18

Mathematical Mapping from Mercator to the Millennium

Robert Osserman
Mathematical Sciences Research Institute

Note: The material presented here ranges from elementary descriptive material all the way to recently developed ideas in complex analysis. It is written throughout in a manner designed to convey the intuitive and geometric ideas behind the mathematics, so that readers may be able to get something out of even the parts that they cannot follow in detail. By way of specific background, section 1 uses only algebra and trigonometry plus the definition of a derivative. Section 2 adds elementary facts about linear algebra and 2×2 matrices. Section 3 deals with maps of the plane, while the remaining sections focus on functions of a complex variable. Readers should not feel discouraged if they find that later sections require more mathematical experience than they currently possess.

It came as something of a revelation, after years of working in and around the subject, to discover that the single, simple, intuitive notion of the *scale* of a map underlies an astonishingly wide swath of basic mathematics—from differential calculus and linear algebra to conformal and quasiconformal mapping and functions of a complex variable. Furthermore, the process of constructing maps with given properties based on scale leads directly to the integral calculus. In the particular case of the Mercator map, finding an explicit formula for its construction led to the formulation and solution of a problem in calculus, as so beautifully told in the article by Rickey and Tuchinsky on "An Application of Geography to Mathematics" [1980].

In view of these multiple connections, one might rightly suspect that the notion of the scale of a map, although indeed intuitive, is not actually all that simple. Our goals, then, will be

first: define exactly what is meant by the scale of a map, both in its simplest form and in its more refined senses,

second: describe a number of historically important geographical maps, many of which are defined in terms of certain scaling properties,

third: explain how a number of purely mathematical notions are related to the concept of scaling, and

fourth: review some of the major mathematical developments of the past 400 years where these mathematical notions are involved.

The term "mathematical mapping" in the title will be used in two ways. First, among geographical maps, some are of the freeform variety, giving the general "lay of the land" but not purporting to convey precise information about shapes and sizes. The maps that concern us have the feature that they are based on some specific mathematical principle. Oddly enough, the use of the word "map" in mathematics itself is extrapolated from the former, less "mathematical" kind of maps, which allow any method at all of assigning to each point of the original—say an area in the countryside—a point in our image: the "map" of that countryside, allowing some parts that interest us to be enlarged and others of less interest to be contracted or even shrunk to a point. The particular mathematical maps that we will deal with here will all be connected in some way to our central theme: scale.

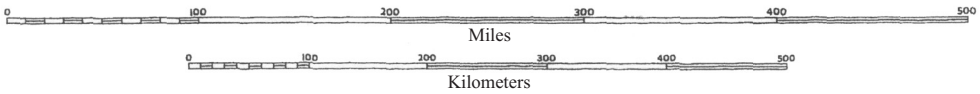
1 Scale

We start with the simplest example of scaling: an architect's drawings showing, say, the floor plan or the front view of a house. The plans will always be drawn to a certain *scale*: the ratio of the size of any object shown in the drawing to the actual size of the object it represents.

Exactly the same idea is used for making maps of cities, states, or other geographical entities. There are three ways commonly used to indicate the scale of a map: arithmetical, geometrical, and verbal.

Arithmetical: The scale is indicated in the form 1:6,000,000, meaning that the ratio of distances on the map to distance on the ground is 1/6,000,000.

Geometrical:



Verbal: 94.7 miles to the inch.

Of course, this last is an approximation. The exact scale is 6 million inches to an inch, but since a mile is 5,280 feet times 12 inches to the foot, we get immediately the rough estimate that 6 million inches is somewhat under a hundred miles. It goes without saying that the same scale is far more easily expressed verbally as 60 kilometers to the centimeter.

Whatever notation one uses, the meaning is the same: saying that the scale s of a map is $1 : n$ or $1/n$ means that if any two points on the map are a distance d apart, then the corresponding points on the region being mapped are a distance nd apart.

There is only one difficulty with this beautifully simple concept; a mathematical theorem states that, for the surface of the earth, no such map exists. In its simplest form, where we take the surface of the earth to be a sphere, the theorem goes back to Euler.

Theorem 1 (Euler [1775]) *It is impossible to make an exact scale map of any part of a spherical surface.*

To risk stating the obvious, when we speak of a map in this context, we refer to a map drawn on a flat sheet of paper. One can always make an exact scale "map" of the earth in the form of a globe. What Euler's Theorem says is that if we draw on a flat piece of paper a map of some region on a spherical surface, there is bound to be some distortion. One sometimes sees the statement that "one can preserve the size or shape, but not both." In fact, Johann Lambert [1772] gave the first general mathematical treatment of cartography, defining precise versions of preserving "size" and "shape." He also gave many new constructions, including some of the most widely used maps since that time. However, the intuitive notions of preserving size or shape had long been known,

together with the empirical fact that one could not do both; one could not construct an exact scale map.

Mercator, in constructing his famous map in 1569, opted for very good reasons to abandon size in favor of shape. We shall come back later to give exact definitions, but we start by explaining the intuition behind Mercator's map. The key idea is that we cannot have a fixed scale for the map, but we *can* have a fixed scale in certain directions.

To make these ideas more concrete, let us examine a particular class of maps known in cartography as "cylindrical projections." In order to define them, we first recall some standard terminology.

The **equator** is the great circle equidistant from the North and South Poles.

The **meridians** are the circular arcs joining the North and South Poles. They are perpendicular to the equator, and are the curves one traverses when traveling due north or south from any point.

The **parallels of latitude**, or **parallels** for short, are the circles perpendicular to the meridians. They are also the circles at fixed distance from the North or South Pole, and they are the curves one traverses when traveling due east or west from any point (other than the poles.)

A **cylindrical projection** is a map constructed as follows: The equator is represented by a horizontal line segment. The length of the segment determines *the scale of the map along the equator*; that is, if L is the length of the equator, and w the width of the map—the length of the horizontal segment representing the equator—then all distances along the equator are represented by the fixed factor w/L. The meridians are represented by vertical lines whose length may be either finite or infinite, and the parallels of latitude by horizontal line segments of the same fixed length w as the equator. As a result, all cylindrical projections have the following properties:

1. The map is in the form of a rectangle or infinite vertical strip representing all of the earth except for the poles, with the two vertical sides corresponding to a single meridian, and every other meridian corresponding to a unique vertical line. (In the infinite case, this is of course the theoretical map, with the actual finite map cut off to represent the portion of the earth between two fixed parallels of latitude.)

2. The map has a fixed scale along each parallel of latitude. Namely, the quarter circle along a meridian from the equator to the North or South Pole is divided into ninety degrees, and the *latitude* of any point is the number of degrees along the meridian north or south of the equator. It follows that at latitude φ north or south, the parallel is a circle of radius $R\cos\varphi$ where R is the radius of the earth and the length of the equator is $L = 2\pi R$. Hence, at latitude φ, the parallel is mapped with fixed scale

$$\frac{w}{2\pi R\cos\varphi} = \frac{w}{L\cos\varphi} = s\sec\varphi,$$

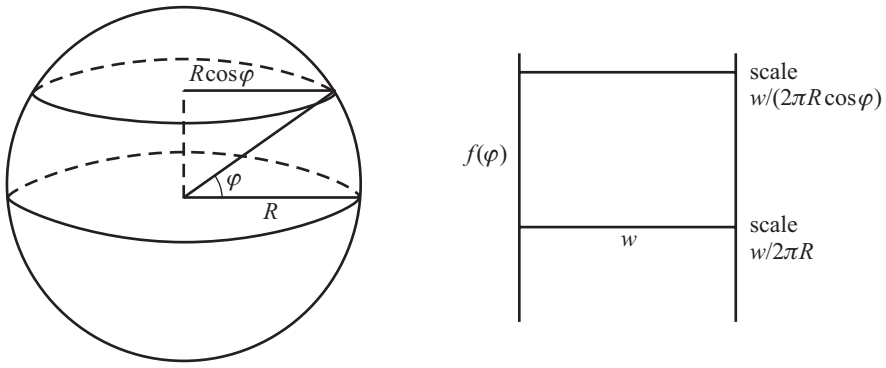


Figure 18.1.

where $s = w/L$ is the scale of the map along the equator.

These two properties illustrate Euler's Theorem in action. The whole familiar class of maps in which "north" is the vertical direction on the map and east-west is horizontal, and that have a fixed scale along some east-west line, are of necessity a portion of a cylindrical projection and cannot have a fixed scale for the whole map. The reason such maps are able to indicate a "scale" is that the factor $\sec\varphi$ in the expression $s\sec\varphi$ for the scale will not vary enough over a small portion of the earth's surface to make any practical difference. (Other inaccuracies in the map are bound to be far greater.)

Among the so-called "cylindrical projections" is one that is a true "projection" in the mathematical sense. It can be defined geometrically as follows. Let S be a globe: a sphere depicting the surface of the earth, and C a circular cylinder tangent to S along the equator. Project S onto C along rays from the center O of S.

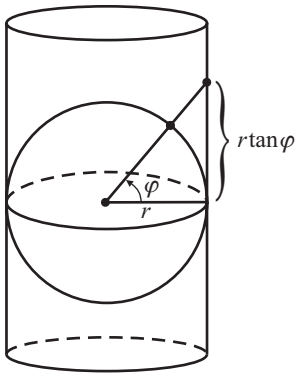


Figure 18.2. Central cylindrical projection

Each meridian on the sphere will clearly map onto a vertical line, and each parallel of latitude will map onto a circle on the cylinder parallel to the equator.

Cut the cylinder along a vertical line and unroll it onto a vertical strip in the plane. The result is a particular case of cylindrical projection known as a "true cylindrical projection" or "central cylindrical projection." When reading about map-making one must be careful to distinguish the way the word "projection" is used there, to mean any systematic form of representation, from its more narrow use in mathematics, or for that matter in art and in everyday parlance, where one thinks of a "projector" such as a slide projector or of a shadow projected on a wall. In cartography, these "true" projections are sometimes referred to as "perspective projections."

In order to give a precise formula for a general cylindrical projection, let us introduce rectangular coordinates with the origin at the point corresponding to the equator on the left edge of the map. The northern hemisphere will be represented by

$$0 \le x \le w, \quad 0 \le y \le H, \qquad \text{with } H \le \infty.$$

An analogous discussion will hold for the southern hemisphere.

A point in the northern hemisphere is traditionally assigned a pair of coordinates: the latitude φ and longitude θ. We have already defined φ as the angular distance above the equator as viewed from the center of the sphere. The equator is divided into 360° starting at some point. That assigns to each point on the equator an angle θ with $0 \le \theta < 360^\circ$. (One actually uses values of θ up to 180° east or west of a given point, but that is equivalent, and mathematically more awkward.) The *longitude* of any point is the value of θ where the meridian through the point hits the equator.

Thus, every cylindrical projection is given explicitly by the equations

$$x = \frac{w\theta}{360}, \quad y = f(\varphi) \tag{1}$$

for some monotonically increasing function f, with $f(0) = 0$, $f(90) = H$. For example, it follows immediately from the geometric definition of a true cylindrical projection that if we want a map of width w, we choose a globe of radius $r = w/2\pi$, and the equations become

$$x = \frac{w\theta}{360}, \quad y = \frac{w}{2\pi}\tan\varphi. \tag{2}$$

As we saw earlier, the scale of every such map along the parallel at latitude φ is

$$s_\varphi = s\sec\varphi \tag{3}$$

where

$$s = \frac{w}{L} \tag{4}$$

is the scale along the equator.

We can now describe exactly what it was that Mercator was trying to do, and how he went about doing it.

Mercator wanted his map to have two properties: first, it should be a cylindrical projection so that at any point of the map, the vertical direction represents north, and second, the map should not distort shapes. Now it is clear intuitively that if a map has different scales in the horizontal and vertical directions, then shapes will be distorted. (Think of looking at yourself in a funhouse mirror.) Since Mercator knew the horizontal scaling factor at each latitude, either given trigonometrically as in equations (3), or else geometrically via the ratio of the two horizontal segments in Figure 18.3, he simply had to adjust the vertical scale accordingly.

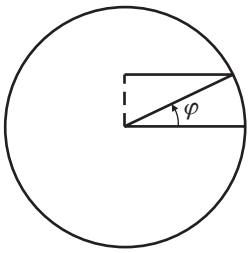


Figure 18.3.

Mercator did not divulge the exact procedure he used to construct his map, but it seems most likely that he proceeded somewhat as follows: divide up the region of the earth between two fixed latitudes into thin strips, each bounded by a pair of nearby parallels of latitude. The horizontal scale would be approximately constant in each strip, and one could use, for example, the exact value at the center parallel of the strip. Map that strip onto a horizontal strip in the plane using the same scale in the vertical direction. Stacking these strips one on top of another would give Mercator the result he was seeking.

Needless to say, this gives only an approximation to a "true" Mercator map, where the vertical scale exactly equals the horizontal scale at each point. However, all actual cartographic maps are only approximate. The real problem was that anyone else wanting to make a "Mercator map" of a part of the earth's surface would have to either copy the original or else repeat the whole tedious procedure. What was wanted was the actual function $f(\varphi)$ in the equations (1) that resulted in

Mercator's map; that is, the function $f(\varphi)$ for which horizontal and vertical scaling are everywhere equal. In order to find such a function, we have to make precise what we mean by the "scale" of a map in the vertical direction at a given point when that scale is constantly changing.

In the case of a cylindrical projection given by equations (1), an arc of a meridian is defined by an interval of latitude, $a < \varphi < b$, while the image of this arc on the map will be a vertical line segment $f(a) < y < f(b)$. Assuming the earth is a perfect sphere, the length of the arc will be $L(b-a)/360$, and the image on the map will have length $f(b) - f(a)$. The overall scale factor for this arc of the meridian is therefore

$$\frac{360}{L}\frac{f(b)-f(a)}{b-a}.$$

The value of this scale factor over smaller and smaller intervals of arc will be closer and closer to the exact scale factor at a point, leading us inevitably to the

Definition. The *vertical scale* of a cylindrical projection given by equation (1) at a point at latitude $\varphi = a$ is equal to

$$\lim_{b\to a}\frac{360}{L}\frac{f(b)-f(a)}{b-a}.$$

In other words, the notion of the *scale at a point* when the scale is continually changing is precisely the notion of a derivative:

$$f'(a)\cdot\frac{360}{L}.$$

The extra factor $360/L$ arises because we have measured distance along the meridian in degrees, rather than arc length, with one degree of latitude having length $L/360$.

In fact, given any monotonically increasing function $y = f(x)$, we may picture it either via a graph or as a map of an interval I of the x-axis onto an interval J of the y- axis. The two interpretations are connected via the picture in Figure 18.4.

The scale factor of the map $I \to J$ at any point p is exactly equal to the derivative $f'(p)$. In particular, the map *shrinks* distances if the scale factor $f'(p)$ is less than 1, and *stretches* them if $f'(p) > 1$.

Returning to our case of cylindrical projections, it is easier to work with them mathematically if we express the latitude φ and longitude θ in radians rather than degrees. The equations (1) then take the form

$$x = \frac{w\theta}{2\pi}, \quad y = F(\varphi), \tag{5}$$

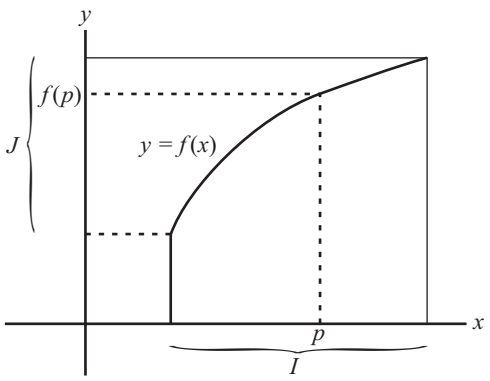


Figure 18.4.

where again, w is the width of the map, and the scale along the equator is $s = w/L$, with L the length of the equator. Since arc length along the meridian is given by $\varphi L/2\pi$, the vertical scale at latitude φ is

$$s_v(\varphi) = F'(\varphi) \cdot \frac{2\pi}{L}. \tag{6}$$

As we saw earlier in equations (3) and (4), the horizontal scale at latitude φ is

$$s_h(\varphi) = s \sec \varphi, \quad s = \frac{w}{L}. \tag{7}$$

Mercator's goal was to have

$$s_v(\varphi) = s_h(\varphi) \qquad \text{for all } \varphi,$$

which reduces to

$$F'(\varphi) = \frac{w}{2\pi} \sec \varphi. \tag{8}$$

The procedure we have outlined that Mercator presumably used in constructing his map was precisely a numerical integration of this equation. That was made explicit by English mathematician Edward Wright who used the method to construct a set of tables [1610] that would allow anyone to draw a much more accurate "Mercator" map than Mercator himself was able to do.

We now know the exact solution to equation (8). With $F(0) = 0$, it is

$$F(\varphi) = \frac{w}{2\pi} \log(\sec \varphi + \tan \varphi)$$

which Mercator could not possibly have known, since logarithms had yet to be invented, to say nothing of the derivatives and integrals used to derive the equation. Again we refer to the article by Rickey and Tuchinsky [1980] for the many steps leading to this result.

In general, constructing a map in which the pointwise vertical scaling is given in advance amounts precisely, according to equation (6), to carrying out an integration. In other words, *the two fundamental operations, differentiation and integration, correspond precisely to determining the scale of a variable scale map and constructing a map when given the (variable) scale.*

As another example, if our goal was to preserve size rather than shape, then instead of having the horizontal and vertical scaling factors be equal, we would make them reciprocal, so that the stretching in one direction would match the shrinking in the other. Going back to equations (6) and (7), we see that we need to make

$$F'(\varphi) \sec \varphi \equiv c, \quad \text{a constant}$$

or

$$F'(\varphi) = c \cos \varphi$$

so that

$$F(\varphi) = c \sin \varphi.$$

We can choose the constant c to make the horizontal and vertical scales equal at any given latitude and hence have a very good map near that latitude. For example, if we let $c = w/2\pi$, then by equations (6) and (7), the vertical and horizontal scales will be equal at the equator, and we get one of Lambert's maps: the cylindrical equal-area map (see Figure 18.5).

As a final example of a cylindrical projection, we note that one of the most simple-minded of all goes back to antiquity and was commonly used in the 16th century under the name "plate carrée." It uses a fixed scale along each meridian—the same scale as that along the equator. It has the advantage that the distance between any pair of points on the same meridian can be read off directly from the map. Also, even though it preserves neither shape nor size, it does not have some

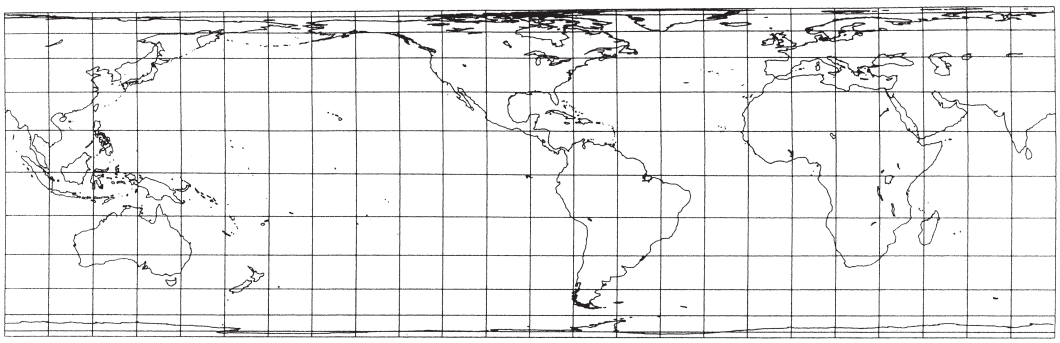


Figure 18.5. Lambert Cylindrical Equal-Area projection with shorelines, 15° graticule. Standard parallel 0°. Central meridian 90° W.

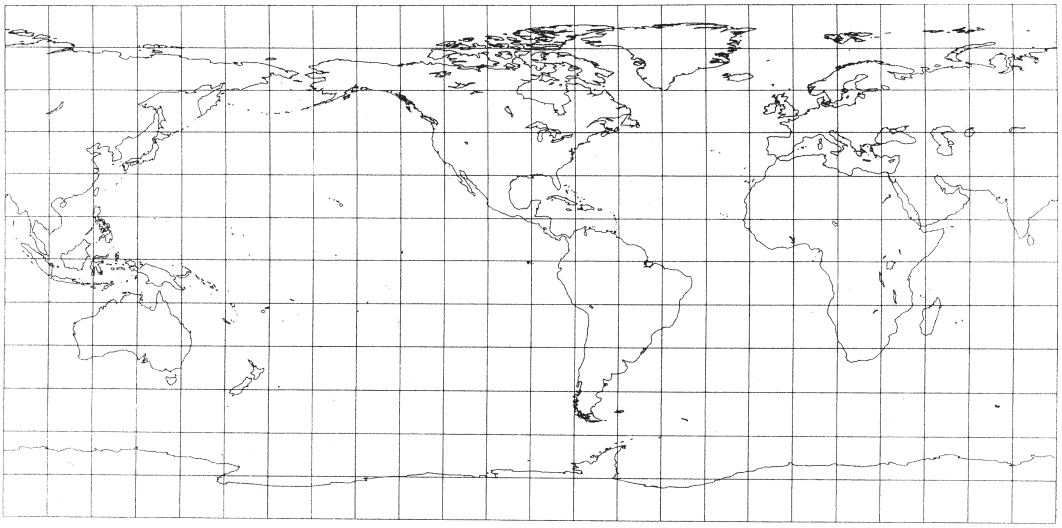


Figure 18.6. Plate Carrée projection with shorelines, 15° graticule. Central meridian 90° W.

of the extreme distortions of the Mercator or Lambert equal-area maps, but serves as a kind of compromise between the two (see Figure 18.6).

The equations for the plate carrée could not be simpler:

$$x = c\theta, \quad y = c\varphi.$$

2 Maps of the plane

Suppose we use any of the cylindrical projections described above to make a map of Australia. If we compose that map with a map of the plane into the plane, we then get a new map of Australia. Conversely, if we make any two different maps of Australia, then they are related to each other by a map of the plane into the plane. Our goal, then, will be to look more closely at maps of the plane into the plane, with our focus again on questions related to scale. That will also allow us to make precise the intuitive notions of preserving shape and preserving size.

We start with the simultaneously simplest and most important case: that of linear maps. We let T be a linear transformation of the x, y-plane into the u, v-plane, which we may write in the form

$$\begin{aligned} u &= ax + by \\ v &= cx + dy \end{aligned} \tag{9}$$

where a, b, c, d are fixed real numbers. An arbitrary line through the origin in the x, y-plane may be given parametrically by expressing x and y as constant multiples of a parameter t. Substituting these expressions in equations (9) gives u and v as constant multiples of t, hence defines a line through the origin in the u, v-plane, the image of the original line under the transformation T. This mapping of a line into a line will have a fixed scale s, which represents the ratio of the distances between any two points on the image line and the distances of their preimages. In general we will have $s > 0$, but we may have the degenerate case $s = 0$ when the entire first line maps onto the origin. As we rotate the original line around the origin, the scale s will attain a maximum s_1 and a minimum s_2 with $0 \leq s_2 \leq s_1$. A basic result of linear algebra is the following.

Lemma 1 *One can choose new axes X, Y in the x, y-plane and U, V in the u, v-plane such that the transformation T takes the form*

$$U = s_1 X, \quad V = s_2 Y. \tag{10}$$

Said differently, the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

of the linear transformation T can be reduced to a diagonal matrix by pre- and post-multiplication by orthogonal matrices. We also have

$$|\det A| = |ad - bc| = s_1 s_2. \tag{11}$$

In particular, the scale factor is always positive whenever A is nonsingular.

To be a bit more specific, we can always choose new axes in the x, y-plane by a rotation of the axes, and if $\det A > 0$, then we may do the same in the u, v-plane. If $\det A < 0$, then we must make an additional reflection in order to put the transformation in the form (10), with the scale factors both positive.

One immediate consequence of equations (10) is that distances get scaled by the factors s_1 and s_2 in two orthogonal directions, giving us the result:

Corollary 1.1 *Under the transformation T, all areas are multiplied by the factor $s_1 s_2$. In particular, areas are preserved under T if and only if $s_1 s_2 = 1$.*

A second consequence follows by noting that a line making an angle α with the X-axis is given parametrically by $X = t\cos\alpha$, $Y = t\sin\alpha$, while its image under T has the equations $U = ts_1\cos\alpha$, $V = ts_2\sin\alpha$ and makes an angle β with the U-axis, where $\tan\beta = (s_2/s_1)\tan\alpha$. We therefore conclude:

Corollary 1.2 *Angles are preserved under T if and only if $s_1 = s_2$.*

The unit circle C is given parametrically by $X = \cos t$, $Y = \sin t$, and its image under T is given by $U = s_1\cos t$, $V = s_2\sin t$, which is a circle if $s_1 = s_2$ and an ellipse with major and minor axes along the U and V axes and area $\pi s_1 s_2$ if $s_1 > s_2$.

That leads us to look at two separate cases.

Case 1: $s_1 > s_2$. The image of the unit circle under the transformation T is an ellipse whose major and minor axes determine the U and V axes respectively. The pre-images of the U and V axes are the X and Y axes: the directions of maximum and minimum scaling. The key properties of the mapping are therefore made graphically clear by drawing the unit circle in the x, y-plane with the directions of the X and Y axes indicated, and the image ellipse in the u, v-plane. *The size of the ellipse shows the area distortion and the eccentricity of the ellipse indicates the shape distortion.*

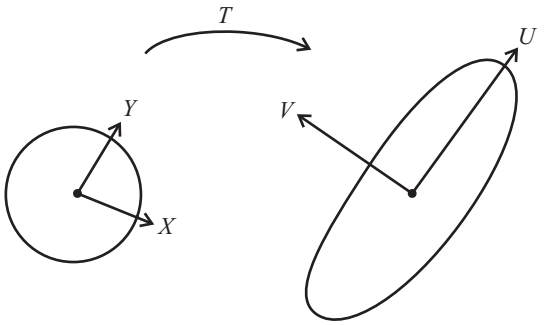


Figure 18.7.

Precisely this device is used by cartographers to give map-viewers an instantaneous overview of the nature of the distortion for a given map. For example, Figures 18.8 and 18.9 are the pictures for the plate carrée and the Lambert equal-area map shown above.

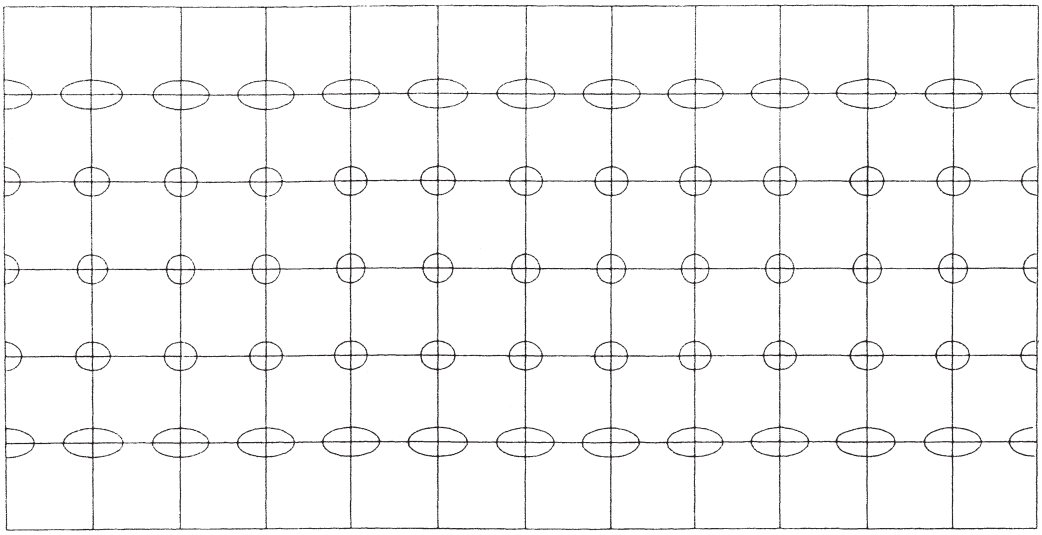


Figure 18.8. Plate Carrée projection with Tissot indicatrices, 30° graticule.

Notice that in both cases there are circles along the equator, indicating no shape distortion —that is, equal scaling in the horizontal and vertical directions. Also in both cases, at all points off the equator the ellipses have major axes in the horizontal direction, which is therefore the direction of maximum stretching, but in the case of the plate carrée the ellipses grow larger and larger toward the poles, indicating area distortion also, whereas in the equal-area map, the ellipses all have the same area as the circles along the equator. The ellipses used in these distortion diagrams are called

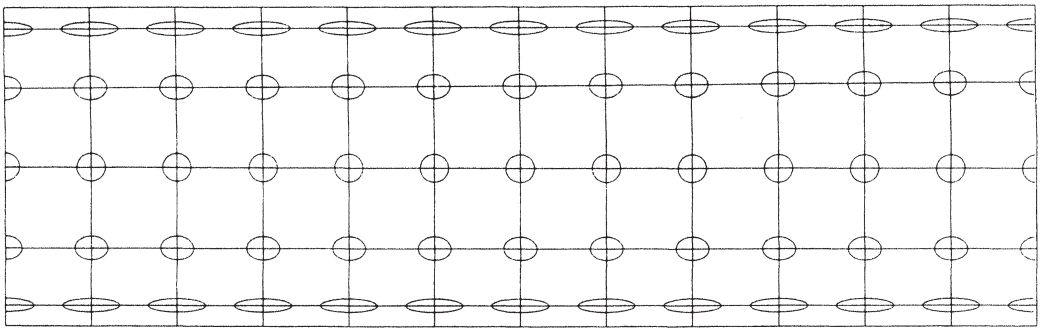


Figure 18.9. Lambert Cylindrical Equal-Area projection with Tissot indicatrices. 30° graticule. Standard parallel 0°. All ellipses have the same area, but shapes vary.

"Tissot indicatrices." At any point of a map, the Tissot indicatrix shows the directions and relative amounts of maximum and minimum scaling.

In contrast to the figures above, the Tissot indicatrices show that for the central cylindrical projection (Figure 18.10) there is distortion in the vertical direction, whereas for the Mercator projection (Figure 18.11), there is less size distortion and no distortion of shape: each Tissot indicatrix is a circle.

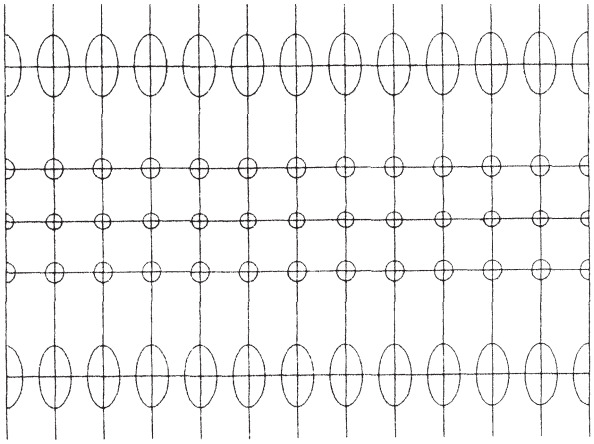


Figure 18.10. Central Cylindrical projection with Tissot indicatrices, 30° graticule.

Case 2: $s_1 = s_2$.

Definition. A linear transformation is called a *simple scaling* or a *homothety* if it is of the form $u = sx$, $v = sy$.

A linear transformation is said to "preserve shape" if it is a *similarity transformation*; that is, it maps every triangle onto a similar triangle.

From the above discussion, together with a bit more argumentation, we can give an extended characterization of these transformations.

Proposition 1 *Let T be a linear transformation (9) with matrix A, and assume that T is orientation-preserving; that is,* $\det A > 0$. *(If T is orientation reversing:* $\det A < 0$*, then we may apply the statements below to the transformation consisting of T followed by a reflection—for example replacing v by $-v$.) The following are equivalent:*

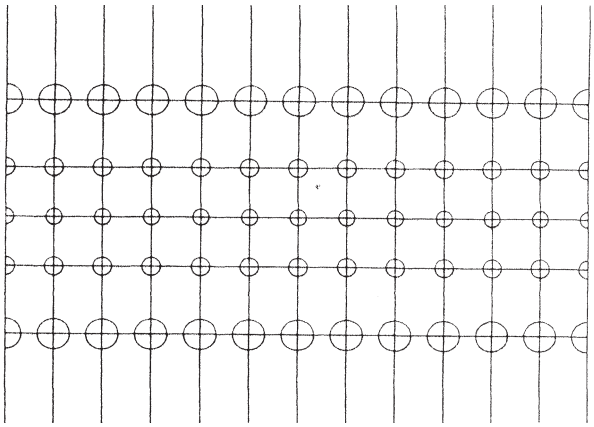


Figure 18.11. Mercator projection with Tissot indicatrices, 30° graticule. All indicatrices are circular (indicating conformality), but areas vary.

(a) *T is a similarity transformation,*

(b) *all distances are multiplied by a fixed factor s,*

(c) *all angles are preserved,*

(d) *the coefficients of the matrix A satisfy the equations: $a = d$, $b = -c$,*

(e) *T is a composition of a rotation and a simple scaling,*

(f) *the equations (9) for T can be written in the form*

$$u = s(x\cos\alpha + y\sin\alpha)$$
$$v = s(-x\sin\alpha + y\cos\alpha)$$

where $s = \sqrt{a^2 + b^2}$, $\cos\alpha = a/s$, $\sin\alpha = b/s$.

The reason for going into this much detail on linear mappings is that they govern the behavior near each point for arbitrary differentiable maps, to which we now turn.

3 Smooth maps

Let F be a continuously differentiable map of the x, y-plane into the u, v-plane. The differential dF of F at a point P is the linear transformation T whose matrix consists of the partial derivatives of F at P; that is,

$$a = u_x(P),\ b = u_y(P),\ c = v_x(P),\ d = v_y(P). \tag{12}$$

There are two common interpretations of the differential. One is that it is the best linear approximation to the map F in a neighborhood of P. The other is that it is the *tangent map* to F; that is, if C is any smooth curve through P, then the tangent vector to the image of C under F at the point $F(P)$ depends only on the tangent vector to C at P, and this induces a linear map of tangent vectors at P to tangent vectors at $F(P)$, which is precisely the differential dF at P. Since the angle between a pair of smooth curves intersecting at P is by definition the angle between their tangent vectors, it follows that the map F preserves angles at P if and only if dF is a similarity transformation.

Definition. A map F is a *conformal map* if it is a diffeomorphism that preserves angles at every point; that is, F is a one-to-one continuously differentiable map with a differentiable inverse, and dF is a similarity transformation at every point.

We shall come back in the next section to a more detailed look at conformal maps in the plane. We note here that the cylindrical projections described in section 1 are all maps of the sphere into the plane, and the differential of any such map can be defined exactly as in the case of plane maps as maps of tangent vectors of curves on the sphere at a point to tangent vectors of their image curves. From Proposition 1 it follows that the angle between any two curves at a point on the sphere equals the angles between their image curves if and only if the maximum and minimum scaling factors are equal; that is, the scaling factor is the same in all directions at the point. But the way we constructed Mercator's map was precisely from that property. We therefore conclude:

Proposition 2 *Mercator's map is the unique cylindrical projection that preserves angles.*

It follows that Mercator's map is the unique map with the two key properties

(i) the vertical direction on the map corresponds to the north/south direction,

(ii) given any two points on the map corresponding to a pair of locations on the earth, if the straight line joining them on the map makes a given angle with the vertical, then starting at the first location and following the fixed compass direction determined by that angle will lead to the second location.

It was this second property that made Mercator's map indispensable for navigation for a very long time.

It is the combination of angle-preserving and the fixed vertical direction for north that gives the second property above. If one does not specify vertical for north, then there are many other possibilities for angle-preserving maps. Indeed, one of them goes back to antiquity. It is called *stereographic projection.* It is a true projection, in which the sphere is projected from some point on it—often chosen to be the North Pole—onto either the plane tangent to the sphere at the antipodal point—say the South Pole—or else the plane parallel to that one through the center of the sphere; which of those two planes one projects onto is immaterial, since the resulting maps will differ by a similarity transformation between the two planes.

For stereographic projections from the North or South Pole, one has the property that the parallels of latitude map onto concentric circles about the origin, and the meridians map onto rays extending outward from the origin. The fact that stereographic projection preserves angles appears to have first been pointed out and proved by Edmund Halley, of comet fame, in 1695.

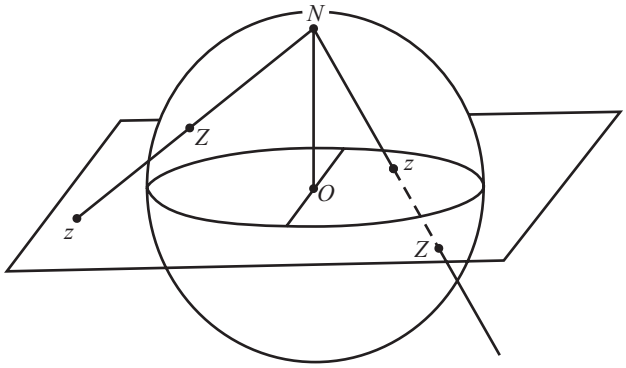


Figure 18.12. Stereographic projection

Our principal interest will be in conformal maps of the plane into the plane. By virtue of condition (d) in Proposition 1 and equations (12), we have the following elementary but fundamental result.

Proposition 3 *A smooth map F defined in a domain D in the x, y-plane and mapping D into the u, v-plane preserves angles if and only if dF is nonsingular everywhere and satisfies the equations*

$$u_x = v_y, \quad u_y = -v_x. \tag{13}$$

These equations are called the *Cauchy-Riemann equations*. They are simply the statement that the tangent map to F at each point is a similarity transformation. In view of condition (e) of Proposition 1, the geometric content of the Cauchy-Riemann equations is that near each point, the map F behaves like a rotation composed with a simple scaling. This is an important picture to keep in mind in the following sections.

4 The complex plane

Solving algebraic equations is one of the oldest and most fundamental problems in mathematics. A particular case of importance is that of polynomial equations, such as

$$x^2 = 1, \quad x^2 + 1 = 0, \quad x^3 = 15x + 4, \quad x^5 = 1.$$

The first of these has the two obvious solutions, $x = 1$ and $x = -1$. The second has no real solutions, but has two solutions if one introduces the imaginary number $i = \sqrt{-1}$, namely, $x = i$ and $x = -i$. The third is of interest because it has three real solutions, but if one uses the general formula for solving cubics developed by a series of Italian mathematicians in the 16th century (see chapter 1 of Nahin [1998] for an excellent review of this subject) then one finds that the expressions for two of these real roots involve imaginary numbers too. This led to the introduction of complex numbers: expressions of the form $a + bi$ where a and b are real numbers. We denote $|a + bi| = \sqrt{a^2 + b^2}$. The last equation above has the one real root $x = 1$ and four complex roots.

A number of facts and a number of questions soon emerged regarding general polynomial equations—that is, an equation with a finite number of terms, each consisting of a constant times a power of the unknown x. The *degree* of the equation is the highest power of the unknown that occurs. The facts were

1. An equation of degree n can have at most n solutions.
2. Solutions may be real or complex or some of both.

The questions were

1. Is there a formula, or a general procedure for solving a polynomial equation of degree n, such as the familiar quadratic formula when $n = 2$?
2. In the absence of a formula, can one at least guarantee that any polynomial equation does have at least one solution?

The answer to the first question is well known. As already mentioned, Italian mathematicians found the solution of the general cubic, and also, shortly afterward, all fourth degree equations. Degree five stumped all comers, until Abel proved in 1824 that the general quintic equation had no such solution.

As to the second question, the answer was generally believed to be "yes" but various attempts at proofs were not considered very satisfactory until Gauss came along and devoted his PhD thesis of 1799 to the subject. According to Gauss' own description in a letter, about a third of the thesis is

devoted to a proof of the theorem that every polynomial can be written as a product of linear and quadratic factors, while the rest is devoted to history and criticisms of previous "proofs" including those of d'Alembert, Euler, and Lagrange.

Gauss returned repeatedly throughout his life to this question, providing a number of different proofs. The result became known as "the fundamental theorem of algebra." In fact, he returned to it in 1849, giving a variant of his first proof, but making more explicit use of the complex plane.

The idea of representing complex numbers by points in the plane appears to have occurred to several people independently towards the end of the 18th century. One gets a geometric picture of the purely algebraic (and abstract) entities—complex numbers—by associating to each number $a+bi$ the point (a, b) in the plane. Or viewed the other way around, the "complex plane" is simply the ordinary Euclidean plane, where to each point (a, b) one assigns the complex "coordinate" $a+ib$. Finding real roots of a real polynomial equation—that is, real values of x such that $P(x) = 0$, where $P(x)$ is a polynomial with real numbers as coefficients—can be pictured geometrically as finding a point where the graph of the equation $y = P(x)$ crosses the x-axis. How does one picture geometrically a complex root of the equation? The answer, not surprisingly, is by means of mappings.

Let x and y be real variables, and let $z = x+iy$. Then a polynomial P assigns to every complex number z another complex number $w = P(z)$, and so P defines a map from the complex z-plane to the complex w-plane. Equivalently, letting $w = u + iv$, P defines a smooth map from the x, y-plane to the u, v-plane. There is always the trivial case to consider—a polynomial of degree zero, which has only a constant term, and does not actually depend on z. Considered as a map, such a polynomial maps the whole z-plane onto a single point—the value of the constant term. What the fundamental theorem of algebra states is that *for every polynomial P of degree > 0, $P(z)$ maps the z-plane onto the entire w-plane.* This says that for any complex number c, the equation $P(z) = c$ has a (complex) solution. It is obvious, but worth stating, that the solvability for every polynomial of degree $n > 0$ for every value of c is completely equivalent to solving $P(z) = 0$ for every such polynomial, since we can just transfer the value c to the left side of the equation.

This picture of a polynomial mapping the z-plane onto the w-plane became the impetus for 200 years of further developments, some of them spectacular, that will be the subject of the remainder of our discussion.

5 Analytic functions

An obvious next step up from polynomials, which have a finite number of terms, is to a kind of "infinite polynomial": $\sum_{j=0}^{\infty} c_j z^j$. Such an infinite sum will define a unique complex number providing it converges, which is equivalent to saying that the infinite sums of the real and imaginary parts of each term converge. In general, any such power series will converge for all z satisfying $|z| < R$ for some $R > 0$, and fail to converge for $|z| > R$, in which case R is called the *radius of convergence* of the series. There are also two extreme cases: the series may not converge for any $z \neq 0$, in which case one says $R = 0$, or it may converge for every value of z, so that $R = \infty$. In this last case, the infinite series will define a function $w = F(z)$ that can again be pictured as a map F of the z-plane into the w-plane. Such functions are called *entire functions.* We shall return to them later for a closer look. The fundamental result at the heart of the subject is the following:

Proposition 4 *Let F be a smooth map of a domain D in the x, y-plane into the u, v-plane. Let $w = f(z)$ be the complex function defined by the map F, where $z = x + iy$ and $w = u + iv$. Then the following are equivalent:*

(a) *u and v satisfy the Cauchy-Riemann equations* (13) *at every point of D,*

(b) *$f(z)$ has a complex derivative at every point of D,*

(c) *for every point c of D, the function f can be written as a power series in $z-c$ with a radius of convergence $R>0$.*

Definition. The complex function $f(z)$ is said to be *analytic* in a domain D if any (hence all) of conditions (a), (b), or (c) holds.

The equivalence of three so diverse-appearing conditions has no analog in the theory of real functions, and is a signal that one can expect complex analytic functions to enjoy many special and sometimes surprising properties. The first of those is that by Proposition 3 complex analytic functions define angle-preserving maps at all points where the differential is not zero. But the complex derivative of a complex function $f(z)$ is given in terms of the real and imaginary parts by $f'(z)=u_x+iv_x$, so that combined with the Cauchy-Riemann equations, $f'=0$ at a point z if and only if all partial derivatives of u and v with respect to x and y are zero at that point. One consequence of condition (c) above is that if an analytic function is not constant, then its derivative can vanish only at isolated points; everywhere else, it defines a conformal mapping.

One of the most important analytic functions is the *exponential function* define by

$$\exp z = 1+z+\frac{z^2}{2}+\frac{z^3}{3!}+\cdots+\frac{z^k}{k!}+\cdots$$

which has the following properties:

1) the series converges for all z, so that $\exp z$ is an entire function
2) $\exp 1 = e$
3) $\exp x = e^x$ for x real
4) $\exp iy = \cos y + i\sin y$ for y real
5) $\exp(a+b) = \exp a \exp b$
6) $\exp(z+2\pi ni) = \exp z$ for every integer n
7) $|\exp z| = e^x \neq 0$ for all z.

One form of the fundamental theorem of algebra is that if $P(z)$ is a polynomial of degree n, and c any complex number, then the polynomial $P(z)-c$ is also of degree n and can be written as the product of n linear factors, each of which contributes one solution to the equation $P(z)=c$, and some of which may be equal. Hence, there are at most n solutions to the equation $P(z)=c$ for any value of c. Furthermore, there are "in general" exactly n distinct solutions, the exceptions being those (at most $n-1$) values c of the form $P(a)$ where $P'(a)=0$. Looked at in terms of mappings, a complex polynomial P of degree n defines a mapping of the complex z-plane to the complex w-plane with the property that for every point in the w-plane, its inverse image consists of exactly n points, with at most a finite number (in fact, at most $n-1$) exceptions where the inverse image consists of fewer than n points.

If we think of entire functions as "polynomials of infinite degree" we might expect something analogous to be true. We shall examine that question more closely in the following section, but it is instructive to examine the case of the exponential function in more detail. We see immediately one important difference. By virtue of property 7 above, the equation $\exp z = 0$ has *no* solutions. For any complex number $c\neq 0$, to find all solutions of the equation $\exp z = c$, we write c in polar form as $c=r(\cos\theta+i\sin\theta)$, where $r=|c|$. Combining properties 3, 4, and 5 of the exponential function, we find that $\exp z = \exp(x+iy) = e^x(\cos y + i\sin y)$ so that

$$\begin{aligned}\exp z = c &\iff e^x = r \text{ and } y=\theta+2\pi ni \text{ for an arbitrary integer } n \\ &\iff x=\log r \text{ and } y=\theta+2\pi ni \text{ for an arbitrary integer } n.\end{aligned} \tag{14}$$

In other words, *the equation* $\exp z = c$ *has an infinite number of solutions for every* $c \neq 0$*, and no solutions for* $c = 0$.

It is clear from equation (14) that the way to visualize the function $w = \exp z$ as a map is to use rectangular coordinates x, y in the z-plane and polar coordinates r, θ in the w-plane. We note first that the image of the entire x-axis is the positive u-axis or the ray $\theta = 0$, while the mapping itself is given by the ordinary exponential function $u = e^x$. Each horizontal line $y = b$ maps in the same way, by $r = e^x$ onto the ray $\theta = b$. We should picture the effect dynamically, as the horizontal line in the z-plane moves upward from $y = 0$ to $y = 2\pi$, the image ray rotates once around from $\theta = 0$ to $\theta = 2\pi$. The effect is that the infinite horizontal strip $0 \leq y < 2\pi$ maps onto the whole w-plane minus the origin. The same process is repeated for $2\pi \leq y < 4\pi$ and so on, with each horizontal strip of width 2π mapping onto the whole plane minus the origin. Said differently, as the horizontal lines in the z-plane sweep out the plane, the image rays rotate around and around infinitely often.

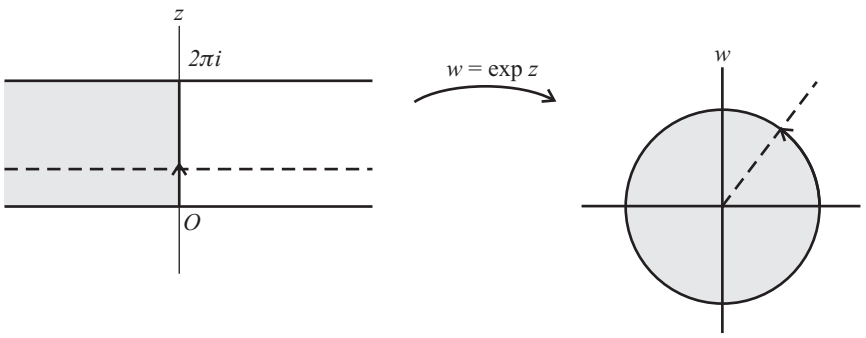


Figure 18.13.

If we restrict to a single point moving vertically, tracing out the line $x = a$, say, the image will trace out the circle of radius e^a; in particular, the y-axis maps onto the unit circle, the left half-plane maps onto the interior of the unit circle except for the origin, and the right half-plane maps onto the exterior of the unit circle.

Next, we may view equations (14) in the reverse direction, as defining a map of the whole w-plane minus the origin onto the horizontal strip $0 \leq y < 2\pi$ in the z-plane. The map is given explicitly in terms of polar coordinates in the w-plane, by

$$x = \log r, \quad y = \theta \tag{15}$$

and is called the complex logarithm, written $z = \log w$. It maps rays emanating from the origin onto horizontal lines, and circles centered at the origin onto vertical line segments.

The application of all this to geographic maps is gradually becoming better known, but not nearly as well-known as it should be. Namely, if we map the globe by stereographic projection from the North Pole onto the plane, the South Pole maps onto the origin, the meridians map onto rays emanating from the origin, and the parallels of latitude onto circles centered at the origin. If we compose this map with the complex logarithm defined by (15), then the meridians map onto horizontal lines and the parallels onto vertical line segments. Furthermore, both stereographic projection and the complex logarithm preserve angles, hence so does the composition. If we now follow by a rotation of 90° in the positive direction, then meridians map onto vertical lines, and parallels onto horizontal line segments, so that we have an example of a cylindrical projection. But it is also a conformal map, and as we have seen, that determines it uniquely—it must be our good old standby, Mercator's projection. Summing up, *to give explicit equations for the Mercator*

map, we simply make a stereographic projection and follow it by the complex logarithm, (15), and rotation through 90°.

One final note on the complex logarithm. What we have defined is actually just a branch of the logarithm, defined by restricting θ to the interval $0 \leq \theta < 2\pi$. However, if we allow all possible values of θ in the polar representation of a complex number w, then equations (15) define the general complex logarithm as a multiple-valued function whose different values all differ by integer multiples of $2\pi i$, just as the inverse trigonometric functions are multiple-valued, with values differing by multiples of π or 2π. However, that is a whole other story that we need not go into here.

For the geometric view of the complex exponential and logarithm, and of complex functions in general, with many beautiful and instructive pictures, we recommend the book *Visual Complex Analysis* [1997] by Tristan Needham.

6 Nineteenth century highlights

Gauss' 1799 proof of the fundamental theorem of algebra was a fitting culmination of eighteenth-century mathematics. Both Gauss and his theorem went off in unanticipated new directions as the nineteenth century got underway.

To start with Gauss, he published two papers in the 1820's devoted to questions regarding mappings. The second of the two, from 1827, is the more fundamental. It lays the foundation for much of the subsequent work in the field of differential geometry. The most famous result in the paper is known as Gauss' *Theorema Egregium* or "most excellent theorem." One corollary of the theorem is a far-reaching generalization of Euler's theorem that there is no exact scale map of any region of a sphere onto a plane. What Gauss proved was that the sphere in this theorem is the rule rather than the exception. Namely, if S is *any* surface, with a very few exceptions, then *there is no exact scale map of any region of* S *onto the plane.* The exceptions are the so-called "developable" surfaces, which may be obtained by simply rolling up a sheet of paper in various forms; for instance, cylinders and cones and a class of surfaces known as "tangent developables."

The other of Gauss' papers, from 1822, contains a positive result about mapping. It says that *any sufficiently smooth surface has a mapping into the plane that is locally conformal.* Said differently, all small regions on the surface can be represented by a plane map that preserves angles. Of course, if the surface is a sphere, then Mercator's map and stereographic projection are examples, but Gauss' theorem states that conformal maps exist "in general."

Gauss' disciple and successor Bernhard Riemann also made two major contributions of relevance to us. The first is known as the *Riemann Mapping Theorem.* It was a complete departure from earlier work, in that instead of asserting that there was *some* conformal mapping in a given situation, it said that you could actually prescribe the shape of the image. So for example, if you took any two plane domains each bounded by a simple closed curve—one might be an ellipse, and the other a rectangle—then Riemann's theorem states that there exists a one-to-one angle-preserving map between the two domains. The way the theorem is usually stated, one of the two domains is a circular disk, which is sufficient, since if you can map each of the two domains conformally onto a circular disk, then you can by composition map them conformally onto each other.

Riemann stated the theorem in his PhD thesis of 1851 and gave what he thought was a proof, but there turned out to be a gap in his reasoning. It took much of the remainder of the century to provide a complete proof and also to find ways to construct explicit mappings for simple cases, such as the ellipse and a rectangle. A key figure in that work was H. A. Schwarz [1869–70].

We will come back to the second of Riemann's contributions shortly, but first jump ahead to one of the biggest surprises of the century in the theory of complex functions. It was proved in

1879 by the young French mathematician, Émile Picard. It may be viewed as the direct successor, 80 years on, to the fundamental theorem of algebra.

Theorem 2 (Picard's Theorem [1879]) *Let $f(z)$ be a nonconstant entire function. Then the equation $f(z) = c$ has a solution for every complex number c with at most one exception.*

It is probably safe to say that up to the time that Picard proved his theorem, there was no evidence at all that such a result would be true. Looked at from the point of view of a mapping, it said that every "infinite polynomial" maps the whole plane either onto the whole plane or onto the plane minus a single point. The fact that there may be an exceptional point was of course well known from the example of the exponential function, where $\exp z = 0$ has no solution.

Again thinking in terms of the fundamental theorem of algebra, as the degree of the polynomial goes up, so does the number of solutions to the equation $P(z) = c$, and so one might expect that an "infinite polynomial" would have an infinite number of solutions. Picard indeed went on to show that a much stronger version of his first theorem was true.

Theorem 3 (Picard's "big" Theorem [1879]) *Let $f(z)$ be an entire function that is not a polynomial. Let R be an arbitrary positive number. Then for every complex number c with at most one exception, the equation $f(z) = c$ has a solution with $|z| > R$.*

Corollary 3.1 *For an entire function $f(z)$, not a polynomial, the equation $f(z) = c$ has an infinite number of solutions for all values of c with at most one exception.*

Proof. Case 1. For every $R > 0$ the equation $f(z) = c$ has a solution with $|z| > R$ for every c, with no exceptions. Let z_1 be any solution. Choose $R > |z_1|$ and choose a corresponding solution z_2 with $|z_2| > R$. Proceeding in the same way gives an infinite number of solutions.

Case 2. For some $R > 0$ there is a value of c such that the equation $f(z) = c$ does not have a solution with $|z| > R$. Then apply the same reasoning above to any complex number $\neq c$.

These theorems of Picard set the stage for much of the research in the theory of functions of a complex variable for the next hundred years and beyond. That will be the subject of our next section. Let us mention here just one immediate corollary of Picard's Theorem that requires the introduction of a new concept.

Definition. A function f is called *meromorphic* in a domain D if, for every point a in D, there is a neighborhood of a in which $f(z)$ can be represented by a power series in $(z - a)$ plus a polynomial in $1/(z - a)$. A point a at which positive powers of $1/(z - a)$ occur is called a *pole* of f.

Examples of meromorphic functions in the whole plane are *rational functions*: quotients $f(z) = P(z)/Q(z)$ of two polynomials. If P and Q have no factors in common, then the poles of f are simply the zeros of the denominator.

For a rational function f, the equation $f(z) = c$ has a solution for every value of c, since one can multiply through by the denominator and apply the fundamental theorem of algebra. However, one cannot view a rational function, or a meromorphic function, as a map into the complex plane, since it has no finite value at a pole. One traditionally talks about a map into the *extended plane* consisting of the ordinary complex plane plus a single point at infinity. Then if a is a pole of f, we write $f(a) = \infty$.

Riemann's second contribution, referred to earlier, was to give a beautiful geometric interpretation of the extended plane. He simply imported stereographic projection from cartography to map the ordinary plane onto a sphere minus a point, and then the point at infinity fills in the missing point on the sphere. Using the standard stereographic projection from the North Pole, it is the North

Pole on the sphere that corresponds to the point at infinity. It is easy to show that if a is a pole of f, then as $z \to a$, $f(z)$ composed with the inverse of stereographic projection tends to the North Pole, so that a meromorphic function in a plane domain D can be considered as a continuous map of D into the sphere. A closer look at this map shows that at the North Pole the map is not only continuous but has exactly the same behavior as at any other point on the sphere. In particular, at a *simple pole,* where the term $1/(z-a)$ occurs, but no higher powers, the map into the sphere is conformal at a. If higher powers of $1/(z-a)$ occur, then the map into the sphere behaves exactly the same as at ordinary points where f is analytic, but $f' = 0$.

The unit sphere viewed this way as the extended complex plane via stereographic projection is called the *Riemann sphere.* It has the effect of taming the "point at infinity" in the complex plane and making it essentially no different from any other point. Another way of thinking about it is that if $f(z)$ is a meromorphic function with a pole at a, then one can compose f with stereographic projection taking $f(a)$ to the North Pole, and follow with a stereographic projection from any other point on the sphere taking the North Pole onto a finite point. Then the composed map of the plane into the plane will be an ordinary analytic function in a neighborhood of the point a.

Picard's two theorems have an immediate extension to meromorphic functions:

Theorem 4 *Let $f(z)$ be a nonconstant meromorphic function in the entire plane. Then viewed as a map into the Riemann sphere, the image of f covers the entire sphere with at most two exceptions. Furthermore, if f is not a rational function, then the same is true for $f(z)$ with $|z| > R$ for any $R > 0$.*

Proof. If the image omits three points on the sphere, and if $c \neq \infty$ is one of them, then the function $g(z) = 1/\big(f(z)-c\big)$ will be an entire function omitting two points, contradicting Picard's Theorem.

Another way to think of it is that if the image omits three points, then make a stereographic projection from one of them onto the plane; that will again give an entire function omitting two points.

When these results were first announced, they probably appeared to be the culmination of a century's work in the subject, and in fact, Picard's proofs used some of the most sophisticated developments of the previous years. However, as is so often the case, Picard's theorems turned out to be just the starting point for a whole array of further investigations. They will be the subject of our final section.

7 The twentieth century

The years following the publication of Picard's theorems saw many new proofs, generalizations, and refinements, but nothing to compare with the sweep and depth of a paper by a young Finnish mathematician, Rolf Nevanlinna, in 1925. That paper inaugurated a whole branch of complex function theory called "value distribution theory" or simply "Nevanlinna theory." What Nevanlinna did was to look at the range of values taken on by an analytic or meromorphic function, and introduce highly refined measures of the relative frequency with which the function took on those values. Now for entire functions or meromorphic functions in the whole plane other than rational functions, all values are assumed infinitely often, with one or two possible exceptions, so that what one is comparing is not simply the size of these sets, but rather a kind of measure of their density. Roughly speaking, one compares the number of solutions of the equation $f(z) = c$ inside a circle of radius R for different values of c, and then sees what happens as R tends to infinity. What Nevanlinna proves is

(i) for "almost all" values of c, in a very strong sense, the solutions of $f(z) = c$ have "the same order of magnitude" or "the same density" in a very precise sense.

(ii) for the exceptional values of c where the equation has "fewer solutions" there is a precise measure of the size of the solution set, called the *defect* of c, and the sum of all the defects is at most 2.

The second of these results is Nevanlinna's far-reaching generalization of Picard's theorem. It follows immediately from the definition of the defect that if a meromorphic function in the plane omits a value altogether, then the defect of that value is equal to 1. Hence there can be at most two such values. For an entire function, the point at infinity is omitted, hence has defect equal to 1. It follows that the sum of all the other defects can be at most equal to 1, and in particular, at most one finite value can be omitted altogether.

Among the many other results proved by Nevanlinna in this groundbreaking paper, we note one of particular interest. For every value c, Nevanlinna introduces in addition to the defect, a "ramification index" of c, and obtains a result similar to (ii) above for the sum of the ramification indices. He also makes the following definition.

Definition. Let $f(z)$ be an analytic or meromorphic function in a domain D. A value c of $f(z)$ is *totally ramified* if whenever $f(b) = c$, $f'(b) = 0$.

In other words, if f is viewed as a mapping from the z-plane to the w-plane, then at none of the points that maps onto c does f define locally a one-to-one conformal map, as would be the case if the derivative were different from zero. Rather, at each of the pre-images of c, f behaves like the function z^n in a neighborhood of the origin, for some $n > 1$. The image is then said to be "branched" or "ramified" in a neighborhood of c.

Theorem 5 (Nevanlinna [1925]) *An entire function can have at most two totally ramified values. A meromophic function in the plane can have at most four totally ramified values.*

Both numbers "two" and "four" in this theorem are sharp. For example, the complex sine and cosine are entire functions defined by

$$2\cos z = \exp iz + \exp(-iz), \quad 2i \sin z = \exp iz - \exp(-iz).$$

It follows that just as for real values of z, $\sin^2 z + \cos^2 z \equiv 1$ and $\cos z$ is the derivative of $\sin z$. Hence $\sin z = \pm 1 \Leftrightarrow \cos z = 0$, which means that the two values 1 and -1 are totally ramified for $f(z) = \sin z$. Similarly, Nevanlinna points out that examples of meromorphic functions with exactly four totally ramified values include classical elliptic functions such as the Weierstrass $\wp$-function. We give a more geometric description of such a function that will be of particular interest in the sequel.

Let a regular tetrahedron be inscribed in the unit sphere, with one vertex at the North Pole. The four vertices of the tetrahedron will form four equally-spaced points on the sphere. Project the tetrahedron onto the sphere from the center of the sphere. The result will be a tiling of the sphere by four congruent spherical triangles. We want to construct a meromorphic function in the plane with the property that each of the four vertices will be totally ramified, and everywhere else the map will be an unramified conformal map. To do so, we make a stereographic projection from the North Pole, under which the four spherical triangles map onto one domain D bounded by three circular arcs meeting at 120°, together with three unbounded domains each having one side in common with a side of D, and the other two sides consisting of rays from the endpoints of that side out to infinity (see Figure 18.14).

We assume that configuration to lie in the w-plane, and we use the Riemann mapping theorem to define a conformal map $f(z)$ of the interior of an equilateral triangle in the z-plane onto the domain D. That can be done in a way that takes the vertices into the vertices and the center

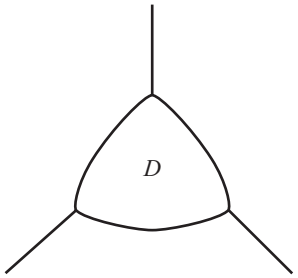


Figure 18.14. 3 circular arcs meeting at 120°

into the center. A fundamental result of H. A. Schwarz that he obtained while investigating the Riemann mapping theorem tells us that the map $f(z)$ can be extended to a meromorphic function in the whole z-plane by a process known as the "Schwarz reflection principle" [1869–70 (a)]. If one pictures the z-plane tiled by the equilateral triangles obtained by successive reflection across the sides of the original triangles and their images, then the extended function $f(z)$ will satisfy $f'(z) \neq 0$ everywhere except at the vertices of the tiling, where the mapping will behave like z^2.

We turn next to a 1926 paper of André Bloch. Bloch's goal was two-fold. First, to give an elementary proof of Picard's Theorem, whose original proof and subsequent refinements tended to be anything but elementary. Second, to implement in this case a general policy that has come to be known as "Bloch's Principle." The idea is that if one has a theorem that applies to functions of a certain class defined in the whole plane, then one should seek a finite version—say in a disk of radius R—that yields the original theorem in the limit as $R \to \infty$.

Theorem 6 (Bloch [1926]) *There exists a positive constant B with the following property. Let $w = f(z)$ be analytic in the unit disk and be normalized so that $f'(0) = 1$. Then for every $r < B$ there exists a disk of radius r in the w-plane that is the one-to-one conformal image under f of a domain inside the unit disk of the z-plane.*

The largest value of B for which Bloch's Theorem holds is known as *Bloch's constant.* Its precise value is not known.

Corollary 6.1 *The mapping of the z-plane into the w-plane defined by an arbitrary nonconstant entire function $w = f(z)$ has the property that for every $R > 0$, there is a disk of radius R in the w-plane that is the one-to-one image under f of some domain in the z-plane.*

Proof. We can first normalize f so that $f'(0) = 1$. Choose $\lambda > R/B$, where B is Bloch's constant. Then apply Bloch's Theorem to the function $g(z) = f(\lambda z)/\lambda$ in the unit disk. Let $r = R/\lambda$. Since $r < B$, we conclude that the image of the disk of radius λ under f includes a disk of radius $\lambda r = R$.

Note. Corollary 6.1 had been proved earlier by Valiron [1926].

Corollary 6.2 *Picard's Theorem.*

To derive Picard's theorem from Corollary 6.1, suppose that the image of a nonconstant entire function f omits two points a, b. Then the function $g(z) = (f(z) - a)/(f(z) - b)$ is an entire function that omits the values 0, 1. We can then compose g with the complex logarithm to get an entire function that omits the points $2\pi ni$ for all integers n. Another slightly more complicated, but still elementary composition yields an entire function that omits a rectangular lattice-type array of points with the property that for some R sufficiently large, every disk of radius R contains one

of those points. This gives an entire function that violates the conclusion of Corollary 6.1. Hence the assumption that the original function could omit two distinct values is false.

The 1930's saw a series of dazzlingly original papers on Nevanlinna theory by another Finnish mathematician, Lars V. Ahlfors, for which he received one of the first two Fields Medals awarded in 1936. In those papers, Ahlfors re-frames, re-formulates, and re-proves the main results of Nevanlinna theory in far more geometric fashion than in the original papers. One of those papers in particular, "On the theory of covering surfaces" from 1935 was singled out by the committee choosing the Fields medalists and was later described by Ahlfors himself as a "much more radical departure from Nevanlinna's own methods" which is indeed the case. We cite just one of the most striking results from that paper.

Theorem 7 (Ahlfors [1935]) *Let $w = f(z)$ be a nonconstant complex function defined in the whole z-plane.*

1) *If f is entire, then given any two disjoint disks in the w-plane, the interior of at least one of them is the image under f of some domain in the z-plane;*
2) *If f is entire, then given any three disjoint disks in the w-plane, the interior of at least one of them is the one-to-one conformal image under f of some domain in the z-plane;*
3) *If f is meromorphic in the whole plane, then the same holds for any five disjoint disks on the Riemann sphere.*

The third statement here is known as Ahlfors' "five islands" theorem.

This three-part theorem of Ahlfors has some of the same aspects of astonishing simplicity of statement and totally unanticipated result that characterizes Picard's original theorem. Part 1 of the theorem is of course a far-reaching generalization of Picard, since if a nonconstant entire function were to omit two values, one could choose a disk about each in contradiction to Ahlfors' result. In the other direction, if one starts with the two disks, one knows from Picard's Theorem that one of them at least must be completely covered by the image of the function, but it might well be in many bits and pieces, whereas Ahlfors' theorem says it is the image of a single connected domain (an "island").

Similarly, part 2 of Ahlfors' theorem implies Corollary 6.1 to Bloch's theorem above, since one can make all three of the given disks as large as one wants. But nothing in Bloch's theorem and its corollaries implies that one can pick specific disks in advance, only that *somewhere* in the image is a disk with the specified property.

Finally, part 3 is an equally surprising generalization of Nevanlinna's Theorem about the maximum number of totally ramified values for a meromorphic function in the plane, since if there were five totally ramified values one could choose five disjoint disks about them and obtain a contradiction to Ahlfors' result.

If Ahlfors had stopped there he probably would still have been awarded the Fields Medal. But in answer to "can you top this?", he did. To fully understand the icing on this cake, one must take note of a particular property of analytic and meromorphic functions that was generally understood to account for the vast majority of special properties that they enjoy. That is the property known as "rigidity." What that means in this context, is that if you change an analytic function in some neighborhood—no matter how small—of a point, then it changes everywhere. Said differently, the values of an analytic or meromorphic function are determined over its whole domain of definition by its values in an arbitrarily small neighborhood of any point in that domain. That property is not shared by even infinitely differentiable real functions, which may be pushed and pulled locally without affecting them elsewhere and still kept infinitely differentiable. The aspect of Ahlfors' paper that must have been the most counterintuitive based on all that came before—where all the elaborate machinery developed specifically for analytic and meromorphic functions had been

invoked—was that his methods showed that none of the special properties of those classes of functions were needed, but rather that the results and all their corollaries actually remain true for an enormously broader class of mappings, with no rigidity properties whatever.

Definition. A smooth map between plane domains, or more generally between surfaces (such as the plane and the sphere) is *quasiconformal* if there is a uniform bound to the ratio of maximum to minimum scaling factors at each point.

To understand the significance of this condition, recall from sections 2 and 3 that the differential of a smooth map at a point is a linear transformation that maps a circle onto an ellipse, and the ratio of major to minor axes of that ellipse is precisely the ratio of maximum to minimum scaling factor at the point. For a conformal map, the ellipse reduces to a circle at each point. In general, as for example, in Lambert's equal area map, or the plate carrée, the ellipses get more and more distorted toward the poles, so that there is no uniform bound on the ratio of maximum to minimum scale factor. Those maps are therefore not quasiconformal. However, the class of quasiconformal maps is far larger than that of conformal maps. What Ahlfors proved was

Theorem 8 (Ahlfors [1935]) *The conclusions of Theorem 7 are all valid for arbitrary quasiconformal maps of the plane into the sphere.*

After many detours, this whole circle of ideas reached its culmination in the final year of the twentieth century with a theorem of Mario Bonk and Alexandre Eremenko [2000]. In order to state the theorem, we recall the example we gave of a meromorphic function that is totally ramified over the four vertices of an equilateral tetrahedron inscribed in the unit sphere. Let C' be a circle passing through three of those vertices, and let D' be the circular disk on the sphere bounded by C'.

Theorem 9 (Bonk and Eremenko [2000]) *Let D be any circular disk on the Riemann sphere smaller than the disk D' described above. Then for any (nonconstant) meromorphic function $f(z)$ in the plane, there is a domain in the z-plane mapped one-to-one conformally by f onto a disk of size D.*

One of the many remarkable features of this theorem is that unlike Bloch's theorem and other similar ones, there is no normalization required. The bound given on the size of the image holds for *all* meromorphic functions in the plane. Furthermore, also unlike the original Bloch's Theorem, where the precise value of Bloch's constant remains unknown, the bound given here is best possible, since any disk larger than D' would include one of the vertices of the tetrahedron in its interior, and the example we have constructed would contradict the conclusion.

But the most remarkable feature of this theorem is that, as the authors show, it implies all the previous theorems of Nevanlinna, Bloch, and Ahlfors described above. The proof employs an eclectic array of tools, from classical spherical geometry to quasiconformal mappings. And it may be worth a passing comment that in order to get a sharp result, the authors resort to the 2,000 year-old device of using stereographic projection from the plane onto the sphere.

Acknowledgement The maps and Tissot indicatrices for maps in Figures 18.5, 18.6, 18.8, 18.9, 18.10, and 18.11 were all taken from "An Album of Map Projections" by John P. Snyder and Philip M. Voxland (U.S. Geological Survey Professional Paper 1453, 1989).
kern-2pt

References

1569. G. Mercator, *Nova et aucta orbis terrae descriptio ad usum navigantium emendate accomodata* (A new and enlarged description of the earth with corrections for use in navigation).

1610. E. Wright, *Certain Errors in Navigation,* London.

1695. E. Halley, An easy Demonstration of the Analogy of the Logarithmic Tangents, to the Meridian Line, or sum of the secants: with various Methods for computing the same to the utmost Exactness, *Philosophical Transactions* 19, 202–214.

1772. J. H. Lambert, Anmerkungen und Zusätze zur Entwerfung der Land- und Himmelskarten (English translation: Notes and Comments on the Composition of Terrestial and Celestial Maps, Ann Arbor, University of Michigan 1972.)

1775. L. Euler, On representations of a spherical surface on the plane, in *Collected Works,* Series 1, Vol. 28, 248–275. (See pp. 251–253.)

1799. C. F. Gauss, *Demonstratio nova theorematis omnem functionem algebraicam rationalem integram unius variabilis in factores reales primi vel secundi gradus resolvi posse,* in *Collected Works,* Vol. III, 3–30.

1822. C. F. Gauss, Allgemeine Auflösung der Aufgabe die Theile einer gegebenen Fläche auf einer andern gegebnen Fläche so abzubilden, dass die Abbildung dem Abgebildeten in den kleinsten Theilen ähnlich wird, in *Collected Works,* Vol. IV, 189–216.

1824. N. H. Abel, *Mémoire sur les équations algébriques,* Christiana.

1827. C. F. Gauss, *Disquisitiones Generales Circa Superficies Curvas,* in *Collected Works,* Vol. IV, 219–258. (Original and a translation in P. Dombrowski, *Astérisque* 62, Soc. Math. de France, Paris 1979.)

1851. B. Riemann, Grundlagen für eine allgemeine Theorie der Functionen einen veränderlichen complexen Grösse. (Ph D thesis, Göttingen), in *Collected Works,* 3–45.

1869–70. H. A. Schwarz, (a) Ueber einige Abbildungsaufgaben, J. reine u. angewandte Math. 70, 105–120; *Collected Works II,* 65–83.
(b) Zur Theorie der Abbildung; *Collected Works II,* 108–132.
(c) Ueber eine Grenzübergang durch alternirendes Verfahren, Vierteljahrschrift der Naturforschungen Gesellschaft in Zürich, 15. Jahrgang, 272–286, *Collected Works II,* 133–143.

1879. E. Picard, Sur une propriétée des fonctions entiéres, *C. R. Acad. Sci. Paris* 88, 1024–1027.

1925B. A. Bloch, Les Théorèmes de M. Valiron sur les fonctions entières et la théorie de l'uniformisation, *Ann. Fac. Sci. Univ. Toulouse III,* 17, 1–22.

1925N. R. Nevanlinna, Zur Theorie der meromorphen Funktionen, *Acta Math.* 46, 1–99.

1926. G. Valiron, Sur les Théorèmes des MM. Bloch, Landau, Montel et Schottky, *C.R. Acad. Sci. Paris* 183 728–730.

1935. L. V. Ahlfors, Zur Theorie der Überlagerungsflächen, *Acta Math.* 65, 157–194.

1980. F. V. Rickey and P. M. Tuchinsky, An Application of Geography to Mathematics, *Math. Magazine* 53, 162–166.

1997. T. Needham, *Visual Complex Analysis,* Clarendon Press, Oxford.

1998. P. J. Nahin, *An Imaginary Tale: The Story of* $\sqrt{-1}$*,* Princeton University Press.

2000. M. Bonk and A. Eremenko, Covering properties of meromorphic functions, negative curvature and spherical geometry, *Annals of Math.* (2) 152, 551–592.