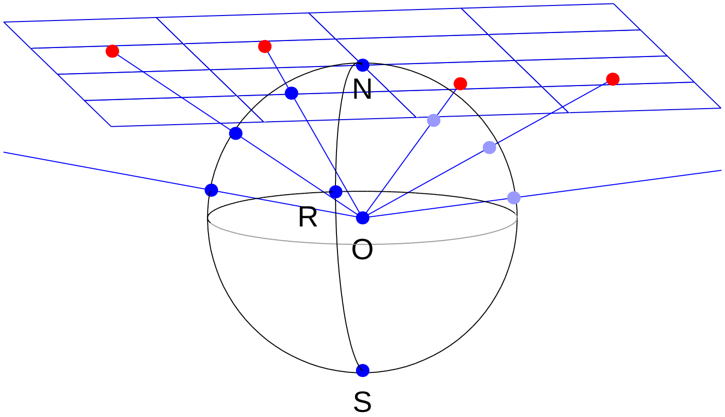

Figure 4. The gnomonic projection.

We should say a word about nomenclature. In cartography, the word ***projection*** is used synonymously with the word map. Central projections are a special type of maps defined as in the previous paragraph. This may seem confusing. Central projections clearly deserve the name projection, since they may be considered to be formed by projecting light from the center of projection, and collecting the image of the sphere on the image plane. However, we will see other maps (also called projections) which are not central projections.

For our first projection, the ***gnomonic projection***, we will take the center of the projection to be the center of the sphere, and the image plane to be the plane tangent to the sphere at some point. Usually this will be the north pole, but it really does not have to be. The gnomonic projection is illustrated in Figure 4. The image of a point on the sphere is the intersection of the line through the point and the center of the sphere with the image plane. In Figure 4 the images are shown in red.

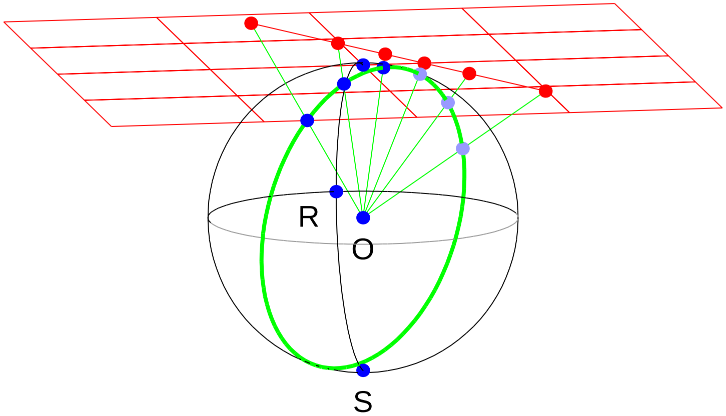

Figure 5. The image of a great circle is a straight line.

By definition, a great circle is the intersection of a plane through the center of the sphere with the sphere. Every line through the center of the sphere and a point on the great circle lies in that plane. Hence the image of the great circle under the projection is the intersection of the same plane with the image plane (see Figure 5). So we see that great circles on the sphere are mapped by the gnomonic projection into straight lines. This is the most important property of the gnomonic projection, and it is why the gnomonic projection has become increasingly important as long distance airplane travel has become more common.

We will analyse the case where the image plane is tangent at the north pole, N. In this case the projection will map the northern hemisphere onto the entire tangent plane, and it is not defined for points that are not in the northern hemisphere. Notice that if we look at the image plane from above, it will have North America on the bottom, and Eurasia on the top, with the north pole in the middle.

You can find out more about the gnomonic projection and an example here.

Distortion in maps

The inevitable distortion in a map differs from point to point on the sphere, and from map to map. Here we will discuss this in a general framework. However, since we have introduced the gnomonic projection in the previous section, we will be able to use it as an example of what we will ultimately do for each of the maps that we discuss.

The best way to discuss distortion is to use the calculus. We want to avoid the use of the calculus for the time being. Instead we will rely on a geometric analysis. the basic idea is to see what the map does to a disk which is tangent to the sphere at the point in question. (The discerning reader will notice in our analysis the point where we invoke the use of the terms “very small” or “tangent,” which signal that we are using the ideas of the calculus.) For any map there is a direction of maximal expansion and another of minimal expansion. The image of the disk will be an ellipse with these directions as the major and minor axes. This ellipse is called the ***distortion ellipse.***

For the maps we consider, the axes of the distortion ellipse are in the north/south and the east/west directions. This makes the analysis somewhat easier. We will illustrate the geometric process first for the gnomonic projection.

For a point A on the sphere with latitude ϕ and longitude θ, let A' denote the image under the projection. The plane containing the center of the sphere 0, A, A', and the north pole N is shown in Figure 6 and from this figure we see that the distance $|OA'|$ between 0 and A' is $|OA'| = 1/\sin\phi$.

Next consider a disk D of very small radius r which is tangent to the sphere at A. We will analyse the distortion by examining what happens to this disk under the projection. We will consider the projection as occuring in two steps. In the first step, which might be called the expansion phase, the disk D is subjected to a central projection

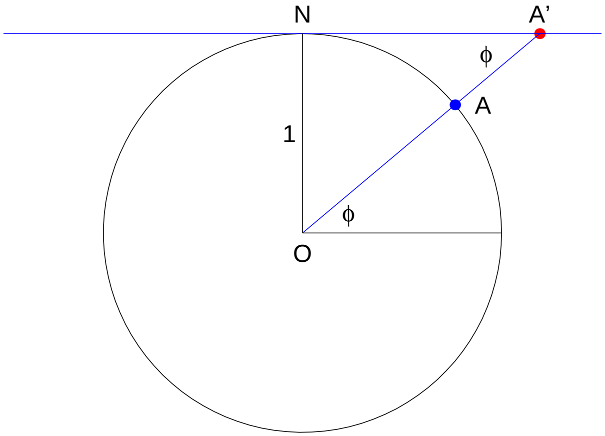

Figure 6. Expansion in the gnomonic projection.

from O onto the plane through A' which is parallel to the disk D. The image is again a disk, D' of radius r'. By the similar triangles in Figure 7, we see that

$$r'/r = |OA'|/1.$$

Thus $r' = r/\sin\phi$.

The second step projects the disk D' onto the image plane P. This time we expect that the disk will be distorted. The image will no longer be a disk. In fact in the east/west direction, the disk D' intersects the image plane, so there is no change under this last projection. In the north/south direction, we have the situation in Figure 8. Since the radius r' is very small in comparison to the distance to the center of projection, the angle $A'BC$ is close to a right angle. Thus if r'' is the distance between A' and C, we have

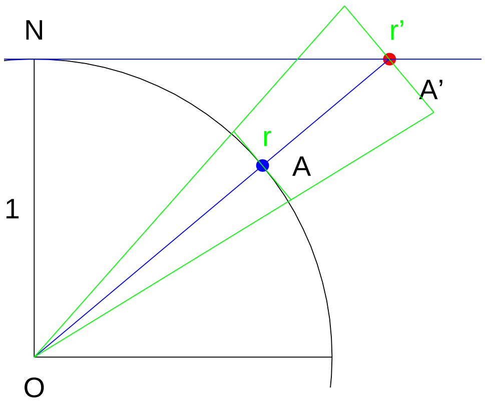

Figure 7. Expansion phase for the gnomonic projection.

$$r'' = r'/\sin\phi = r/\sin^2\phi.$$

Thus the image D'' of D is not a disk, but an ellipse with semi-axes $r/\sin\phi$ and $r/\sin^2\phi$ (see Figure 9).

What we have just done with the gnomonic projection we will do with the other projections that we will consider. We will find that a small tangent disk is mapped into an ellipse, and we will be able to determine the semi-axes of this ellipse. The picture for any of the projections which we will study is very much like Figure 9. The ratio of the semiaxis r'' to r will measure the expansion in the north/south direction, and the ratio of r' to r will measure that in the east/west direction.

The information about distortion is summed up in the ellipse at the right in Figure 9. This ellipse is therefore called the ***distortion ellipse.*** Notice that this ellipse varies

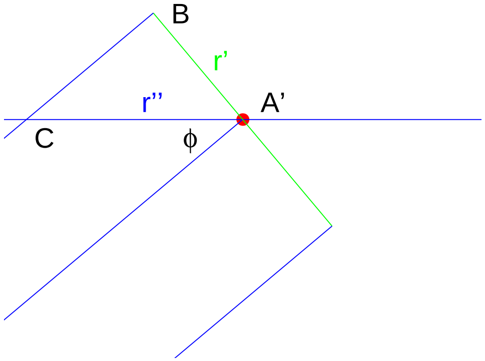

Figure 8. Distortion phase for the gnomonic projection.

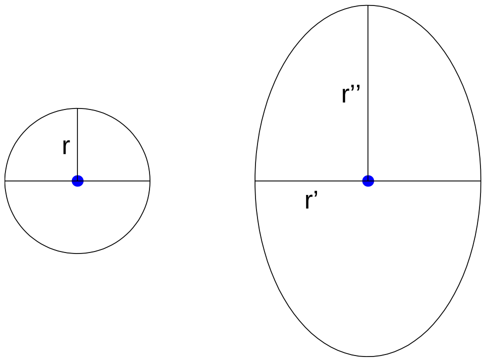

Figure 9. Distortion in the gnomonic projection.
$r' = r/\sin\phi$ and $r'' = r/\sin^2\phi$

from point to point on the sphere. It might be a circle at some points and very elongated

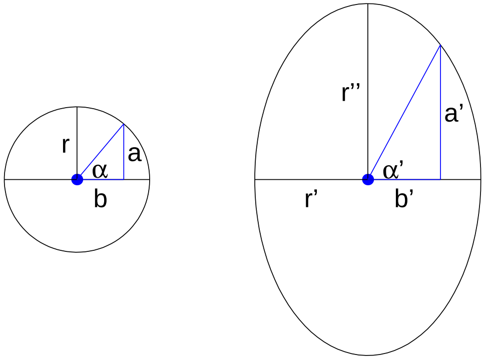

Figure 10. The effect of distortion on angles.

at others.

Now let's examine what the distortion does to the angles between curves. Consider the case of the angle α in the disk D and its image α' in D'', as indicated in Figure 10. We have

$$\tan\alpha = a/b \quad \text{and} \quad \tan\alpha' = a'/b'.$$

Because the expansion is different in different directions we see that

$$a' = a\frac{r''}{r} \quad \text{and} \quad b' = b\frac{r'}{r}.$$

Hence

$$\tan\alpha' = \frac{r''}{r'}\tan\alpha.$$

Thus all such angles will be the same in the image if and only if $r'' = r'$, i.e., when the semi-axes of the image ellipse are equal. This happens when the image ellipse is actually a circle. In general the image ellipse will not be a circle, and we will be able to conclude that the mapping does not preserve the angles between curves. On the other hand, if the ellipse is a circle at every point, then the mapping does preserve the angles between curves. Such a mapping will be called a ***conformal*** mapping.

In the case of the gnomonic projection, we have

$$\frac{r''}{r'} = 1/\sin\phi,$$

and we conclude that in general the image of the angle α is not equal to α. Thus the gnomonic projection is not conformal. The next map we consider will have this property.

Look again at Figure 9. The area of the circle is πr^2, and the area of the image ellipse is $\pi r' r''$. These two will be the same if and only if

$$r' r'' = r^2.$$

It takes some doing, but it can be shown that if this condition is true at every point, then the image of any set is equal to the area of the set itself. Such a projection is called an ***equi-area projection***, or an ***area preserving projection***.

Since for the gnomonic projection

$$r' r'' = r^2/\sin^3\phi,$$

we see that the gnomonic projection is not area preserving. We will later look at a projection which is area preserving.

The stereographic projection

Like the gnomonic projection, the stereographic projection is a central projection. However, this time the center of the projection is a point on the sphere and the image plane is the tangent plane through the point which is antipodal to the center of the projection. We will consider the case when the center of the projection is the north pole N. Then the image plane is the plane tangent to the sphere at the south pole. See Figure 11. The stereographic projection maps the sphere with the north pole deleted onto this plane.

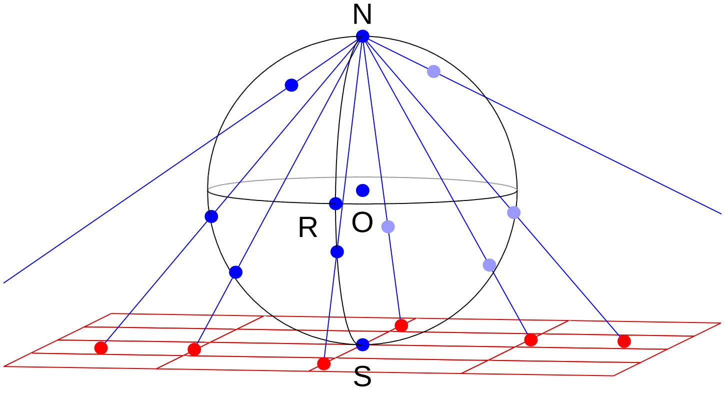

Figure 11. The stereographic projection.

To determine the distortion inherent in the stereographic projection we proceed exactly as we did with the gnomonic projection. We break the effect of the mapping on a small tangent disk into an expansion phase and a distortion phase. Let A be a point on the sphere with latitude ϕ and longitude θ. Let A' denote the image of A under the stereographic projection. The plane through the center of the sphere containing A and A' is shown in Figure 12.

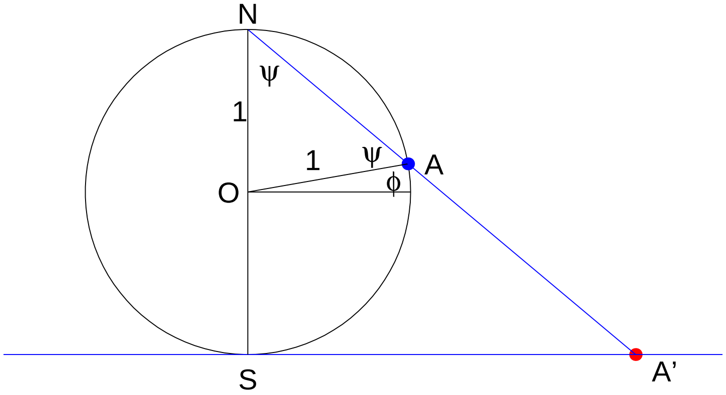

Figure 12. The stereographic projection.

From this figure we see that the triangle NOA is isoceles, with two sides of length 1. Thus the angles ONA and OAN are equal. If we denote this angle by ψ, then the sum of the angles formula becomes

$$\psi + \psi + (\pi/2 - \phi) = \pi.$$

Thus $\psi = \phi/2 + \pi/4$. In terms of this angle we find that the distance $|NA|$ between N and A is given by

$$|NA| = 2\cos\psi.$$

Much easier is the fact that the distance $|NA'|$ between N and A' satisfies $|NA'|\cos\psi = 2$. Thus

$$|NA'| = 2/\cos\psi.$$

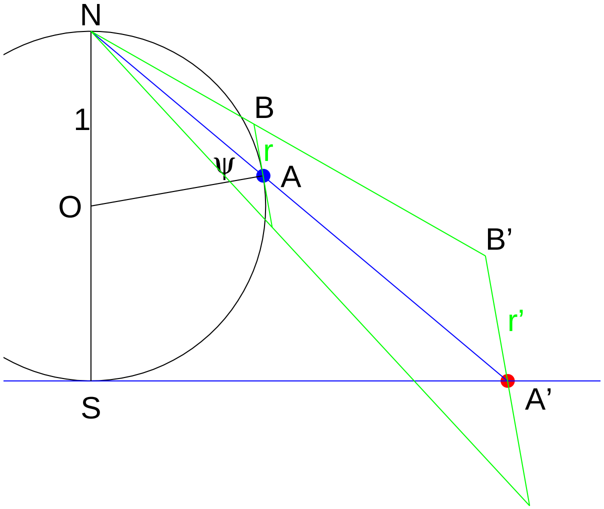

Figure 13. Expansion phase for the stereographic projection.

Now we can examine the expansion phase using Figure 13. We see that a disk D of radius r which is tangent to the sphere at A is sent into a parallel disk D' through A' of radius r', and similar triangles show immediately that $r'/r = |NA'|/|NA|$. Thus

$$r' = r/\cos^2\psi = r\sec^2(\phi/2 + \pi/4).$$

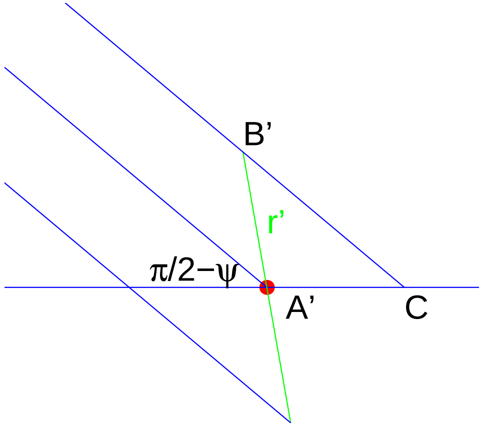

Figure 14. Distortion phase for the stereographic projection.

The distortion phase is illustrated in Figure 14. First a look at Figure 11 convinces us that $\angle SA'N = \pi/2 - \psi$, and it is so labeled in Figure 14. The disk D' with radius r' is projected onto the image plane. Again in the east/west direction, the plane of the disk D' and the image plane intersect, so there is no change. In the north/south direction we have the situation illustrated in Figure 14. The radius r' of D' is the segment $A'B'$. Considering that r' is very small in comparison to the distance $|NA'|$, we may assume that the projection is parallel. Then the segment $A'B'$ is projected into the segment $A'C$, and the segment $B'C$ is parallel to NA'. Using this fact we see that $\angle A'B'C = \pi/2 - \psi$, and that $\angle A'CB' = \pi/2 - \psi$. This means that the triangle $B'A'C$ is isoceles, and that $|A'C| = |A'B'| = r'$.

Consequently the stereographic projection expands a tangent disk of radius r into a disk of radius $r' = r\sec^2(\phi/2 + \pi/4)$ (see Figure 15). Thus the distortion ellipse is

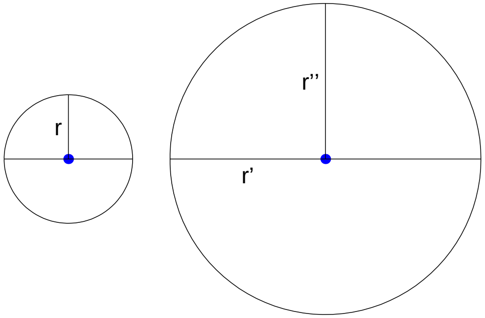

Figure 15. Distortion in the stereographic projection.
$r' = r'' = r \sec^2(\phi/2 + \pi/4)$.

a disk, and by the argument given at the end of the previous section, we know that the stereographic projection is conformal. This property makes the stereographic projection very important mathematically. It is occasionally used in cartography as well, although it is not used as frequently as other types of maps, especially the one to be described next.

The Mercator projection

The Mercator projection is probably the map that is most familiar to all of us. It was invented by the Dutch mathematican and cartographer Gerhardus Mercator in 1569. Its basic properties made it very useful to navigators, and it became popular immediately. These properties are three in number:

- The Mercator projection maps the meridians of longitude into parallel straight lines.
- It maps the parallels of latitude into parallel straight lines which meet the meridians at right angles.
- the map is conformal.

The immediate result of these three properties is that the Mercator projection maps rhumb lines on the sphere into straight lines, and vice versa. As a result if a navigator wants to find the course to steer to get from point A to point B, he needs only to find these points on his Mercator projection, and to draw the straight line between them. The Mercator projection is still the most important map for navigational purposes.

Mercator never published a mathematical description of his projection. He simply published and sold the maps. The first mathematical description was published in 1599 by Edward Wright, a mathematican at Caius College, Cambridge, in a book entitled ***Certaine Errors in Navigation***. His description is quite intuitive and easily displays the major features of the map. Imagine that the sphere is a balloon, contained in a cylinder which is tangent to the sphere at the equator. Now blow up the balloon. As the balloon expands, its expansion is limited by the cylinder. As each part of the balloon expands enough so that it reaches the cylinder, the expansion of that part stops. After enough of the balloon has been applied to the cylinder, it is cut along a meridian, and unrolled onto a plane. The result is the Mercator projection.

While Wright's description is graphic, we will have to do some work to put it to use. Following Wright, we will discuss the distortion first and use that to figure out how to make a Mercator projection. Consider what happens under the projection to a small disk tangent to a point of the sphere at latitude ϕ. Since as the balloon expands, it expands the same in every direction, the disk is expanded, but it always remains a disk. Then when the center of the disk hits the cylinder the expansion stops immediately. Thus under the Mercator projection, a tangent disk is sent into a disk. Consequently the map is conformal.

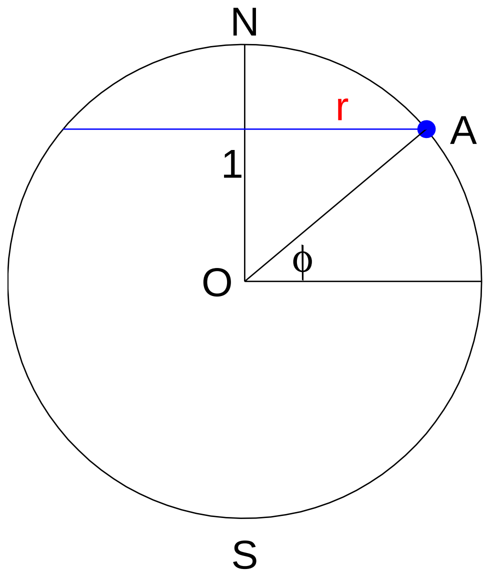

Figure 16. Expansion in the Mercator projection.

We have yet to determine the expansion factor. To determine how much the tangent disk is expanded, notice that the parallel of latitude ϕ is a circle in space. The radius of this circle can be found using Figure 16 to be $r = \cos\phi$. Thus each distance along the parallel must be multiplied by $1/\cos\phi = \sec\phi$. By the conformality, this is the expansion factor for the tangent disk. The distortion ellipse is a circle, as indicated in Figure 17.

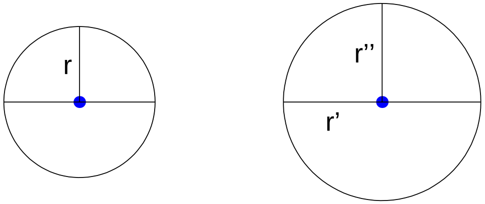

Figure 17. Distortion in the Mercator projection.
$r' = r'' = r \sec\phi$.

Because of the conformality, the tangent disk is expanded by the factor $\sec\phi$ in all directions. In particular this is true in the north/south direction, and this means that if $\Delta\phi$ is a small increment of latitude beginning at the latitude of ϕ, then the image under the projection of this increment has length Δy, which is approximately equal to $\Delta\phi \cdot \sec\phi$, i.e.

$$\Delta y \approx \Delta\phi \cdot \sec\phi. \tag{1}$$

In Figure 18 A is a point with latitude ϕ and B has latitude $\phi + \Delta\phi$. Their images are A' and B', and the difference in the the north/south direction is Δy.

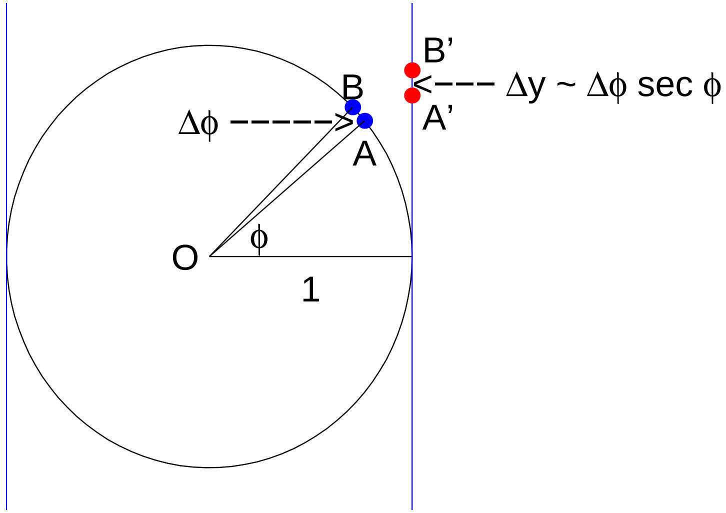

Figure 18. Expansion of latitude in the Mercator projection.

Wright used equation (1) to construct a table of what he called ***meridional parts***. He divided the meridian into a large number of very small pieces, each of the same size $\Delta\phi$ (he used $\Delta\phi = 1' = (1/60)^\circ$). For each of these increments, he calculated the corresponding approximate increment in y using equation (1) . Then to calculate the proper position for a particular latitude ϕ he simply added together all of the increments corresponding to latitudes between 0 and ϕ. In his book he published a table of the resulting values. With this information anyone could construct a Mercator projection.

In the introduction to his book, Wright is very careful to say that Mercator's chart inspired him, but that neither Mercator nor anyone else had previously shown how to construct the projection. He goes on to tell about a Dutch cartographer (Jodocus

Hondius) who had visited and worked with him in Cambridge. During his stay Wright told him of his discovery of the secret to the Mercator projection. Hondius returned to Holland and promptly published it himself without giving any credit to Wright.

An area preserving map

The area of a sphere was first computed by Archimedes. He did it by examining the map that we will discuss next. According to legend, He was so proud of this accomplishment, he directed that a diagram much like Figure 19 should be inscribed on his tomb. We will therefore call it the Archimedes projection. It also goes by the name of the Lambert equal-area projection.

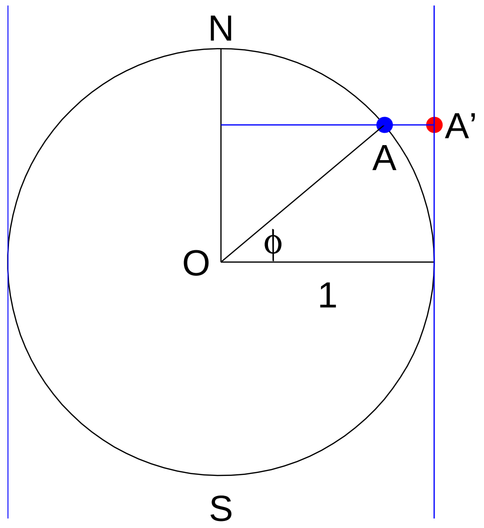

Figure 19. Archimedes projection.

The Archimedes projection is again a cylindrical projection, i.e. the image is a cylinder which encloses the sphere, and touches it along the equator. Like the Mercator projection, this one is not a central projection. Rather it is the projection from the line connecting the poles, and parallel to the equatorial plane. Thus a point on the sphere with latitude ϕ and longitude θ, is mapped into the point on the cylinder with the same longitudinal angle θ and the same height above or below the equatorial plane. Clearly this height is $\sin\phi$.

Notice that our map is defined on the sphere with the poles deleted and maps that set onto a cylinder C of height 1, and radius 1. It is no accident that the area of C is 4π, the same as that of the sphere. Our new map has the very interesting property that it maps any region on the sphere into a region in the plane which has exactly the same area.

Now let's check the distortion. Consider a small disk D of radius δ which is tangent to the sphere at a point A at latitude ϕ. It should be clear that the distortion does not depend on the longitude. Along the parallel of latitude ϕ, the stretching is the same as it was in the Mercator projection. I.e., the semiaxis of the distortion ellipse in the east/west direction is $r' = r\sec\phi$.

In the north/south direction we use the fact the projection is parallel to the equatorial plane. See Figure 20. By elementary trigonometry, the semi-axis of D' in the north/south direction must be $r'' = r \cdot \cos\phi$. Figure 21 shows the distortion ellipse for this case.

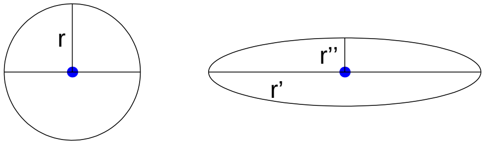

Figure 21. Distortion in the Archimedes projection.
$r' = r\sec\phi$ and $r'' = r\cos\phi$.

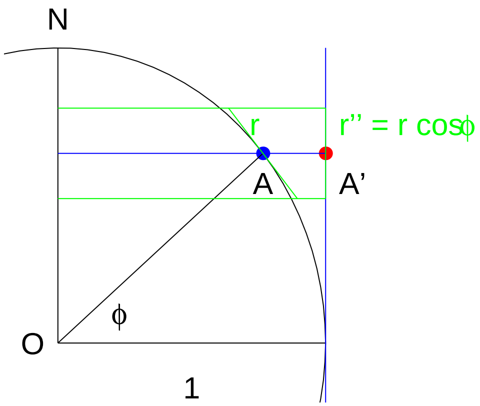

Figure 20. Distortion in the north/south direction in the Archimedes projection.

Notice that we have

$$r'r'' = (r\sec\phi)(r\cos\phi) = r^2.$$

Thus by the previous discussion we know that our map is area preserving. The area ofthe image of a more general region on the sphere is equal to the area of the region itself. This is probably the simplest map with this property, but it is not the only one. The London Times Atlas is particularly fond of area preserving maps, or equal-area maps, as they are called in that atlas.

Exercise: Suppose that T is a triangle with one vertex at a pole and the other two vertices on the equator. Show by direct calculation that T and its image under this map have the same area.