CHAPTER 3

Spatial Data

It is impossible not to feel stirred at the thought of the emotions...at certain historic moments of adventure and discovery: Columbus when he first saw the Western shore, Pizarro when he stared at the Pacific Ocean, Franklin when the electric spark came from the string of his kite, Galileo when he first turned his telescope to the heavens. Such moments are also granted to students in the abstract regions of thought, and high among them must be placed the morning when Descartes lay in bed and invented the method of coordinate geometry.

A. N. Whitehead, quoted in Maling (1973)

Spatial information is comprised of data that can be viewed or located in two or three (or more) dimensions. As we have seen in Chapter 1, when the spatial arrangement of the data is important to their understanding, an analysis that explicitly uses spatial information can be very informative. In this chapter we explore the essential features of spatially referenced data, including location, map projections, and support. We also review the types and sources of spatial data pertaining to public health and give an overview of geographic information systems (GISs) that provide computational tools for managing, merging, and displaying spatial data.

3.1 COMPONENTS OF SPATIAL DATA

There are three components to spatial data: *features, supports,* and *attributes.* A *feature* is an object with a specific spatial location and distinct properties. There are several types of spatial features:

1. *Point*: a precise location, $\mathbf{s}$, in space; a dot on a map. For example, a point could be the geographic location of your house or the location of an air monitoring station.

Applied Spatial Statistics for Public Health Data, by Lance A. Waller and Carol A. Gotway
ISBN 0-471-38771-1 Copyright © 2004 John Wiley & Sons, Inc.

2. *Line*: a sequential collection of connected points. Roads, rivers, and geographical boundaries are examples of linear features.
3. *Area*: a region enclosed by lines. Counties, states, and census tracts are all examples of areal spatial objects.
4. *Volume*: a three-dimensional object having height or depth (vertical extent) as well as horizontal extent. The most common examples of volumetric features are geologic formations such as aquifers.

A collection of features of the same type is called a *feature class*. For example, if we know the locations of several air monitoring stations, each station is a point feature and the collection of all locations is a point feature class.

Each feature is of a certain size and shape and has a specific spatial orientation. Taken together, these properties form the *support* of the data. Points, or spatial locations, have the smallest support. They have zero size, no shape, and no orientation. Lines have length and can indicate direction. Regions have area and boundaries that may impose properties on the associated features. For example, a circle and a rectangle are both areal features, yet they are inherently different spatial objects even if they have the same area.

Attributes are observations or measured values associated with features (e.g., the NO_x concentrations recorded at air monitoring stations, the racial composition of counties, the salinity of rivers). Attributes provide the data with which statisticians are most familiar [i.e., most (nonspatial) statistical analyses examine attribute data without regard to location or support]. In any given analysis we may have a single type of attribute of interest, or several types of attributes associated with each feature (e.g., we may have ozone, particulate matter, and sulfur dioxide measurements at each monitoring station; or the percentage of the population in each county self-identifying in each of several race categories). When several types of attributes are associated with the spatial features, the data are called *multivariate*. Some authors do not always distinguish spatial (*location space*) from multivariate (*variable space*), since spatial data can be referenced by two coordinates in the plane that may also be considered "variables." However, statisticians need to distinguish the two terms since the body of methods presented in courses on multivariate statistics is often very different from the multidimensional methods used for spatial data analysis. The distinction is simple if we think in terms of attributes and features. *Multivariate* refers to more than one type of attribute; *multidimensional* refers to more than one coordinate axis in space. Since this book is introductory, our primary focus will be on the analysis of a single type of attribute in two-dimensional space, although we provide references to statistical methods for multivariate and three-dimensional spatial data.

Thus, spatial data consist of features indexed by spatial locations and with specified supports, and attributes associated with those features. Spatial statistical analysis will be based not only on the attribute data, but will also depend on the spatial locations and the features associated with these locations. To get started, we need ways to reference spatial location, as described in the next section.

3.2 AN ODYSSEY INTO GEODESY

Webster's Collegiate Dictionary defines *geodesy* as "a branch of applied mathematics concerned with the determination of the size and shape of the earth and the exact positions of points on its surface." To work with spatial data, we need a way to reference spatial location. We also need methods for measuring distances between locations and for describing complex shapes and their properties. In this section we provide a very brief overview of the science of geodesy, drawing on concepts from geometry and topology.

3.2.1 Measuring Location: Geographical Coordinates

Many different *coordinate systems* have been developed to reference a point uniquely on Earth's surface. Most of these involve approximating Earth by a sphere or ellipsoid in order to use the geometrical properties of these objects to form the basis of the coordinate system. Earth is not a perfect sphere, nor a perfect ellipsoid, and its surface is not smooth, complicating the calculation of precise locations. In this book we make the simplifying assumption that latitude and longitude (a spherical coordinate system described in some detail below) provide enough location accuracy for our purposes, and refer interested readers to the geodesy literature for more detailed discussions (cf. Smith 1997).

The system of latitude and longitude provides a means of uniquely referencing any point on the surface of a sphere (Figure 3.1). Lines of longitude circle the Earth passing through the north and south poles. All places on the same meridian have the same longitude. The line of longitude passing through the Greenwich Observatory in England has the value 0°. Thus, longitude measures the horizontal angle formed between the line drawn from a given point to the center of the sphere and a line drawn from the center of the sphere to the 0° line of longitude (Figure 3.2). Due to its rotational nature, we report longitude in degrees (0° to 180°) east or west from the 0° meridian, with meridians west of 0° longitude termed *west longitude* and those east of 0° termed *east longitude*. Since the surface of Earth curves, the distance between two meridians depends on where we are on Earth: the intermeridian distance is smaller near the poles and larger near the equator.

To reference north–south positions, lines of latitude (called *parallels*) are drawn perpendicular to the lines of longitude, with the equator designated as 0° latitude (the largest circle defined by a plane perpendicular to the axis of Earth's rotation). Latitudes in the northern hemisphere are termed *north latitudes* and those in the southern hemisphere are called *south latitudes* (see Figure 3.1). Thus, on a spherical Earth, latitude measures the vertical angle (in degrees) between two line segments: one going from the location of interest to the center of the sphere, the other joining the equator with the center of the sphere. Figure 3.2 indicates that on a spherical Earth, these segments intersect at the center of the Earth, but since Earth is actually flattened somewhat at the poles, the true point of intersection is offset somewhat from the center (cf. Longley et al. 2001, p. 88). True to their name and

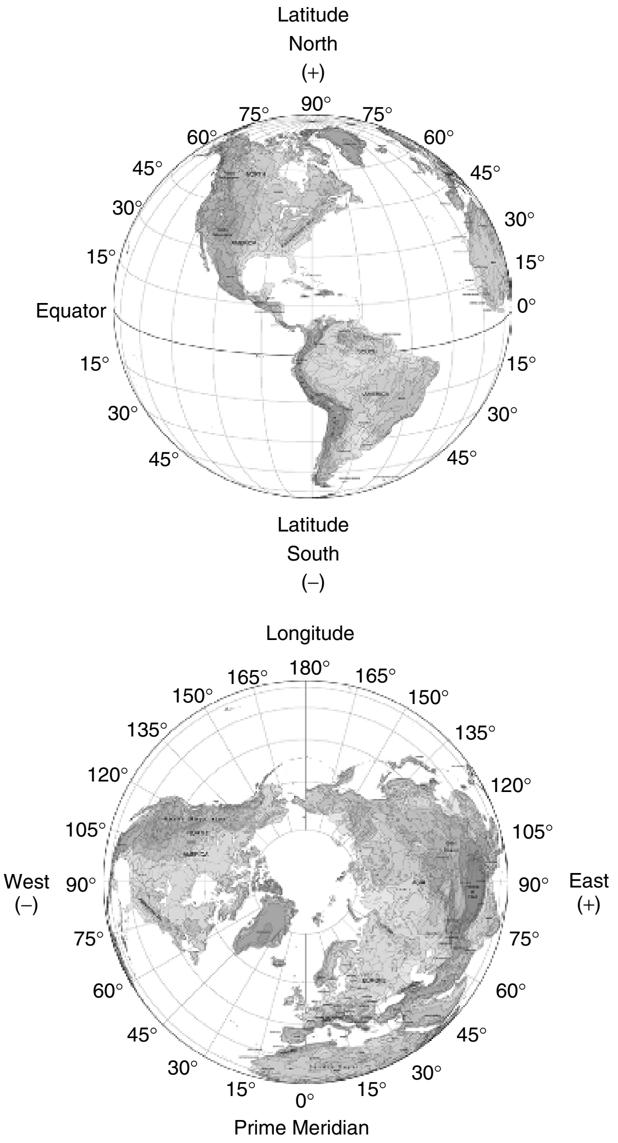

FIG. 3.1 Geometric definitions of latitude and longitude on a spherical Earth.

unlike meridians, parallels are parallel to one another, and differences in latitude are constant over Earth's surface. One degree latitude is approximately 69 miles.

Any point on Earth, s, can be georeferenced by the coordinate pair (longitude, latitude). Each coordinate can be as finely measured as we need it to be by dividing degrees into 60 minutes and each minute into 60 seconds. For example, a longitude value written as $46°22'38''$W denotes a point that is located 46 degrees, 22 minutes,

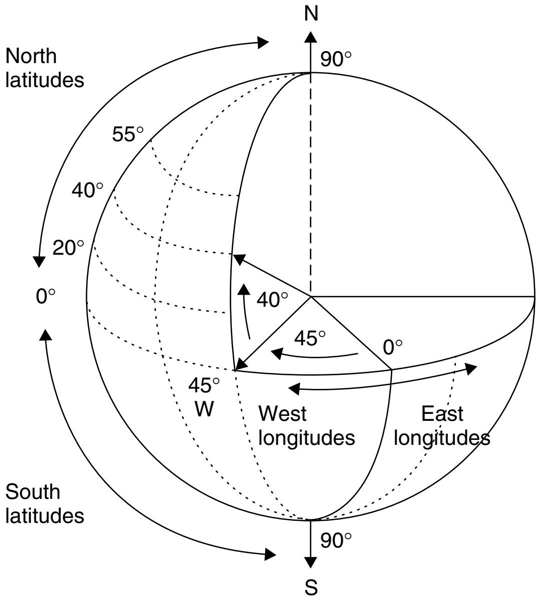

FIG. 3.2 Latitude and longitude system of coordinates.

and 38 seconds west of 0°. For calculations, this specification is often translated into *decimal degrees* in much the same way as we would translate 1 hour and 30 minutes into 1.5 hours. Also, "E" and "W" are designated by "+" and "−" so $46°22'38''$W is the same as $-46.3772°$.

3.2.2 Flattening the Globe: Map Projections and Coordinate Systems

A three-dimensional globe often is not as convenient as a two-dimensional map. A *map projection* is a mathematical transformation used to represent a spherical (or ellipsoidal) surface on a flat map. The transformation assigns each location on the spherical Earth to a unique location on the two-dimensional map. However, we cannot fit the curved surface of Earth to a plane without introducing some distortion. Map projections differ in the degree of distortion introduced into areas, shapes, distances, and directions. *Conformal* (e.g., Mercator) projections preserve local shape. Typical uses of such maps involve measuring angles (e.g., navigation charts and topographic maps), since a line drawn in a particular direction will appear straight in a conformal projection. Small areas are relatively undistorted, but conformal projections are unsuitable for large regions because areas are distorted. *Equal-area* (e.g., Albers' equal-area) *projections* preserve area, so regions will maintain their correct relative sizes after projection. Equal-area projections are useful for representing distributions of attributes (e.g., population size/density and land use) over large areas. Projected maps of these types of attributes produced using an equal-area projection will maintain the relative sizes of each region, and

Table 3.1 Distances (in Miles) between Route Locations for Various Projections

	Projection			
Route	Unprojected	Albers'	Mercator	Equidistant
Atlanta–Seattle	2185	2098	2930	2180
Atlanta–Chicago	588	617	751	601
Atlanta–New York	754	737	931	742
Atlanta–Knoxville	161	171	175	151

hence the relative extents of the associated attributes. These types of maps can be misleading when using another projection in which the areas are distorted. *Equidistance projections* preserve distance relationships in certain directions along one or a few lines between places on the map. These projections allow accurate measures of surface distances by the corresponding measured distances on the map.

Each type of projection can preserve only one property. Thus, a conformal mapping distorts areas, and an equal-area projection distorts shape. There are many compromise projections that are not conformal, equal-area, or equivalent, and each can be thought of as providing a projection providing *minimum total error* as defined by a summary of resulting distortions in area, shape, distance, and direction. Snyder (1997) provides a summary of a wide variety of map projections and discusses the strengths and weaknesses of each. Figure 3.3 shows four different maps of the continental United States. Notice how the relative sizes and shapes of the states vary among the maps. Table 3.1 shows how distances can vary as well.

The first step in a projection is the definition of the shape of the Earth we plan to project and the relationship between this shape and locations on Earth. As noted in Section 3.2.1, Earth is not a perfect sphere and is somewhat flattened at the poles. As a result, an ellipsoid (or spheroid) provides a better starting approximation, and several standard ellipsoids exist (with different ones providing different levels of accuracy). The position and orientation of the ellipsoid relative to Earth also need to be defined. When an ellipsoid is fixed at a particular orientation and position with respect to Earth, it is called a *geodetic datum*. With a *local datum*, the ellipsoid more closely approximates Earth for a particular area. For example, the NAD27 datum has the location (98°32′30″W, 39°13′30″N), corresponding to Meades Ranch, Kansas, as the reference point. At this point, the ellipsoidal model of Earth and true Earth coincide exactly. The NAD83 datum is an example of one widely used Earth-centered geodetic datum, calculated using the center of the Earth as a reference point. The WGS84 (World Geodetic System of 1984) is not referenced to a single datum, but instead, defines an ellipsoid whose placement, orientation, and dimensions best fit Earth's surface. Longley et al. (2001, p. 88) list several other standards. Different datums have different coordinate values for the same location, so two maps referenced to different datums can give locations for the same point that differ by several hundred meters.

Having chosen our standard ellipsoid, we can geometrically project locations on the ellipsoid Earth onto any of three types of surfaces, called *developable surfaces,*

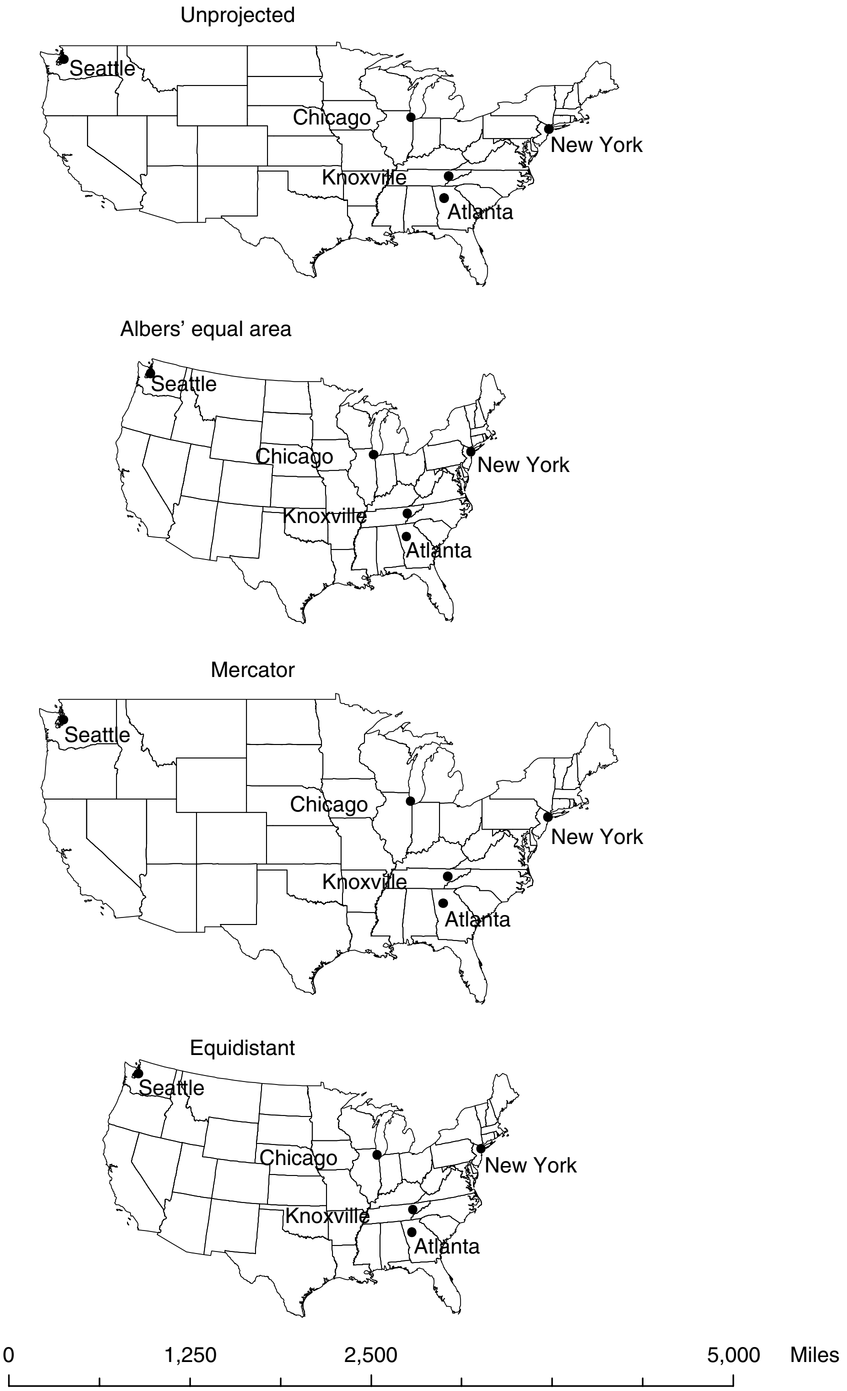

FIG. 3.3 Comparative view of different map projections.

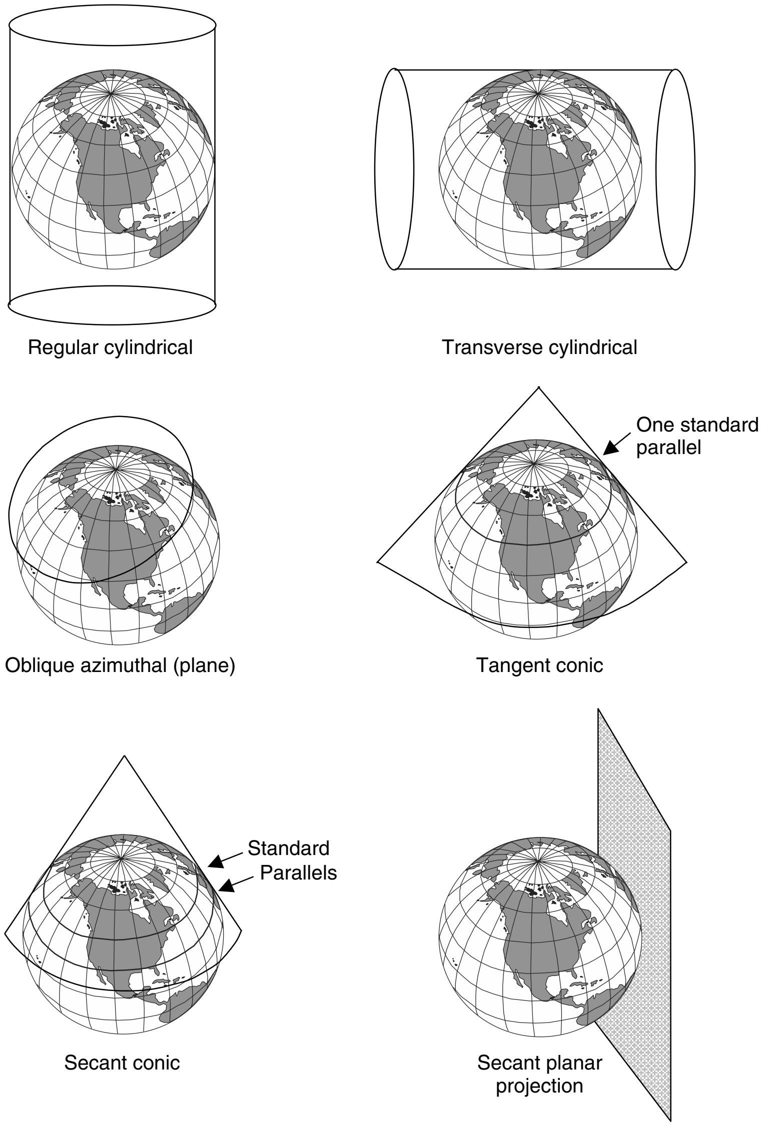

FIG. 3.4 Developable surfaces used in map projections.

that have the property that they can be flattened without distortion. Developable surfaces include planes, cones, and cylinders, resulting in *planar* or *azimuthal*, *conic*, and *cylindrical projections*, respectively (see Figure 3.4). Thus, each type of projection described above (e.g., conformal, equidistant) can also be classified by the appropriate developable surface and whether that surface is tangent to or intersects the surface of the ellipsoid. For example, for conic projections, the point of the cone falls along the axis of rotation and the cone is either tangent to the ellipsoid along a circle (one *standard parallel*) or intersects the ellipsoid at two circles (two standard parallels). One popular example is Lambert's conformal conic projection, with two standard parallels. Azimuthal projections preserve direction from one point to all other points, but preserve distance only along the standard parallel(s). As mentioned above, azimuthal projections can be combined with equal-area, conformal, and equidistant projections, creating, for example, the Lambert equal-area azimuthal and the azimuthal equidistant Projections.

A map *scale* is the relationship between a distance on the map and the corresponding distance on the ground. It is expressed as a fraction such as 1:24,000, meaning that 1 unit on the map corresponds to 24,000 units on the ground. If the units are in miles, then 1 mile on the map represents 24,000 miles on the ground. Because of the projection, the scale actually varies over the flattened map. Thus, an average approximation is usually given in the map legend to give the map interpreter some idea of distance. Small-scale maps (e.g., 1:1,000,000) show little detail but great extent and thus have low spatial resolution. Large-scale maps (e.g., 1:1000) are much smaller in extent, but show greater detail and thus have high spatial resolution.

Once we have projected points on Earth to a two-dimensional, flat surface, we need to set up a grid system to reference each point. Thus, we need to designate the center of the grid, the units, the central meridian, and the *scale factor* used in the projection. The scale factor (usually, a value ≤ 1.0) is applied to the scale of the centerline of a map projection where the developable surface intersects the ellipsoid (usually, the central meridian or a standard parallel). Scale values less than 1.0 are used to reduce the overall distortion of a projection. Most coordinate systems have already specified these parameters for us. For example, one commonly used coordinate system is the Universal Transverse Mercator (UTM) coordinate system. It results from a conformal mapping onto a cylinder wrapped around the poles of the Earth instead of around the equator as with the ordinary Mercator projection (cf. Longley et al. 2001, pp. 92–94). This projection is very accurate in narrow zones around the meridian tangent to the cylinder. The globe is subdivided into narrow longitude zones, 6° wide, each projected with a transverse Mercator projection. These zones are numbered with zone 1 between 180° and 174° west longitude and moving eastward to zone 60 between 174° and 180° east longitude. The lower 48 U.S. states are covered by zone 10 on the west coast through zone 19 on the upper east coast (Figure 3.5). In each zone we report UTM coordinates in meters north (*northings*) and east (*eastings*). To avoid negative numbers for locations south of the equator, the value 10,000,000 meters represents the equator. Each zone contains a central meridian assigned a value of 500,000 meters. Grid values to the west of this central median are less than 500,000, and to the east they are greater than 500,000. More complete descriptions of the UTM coordinate system as well as other map projections can be found in Snyder (1997) and Clarke (2001).

With all the different projections and coordinate systems, how do we know which one to use? Quantitative uses of a map (e.g., measurement of distances, areas, and angles) are more likely to reflect projection distortions than are visually

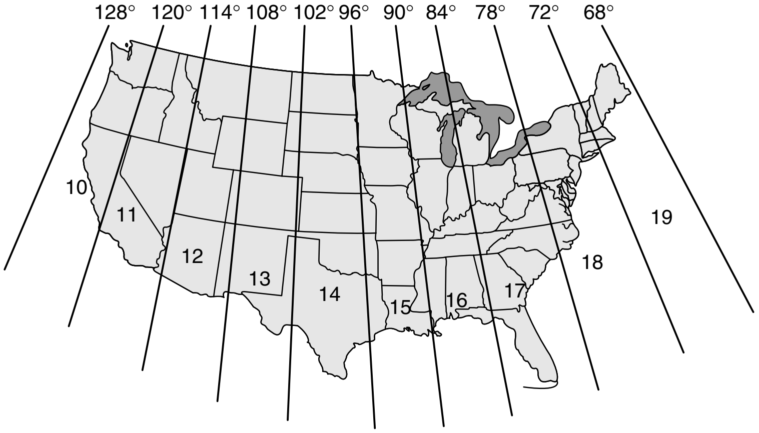

FIG. 3.5 UTM zones of the continental United States (from the U.S. Geological Survey). The upper set of numbers gives the longitude value, and the center numbers denote the UTM zones.

based subjective determinations. Thus, if we do not require a high level of accuracy for locations (e.g., we will not be performing queries based on location and distance, or we just want to make a quick map), we may not need to transform our data to a projected coordinate system; a planar display based on latitude and longitude coordinates (the default in most automated mapping software) will suffice. Scale, distance, area, and shape are all distorted, with the degree of distortion increasing with distance from the equator. If, however, we need to make precise measurements on our map, or we want to preserve one or more of these properties (area, shape, distance, or direction) for calculations or accurate visual depiction, we should choose a projected coordinate system. The system you should use depends on what you want to display. The amount of distortion resulting from a projection depends on the location, size, and shape of the region of interest. Distortion is least for small, compact regions and greatest in maps of the world. Popular projections include the Robinson projection (a compromise projection) for maps of the world, and the Albers equal-area conic and Lambert azimuthal or polyconic projections for maps of continental extent in the middle latitudes (e.g., North America). UTM coordinates (comprised of multiple local projections as outlined above) provide another common projection system for maps of the United States. Individual states (within the United States) often use the state plane coordinate system for regional mapping. This system is philosophically similar to the UTM system, but each zone may be projected differently than the others. More specific details, examples, comparisons, and mathematics that may be useful when choosing a particular coordinate system are given in Maling (1973), Snyder (1997), and Clarke (2001).

3.2.3 Mathematics of Location: Vector and Polygon Geometry

Once we have a set of locations projected onto a plane, we use vector notation and linear algebra to summarize common quantities such as distances, directions, and paths through locations. This notation provides the symbolic language for the statistical methods developed in subsequent chapters, and we provide a brief overview here.

Vectors Some physical quantities, such as location and length, are completely determined when their values are given in terms of specific units. For other quantities, such as velocity, direction is also important. Such quantities are called *vectors*. It is common to represent a vector visually by a directed line segment whose direction represents the direction of the vector and whose length represents its magnitude. Mathematically, vectors in two-dimensional space are often represented by a matrix with two rows and one column, whose elements give the u and v coordinates of the vector starting from $(0, 0)$ and ending at location (u, v) in the plane. To add two vectors, say $\mathbf{h}_1 = (u_1, v_1)'$ and $\mathbf{h}_2 = (u_2, v_2)'$, simply add the corresponding components together: $\mathbf{h}_1 + \mathbf{h}_2 = (u_1 + u_2, v_1 + v_2)'$. Subtraction can be done analogously. The *length* or *magnitude* of a vector $\mathbf{h} = (u, v)'$ is denoted by $\|\mathbf{h}\| = \sqrt{u^2 + v^2}$. This is also called the *norm* of the matrix $(u, v)'$. The *zero vector*, denoted $\mathbf{0}$, is a vector whose length is zero.

Polygons A polygon is a closed planar figure with three or more sides and angles. We are probably most familiar with rectangles, which are one of the simplest polygonal forms. We are probably also familiar with how to find the center of a rectangle and its area. However, many of the polygons in spatial analysis (e.g., the boundaries of census blocks, counties, and states) are more complex than simple rectangles. In some applications, we may even want to specify our own polygons, which describe boundaries of regions of interest to us. We may also want to specify the "location" of a polygon and determine its area.

In the plane, a polygon is specified mathematically by an ordered set of points, $\{u_i, v_i,\ i = 0, \dots, n\}$, connected by line segments. These points define the *vertices* of the polygon. We adopt a notation for which the first and last vertex are equivalent, so that $u_0 = u_n,\ v_0 = v_n$. In some instances, we wish to define the center of a polygon in order to define some notion of distance between polygons. One definition of *center* is the *central value* of a polygon, obtained by averaging the u values and the v values to obtain location (u_m, v_m). Another definition of center is the *centroid* of a polygon defined by the center of mass (or balancing point) of the polygon. The coordinates of the centroid of a polygon, R are given by

$$c_u = (1/A) \iint\limits_{\mathrm{R}} u \, du \, dv$$

$$c_v = (1/A) \iint\limits_{\mathrm{R}} v \, du \, dv,$$

where A denotes the area of polygon R. Since the centroid depends on the vertices only through their definition of the perimeter of R, the centroid is less influenced than the central value by the number of vertices along any boundary of R. For example, consider a polygon with one edge defined by a curving feature such as a river. If we attempt to represent the curve through many line segments (hence many vertices), we will drive the central value toward the edge containing many vertices, while the centroid value remains relatively stable.

At first glance, the centroid seems more difficult to compute than the central value, and the area of a general polygon can be difficult to derive mathematically. However, many simple algorithms for computing the area and centroid of a polygon do exist. One that we have found useful is

$$A = \left|(1/2)\sum_{i=0}^{n-1} u_i v_{i+1} - u_{i+1} v_i\right|$$

$$c_u = [1/(6A)]\sum_{i=0}^{n-1}(u_{i+1} + u_i)(u_i v_{i+1} - u_{i+1} v_i)$$

$$c_v = [1/(6A)]\sum_{i=0}^{n-1}(v_{i+1} + v_i)(u_i v_{i+1} - u_{i+1} v_i).$$

Both central values and centroids can fall outside the polygons if the polygons have very unusual shapes (e.g., crescent, donut). In these cases, we may use known point locations to reference the polygons spatially.

For many analysis applications, we will not need the central values or the centroids of the polygons. However, many spatial analyses rely on distances (and directions) among locations to describe spatial relationships. In such cases we calculate distances between central values or centroids and use these to infer distances between polygons. For most applications, it matters little if we use the central value or the centroid to indicate the location of a polygon as long as we use the same definition consistently throughout the analysis. However, spatial analyses based on distances computed using central values may differ from those using distances computed using centroids since distances among central values probably differ from distances among centroids. Summarizing the location of a polygon by any one point in space necessarily introduces uncertainty in the analysis that we may want to adjust for when interpreting the results.

How Far? Distance Measures and Proximity As we mentioned in Chapter 1, one of the key concepts in spatial statistics is the idea that attribute values measured on features near one another tend to be more similar than those measured on features farther apart. Thus, to quantify this for use in statistical analysis, we need mathematical descriptions of *near* and *far*. We can quickly think of a very easy description: the distance between two features. However, there are many ways to measure distances, and as we saw for polygons, sometimes the idea of the location of a feature can be a bit vague. In this section we describe several different measures of distance that quantify the degree of closeness between two spatial features.

As the World Turns: Great Arc Length Suppose that we are using the longitude/latitude coordinate system to pinpoint locations on Earth's surface and we have two such locations, $\boldsymbol{s}_1 = (\lambda_1, \phi_1)$ and $\boldsymbol{s}_2 = (\lambda_2, \phi_2)$, where λ denotes the longitude coordinate and ϕ denotes the latitude coordinate. Then, the shortest distance between these two locations along the surface of a spherical Earth is given by

$$d(\boldsymbol{s}_1, \boldsymbol{s}_2) = (6378) \cdot \arccos[\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\lambda_1 - \lambda_2)], \qquad (3.1)$$

where 6378 kilometers is the radius of the (spherical) Earth.

As the Crow Flies: Euclidean Distance Suppose, instead, that we are working with a projected coordinate system and we have two locations, $\boldsymbol{s}_1 = (u_1, v_1)$ and $\boldsymbol{s}_2 = (u_2, v_2)$, in a two-dimensional plane. Then the shortest distance between these two locations on a flat map is given by

$$d(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sqrt{(u_2 - u_1)^2 + (v_2 - v_1)^2}. \qquad (3.2)$$

Using the notation for vectors, this distance can also be referred to as $\|\boldsymbol{s}_2 - \boldsymbol{s}_1\|$. This is called the *Euclidean norm*, and the distance measure in equation (3.2) is called *Euclidean distance*. We could also use the longitude and latitude coordinates, $\boldsymbol{s}_1 = (\lambda_1, \phi_1)$ and $\boldsymbol{s}_2 = (\lambda_2, \phi_2)$, in this formula, but the resulting distance would not take into account the curvature of the Earth. In general, the Euclidean distance measure should not be used to compute distances between sets of longitude and latitude coordinates, particularly if the distances are over a large area. Since no adjustment is made for the curvature of the Earth, distances affected by this curvature will be distorted.

As the Person Walks: City-Block Distance In some situations, measuring the shortest distance is not at all meaningful. For example, in urban areas where there are one-way streets and buildings between blocks, we cannot travel a straight line to our destination. Thus, we have to go "around the block," and to do this we drive or walk along a series of perpendicular segments. This gives rise to the idea of the "city-block" distance between two locations, $\boldsymbol{s}_1 = (u_1, v_1)$ and $\boldsymbol{s}_2 = (u_2, v_2)$:

$$d(\boldsymbol{s}_1, \boldsymbol{s}_2) = |(u_2 - u_1)| + |(v_2 - v_1)|. \qquad (3.3)$$

There are other distance measures for analogous situations [e.g., "as the fish swims" (Little et al. 1997) and "as the water flows" (Cressie and Majure 1997)]. More details on using these distance measures in spatial analysis are given in Section 8.4.6.

Across the Picket Fence: Adjacency When we have polygonal features instead of point locations, we can index the location of each polygon by its centroid and then use any of the distance measures described above to compute distances between centroids. This gives one measure of the distance between two polygons. However, sometimes a meaningful measure of the "closeness" of two polygonal features is simply whether or not they share a boundary (i.e., whether or not they are adjacent). This gives rise to a binary proximity measure

$$w_{ij} = \begin{cases} 1 & \text{if polygons } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise.} \end{cases}$$

We use the term *proximity* here rather than *distance* since the distance measures described above are actually topological *metrics* that must satisfy certain relationships not necessarily satisfied by proximity measures. There are many different proximity measures that can be used to define the closeness between two polygonal features, and we discuss these in more detail in Sections 4.4.1 and 7.4.2. As we shall see throughout this book, both distance measures and proximity measures have their useful place in spatial data analysis.

3.3 SOURCES OF SPATIAL DATA

The availability of spatially referenced data continues to increase at a rapid pace. Based on our experience, we focus our outline of available resources on data for the United States and Canada. However, similar data exist for other countries (or collections of countries). The following references or organizations offer access or descriptions of other geographically referenced data sets: Lawson (2001) (United Kingdom), Statistics Finland (Finland), Pan American Health Organization (Central and South America), World Health Organization, and the United Nations. Geographic scope and support (e.g., enumeration districts, counties, states, nations) of particular data sets vary widely.

In Canada and the United States, digital spatial features (i.e., spatial data that can be described by numbers) are produced by the national mapping agencies, agencies responsible for the decennial census, and other national organizations. Considerable effort has been made to coordinate and standardize the production and distribution of digital geographic data and most of these data are now readily available free on the Internet. In the following sections we describe some of the types of spatial data available that are of potential interest to public health professionals. We avoid giving actual Internet addresses, since such addresses change over time. Instead, we provide names of organizations, programs, and surveys and enough detail to enable Internet search engines to locate particular data sources. With the exception of some health data, most of the data that we describe are also available on CD-ROM for minimal processing fees.

3.3.1 Health Data

Data regarding health events vary widely in terms of purpose, ranging from specific clinical trials and localized observational studies to national (and international) disease surveillance efforts. Many spatial analyses of public health data utilize health outcome data collected and summarized by governmental agencies, often state health departments, and then released for public use.

Stroup et al. (1994) provide a thorough overview of and comprehensive bibliography to data sources relating to public health. We summarize several types of health data collected by various national and local governments, with particular attention to spatial aspects of various data sources. Many of the data sources listed below are publicly available through contact with the agency or organization responsible for the data (although we note that "publicly available" need not always equate with "easily available"). The different types of data involve a variety of collection methods and purposes, and accordingly, vary in terms of availability, spatial coverage, and spatial support.

As noted in Chapter 2, confidentiality affects the availability and spatial resolution of public health data. Reporting agencies must balance the individual's right to privacy with the public's right to know. This balance tips in different directions at different times under different governmental domains and regulations, so the spatial support available for similar types of data sources may vary between reporting units.

Vital Statistics Certification of death (issuance of death certificates) represents one of the earliest attempts to routinely gather and summarize health-related information. Parish records in western Europe dating from the fifteenth century and London's Bills of Mortality, a weekly report of deaths categorized by causes beginning in 1537, are early examples of data collection and reporting efforts relating to vital (birth and death) statistics. Stroup et al. (1994, pp. 38–44) provide a thorough overview of the history and development of the collection of vital statistics noting that vital statistics are the only health-related data available from many countries in a standard format. About 80 countries or areas currently report vital statistics to the World Health Organization, coded according to the *International Classification of Disease* (ICD) (ninth or tenth edition).

In the United States, the National Center for Health Statistics (part of the Centers for Disease Control and Prevention) collects, coordinates, and maintains vital statistics, including mortality data, some published recently in atlas form (Pickle et al. 1996; Devesa et al. 1999; Casper et al. 2000; Barnett et al. 2001). In general, spatial resolution is no finer than the county level, and for particularly rare causes of death in sparsely populated counties, numbers may be suppressed due to confidentiality concerns. The National Cancer Institute's *Atlas of Cancer Mortality in the United States, 1950–94* (Devesa et al. 1999) reports mortality for counties and for state economic areas, collections of counties within states based on demographic and economic variables as measured in 1960. The atlas does not include maps when many of the small areas (counties or state economic areas) contain sparse data, so maps are not reported for some disease/race/gender combinations. The National Center for Health Statistics' *Atlas of United States Mortality* (Pickle et al. 1996) reports for health services areas, another aggregation of counties based on a cluster analysis (classification algorithm) linking counties based on where residents aged 65 years and over obtained short-term hospital care in 1988. Unlike state economic areas, health service areas cross state boundaries.

Notifiable Diseases Due to their devastating impact and high infectivity, certain diseases often motivate surveillance by governmental public health units in order to stem outbreaks, monitor trends, and plan intervention strategies. *Notifiable diseases* are those associated with regulatory reporting requirements (i.e., by law, each incident case must be reported to a reporting agency or system upon diagnosis or laboratory verification). Stroup et al. (1994) provide a history of notifiable disease-reporting systems for the United States and internationally.

In the United States, the Centers for Disease Control and Prevention (CDC) maintains the Nationally Notifiable Disease Surveillance System to manage information on 54 reportable diseases. Some states add to the list of notifiable diseases. The National Electronic Telecommunications System for Surveillance (NETSS) provides the CDC with weekly data regarding each of the nationally reportable diseases. The data include the date of diagnosis, age, gender, race/ethnicity, and county of residence of each reported case, but no personal identifiers (e.g., name of the case). The NETSS provides information summarized in the CDC's *Morbidity and Mortality Weekly Report* (MMWR), but the CDC is currently upgrading from the NETSS to the National Electronic Disease Surveillance System (NEDSS) to address some of the reporting difficulties experienced by NETSS users.

Registries *Disease registries* differ slightly from vital statistics and notifiable disease data collection mechanisms in that registries link multiple sources of information for each case. Examples of information sources include hospital-discharge reports, death certificates, pathology reports, billing records, and in some cases the medical charts themselves. Registries attempt to consolidate information by patient so that each case appears only once in the registry.

Stroup et al. (1994) contrast *case series* and *hospital-based registries* from *population-based registries*. Case series and hospital-based registries attempt to provide information to improve patient care, and often do not provide accurate estimates of incidence rates (proportions) for the population. In comparison, population-based registries seek broader coverage in order to provide accurate estimates of overall incidence and (possibly) for local areas.

Individual registries focus on particular diseases, such as cancer and birth defects. Several cancer registries operate internationally. In the United States, the North American Association of Central Cancer Registries (NAACCR) serves to coordinate efforts and monitor quality among population-based cancer registries. NAACCR recently produced a report providing an overview and introduction to basic geographic information system (GIS) practices for cancer registries that addresses issues directly relating to spatial coverage, support, and analysis of cancer registry data in the United States (Wiggins 2002). For birth defects in the United States, the Birth Defects Monitoring Program (BDMP) links individual states, the CDC, the National Institute of Child Health and Human Development, and two nonprofit organizations (the March of Dimes and the Commission on Professional and Hospital Activities) to provide ongoing surveillance of the incidence of birth defects through registry programs (Oakley et al. 1983).

Different registries offer different spatial coverages and support. For instance, the National Cancer Institute's Surveillance, Epidemiology, and End Result (SEER) cancer registry is most likely the largest population-based cancer registry in the Western world. The SEER registry provides comprehensive incidence data since 1973, but only for the 11 SEER sites, which include state registries in Utah, New Mexico, Iowa, Hawaii, and Connecticut, and regional registries in Detroit, Seattle–Puget Sound, Los Angeles, San Francisco–Oakland, San Jose–Monterey, and Atlanta. The SEER registry also includes supplemental sites, including a Native American registry for the state of Arizona, a collection of rural Georgia counties, and the Alaska Native Tumor Registry. Planned expansion to the SEER program include registries in New Jersey, greater California, Kentucky, and Louisiana. The National Program of Cancer Registries (NPCR), established in 1992 by a U.S. congressional mandate and managed by the CDC, offers coordination and certification for registries covering non-SEER areas and states. For those seeking information on international cancer registry activities, the International Agency for Research on Cancer (IARC), part of the World Health Organization (WHO), collects information from cancer registries around the world, and provides a point of contact.

Health Surveys In contrast to registries and surveillance that seek to record all incident events within a given time period, health surveys use population-based statistical samples to draw inference regarding incidence and prevalence of health outcomes and related demographic and risk factor information. Analysts typically use design-based estimation to provide inference for the reference population, typically the aggregate population from which researchers draw the sample.

The National Health and Nutrition Examination Survey (NHANES) and the National Health Interview Survey (NHIS) are two examples of national health surveys in the United States (Korn and Graubard 1999). NHANES involved three separate data collection efforts. NHANES I collected data from civilian, noninstitutionalized individuals aged 1 to 74 years between the years 1971 and 1974 (with some follow-up in 1974 and 1975). NHANES II collected similar data from 1976–1980 (with the minimum age decreased to 6 months), and NHANES III collected data from 1988–1994 (with the minimum age decreased to 2 months and removal of an upper bound on age). The NHANES design involved primary sampling units of counties (or collections of contiguous counties). Within sampled areas, researchers collected data from household interviews and medical examinations performed in mobile examination centers (thereby providing some clinical measures linked to interview data). In comparison, the NHIS uses counties or metropolitan areas as primary sampling units, then conducts household interviews within selected units. The NHIS has been in continuous operation since 1957, with some modifications from time to time. A key difference between NHANES and NHIS data is the presence of some clinical measures (e.g., blood pressure and blood lead) in the NHANES data, collected by mobile examination centers in addition to the household interview data.

The Behavioral Risk Factor Surveillance System (BRFSS) provides an example of an annual health survey conducted by each state within the United States (Centers for Disease Control and Prevention 1998). Based on a structured telephone interview, the BRFSS collects information on health outcomes and behavioral and lifestyle factors such as exercise and diet. States collect data sufficient for obtaining statewide estimates and within any particular year, states with large numbers of counties (e.g., Texas and Georgia) will have several counties not contributing to the sample.

Surveys are designed for estimation at a particular level of aggregation (e.g., a nation for NHANES and NHIS, a state for BRFSS). As we might expect, there is often interest in obtaining sample-based estimates for local regions within the entire study area (e.g., states within a nation, or counties within a state). Cost often rules out obtaining adequate sample sizes to support design-based estimation within each local administrative unit, so researchers often use *small-area estimation* to combine data from regions statistically to stabilize estimates (Ghosh and Rao 1994, Schaible 1996, Malec et al. 1997). Only recently have small-area estimation procedures included spatial correlations, and Ghosh et al. (1998) provide an example using county-level lung cancer mortality rates in Missouri.

3.3.2 Census-Related Data

In the United States and Canada, the agencies responsible for collecting and disseminating census data provide a number of digital data sets that can be useful in public health applications. In the United States, in preparation for the 1990 census, the U.S. Bureau of the Census developed the TIGER (Topologically Integrated Geographic Encoding and Referencing) system to standardize, encode, and aid in processing of census questionnaires. TIGER/Line files cover the 50 states, the District of Columbia, Puerto Rico, the Virgin Islands, and the outlying areas of the Pacific Ocean over which the United States has jurisdiction. The spatial data contained in these files includes street networks, address ranges for street segments, railroads, political boundaries (including digital boundaries of census block groups, census tracts, counties, and states), and the boundaries of major hydrographic features, all geographically referenced by longitude and latitude coordinates. TIGER/Line files do not contain demographic attribute data, but they do contain region identifiers allowing one to link attributes from the associated census with TIGER polygons. Typical linked attribute data include feature names and codes (e.g., codes for state, county, census tract, and block) and may include population and housing unit counts, income, racial classification, and housing values. Not all attributes are available for every geographic level; for example, only the total population and housing unit counts are available at the census block level. Many local governments and software development companies have enhanced, reorganized, or simplified TIGER files for their own use. Statistics Canada produces similar spatial data sets.

3.3.3 Geocoding

Much of the public data collected in the United States is comprised of individual records, often specified by a person's address. If you have an emergency and call "911," one of the first pieces of information the operator will ask you for is your address. This address provides definitive location information for the purposes of dispatching emergency care. However, it is not definitive for automated, computer-based cartography. For this, we need geographic coordinates of the address. One way to obtain these coordinates is through *geocoding:* the process of assigning a spatial location to an address record.

Geocoding involves matching records in (at least) two databases: the database containing the address information and a reference geographic base file that contains both addresses and the geographic coordinates of those addresses. This assumes that a complete, easily available, accurate geographic base file exists, which is usually not the case except at a very local level using digital municipal data files. More commonly, the Census TIGER/Line data described in Section 3.3.2 provide the geographic base file, address ranges, and street segment records for geocoding. Some mode of interpolation then provides the actual location of the address within the street segment. Some location error remains in most geocoded addresses, and the error could be quite large for rural areas containing few street segments. Thus, the geographic base file providing address and/or street location is critical to achieving accurate locations. A variety of geographic base files are now available from both public-domain and private-sector publishers. If you provide addresses that are complete, specific (i.e., have ZIP + 4 designation), and accurate (i.e., contain no typographical errors and record address elements such as street names in a standard format), it will usually be possible to match 80–90% of your addresses to the addresses and associated geographic coordinates in the base file.

3.3.4 Digital Cartographic Data

In the United States, the U.S. Geological Survey (USGS) has long been a source of maps: topological maps, detailed quadrangle maps, geological maps, and so on. Most of these maps are now available in digital formats. The USGS's Digital Line Graph (DLG) data set includes transportation lines, hydrography, political boundaries, and elevation contours for the entire United States. The USGS land use/land cover data set delineates urban areas, agricultural lands, forests, and wetlands. The USGS has also enhanced the comprehensive digital elevation data first produced by the U.S. Defense Mapping Agency. These data sets provide an elevation value for any location in the United States and provide necessary information for many engineering and urban planning applications.

3.3.5 Environmental and Natural Resource Data

As with the health surveys (cf. Section 3.3.1), the U.S. government also conducts a number of different environmental surveys, designed to monitor the status and trends of ecological and natural resources. Many of these are national, long-term monitoring and assessment programs. Some collect data at monitoring stations (point locations), others collect information over small areal units, and some programs operate at state or regional levels. The type of attribute data collected varies widely across programs. Many use probability sampling to select the units for measurement; others select units based on judgment/convenience. Below we provide a brief overview of some of the major national environmental monitoring programs that provide public-domain data that might be useful in public health studies. Much of our information is based on the work of Olsen et al. (1999), and more detailed, statistical information about many of the surveys we describe can be found in this work.

Agriculture and Natural Resources The National Resources Inventory (NRI), part of the Natural Resources Conservation Service (NRCS) within the U.S. Department of Agriculture (USDA), collects data on land use, wetlands, soil erosion, conservation practices, and habitat diversity. The primary sampling unit is a square plot, containing approximately 160 acres. Some data are collected on these units (e.g., land use, habitat diversity), while more specific information is recorded at individual locations within each unit. NRI data provide estimates of natural resource conditions and changes in these conditions that are used to develop natural resource conservation programs.

The National Agricultural Statistics Service (NASS), also within the USDA, collects data on agricultural lands. It maintains a huge database of agricultural statistics such as crop acreage and production. Some of the information collected pertains to environmental monitoring on agricultural lands. For example, NASS maintains an agricultural chemical-use database that includes information on the type of chemical applied (e.g., specific type of fertilizer, insecticide, or herbicide), the total amount applied, the percentage of cropland treated, and so on. It is also a probability-based survey, based on a stratified, two-stage random sample of segments in land-use strata, combined with a list frame sample of individual farms. The spatial resolution depends on the information collected, most of which is at the state or county levels. NASS is one of the oldest and largest national survey organizations; the survey was mandated by Congress in 1839. Today, NASS publishes 400 national and 9000 state reports each year (Olsen et al. 1999).

Water Quality The Environmental Protection Agency (EPA)'s Environmental Monitoring and Assessment Program (EMAP) is another national probability-based survey. It was initiated to provide information on the status and trends in environmental quality and to identify emerging environmental problems by developing reliable and specific ecological indicators. It is based on a triangular grid covering the entire United States and uses systematic random sampling to determine units for measurement. EMAP Estuaries monitors all U.S. coastal waters measuring indices of ecological condition (e.g., the benthic index, based on surveying benthic invertebrates and combining measures of their abundance and diversity into a single index) and exposure indicators (e.g., dissolved oxygen). EMAP Surface Waters monitors rivers, streams, reservoirs, and lakes (except the Great Lakes). This program collects measurements on a variety of indicators that can be used to infer the "health" of the stream or river, including water quality measurements (e.g., pH), sediment toxicity, and chemical contaminants in fish. More specific information about EMAP design and component programs is provided in Stevens (1994) and Olsen et al. (1999).

The U.S. Geological Survey (USGS) implements several water quality assessment programs. One of the largest, the National Water-Quality Assessment (NAWQA) program, was initiated to collect information on the quality of the nation's ground and surface waters. The sampling design is not probability-based, but instead, study units were selected by a linear optimization algorithm, and the units now participating in the program represent 50 major hydrogeologic basins that comprise the majority of the nation's water use. The data collected on each study unit depend on the characteristics of each particular unit, but often include measurements on water pH, temperature, dissolved oxygen, and nutrient concentrations. Different programs within NAWQA focus on more specific water quality characteristics. For example, the NAWQA Pesticide National Synthesis Project aims to provide a national assessment of pesticides in surface and ground waters and supplies important information to the U.S. EPA for regulations concerning pesticide use and biodegradability requirements.

Air Quality In addition to environmental programs that focus on terrestrial and aquatic ecosystems, the U.S. government also directs many national atmospheric monitoring programs. Under the guidance of the EPA, the National Atmospheric Deposition Program (NADP) and the Clean Air Status and Trends Network (CASTNET) were developed to provide data necessary to assess the effectiveness of air pollution control efforts. Such effectiveness can be assessed by monitoring changes in atmospheric deposition, primarily acid deposition levels, high levels of which are caused by industrial and automobile emissions. The NADP collects weekly wet acid deposition samples from almost 200 sites across the United States. Each site in this network measures important components of precipitation chemistry such as sulfate, hydrogen ion, and chloride. CASTNET consists of over 70 monitoring stations across the United States that provide information on dry acid deposition, ground-level ozone, and other forms of atmospheric pollution. Sites in this network measure weekly average atmospheric concentrations of sulfate and nitrate and hourly concentrations of ozone and meteorological data used to compute acid deposition rates. The monitoring sites in both the CASTNET and NADP programs are located in rural areas and so provide information on natural background pollution concentrations. Other monitoring networks [e.g., the National Air Monitoring Stations (NAMS) network] have stations located in urban areas, and data from these networks can be combined with CASTNET information for a more comprehensive national air quality assessment.

Climate The National Climatic Data Center (NCDC), part of the National Oceanic and Atmospheric Administration (NOAA), is the world's largest archive of global climate data. It collects weather data from the National Weather Service, the Federal Aviation Administration, the U.S. military, and from several international agencies as well. NCDC data are used to provide short- and long-term national and regional climate forecasts. They also provide a complete historical record that can be used to measure global climate change. In environmental health studies, data from the NCDC are often used in modeling the effects of climate on human health and in adjusting models quantifying the effects of air pollution on human health for climatic effects.

3.3.6 Remotely Sensed Data

Remotely sensed data are data collected from a distance. The most common example of remotely sensed data is the aerial photograph. Such photographs can provide reliable spatial measurements such as elevation, soil type, and land use, but the exercise often is not as simple as snapping a photograph: The entire scientific discipline of *photogrammetry* revolves around this endeavor (Jensen 1996). Another approach to collecting remotely sensed data is through the use of satellite images. These are produced from sensors located on satellites that measure the reflected and emitted radiation from Earth's surface. This type of radiation cannot be detected by ordinary photographic film. The images are obtained directly in digital form and are comprised of cells called *pixels* (picture elements). Each pixel corresponds to an area on the ground, and the size of this area determines the *resolution* of the image. Associated with each pixel is a wave of electromagnetic energy. Different features on the ground will emit different energy waves that can then be analyzed (using our knowledge of light and electromagnetism) and interpreted to infer physical characteristics. Today, satellite imagery produces some of the most accurate and globally comprehensive information on the Earth, and many disciplines, including public health (particularly in vector-borne diseases), are finding creative and cost-effective uses for this technology (Cline 1970; Beck et al. 1994; Washino and Wood 1994; Messina and Crews-Meyer 2000; Xiang et al. 2000).

3.3.7 Digitizing

Any map available in hard copy can be scanned into a computer and digitized to produce an electronic file of digital boundaries. To tie the digitized map to a preexisting georeferenced coordinate system (e.g., longitude and latitude), the digitizer must specify the true coordinates of at least three separate locations on the map. This approach provided the initial means for transferring paper maps based on historical land surveys to the digital spatial data sets currently available.

3.3.8 Collect Your Own!

The global positioning system (GPS) is a system of 24 satellites orbiting the Earth. A GPS receiver locates the nearest satellites and receives a signal from each of them. By knowing the time differential between the signals, the positions of the satellites, and details about their orbit, it is possible to determine the exact location (longitude, latitude, and elevation) of the receiver. GPS receivers are quickly finding their way into all kinds of equipment (including automobiles), and handheld GPS receivers are now a cost-effective scientific tool. In conjunction with a laptop computer and some software, analysts can use a handheld GPS receiver to record the boundaries of the spatial features of interest. We may also measure attributes associated with these features (e.g., administer questionnaires to people; take water, soil, or blood samples and have them analyzed by a laboratory). Of course, we do not need a GPS to collect our own spatial data. All we need is a way to record the location of our features relative to other features. For example, we could designate an arbitrary point somewhere as (0,0) and then record attribute information on a regular grid extending from this origin. Agricultural scientists and ecologists have been using this approach for decades and have developed systematic sampling and mapping strategies designed to collect spatial information quickly and easily (e.g., Seber 1986; Stehman and Overton 1996; Wollenhaupt et al. 1997). Thus, public health practitioners should not rely solely on the large, existing spatial data sets described above. Much scientific discovery and understanding remains to be obtained by conducting our own studies that collect the information, both spatial and nonspatial, that we believe is important and relevant.

3.4 GEOGRAPHIC INFORMATION SYSTEMS

The term *geographic information system* (GIS) means many things to many people and has various definitions. The literature on GISs extends back at least to the mid-1960s and the development of the Canada Geographic Information System for the Canadian Land Inventory (Longley et al. 2001, pp. 10–13). From our point of view, a GIS is a complex, interactive software for the management, synthesis, and display of spatial data. As Bonham-Carter (1994) notes, the word *geographic* means that the spatial locations can be specified by geographical coordinates, latitude and longitude. The term *information* implies that the data input into a GIS can by organized in a useful way facilitating interpretation (e.g., through maps, images, charts, and tables). Finally, the word *system* indicates that a GIS is comprised of several different but interrelated components working together. As noted by Clarke (2001), a GIS allows easy visualization of geographic features comprised of points, lines, and areas and allows us easily to address questions regarding features' respective sizes, shapes, orientation, and spatial distribution, such as: Where? How far? In what direction? How big? Several such aspects of spatial features are illustrated in Figure 3.6. A GIS also allows us to link this information with various *attributes* associated with these features. Thus, we can also answer questions such as: Where are the features associated with large attribute values? How close are features with the same attribute value? These are simple but important questions, not quickly answered without GIS technology. Answers to these questions form the basic building blocks for more complex questions, including those addressed in subsequent chapters.

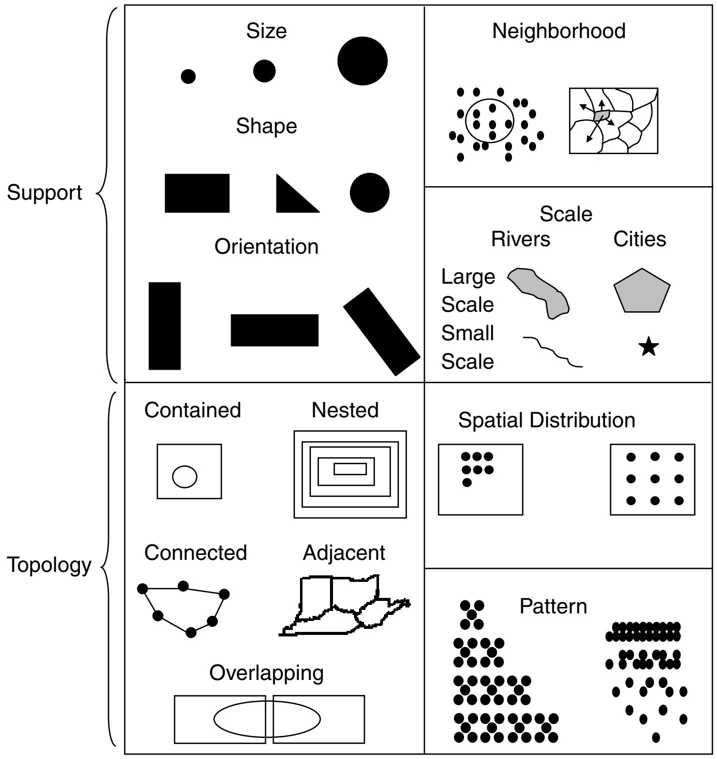

FIG. 3.6 Important aspects of spatial data. [Modified from Clarke (2001).]

3.4.1 Vector and Raster GISs

The literature distinguishes between *vector* and *raster* GISs, depending on whether locations are stored as points/lines/areas or as pixels, respectively. The underlying geographic data structures determine both the computational storage burden and the primary GIS operations of interest. Vector data often involve much less storage since we store attribute data only for points, lines, and areas rather than for every pixel, as in raster data, although various image compression systems significantly reduce the storage requirements of raster data. The speed of computational operations varies between vector and raster data. For instance, we may search through vector data by iteratively referencing each point, line, and/or area from a reference list, while a brute-force search of raster data could involve looping through the grid of pixel points (typically much larger than the vector database). Although many modern GISs manage both vector and raster data, the different storage, searching, and algorithmic strategies associated with the data types merit retention of the distinction.

Different types of data fall more naturally into the vector and raster frameworks. For instance, health or demographic data summarized to census enumeration regions fit well into the vector paradigm, due to the assignment of multiple attributes to each of a set of polygons covering the study area. In contrast, satellite imagery and aerial photography better fit into the class of raster data where each image consists of multiple pixels and their associated values.

3.4.2 Basic GIS Operations

Although particular GIS packages may differ in interface and functionality, there are certain operations central to every GIS. Although these operations do not constitute statistical analysis per se, the operations provide means for querying and linking spatial data.

Spatial Query The operation distinguishing a GIS from any other relational database is the ability to query data elements with respect to their locations as well as their attribute values. The spatial query allows us to request summaries of attribute values in the subset of locations meeting a spatial criterion (e.g., the number of pediatric asthma case residences within 500 meters of a major roadway). The spatial query underlies most GIS functionality and provides the means for sorting, subsetting, and summarizing data with respect to location and distances.

Layering Much as its name implies, the GIS operation known as *layering* consists of overlaying several different spatial data sets, linking values by location. Conceptually, think of a set of maps printed on transparencies, where each map represents a different set of data collected over the same study area (e.g., a map of population density, a map of roads, and a map of land cover). Stacking or layering the maps provides a single map consisting of the composite information from all maps.

Computationally, the GIS matches attribute values based on common locations, allowing one to query the combined attributes by a single location reference file. The term *spatial join* defines the algorithmic operation linking the layered data into a combined (joined) data set.

Layering provides a powerful visualization and data linkage tool. In public health, layering allows one to combine census data providing information on local demographics, disease registry data providing local health outcomes, and environmental monitoring data providing locations of pollution sources, each collected over the same area. As a particular example in a public health setting, Xiang et al. (2000) layered two spatial data sets: a map of maternal residences at the time of birth for infants born between 1991 and 1993 in Weld County, Colorado, and a (raster) map of crop type for 30-meter by 30-meter pixels as determined by satellite imagery. Xiang et al. (2000) next inferred pesticide use based on crop type and a survey of Colorado farmers regarding the type and application patterns of particular pesticides by crop. Through these GIS operations, Xiang et al. (2000) created a combined data set which allowed preliminary assessments of the associations between patterns of low birth weight and pesticide use.

The ability to combine several data sets collected by different agencies for different purposes is both a strength and a weakness of layering. Data merging is a strength in that it allows us to address questions we could not answer using any of the individual data sets alone. In contrast, data merging may be a weakness in that the issue of data quality becomes murky. That is, while each individual layer may meet quality levels sufficient for its intended purpose, the data may not be accurate enough to address the requirements for inference regarding the combined data. For instance, air monitoring stations installed to measure pollution levels near an industrial park may not offer accurate exposure information for subjects throughout the study area. Thus, we emphasize again that the great availability of spatial data and the ease with which such data can be combined and displayed in a GIS are still not substitutes for more focused studies designed to obtain reliable and relevant information.

Buffering *Buffering* refers to a particular type of spatial query, the definition of the area within a specified distance of a particular point, line, or area. For example, Xiang et al. (2000) define areas (*buffers*) of 300 and 500 meters, respectively, around each maternal residence in their study, then assign pesticide values within the buffer based on the remotely sensed satellite data. Other typical applications of buffering in public health studies include the definition of exposure (or *exposure potential*) zones around sources of hazard (e.g., hazardous waste sites). Buffering, too, has advantages and disadvantages. It allows us to combine spatial data collected on different features with differing supports, but in doing so, we have introduced errors and uncertainty into any analysis and the resulting conclusions.

3.4.3 Spatial Analysis within GIS

Many authors note the analytical and predictive capabilities of GISs. Some (e.g., Bailey and Gatrell 1995) distinguish between *spatial analysis*, the study of phenomena occurring in a spatial setting using the basic GIS operations outlined above, and *spatial data analysis*, the application of statistical description and modeling to spatially referenced data. The distinction provides a boundary between standard GIS operations and queries and the application of data analysis algorithms for estimation, prediction, and simulation familiar to many statisticians. Other authors use the term *spatial analysis* quite broadly as a field and include inferential statistics among the tools. For instance, Longley et al. (2001, p. 282) describe six general categories for GIS-based spatial analysis: queries (enabled by the spatial relational database underlying the GIS), measurements (e.g., length, shape, distance), transformations (e.g., spatial joins, and conversion from vector to raster, or vice versa), descriptive summaries (e.g., calculating the mean response in a particular area), optimization (e.g., searching for minima/maxima across a spatial area), and hypothesis testing. The last category specifically involves inferential statistical methods, while the other categories describe a variety of tools for quantifying trends, features, and patterns within a spatial data set. We focus on statistical methods (spatial data analysis) in this book but note that nonstatistical GIS operations often provide necessary precursory elements for the estimation, testing, and modeling approaches outlined in the following chapters.

With the exception of recent software modules implementing the geostatistical methods discussed in Chapter 8, most modern GISs allow little in the way of routine calculations for *spatial data analysis* involving statistical inference. As a result, most analyses described in subsequent chapters involve the use of separate statistical packages and are not analyzed within a GIS setting. Like any set of software packages, the capabilities of a GIS stem from the needs of the users as addressed by software developers. As more users seek spatial statistical modeling capabilities within (or in concert with) GISs, developers will seek to meet the need. *Scripts* (much like macros in SAS and functions in S-Plus or R) can greatly extend the statistical capabilities of a GIS. In addition, a growing variety of *applets* (Web-based application tools) for both statistical and GIS computing provides toolboxes for further development of software tools for spatial data analysis.

3.5 PROBLEMS WITH SPATIAL DATA AND GIS

The apparent availability of spatial databases and the ease with which these can be used and combined within a GIS make it seem easy to obtain relevant and important information for almost any study. However, as alluded to above, there are also common problems, misuses, and limitations associated with the use of spatial databases. Understanding these limitations, and anticipating them a priori, will help to ensure more productive analyses.

3.5.1 Inaccurate and Incomplete Databases

Any database can contain typographical errors and misspellings. Spatial databases are no exception, and such errors can occur in both the location values and in the attribute values. Sometimes, these errors are fairly easy to notice simply by plotting the locations on a map using a GIS. For example, when a location plots outside the domain of interest, it is often the case that the latitude and longitude coordinates were reversed, a negative sign was deleted from the longitude values, or the first digit in one of the georeferenced locations is incorrect. When two spatial databases appear to be offset by a small amount, differing datums could be the culprit.

A quality control program including simple edit checks can help to identify potential problems. For example, we could check to make sure that the date of birth and the age of the person are consistent (i.e., the person cannot be older or younger than the difference between the date of the study and their date of birth), and we can check to see that fields for city, ZIP code, county, and state are all consistent, and so on. These types of checks can be automated easily; it is just a matter of developing a comprehensive system of checks and balances.

Such quality control checks will take us only so far. If we do not know what the attribute values mean, the projection used, and the time frame over which the database was compiled, the database will have limited value. The Federal

Geographic Data Committee (FGDC) has been developing a standard for what is now called *metadata*, or "data about the data." Metadata give us important information about the database, such as who created it, when it was created, when it was last updated, the map projection used, and data quality assessments such as the accuracy of both attribute and locational information. Thus, we should always be sure to look for and understand any metadata before working with a particular spatial database.

It is very difficult if not impossible to find a complete and comprehensive spatial database ideally suited to our interests. Most databases will suffer from one of two problems: lack of extent or lack of resolution. For example, most environmental data are often of point support, but they are usually very localized, pertaining to a county or a region. Some national databases are very sparse, having but one point per state. On the other hand, it is easy to obtain comprehensive, national, state-level health data, but such data are usually not available at finer resolutions (e.g., counties). For any particular study, we will probably have to work very hard to supplement the existing spatial databases with data specific to our needs.

3.5.2 Confidentiality

Many health data sets contain sensitive information. Good ethical practice, and in many cases, federal laws, require us to protect confidential information. The FGDC is developing consistent guidelines that ensure the protection of confidential information, but many institutions have developed their own standards. For example, the U.S. Bureau of the Census will not release any individual-level information, and many U.S. government agencies have "cell suppression" rules for tables (e.g., if a count in a particular cell of a table contains five or fewer individuals, the value will not be released). While the usual patient identifiers such as name and address can obviously be used to identify patients, point locations obtained from geocoding or GPS can be used in the same way. Thus, many institutions refuse to share this type of data or have developed policies that restrict access to such data. For example, to work with some of the data collected by NCHS at the county level, screened users must conduct their research at NCHS, where their use of the data can be carefully controlled and monitored. Unfortunately, these policies and precautions also limit health research studies and the conclusions that can be obtained from them. Recently, *geographical masks* have been designed that preserve the confidentiality of individual health records but also allow analyses that require specific locational information to address important research questions of interest. Armstrong et al. (1999) provide a comprehensive review and discussion of many different types of geographical masks.

3.5.3 Use of ZIP Codes

Since geocoding is expensive or time consuming, many health studies georeference individuals to ZIP codes since patient records often contain a ZIP code field as part of the address. It is very important to remember that ZIP codes were created by the

U.S. Postal Service for delivering mail. Unlike census tracts and blocks, they were not created to be homogeneous with respect to sociodemographic variables, and in many instances, they will not manifest such homogeneity. Sociodemographic information is available by ZIP code, but it is often averaged from census block data. This averaging, over units not necessarily homogeneous, can often lead to misleading conclusions about data mapped at the ZIP-code level. For example, Krieger et al. (2001) report the results of a comprehensive study on whether or not the choice of area-based geographic units really matters when mapping data for public health surveillance. They found that when health outcomes were reported and mapped at the ZIP-code level, ZIP code measures failed to detect gradients or detected trends and patterns that were contrary to those observed with block groups or census tracts.

3.5.4 Geocoding Issues

Address matching works best for completely specified, correctly spelled addresses in urban areas. In many cities, a designation such as "East" or "North" is very important. For example, in Atlanta, North Peachtree Street is in a very different location than Peachtree Street. It is also important to provide all the aliases for a given street (e.g., Peachtree is often abbreviated as "P'tree"). Address matching does not work well in rural areas, and it cannot be used for P.O. boxes or rural route designations.

Achieving a 100% match rate occurs only when we geocode error-free addresses using an error-free base map. Of course, a high match rate does not ensure correct spatial coordinates for each address. Cromley and McLafferty (2002, p. 87) note that it is not uncommon for 7% of locations assigned to addresses within a base map to be incorrect. In addition, Krieger et al. (2001) report variable accuracy in a comparison of four independent geocoding vendors, each assigned the same original set of addresses. Thus, we stress that although many automated geocoders may match most addresses within seconds, geocoding is an iterative process that requires substantial checking and verification to ensure accurate spatial information.

3.5.5 Location Uncertainty

Even a foolproof geocoding approach will not obviate all locational issues. In human health applications, we traditionally assign the residence location to each case. Although we will use residential location for the examples below, such locational assignment may not be entirely satisfactory for some applications. For instance, assigning residence location to each case ignores human mobility and may assign cases to locations far from areas where relevant (e.g., occupational or school-based) exposures occur. Also, people move from place to place during the course of the day and may receive significant environmental exposures at their workplace, in their car, or in other locations. Finally, in studies of chronic diseases (such as various cancers) where disease onset may occur years after the suspected relevant exposure(s), appropriate locational assignment may involve collection of historical housing and occupational data for each case and any relevant noncases collected as a comparison group.

Lilienfeld and Stolley (1984, pp. 138–139) and Cromley and McLafferty (2002, pp. 214–215) both provide an example of location issues based on a study of endemic typhus fever in Montgomery, Alabama, originally published by Maxcy (1926). Maxcy mapped residences of cases revealing little spatial pattern. Maxcy then mapped occupational sites for cases showing a concentration of cases in the city's central business district. Closer examination of additional data associated with type of occupation showed higher incidence for employees of food depots, groceries, feed stores, and restaurants, suggesting a rodent reservoir of the disease with transmission via fleas, mites, or lice. This study illustrates two important points: (1) residence may not be the primary location of interest (e.g., location of the relevant exposure), and (2) additional, nonspatial data (here, type of business) often refine theories linking cases, and eventually, provide more detailed etiologic hypotheses than location alone.

These are some of the problems that we may encounter when working with spatial data. There are undoubtedly more that we can expect with particular applications. It is important to be aware of them, but we should not let them deter us from spatial analysis. We hold spatial data to very high standards: We do not seem to expect other types of data to be so widely and publicly available, nor do we have entire committees ensuring their accuracy and mandating their documentation!

In this chapter we have provided an overview of spatial public health data, from attributes and features, through geocoding, geodesy, and GIS, to sources, surveys, and use of ZIP codes. We turn now to spatial data analysis, and as we will see in the next chapter, this begins with the principles of cartography and the art and science of visualization.