4 Maps, Data Entry, Editing, and Output

Building a GIS Database

Introduction

Spatial data entry and editing are frequent activities for many GIS users. A large number of coordinates is needed to represent features in a GIS, and each coordinate value must be entered into the GIS database. This is often painstakingly slow, even with automated techniques, and spatial data entry and editing take significant time for most organizations.

Most spatial data sources may be categorized as either *hardcopy* or *digital*. Hardcopy forms are any drawn, written, or printed documents, including hand-drawn maps, manually measured survey data, legal records, and coordinate lists with associated tabular data. Most historical spatial data were recorded on maps (Figure 4-1), and although not all maps are suitable for conversion to digital formats, many maps are. Much data were created from hardcopy sources in the early years of GIS via *digitizing*, the process of collecting

Figure 4-1: Maps have served to store geographic knowledge for at least the past 4,000 years. This early map of northern Europe shows approximate shapes and relative locations.

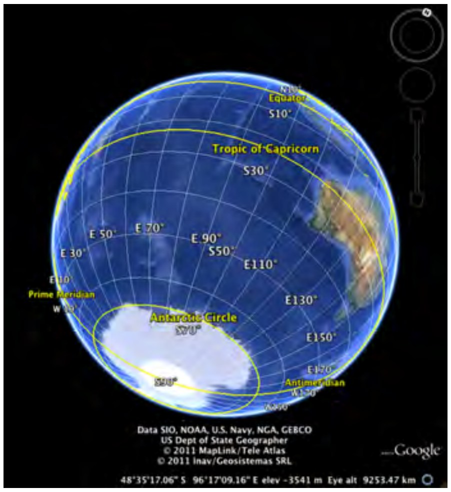


Figure 4-2: An example of commonly produced digital maps (courtesy Google).

digital coordinates. Digitizing is a common data entry method today, although primarily from satellite and aerial images.

Digital maps are an electronic, graphic depiction of spatial data, and are by far the most common map form today (Figure 4-2). Millions of electronic maps are generated each hour, composed on demand in response to web queries, on automobile navigations systems, and for commerce and advertising. These maps are flexible, easily customized, inexpensively distributed, and often dynamic.

Most maps, whether digital or hardcopy, contain several components (Figure 4-3). A *data area* or *pane* occupies the largest part of the map, and contains most of the depicted spatial data. A *neatline* is often included to provide a frame around all map elements, and *insets* may contain additional map elements. *Scalebars, legends*, titles, and other graphic elements such as a *north arrow* are often included. All maps have a *map scale*, defined as the ratio of the distance on the map to corresponding distance on the ground.

Maps often depict coordinate lines (Figure 4-4). When the lines represent constant latitude and longitude, a set of coordinate lines is called a *graticule* (Figure 4-4a). These lines may appear curved, depending on the map scale, the map coordinate system, and the location of the area on Earth's surface. Maps may also depict a *grid* con-

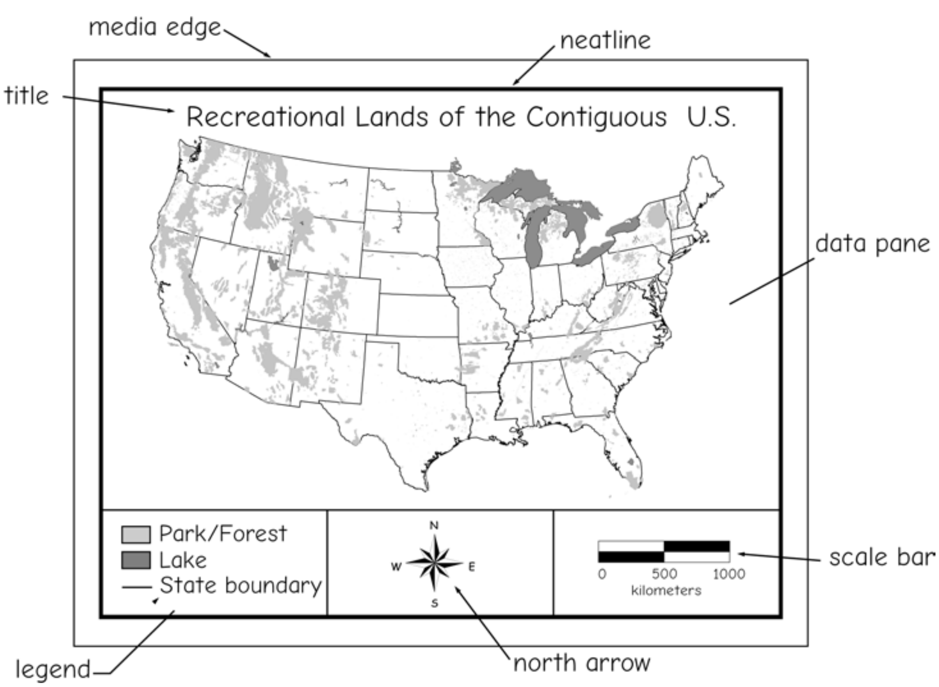


Figure 4-3: An example of a map and its components.

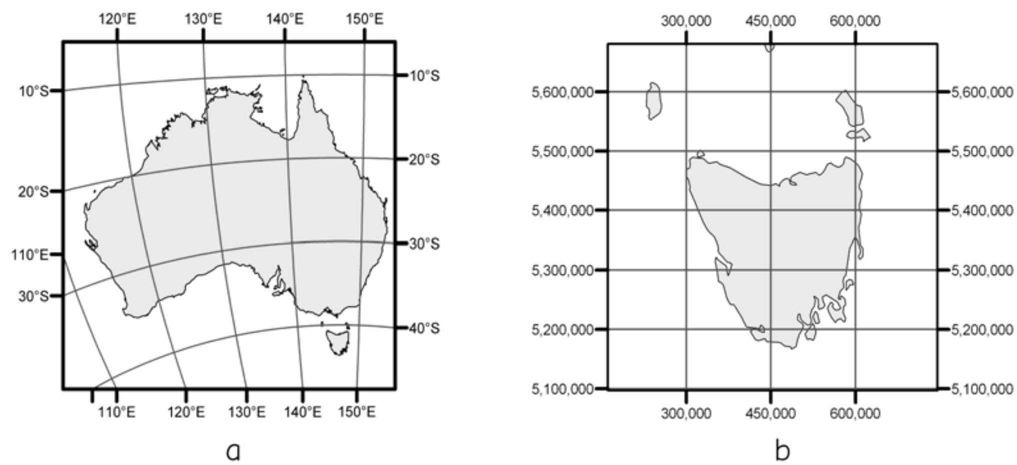


Figure 4-4: Maps often depict lines representing (a) a *graticule* of constant latitude and longitude or (b) a *grid* of constant X and Y coordinates.

sisting of lines of constant coordinates. Grid lines are typically drawn in both the x and y directions, and appear straight on most maps (Figure 4-4b). Graticules and grids are useful because they provide a reference against which location may be quickly estimated. Graticules are particularly useful for depicting the distortion inherent in a map projection, because they show how geographic north or east lines are deformed, and how this distortion varies across the map. Grids may establish a map-projected north, in contrast to geographic north, and may be useful when trying to navigate or locate a position on the map.

Historical and current images are valuable sources of geographic data, and although they are not maps, the line is becoming blurred, as aerial and satellite photographs become common backdrops for digital maps. Photographs do not typically provide an orthographic (flat, undistorted) view unless they are geometrically corrected, and houses, rivers, or features of interest are not explicitly identified. However, images are a rich source of geographic information, and standard techniques may be used to extract features through manual digitizing (described later in this chapter), or through image classification (described in Chapter 6).

Digital spatial data are those provided in a computer-compatible format. These include complete raster and vector data layers, text files, lists of coordinates, and digital images. Files and export formats can be used to transfer them to a local GIS system. Global Navigation Satellite Systems (GNSS), such as the U.S. Global Positioning System (GPS) and coordinate survey devices described in Chapter 5, are direct measurement systems that can be used to record coordinates in the field and report them directly into digital formats. Finally, many digital image sources are available, such as satellite or airborne images that are collected in a digital raster format, or hardcopy aerial photographs that have been scanned to produce digital images.

Hardcopy maps were an important source of geographic information, but now are primarily used for display. Most geographic information produced before 1980 was recorded in hardcopy form. Advances in optics, metallurgy, and industry during the 18th and 19th centuries allowed the mass production of precise surveying devices, allowing much of the world to be plotted on *cartometric* quality maps. These maps faithfully represent the relative position of objects.

Data entry from digital sources now dominates. Coordinates are increasingly captured via interpretation of digital image sources (these sources are described in Chapter 6) or collected directly in the field by satellite-based positioning services (Chapter 5).

Our objective in this chapter is to introduce spatial data entry via digitizing and coordinate surveying. We will also cover basic editing methods and data documentation, and rudimentary cartography and output.

Map Types

Many types of digital and hardcopy maps are produced, and the types are often referred to by the way features are depicted on the map. *Feature maps* are commonly used to map points, lines, or areas and provide nominal information (Figure 4-5, upper left). No attempt is made for symbols to represent true scale. A road may be plotted with a symbol defining the type of road, but the width of the road as plotted is not true to scaled size on the ground.

Choropleth maps depict quantitative information for areas. A mapped variable such as population density may be represented in the map (Figure 4-5, top right). Polygons define area boundaries, such as counties or census tracts. Each polygon is given a color, shading, or pattern corresponding to values for a mapped variable.

Dot density maps show quantitative data (Figure 4-5, bottom left). Dots or other point symbols are plotted to represent values. Dots are placed in the polygon such that the number of dots equals the total value for the

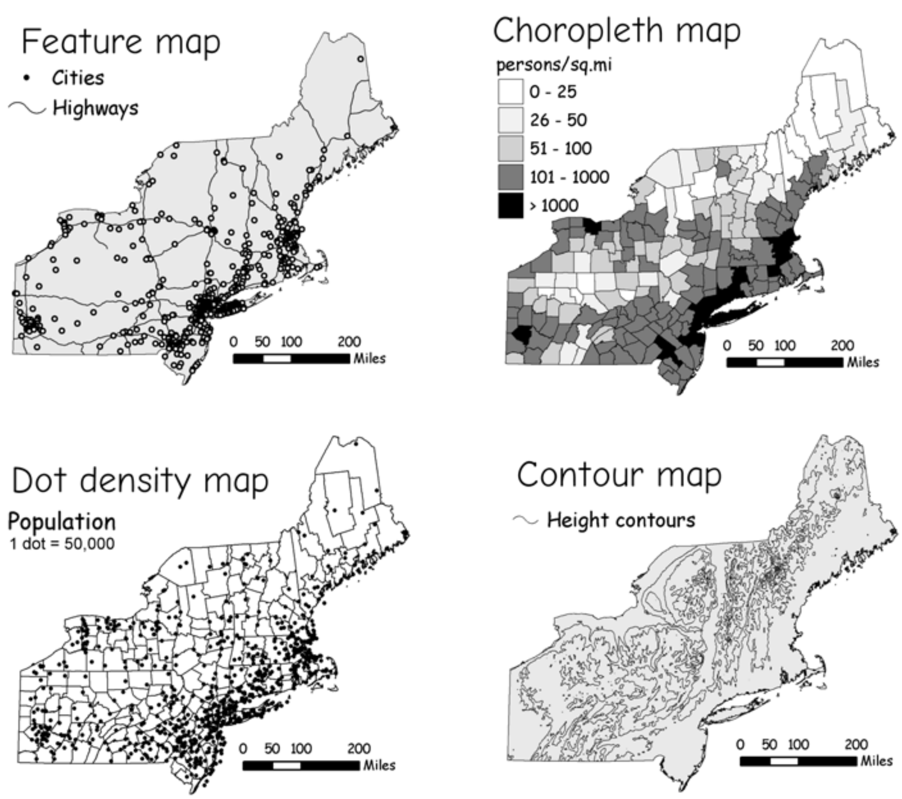


Figure 4-5:
Common hardcopy map types depicting the northeastern United States.

polygon. Note that the dots are typically placed randomly within the polygon area. Each dot on the example map represents 50,000 people; however, each point is not a city or other concentration of inhabitants. Note the position of points in the dot-density map relative to the city locations in the feature map directly above it in Figure 4-5.

Isopleth maps, also known as *contour maps*, display lines of equal value (Figure 4-5, bottom right). Isopleth maps are used to represent continuous surfaces. Rainfall, elevation, and temperature are features that are commonly represented using isopleth maps. A line on the isopleth map represents a specified value, for example, a 10^{o}C isopleth defines the position on the landscape at that temperature. Lines typically do not cross, in that there cannot be two different temperatures at the same location. However, isopleths often depict elevation, and cliffs or overhanging terrain do have multiple elevations at the same location. In this case the lower elevations typically pass "under" the higher elevations, and the isopleth is labeled with the tallest height (Figure 4-6). Note that the isopleths are typically estimated surfaces, with the lines drawn based on measurements at a set of point locations; various methods for estimating isopleth lines from points are described in chapter 12.

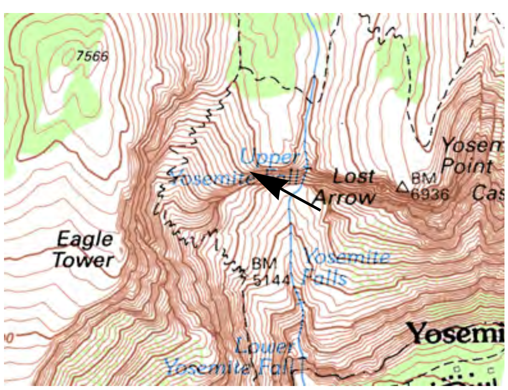


Figure 4-6: Lines on isopleth maps typically do not cross. However, as shown at the arrow in this image, lines may coincide when there is a common value. Here cliffs or overhangs result in converging isopleth lines.

Scale

All maps or digitally displayed data have a scale, a relationship between a distance on the map or screen and a corresponding distance projected on Earth. Map or screen scale is often reported as a distance conversion, such as one inch to a mile, meaning one inch on the map equals one mile on Earth. They may also be expressed as a unitless ratio, such as 1:24,000, indicating a unit distance on the display is equal to 24,000 units on Earth's surface. Digital maps most often use a third method to report scale, as a bar or line of known distance, labeled on the map (Figure 4-7).

Note that depicting map scale was unambiguous when only hardcopy maps were produced. A written ratio or conversion, for example, 1 inch to the mile, was true because the features were fixed on paper or other physical media. When displaying data on-screen with a known dimension, the scale may be re-computed at any display size, because the computer can measure the displayed and data distances, and update the displayed scale in real time. However, displaying a fixed scale may be erroneous on an electronic document such as a pdf without georeferencing, because the scale is altered by zooming. This changes the magnification on an electronic display, without the ability to automatically recompute scale. One inch as displayed may not correspond to a mile. Fixed-scale digital documents should most often include a graphic scale bar, depicting an equivalent surface distance, for example, 1 km, embedded in the map.

The notion of large vs. small scale is often confused because scale implies a ratio or fraction. A larger ratio signifies a large-scale map, so a 1:24,000 scale map is considered large-scale relative to a 1:100,000 scale map. Many people mistakenly refer to a 1:100,000 scale map as larger scale than a 1:24,000 scale map because it covers a larger area. A 1:100,000 scale map that is 50 cm (20 in) on a side covers more ground than a 1:24,000 scale map that is 50 cm on a side. However, it is the size of the ratio or frac-

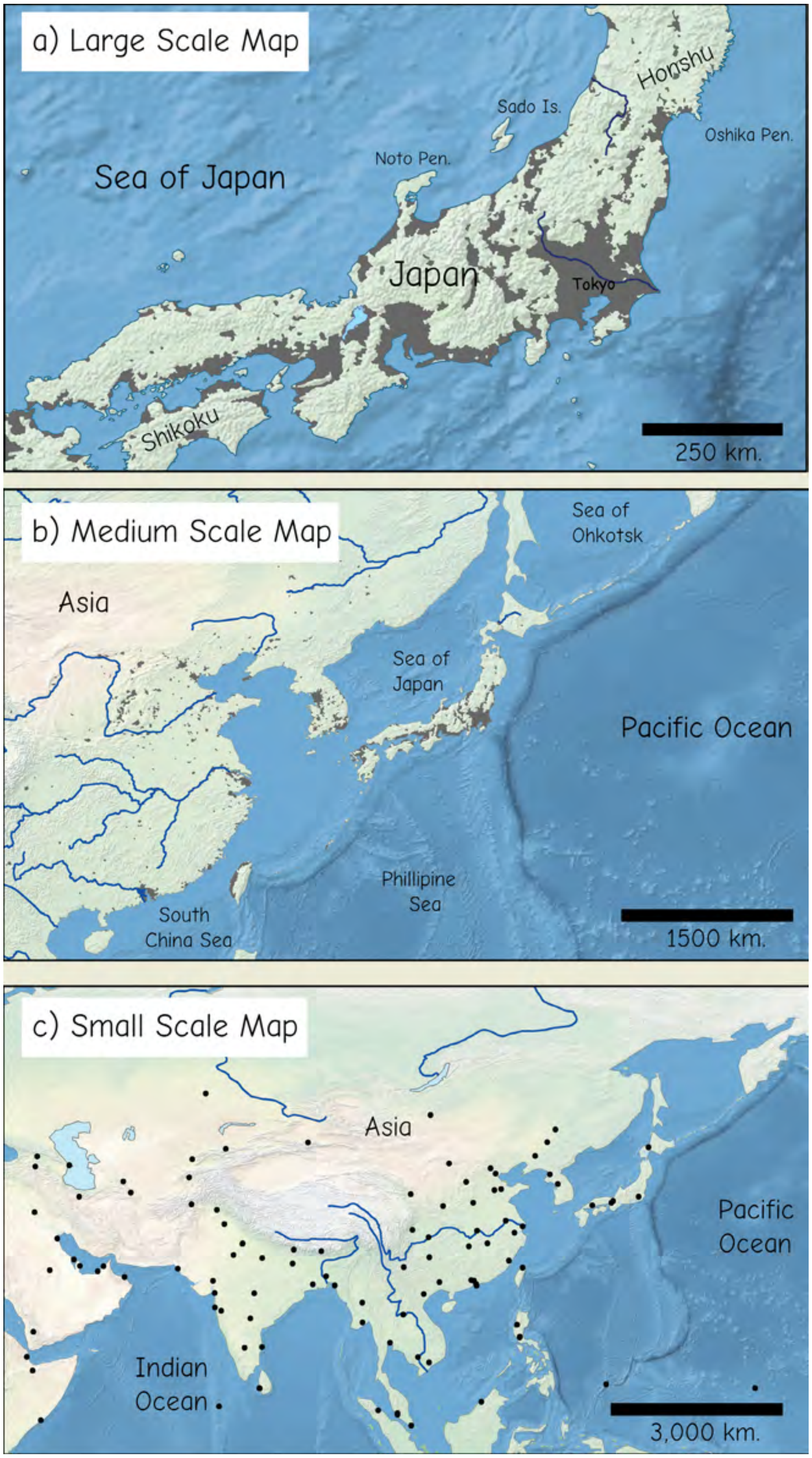


Figure 4-7: Coverage, relative distance, and detail change from larger scale (top) to smaller scale (bottom) maps.

tion, and not the area covered, that determines the map scale. It is helpful to remember that features appear larger on a large-scale map (Figure 4-7). It is also helpful to remember that large-scale maps of a given physical dimension often show more detail, but less area. Notice in Figure 4-7, the larger-scale map at the top shows details of Tokyo city. Tokyo shrinks in the successively smaller-scale maps, but large additional areas are covered. The larger the ratio (and smaller the denominator), the larger the map scale.

Because maps often report an average scale, and because there are upper limits on the accuracy with which data can be plotted on a map, large-scale maps generally have less geometric error than small-scale maps if the same methods were used to produce them. Small errors in measurement, plotting, printing, and paper deformation are magnified by the scale factor. These errors, which occur during map production, are magnified more on a small-scale map than a large-scale map.

Map and Data Generalization

Maps are abstractions of reality, as are spatial data in a GIS database. This abstraction introduces *map generalization*, the unavoidable approximation of real features when they are represented on a map. Not all the geometric or attribute detail of the physical world are recorded; only the most important characteristics are included. The set of features that are most important is subjectively defined and will differ among users. The mapmaker determines the set of features to place on the map, and selects the methods to collect and represent the shape and location of these features on the map.

The choice of data sources and digitizing methods will unavoidably set limits on the size and shape of features that may be represented. Consider mapping lakes, based on image data with a 250-meter cell size (Figure 4-8). The abstraction of the shoreline will not represent bays and peninsulas that are smaller than approximately 250 meters across, by conscious choice of the mapmakers. Small features will be missed, edge detail will be lost, and distances along boundaries will depend on the resolution of the source image.

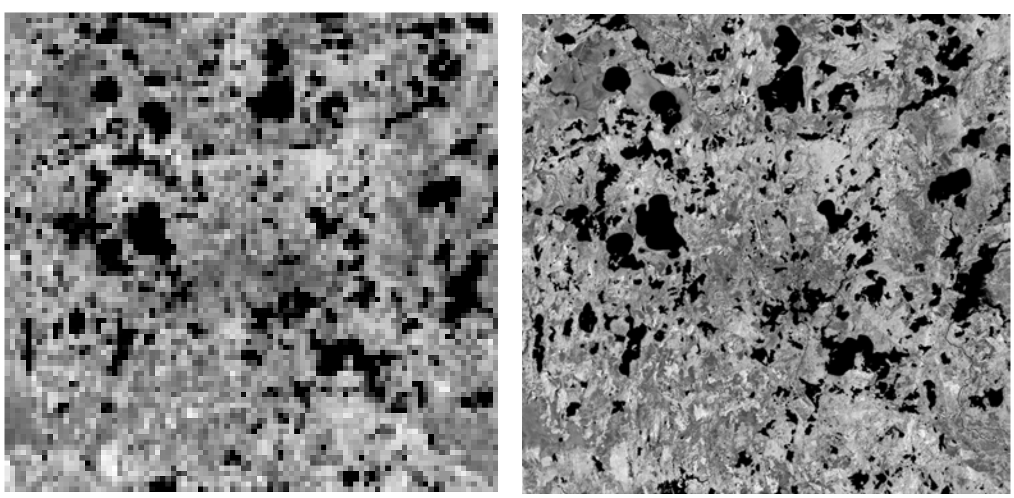


Figure 4-8: A mapmaker chooses the materials and methods used to produce a map, and so imposes a limit on spatial detail. Here, the choice of an input image with a 250 meter resolution (left) renders it impossible to represent all the details of the real lake boundaries (right). In this example, features smaller than approximately 250 meters on a side may not be faithfully represented on the map.

A finer resolution source, such as a 30-meter resolution, may more faithfully depict map detail, but may not be an appropriate choice. The finer resolution may be more expensive, difficult to reproduce, unavailable for the entire mapping area, or inappropriate because it does not show important features, for example, vegetation types or recent developments. Cartographers often must balance several factors in map design, and their choices inevitably lead to some form of map generalization.

Feature generalization is one common form of generalization. Feature generalization is a modification of features when representing them on a map. The geographic aspects of features are generalized because there are limits on the time, methods, or materials available when collecting geographic data. These limits also apply when compiling, displaying, or printing a map. These feature generalizations, depicted in Figure 4-9, may be classed as:

Fused: multiple features may be grouped to form a larger feature,

Simplified: boundary or shape details are lost or "rounded off",

Displaced: features may be offset to prevent overlap or to provide a standard distance between mapping symbols,

Omitted: Small features in a group may be excluded from the map, or

Exaggerated: standard symbol sizes are often chosen, for example, standard road symbol widths, which are much larger when scaled than the true road width.

Generalization is present at some level in almost every data layer or map, and should be recognized and evaluated for each data source in a GIS, including maps (Figure 4-10). If generalization results in omission or degradation of data beyond acceptable levels, then the analyst should switch to a larger-scale map if appropriate and available, or return to the field or original source materials to collect data at the required precision.

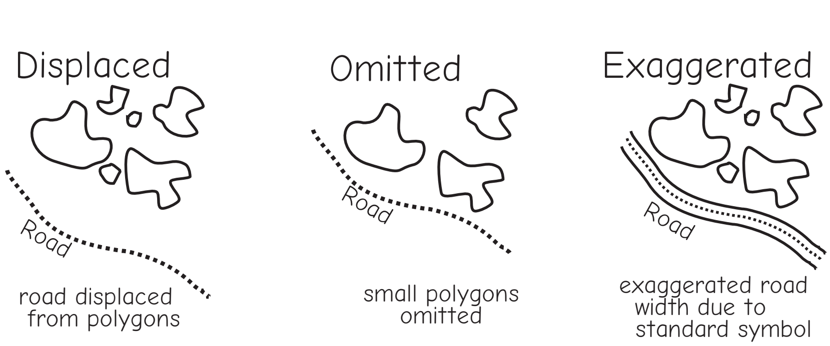


Figure 4-9: Generalizations common in maps and data layers.

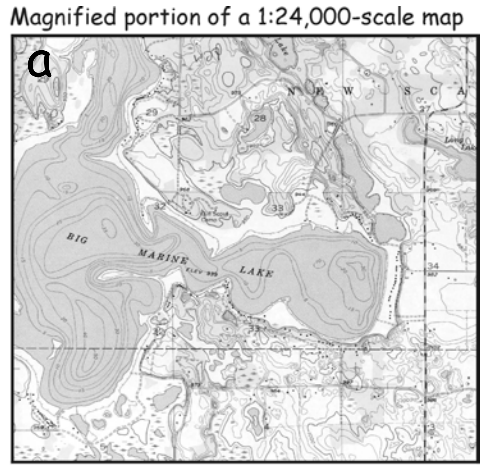
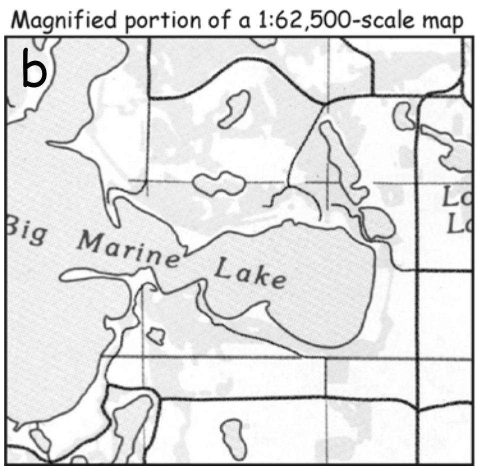
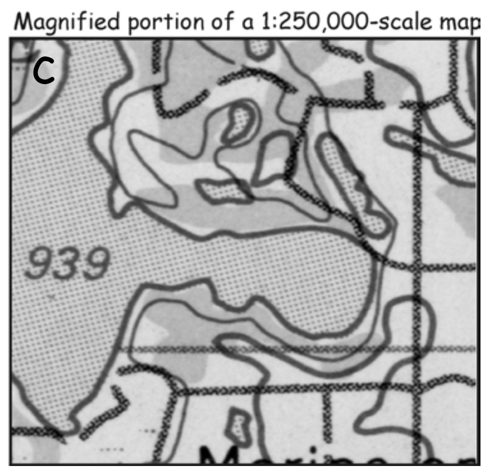


Figure 4-10: Examples of map generalization. Portions are shown for three maps for an area in central Minnesota. Excerpts from a large-scale (a, 1:24,000), intermediate-scale (b, 1:62,500), and small-scale (c, 1:250,000) map are shown. **Note that the maps are not drawn at true scale to facilitate comparison**. The smaller-scale maps (b and c) have been magnified more than a to better show the effects of generalization. Each map has a different level of map generalization. Generalizations increase with smaller-scale maps and include omissions of smaller lakes, successively greater road width exaggerations, and increasingly generalized shorelines as one moves from maps a through c.





Map Boundaries and Spatial Data

One final characteristic of maps affects their use as a source of spatial data: digital data are often created or stored in tiles, and hardcopy maps have edges. Discontinuities often occur at tile or map boundaries. These errors may be lacking in many newer, large-area data collected with digital methods, but edges will be encountered and should be understood.

Large-scale, high-quality maps and digital data sets often cover small areas or come broken into smaller tiles. This is because of the trade-off between data volumes, scale, and area coverage, and because of limits on the practical size of a map. High-density image or field-scanned data often tax the storage or display processing capabilities of common computers, and so data are broken into more manageable pieces for access and display. Hardcopy maps above a meter or so in size are expensive and difficult to print, store, or view. Thus, human ergonomics set a practical limit on the physical size of a map.

Because spatial data in a GIS often span large-scale maps or detailed data tiles, these map boundaries may be apparent in a spatial database. Problems often arise when adjacent maps or data layers are entered into a spatial database because features do not align or have mismatched attributes across map boundaries.

Differences in the time of data collection for adjacent map sheets may be a cause of inconsistencies across tile or map borders. Landscape change through time is a major source of differences across map boundaries. For example, scanned and digital aerial photographs are collected for most of the agricultural lands of the U.S. on an annual basis. Data are collected in blocks, over the growing season, and because of weather, schedules, and equipment failures, several days to weeks may pass between data collection on adjacent blocks. Many features, such as crop type, stage of development, wetland size, or harvest state may be discontinuous or inconsistent across blocks, and hence the data layer.

Different interpreters may also cause differences across collection boundaries. Large-area mapping projects typically employ several interpreters, each working on different areas of a region. All professional, large-area mapping efforts should have protocols specifying the scale, sources, equipment, methods, classification, keys, and cross-correlation to ensure consistent mapping across map sheet boundaries. In spite of these efforts, some differences due to human interpretation occur. Feature placement, category assignment, and generalization vary among interpreters. These problems are compounded when extensive checking and guidelines are not enforced across map sheet boundaries, especially when adjacent areas are mapped at different times or by two different organizations.

Finally, differences in coordinate registration can lead to spatial mismatch across map sheets. *Registration* or *coordinate transformation*, discussed later in this chapter, is the conversion of digitizer or other coordinate data to an Earth-surface coordinate system. These registrations contain unavoidable errors that translate into spatial uncertainty. There may be mismatches when data from two separate registrations are joined along the edge of a data layer or map.

Spatial data stored in a GIS are not bound by the same constraints that limit the physical dimensions of hardcopy maps. Digital storage enables the production of seamless digital maps of large areas. However, the inconsistencies that exist in data collection or on hardcopy maps may be transferred to the digital data. Inconsistencies at tile or map sheet edges need to be identified and resolved in digital formats.

Digitizing: Coordinate Capture

Digitizing is the process by which coordinates from a map, image, or other sources are converted into a vector data layer in a GIS. The point, line, and area coordinate values that define the locations and shapes of entities must be captured, that is, recorded as numbers and structured in the spatial database. There is a wealth of spatial data in existing maps and photographs, and new imagery and maps add to this source of information on a nearly continuous basis.

Manual digitization is human-guided coordinate capture from a map or image source. The operator guides an electronic device over a map or image and signals the capture of important coordinates, often by pressing a button on the digitizing device. Important point, line, or area features are traced on the source materials, and the coordinates are recorded in GIS-compatible formats. Valuable data on historical maps may be converted to digital forms through the use of manual digitizing. On-screen digitizing and hardcopy digitizing are the two most common forms of manual digitization.

On-Screen Digitizing

On-screen digitizing, also known as heads-up digitizing, involves manually digitizing on a computer screen, using a digital image as a backdrop. Most image data produced for spatial data entry since 2010 is delivered with a coordinate system defined for the images, so that data extracted from the images are specified in that coordinate system. Digitizing software allows the operator to trace the points, lines, or polygons that are identified on the image (Figure 4-11) and saves the coordinates and added attribute data into spatial data layers. Digitizing software allows the human operator to specify the type of feature to be recorded, the extent and magnification of the image on screen, the mode of digitizing, and other options to control how data are input. The operator typically guides a cursor over points to be recorded using a mouse, and depresses a button or sequence of buttons to collect the point coordinates. On-screen digitizing can be used for recording information from digital photographs, satellite images, scanned maps, or other image sources.

Hardcopy Map Digitization

Hardcopy digitizing is human-guided coordinate capture from a paper, plastic, or other hardcopy map. An operator securely attaches a map to a digitizing surface and traces lines or points with an electrically sensitized puck (Figure 4-12).The most common digitizers are based on a wire grid embedded in or under a table. Depressing a button specifies the puck location relative to the digitizer coordinate system. Digitizing tables can be quite accurate, with a resolution of between 0.25 and 0.025 mm (0.01 and 0.001 in).

While once a major method for capturing spatial data, hardcopy map digitizing is diminishing in importance as most paper documents have been converted to digital forms. The tables are large, somewhat expensive, and now little used. However, because data from hardcopy sources are

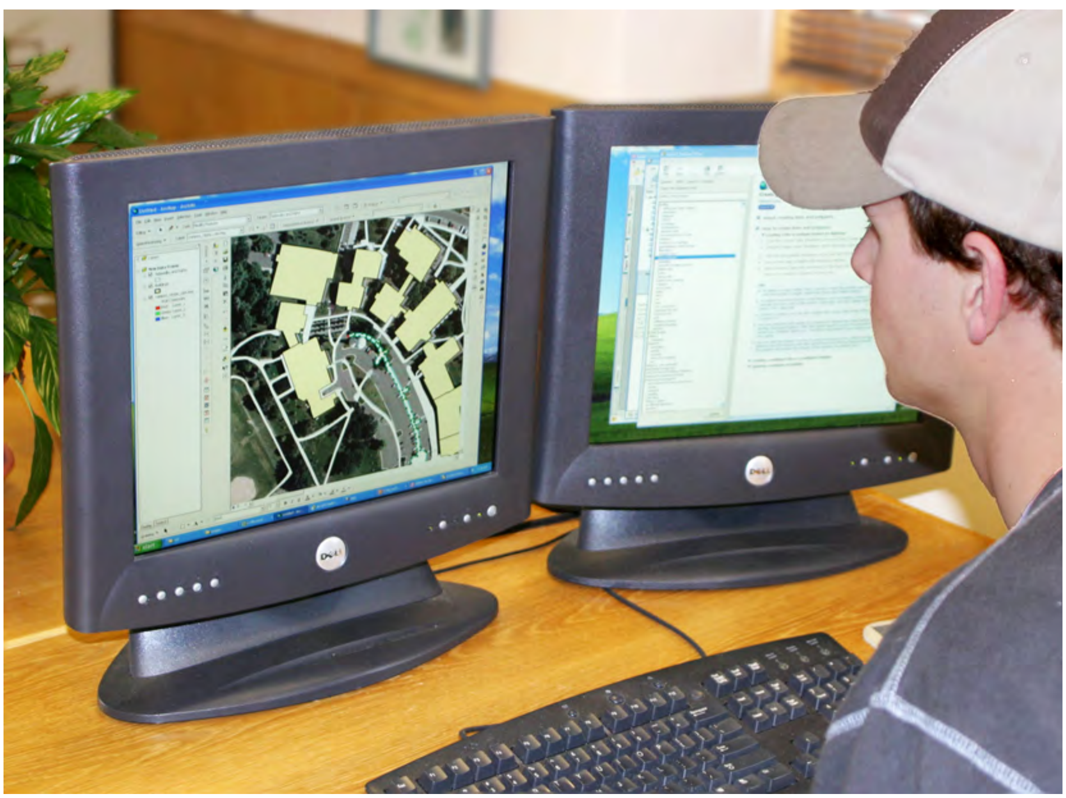


Figure 4-11: An example of on-screen digitizing. Images or maps are displayed on a computer screen and feature data digitized manually. Buildings, roads, or any other features that may be distinguished on the image may be digitized.

likely to persist for many decades, and there are still documents to convert, you should be familiar with the process.

Not all maps are appropriate as a source of information for GIS. The type of map, how it was produced, and the intended purpose must be considered when interpreting the information on maps. Only cartometric maps should be directly digitized, and even though cartometric, a map may not be suitable.

Characteristics of Manual Digitizing

Manual digitizing from digital images or hardcopy sources is common because it provides sufficiently accurate data for most applications. Proper digitizing adds little error to that already in the source materials, requires inexpensive equipment, and is often the best way to provide the needed information. The human ability to interpret images still outpaces that of machines, although artificial intelligence is improving. Humans are usually much better than machines at interpreting the information contained on faded, stained, or poor quality maps and images. Finally, manual digitizing is often best because short training periods are required and data quality may be frequently evaluated. For these reasons, manual digitization is likely to remain an important data entry method for some time to come.

Many factors may affect accuracy while digitizing. As noted earlier, map or image scale and resolution impacts the spatial accuracy of digitized data. This scale may be the production scale for hardcopy maps, or the display scale for digital images or scanned maps. Table 4-1 illustrates the effects of display or scale on data quality. Errors of 1 millimeter (0.039 in) on a 1:24,000 scale map or screen correspond to 24 meters (79 ft) on the surface of the Earth. This same 1 millimeter error on a 1:1,000,000 scale map corresponds to 1,000 meters (3,281 ft) on the Earth's surface. Thus, small errors in image data collection or map production or inter-

Figure 4-12: Manual digitizing on a digitizing table.

pretation may cause significant positional errors when scaled to distances on Earth, and these errors are greater for smaller-scale source materials. Errors due to human pointing ability are reduced for on-screen digitizing, because the operator can zoom in to larger scales as needed. However, this does not overcome errors inherent in original images or scanned documents.

The image source and display scale should be considered when on-screen digitizing, and device precision and map scales should be considered when selecting a hard-copy digitizing. Map scale and repeatability both set an upper limit on the positional quality of digitized data. The most precise digitizers may be required when attempting to meet a stringent error standard while digitizing small-scale maps.

The abilities and attitude of the person digitizing affects the geometric quality of manually digitized data. Operators differ in visual acuity, hand steadiness, attention to detail, and ability to concentrate. The abilities of any single operator will also vary through time due to fatigue or difficulty maintaining focus on a repetitive task. Operators should take frequent breaks from digitizing, and comparisons among operators and quality and consistency checks should be integrated into any manual digitization process to ensure accurate and consistent data collection.

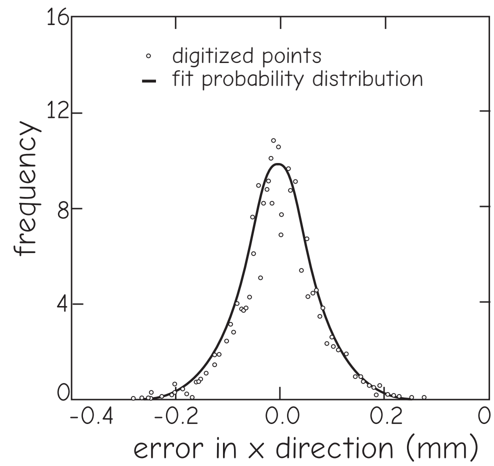


Figure 4-13: Digitizing error, defined by repeat digitizing. Points repeatedly digitized cluster around the true location and follow a normal probability distribution (from Bolstad et al., 1990).

Table 4-1: The surface error caused by a 1 millimeter (0.039 in) digitizing or map error will change as scale changes. Note the larger error at smaller scales.

Map Scale	Error (m)	Error (ft)
1:24,000	24	79
1:50,000	50	164
1:62,500	63	205
1:100,000	100	328
1:250,000	250	820
1:1,000,000	1,000	3,281

The combined errors from both operators and equipment are usually quite small. One test using a high-precision digitizing table revealed digitizing errors averaging approximately 0.067 mm (Figure 4-13). Errors followed a random normal distribution, and varied significantly among operators. These average errors translated to an approximately 1.6-meter error when scaled from the 1:24,000 scale map to a ground-equivalent distance. This average error is less than the acceptable production error for the map, and is suitable for many spatial analyses.

The Digitizing Process

Manual digitizing involves displaying a digital image on screen or placing a map on a digitizing surface, and tracing the location of feature boundaries. Coordinate data are sampled by manually positioning the puck or cursor over each target point and collecting coordinate locations. This position/collect step is repeated for every point to be captured, and all features of interest recorded.

Lines have a *starting node*, then a set of *vertices* defining the line shape, and an *ending node* (Figure 4-14). Hence, lines may be viewed as a series of straight line segments connecting vertices and nodes. Polygons are connected lines that enclose areas.

Digitizing may be in *point mode*, where the operator must depress a button or otherwise signal to the computer to sample each point, or in *stream mode*, where points are automatically sampled at a fixed time or distance frequency, perhaps once each meter. Stream mode helps when large numbers of lines are digitized, because vertices may be sampled more quickly and the operator may become less fatigued.

The stream sampling rate must be specified with care to avoid over- or under-sampled lines. Too short a collection interval results in redundant points. Too long a collection interval may cause the loss of important spatial detail. When using time-triggered stream digitizing, the operator must continuously move the digitizing puck; pausing for a period longer than the sampling interval digitizes multiple points clustered together. These will redundantly represent a line portion, often with overlapping segments. Pausing for extended periods often creates a "rat's nest" of lines that must later be removed.

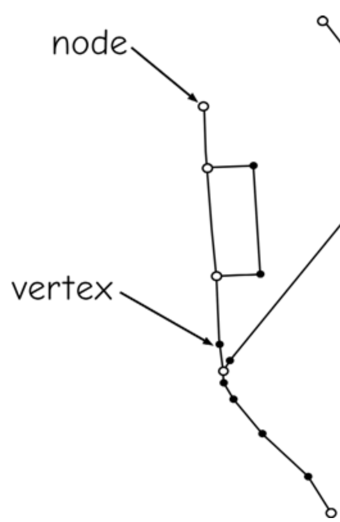


Figure 4-14: Nodes define the starting and ending points of lines; vertices define line shape.

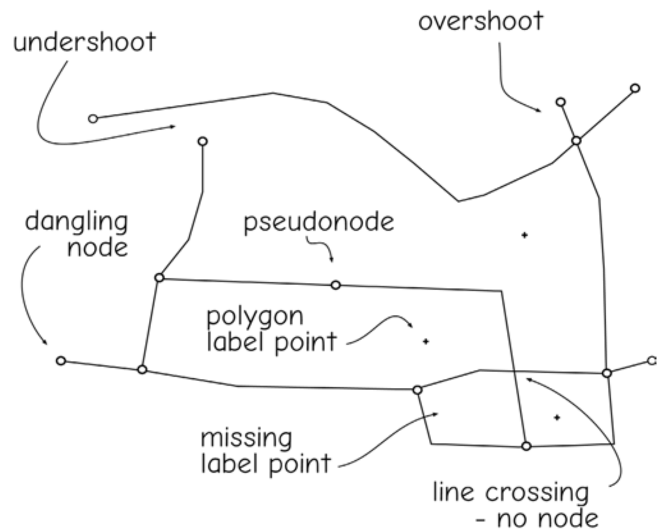
Minimum distance digitizing is a variant of stream mode digitizing that avoids some of the problems inherent with time-sampled streaming. In minimum distance digitizing, a new point is not recorded unless it is more than some minimum threshold distance from the previously digitized point. The operator may pause without creating a rat's nest of line segments. The threshold must be chosen carefully – neither too large, missing useful detail, nor too small, in effect reverting back to stream digitizing.

Digitizing Errors, Node and Line Snapping

Positional errors are inevitable when data are manually digitized. These errors may be "small" relative to the intended use of the data; for example, the positional errors may be less than 2 meters when only 5-meter accuracy is required. However, these relatively small errors may still prevent the generation of correct networks or polygons. For example, a data layer representing a river system may not be correct because major tributaries may not connect. Polygon features may not be correctly defined because their boundaries may not completely close. These small errors must be removed or avoided during digitizing. Figure 4-15 shows some common digitizing errors.

Undershoots and *overshoots* are common errors that occur when digitizing. Undershoots are nodes that do not quite reach the line or another node, and overshoots are lines that cross over existing nodes or lines (Figure 4-15). Undershoots

Figure 4-15: Common digitizing errors.



cause unconnected networks and unclosed polygons. Overshoots typically do not cause problems when defining polygons, but they may cause difficulties when defining and analyzing line networks.

Node snapping and *line snapping* are used to reduce undershoots and overshoots while digitizing. Snapping is a process of automatically setting nearby points to have the same coordinates. Snapping relies on a *snap tolerance* or *snap distance*. This distance may be interpreted as a minimum distance between features. Nodes or vertices closer than this distance are moved to occupy the same location (Figure 4-16). While digitizing, an existing node or vertex becomes "magnetic," and pulls a new node or vertex to it within the snap distance. Node snapping prevents a new node from being placed within the snap distance of an already existing node. Remember that nodes are used to define the ending points of a line. By snapping two nodes together, we ensure a connection between digitized lines.

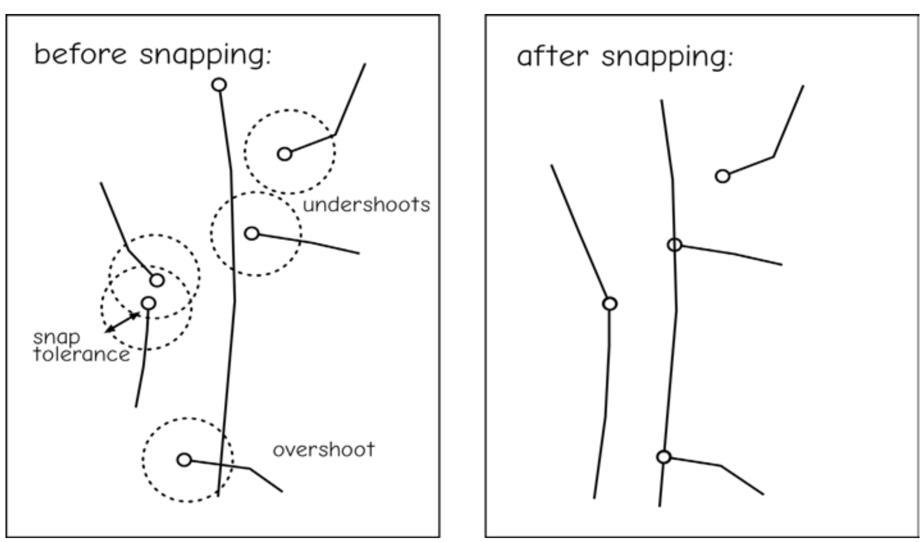


Figure 4-16: Undershoots, overshoots, and snapping. Snapping may join nodes, or may place a node onto a nearby line segment. Snapping does not occur if the nodes and/or lines are separated by more than the snap tolerance.

Line snapping, sometimes called edge snapping, may also be specified. Line snapping inserts a node at a line crossing and clips the end when a small overshoot is digitized. This forces a node to connect to a nearby line while digitizing, but only when the undershoot or overshoot is less than the snapping distance. Line snapping requires the calculation of an intersection point on an already existing line. A snap places a new node at the intersection point, and connects the digitized line to the existing line. This splits the existing line into two new lines. When used properly, line and node snapping reduce the number of undershoots and overshoots. Closed polygons or intersecting lines are easier to digitize accurately and efficiently under node and line snapping.

The snap distance must be carefully selected for snapping. If the snap distance is too short, then snapping has little impact. Consider a configuration where digitized data has better than 5 meter accuracy only 10% of the time. This means 90% of the digitized points will be more than 5-meters from the intended location. If the snap tolerance is set to the equivalent of 0.1 meters, then few points will be snapped. Conversely, we sacrifice accuracy if the snap tolerance is set too large. In our previous example, a snap tolerance of 10 m is higher than our 5 m target, and we may lose significant spatial information contained in the data source. Lines less than 10 meters apart cannot be digitized as separate objects. The snap distance should be smaller than the desired positional accuracy, such that significant detail contained in the digitized map is recorded. It is also important that the snap distance is not below the capabilities of the system used for digitizing. Careful selection of the snap distance may reduce digitizing errors and significantly reduce time required for later editing.

Reshaping: Line Smoothing and Thinning

Digitizing software may provide tools to smooth, densify, or thin points while entering data. One common technique uses *spline* functions to smoothly interpolate curves between digitized points, and thereby both smooth and densify the set of vertices used to represent a line. A spline is a set of polynomial functions that join smoothly (Figure 4-17). Polynomial functions are fit to successive sets of points along the vertices in a line; for example, a function may be fit to points 1 through 5, and a separate polynomial function fit to points 5 through 11 (Figure 4-17). Constraints force these functions to connect smoothly, usually by requiring the first and second derivatives of the functions to be continuous at the intersection point. This means the lines have the same slope at the intersection point, and the slope is changing at the same rate for both lines at the intersection point. Once the spline functions are calculated, they may be used to add vertices. For example, several new vertices may be automatically placed on the line

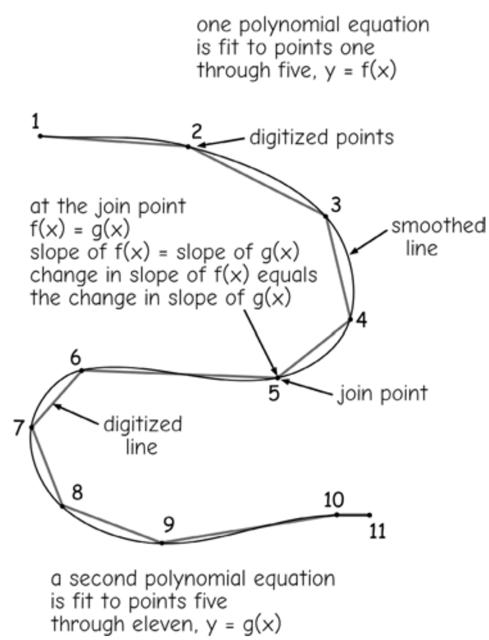


Figure 4-17: Spline interpolation to smooth digitized lines.

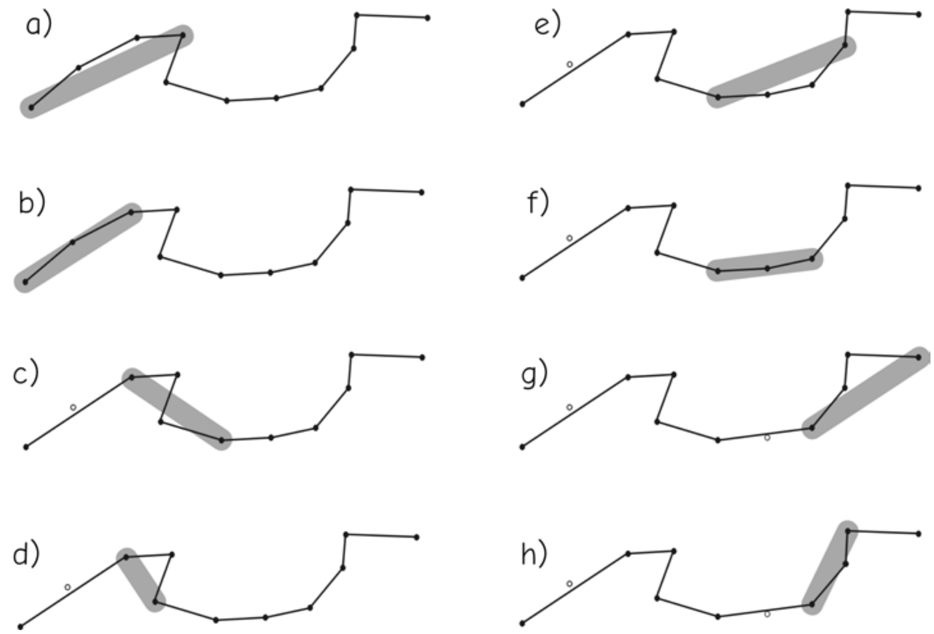


Figure 4-18:The Lang algorithm is a common line-thinning method. In the Lang method, vertices are removed, or thinned, when they are within a weed distance to a spanning line (adapted from Weibel, 1997).

between digitized vertices 8 and 9, leading to the "smooth" curve shown in Figure 4-17.

Data may also be digitized with too many vertices. High densities may occur when data are manually digitized in stream mode, and the operator moves slowly relative to the time interval. High vertex densities may also be found when data are derived from spline or smoothing functions that specify too high a point density. Finally, automated scanning and then raster-to-vector conversion may result in coordinate pairs spaced at absurdly high densities. Many of these coordinate data are redundant and may be removed without sacrificing spatial accuracy. Too many vertices may be a problem in that they slow processing, although this has become less important as computing power has increased. Point thinning algorithms have been developed to reduce the number of points while maintaining the line shape.

Many point thinning methods use a perpendicular "weed" distance, measured from a spanning line, to identify redundant points (Figure 4-18, top). The Lang method exemplifies this approach. A spanning line connects two nonadjacent vertices in a line. A predetermined number of vertices is spanned initially. The initial spanning number has been set to 4 in Figure 4-18, meaning four points will be considered at each starting point. Areas closer than the weed distance are shown in gray in the figure. A straight line is drawn between a starting point and the endpoint that is the fourth point down the line (Figure 4-18a). Any intermediate points that are closer than the weed distance are marked for removal. In Figure 4-18a, no points are within the weed distance, therefore, none are marked. The endpoint is then moved to the next closest remaining point (Figure 4-18b), and all intermediate points tested for removal. Again, any points closer than the weed distance are marked for removal, shown as open circles. Note that in Figure 4-18b, one point is within the weed distance, and is removed. Once all points in the initial spanning distance are checked, the last remaining endpoint becomes the new starting point, and a new spanning line is drawn to connect 4 points (Figure 4-18c, d).

The process may be repeated for successive sets of points in a line segment until all vertices have been evaluated (Figure 4-18e to h). All close vertices are viewed as not recording a significant change in the line shape, and hence are expendable. Increasing the weed distance thins more vertices, and at some upper weed distance, too many vertices may be removed. A balance must be struck between the removal of redundant vertices and the loss of shape-defining points, usually through a careful set of test cases with successively larger weed distances.

There are many variants on this basic concept. Some look only at three immediately adjacent points, testing the middle point against the line spanned by its two neighboring points. Others constrain or expand the search based on the complexity of the line. Rather than always looking at four points, as in our example above, more points are scrutinized when the line is not complex (nearly straight), and fewer when the line is complex (many changes in direction).

Scan Digitizing

Optical scanning is another method for converting hardcopy documents into digital formats. Scanners have elements that emit and sense light. Most scanners pass a sensing element over an illuminated map. This device measures both the precise location of the point being sensed and the strength of the light reflected or transmitted from that point. Reflected light intensities are converted to numbers, producing a raster representation of the map. Values are recorded where points or lines exist on the map, and null or zero values are recorded in the intervening spaces.

Scan digitization usually requires some form of *skeletonizing*, or line thinning, particularly if the data are to be converted to a vector data format. Scanned lines are often wider than a single pixel (Figure 4-19). One of several pixels may be selected to specify the position of a given portion of the line. The same holds true for points. A pixel near the "center" of the point or line is typically chosen, with the center of a line defined as the pixel nearest the center of the local perpendicular bisector of the line. Skeletonizing reduces the widths of lines or points to a single pixel.

Editing Geographic Data

Spatial data may be edited, or changed, for several reasons. Errors and inconsistencies are inevitably introduced during spatial data entry. Undershoots, overshoots, missing or extra lines, and missing or extra points or labels are all errors that must be corrected. Spatial data can change over time. Parcels are subdivided, roads extended or moved, forests grow or are cut, and these changes may be entered in the spatial database through editing. New technologies may be

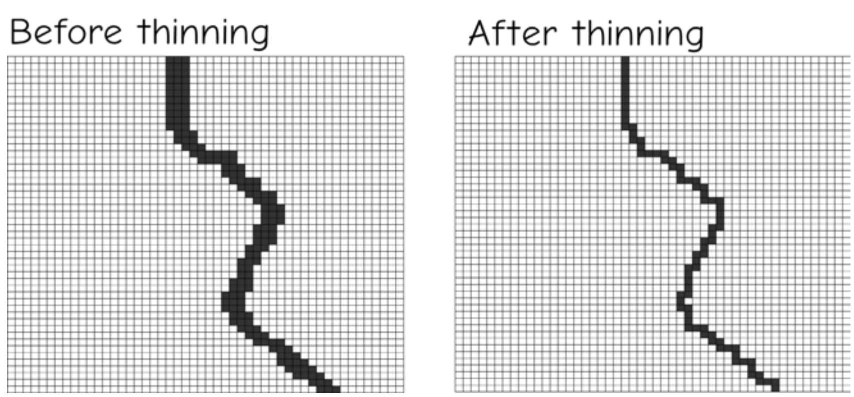


Figure 4-19: Skeletonizing, a form of line thinning that is often applied after scan digitizing.

developed that provide more accurate positional information, and even though existing data may be consistent and current, the more accurate data may be more useful, leading to data editing.

Software helps operators identify potential errors. Line nodes may be classified as connecting or dangling. A connecting node joins two or more lines, while a dangling node is attached to only one line. Some dangling nodes may be intentional, for example, a cul-de-sac in a street network, while others will be the result of under- or overshoots. Dangling nodes can be quickly evaluated and, if appropriate, corrected.

Attribute consistency may also be used to identify errors. Operators note areas in which contradictory theme types occur in different data layers. The two layers are either graphically or cartographically overlain. Contradictory co-occurrences are identified, such as water in one layer and upland areas in a second. These contradictions are then either resolved manually, or automatically via some predefined precedence hierarchy.

Many GIS software packages provide a comprehensive set of editing tools (Figure 4-20). Editing typically includes the ability to select, split, update, and add features. Selection may be based on geometric attributes, or with a cursor guided by the operator. Selections may be made individually, by geographic extent (select all features in a box, circle, or within a certain distance of the pointer), or by geometric attributes (e.g., select all nodes that connect to only one line). Once a feature is selected, various operations may be available, including eras-

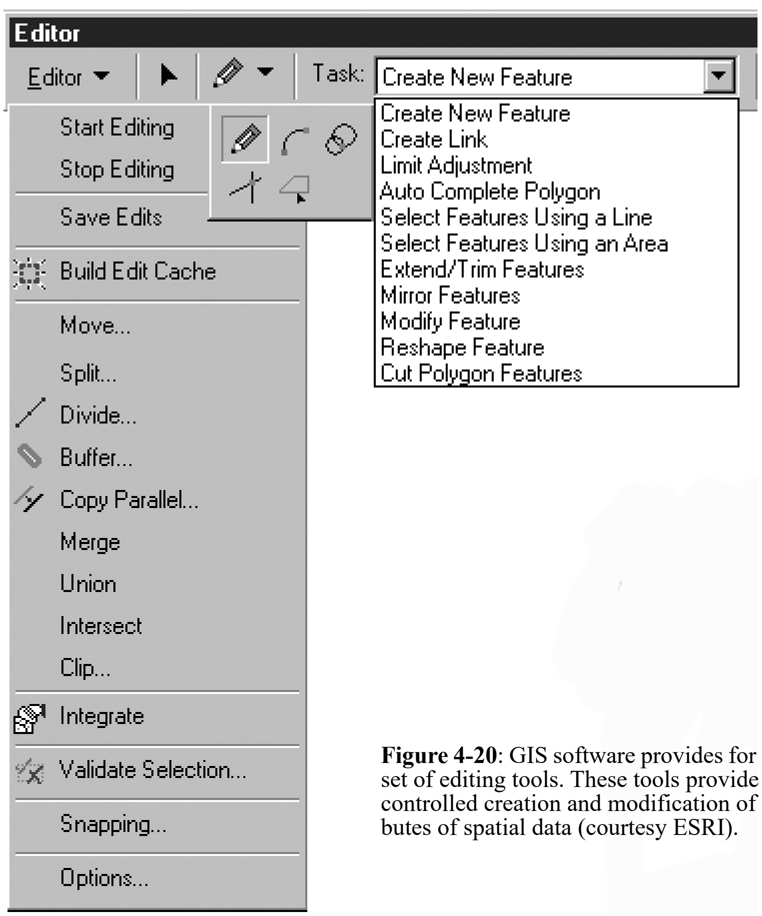


Figure 4-20: GIS software provides for a flexible and complete set of editing tools. These tools provide for the rapid, precise, controlled creation and modification of coordinates and attributes of spatial data (courtesy ESRI).

ing all or part of the feature, changing the coordinate values defining the feature, and in the case of lines, splitting or adding to the feature. A line may be split into parts, either to isolate a segment for future deletion, or to modify only a portion of the line. Coordinates are typically altered by interactively selecting and dragging points, nodes, or vertices to their best shape and location. Points or line segments are added as needed.

Groups of features in an area may be adjusted through interactive *rubbersheeting*. Rubbersheeting involves fitting a local equation to adjust the coordinates of features. Polynomial equations are often used due to their flexibility and ease of application. Anchor points are selected, again on the graphics screen, and other points are selected by dragging interactively on the screen to match point locations. All lines and points except the anchor points are interactively adjusted.

All edits should be made with due attention to distance shifts during editing. On-screen editing to eliminate undershoots should only be performed when the "true" locations of features may be identified accurately, and the new features can be confidently placed in the correct location. Automatic removal of "short" undershoots may be performed without introducing additional spatial error in most instances. A short distance for an undershoot is subjectively defined, but typically it is below the error inherent in the source map, or at least a distance that is insignificant when considering the intended use of the spatial data.

Features Common to Several Layers

One common problem in digitizing derives from representation of features that occur on different maps or images. These features rarely have identical locations on each map or image, and often occur in different locations when digitized into their respective data layers (Figure 4-21). For example, water boundaries on soil survey maps rarely correspond exactly to water boundaries found on USGS topographic maps.

Features may appear differently on different maps for many reasons. Perhaps the maps were made for different purposes or at different times. Features may differ because the maps were from different source materials; for example, one map may have been based on ground surveys while another was based on aerial photographs. Digitizing can also compound the problem due to differences in digitizing methods or operators.

There are several ways to remove this "common feature" inconsistency. One involves removing inconsistencies while redrafting the data from conflicting sources onto a new base map. Redrafting is labor intensive and time consuming, but forces a resolution of inconsistent boundary locations. Redrafting also allows several maps to be combined into a single data layer.

A second, often preferable method involves establishing a "master" boundary that is the highest accuracy composite of the available data sets. A digital copy or overlay operation establishes the common features as a base in all the data layers, and this base may be used as each new layer is produced. For example, water boundaries might be extracted from the soil survey and USGS quadrangle maps, and these data combined in a third data layer. The third data layer would be edited to produce a composite, high-quality water layer. The composite water layer would then be copied back into both the soils and USGS quad layers. This second approach, while resulting in visually consistent spatial data layers, is in many

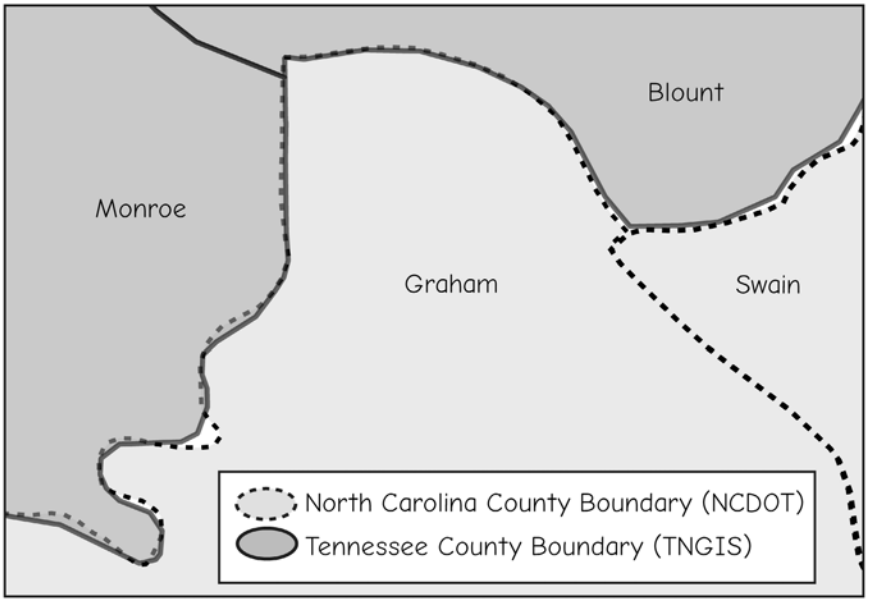


Figure 4-21: Common features may be spatially inconsistent in different spatial data layers. These data were the highest resolution data available from the respective official state distribution nodes. Note the boundaries are inconsistent in several locations, creating gaps and overlaps along their shared margin. These gaps must be addressed if the data sets are to be combined in an analysis.

instances only a cosmetic improvement of the data. If there are large discrepancies ("large" is defined relative to the required spatial data accuracy), then the source of the discrepancies should be identified and the most accurate data used, or new, higher-accuracy data collected from the field or original sources.

Coordinate Transformation

Coordinate transformation is a common operation in the development of spatial data for GIS. A coordinate transformation brings spatial data into an Earth-based map coordinate system so that each data layer aligns with every other data layer. This alignment ensures features fall in their proper relative position when digital data from different layers are combined. Within the limits of data accuracy, a good transformation helps avoid inconsistent spatial relationships such as farm fields on freeways, roads under water, or cities in the middle of swamps, except where these truly exist. Coordinate transformation is also referred to as *registration*, because it "registers" the layers to a map coordinate system.

Coordinate transformation is most commonly used to convert newly digitized data from the digitizer/scanner coordinate system to a standard map coordinate system (Figure 4-22). The input coordinate system is usually based on the digitizer or scanner-assigned values. An image may be scanned and coordinates recorded as a cursor is moved across the image surface. These coordinates are usually recorded in pixel, inch, or centimeter units relative to an origin located near the lower left corner of the image. The absolute values of the coordinates depend on where the image happened to be placed on the table prior to scanning, but the relative position of digitized points does not change. Before these newly digitized data may be used with other data, these "inch-space" or "digitizer" coordinates must be transformed into an Earth-based map coordinate system.

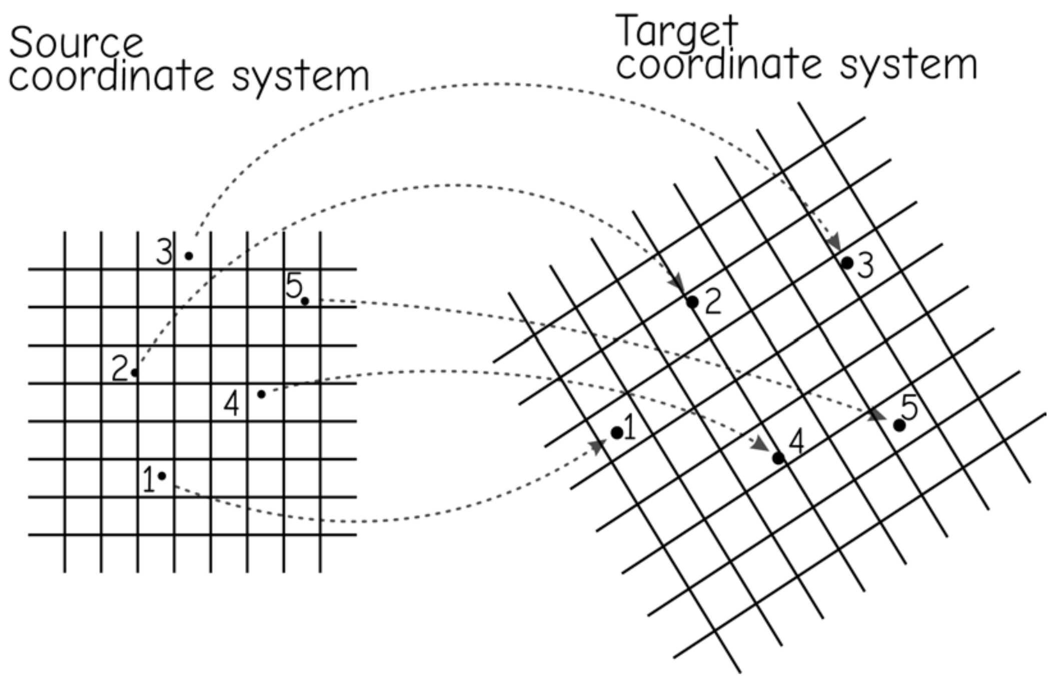


Figure 4-22: Control points in a coordinate transformation. Control points are used to guide the transformation of a source, input set of coordinates to a target, output set of coordinates. There are five control points in this example. Corresponding positions are shown in both coordinate systems.

Figure 4-23 depicts the application of a coordinate transformation in data development. Early surveys were often stored on paper maps; in this instance, the original PLSS surveys. These depict the original PLSS boundary lines, as well as lakes and wetlands, and in some instances forests, grasslands, or other features. We may wish to compare past and current conditions, but the PLSS is a system of land division, with no coordinates or projection associated with the lines. We need to register the paper maps to a projected coordinate system prior to use.

The PLSS line intersections may be used to register the original maps to a projected coordinate system. As noted in Chapter 3, PLSS lines often became property boundaries, and subsequent roads often followed these boundaries. A road between properties shared the imposition among property owners, and served multiple adjacent properties. Road intersections often occur at PLSS corners, particularly in the flatter terrain over much of the eastern and central United States. Section line intersections may be surveyed directly, or extracted

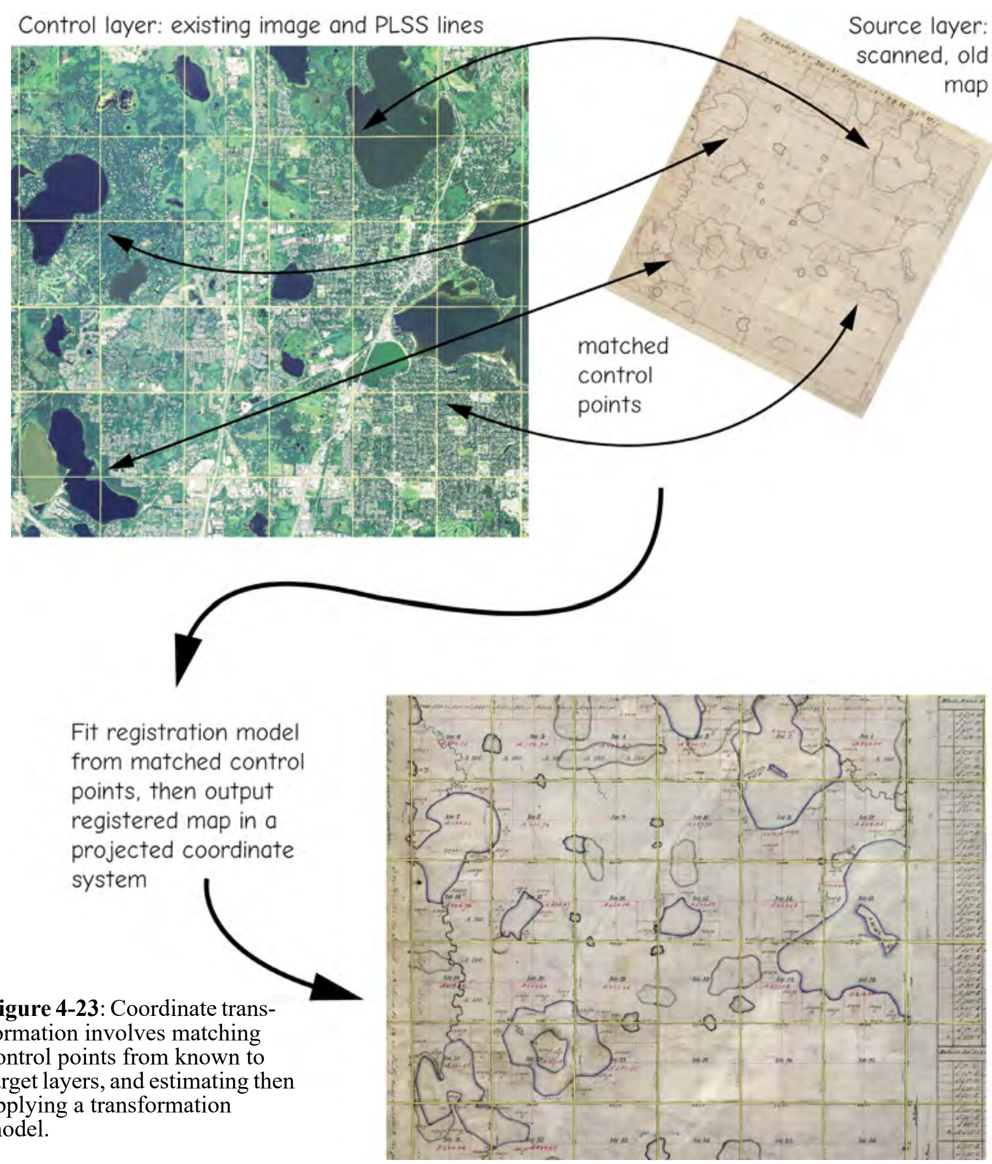


Figure 4-23: Coordinate transformation involves matching control points from known to target layers, and estimating then applying a transformation model.

from other registered data such as aerial images, and hence used to transform the original maps to a projected coordinate system. Features on the original maps, such as lakes or wetlands depicted at the time of the original surveys during the 1800s, may be digitized and compared to current feature locations.

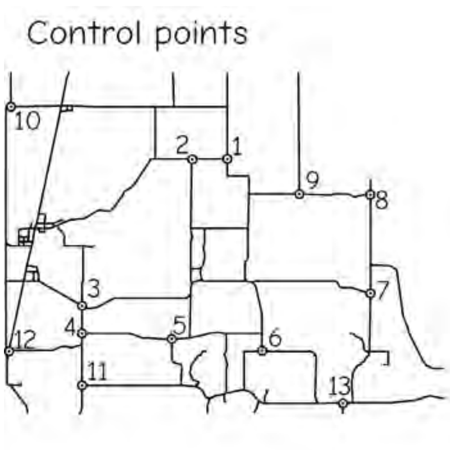
Control Points

A set of *control points* is used to transform the digitized data from the digitizer or photo coordinate system to a map-projected coordinate system. These control points are used to estimate equations that we use for the coordinate transformation (Figure 4-24). Control points are different from other digitized features. When we digitize most points, lines, or areas, we do not know the map projection coordinates for these features. We simply collect the digitizer X and Y coordinates that are established with reference to some arbitrary origin on an image or digitizing table. Control points differ from other digitized points in that we know both the map projection coordinates and the digitizer coordinates for these points.

These two sets of coordinates for each control point, one for the map projection and one for the digitizer system, are used to estimate the coefficients for transformation equations, usually through a statistical, least squares process. The transformation equations are then used to convert coordinates from the digitizer system to the map projection system.

The transformation may be estimated in the initial digitizing steps, and applied as the coordinates are digitized from the map or image. This "on-the-fly" transformation allows data to be output and analyzed with reference to map-projected coordinates. A previously registered data layer or image may be displayed on screen just prior to digitizing a new map. Control points may then be entered, the new map attached to the digitizing table, and the map registered. The new data may then be displayed on top of the previously registered data. This allows a quick check on the location of the newly digitized objects against corresponding objects in the study area.

In contrast to on-the-fly transformations, data can also be recorded in digitizer coordinates, and the transformation applied later. All data are digitized, including the control point locations. The digitizer coordinates of the control point may then be matched to corresponding map projection coordinates, and transformation equations estimated. These transformation equations are then



Digitizer coordinates			Projection coordinates (UTM)	
ID	X	Y	E	N
1	103.0	-100.1	500,083.4	5,003,683.5
2	0.8	-69.1	504,092.3	5,002,499.5
3	-20.0	-69.0	504,907.5	5,002,499.5
4	-60.0	-47.0	506,493.3	5,001,673.5
5	-102.0	-47.2	508,101.3	5,001,651.7
6	-101.7	10.8	508,090.1	4,999,384.0
7	-86.0	75.8	507,475.9	4,996,849.0
8	-40.0	45.7	505,689.2	4,998,022.0
9	11.0	36.8	503,679.2	4,998,368.0
10	63.0	34.0	501,657.9	4,998,479.5
11	63.0	17.7	501,669.1	4,999,116.0
12	63.0	64.3	501,680.3	4,997,296.0
13	106.0	47.7	500,005.3	4,997,943.5

Figure 4-24: An example of control point locations from a road data layer, and corresponding digitizer and map projection coordinates.

applied to convert all digitized data to map projection coordinates.

Control points should meet several criteria. First, they should be from a source with the highest feasible accuracy. Second, control point should be more accurate than the desired overall positional accuracy for the spatial data. Third, control points should be evenly distributed throughout the data area. A sufficient number of control points should be collected, above the minimum to improve the statistically fit transformation functions.

The x, y (horizontal), and sometimes z (vertical or elevation) coordinates of control points are known to a high degree of accuracy and precision. Because high precision and accuracy are subjectively defined, there are many methods to determine control point locations. Subcentimeter accuracy may be required for control points used in property boundary layers, while accuracies of a few meters may be acceptable for large-area vegetation mapping. Common sources of control point coordinates are traditional transit and distance surveys, global positioning system measurements, existing cartometric quality maps, or existing digital data layers on which suitable features may be identified.

The Affine Transformation

The *affine coordinate transformation* employs linear equations to calculate map coordinates. Map projection coordinates are often referred to as eastings (E) and northings (N), and are related to the x and y digitizer coordinates by the equations:

$$E = T_E + a_1 x + a_2 y \qquad (4.1)$$

$$N = T_N + b_1 x + b_2 y \qquad (4.2)$$

Equations 4.1 and 4.2 allow us to move from the arbitrary digitizer coordinate system to the project map coordinate system. We know the x and y coordinates for every digitized point, line vertex, or polygon vertex. We may calculate the E and N coordinates by applying the above equations to every digitized point.

T_E and T_N can be thought of as shifts, or translations, in the origins from one coordinate system to the next. The a_i and b_i parameters incorporate the change in scales and rotation angle between coordinate systems. . The affine is the most commonly applied coordinate transformation because it provides for these three main effects of translation, rotation, and scaling, and because it often introduces less error than higher-order polynomial transformations.

The affine system of equations has six parameters to be estimated: T_E, T_N, a_1, a_2, b_1, and b_2. Each control point provides E, N, x, and y coordinates, and allows us to write two equations. For example, we may have a control point consisting of a precisely surveyed center of a road intersection. This point has digitizer coordinates of $x = 103.0$ centimeters and $y = -100.1$ centimeters, and corresponding Earth-based map projection coordinates of $E = 500{,}083.4$ and $N = 4{,}903{,}683.5$. We may then write two equations based on this control point:

$$500{,}083.4 = T_E + a_1(103.0) + a_2(-100.1) \qquad (4.3)$$

$$4{,}903{,}683.5 = T_N + b_1(103.0) + b_2(-100.1) \qquad (4.4)$$

We cannot find a unique solution to these equations, because there are six unknowns (T_E, T_N, a_1, a_2, b_1, b_2) and only two equations. We need as many equations as unknowns to solve a linear system of equations. Each control point gives us two equations, so we need a minimum of three control points to estimate the parameters of an affine transformation. Statistical estimation requires a total of four control points. As with all statistical estimates, more control points are better than fewer, but we will reach a point of diminishing returns after some number of points, typically somewhere between 18 and 30 control points.

The affine coordinate transformation is usually fit using a statistical method that minimizes the *root mean square error* (RMSE). The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 \ldots + e_n^2}{n}} \qquad (4.5)$$

where the e_i are the residual distances between the true E and N coordinates and the E and N coordinates in the output data layer:

$$e = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2} \qquad (4.6)$$

This residual is the difference between the true coordinates x_t, y_t, and the transformed output coordinates x_d, y_d. Figure 4-25 shows examples of this lack of fit. Individual residuals may be observed at each control point location.

A statistical method for estimating transformation equations is preferred

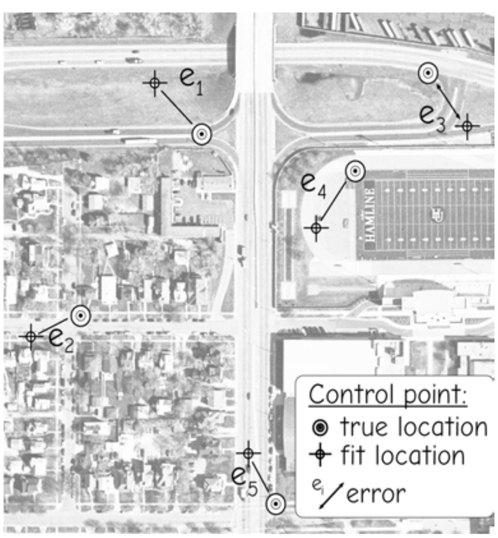


Figure 4-25: Examples of control points, predicted control locations, and residuals from coordinate transformation.

because it also identifies transformation error. Control point coordinates contain unavoidable measurement errors. A statistical process provides an RMSE, a summary of the differences between the "true" (measured) and predicted control point coordinates. It provides one index of transformation quality. Transformations are fit (Figure 4-26). The RMSE will usually be less than the true transformation error at a randomly selected point, because we are actively minimizing the N and E residual errors when we statistically fit the transformation equations. However, the RMSE is an index of accuracy, and a lower RMSE generally indicates a more accurate affine transformation.

Estimating the coordinate transformation parameters is often an iterative process. Control points are rarely exact, and x and y coordinates may not be precisely digitized. Poor eyesight, a shaky hand, fatigue, lack of attention, misidentification of the control location, or a blunder may result in erroneous x and y values. Typically, control points are entered, the affine transformation parameters estimated, and the overall RMSE and individual point E and N errors evaluated (Figure 4-25, Figure 4-26). Suspect points are fixed, and the transformation re-estimated and errors evaluated until a final transformation is estimated. The transformation is then applied to all features to convert them from digitizer to map coordinates.

Model Fit 1:

E = 1.3325289 * x
+ 0.0058654 * y
- 206851.8

N = - 0.002886 * x
+ 1.3296931 * y
- 1660286

RMSE = 9.36

Examine points 15 & 17, adjust noted blunders, refit model

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	3.07
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	8.13
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	4.38
4	518,700.1	5,014,393.4	513,738.9	5,005,831.3	10.99
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	2.79
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	8.18
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	5.66
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	0.88
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	4.05
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	3.37
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.05
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	2.84
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	10.36
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	9.16
15	518,079.9	5,016,484.0	512,912.3	5,008,591.6	19.49
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	7.16
17	517,016.3	5,014,444.0	511,485.6	5,005,903.6	17.05
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	9.51
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.46
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	10.04
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	9.50
22	516,719.0	5,014,555.9	511,106.7	5,006,028.1	14.96

Model Fit 2:

E = 1.3319386 * x
+ 0.0057193 * y
- 205812.1

N = - 0.002462 * x
+ 1.329962 * y
- 1161855

RMSE = 7.72

Examine points 4 & 22, adjust noted blunders, refit model

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	2.56
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	7.22
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	3.21
4	518,700.1	5,014,393.4	513,738.9	5,005,831.3	11.45
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	1.77
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	7.79
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	6.34
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	1.38
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	4.62
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	4.09
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.90
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	2.79
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	9.15
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	8.04
15	518,079.1	5,016,483.3	512,921.1	5,008,596.5	7.71
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	9.22
17	517,015.8	5,014,443.1	511,495.1	5,005,894.9	6.36
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	9.62
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.88
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	8.78
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	10.08
22	516,719.0	5,014,555.9	511,106.7	5,006,028.1	13.68

Model Fit 3:

E = 1.33118637 * x
+ 0.0056629 * y
- 205490.3

N = - 0.003516 * x
+ 1.3297296 * y
- 1160143

RMSE = 6.78

Examine points, no more blunders found.

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	2.48
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	5.63
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	3.84
4	518,699.6	5,014,396.8	513,739.3	5,005,831.0	6.62
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	2.71
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	8.40
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	6.55
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	1.52
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	3.30
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	5.27
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.78
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	3.54
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	9.34
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	8.30
15	518,079.1	5,016,483.3	512,921.1	5,008,596.5	5.67
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	6.73
17	517,015.8	5,014,443.1	511,495.1	5,005,894.9	3.30
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	8.86
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.13
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	9.55
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	9.53
22	516,717.8	5,014,546.4	511,106.4	5,006,028.8	9.01

Figure 4-26: Iterative fitting of an affine transformation. Control points were examined after each fit, to discover blunders in entry or poor matching of points. Control points with large residuals were examined to determine if the cause for the error may be identified. If so, the control point coordinates may be modified, and transformation refit.

Other Coordinate Transformations

Other coordinate transformations are sometimes used. The *conformal coordinate transformation* is similar to the affine, and has the form:

$$E = T_E + cx - dy \qquad (4.7)$$

$$N = T_N + dx + cy \qquad (4.8)$$

The coefficients T_E, T_N, c, and d are estimated from control point data. Like the affine transformation, the conformal transformation is also a first-order polynomial. Unlike the affine, the conformal transformation requires equal scale changes in the x and y directions. Note the symmetry in equations 4.7 and 4.8, in that the x and y coefficients match across equations, and there is a change in sign for the d coefficient. This results in a system of equations with only four unknown parameters, and so the conformal may be estimated when only two control points are available.

Higher-order polynomial transformations are sometimes used to transform among coordinate systems. An example of a 2nd-order polynomial is:

$$E = b_1+b_2x+b_3y+b_4x^2+b_5y^2+b_6xy \qquad (4.9)$$

Note that the combined powers of the x and y variables may be up to 2. This allows for curvature in the transformation in both the x and y directions. A minimum of six control points is required to fit this second-order polynomial transformation, and seven are required when using a statistical fit. The estimated parameters T_E, T_N, a_1, a_2, b_1, and b_2 will be different in equations 4.1 and 4.2 when compared to 4.9, even if the same set of control points is used for both statistical fits. We change the form of the equations by including the higher-order squared and xy cross-product terms, and all estimated parameters will vary.

A Caution When Evaluating Transformations

Selecting the “best” coordinate transformation to apply is a subjective process, guided by multiple goals. We hope to develop an accurate transformation based on a large set of well-distributed control points. Isolated control points that substantially improve our coverage may also contribute substantially to our transformation error.

There are no clear rules on the number of points versus distribution of points trade-off, but it is typically best to strive for the widest distribution of points. We want at least two control points in each quadrant of the working area, with a target of 20% in each quadrant. These goals are often not impossible. The transformation equation should be developed with the following observations in mind.

First, bad control points happen, but we should thoroughly justify the removal of any control point. Every attempt should be made to identify the source of the error, either in the collection or in the processing of field coordinates, the collection of image coordinates, or in some blunder in coordinate transcription. A common error is the misidentification of coordinate location on the image or map, for example, when the control location is placed on the wrong side of a road.

Second, a lower RMSE does not mean a better transformation. The RMSE is a useful tool when comparing among transformations that have the same model form, for example, when comparing one affine to another affine, as in Figure 4-26. The RMSE is not useful when comparing among different model forms, for example, when comparing an affine to a second-order polynomial. The RMSE is typically lower for a second- and other higher-order polynomials than an affine transformation, but this does not mean the higher-order polynomial provides a more accurate transformation. The higher-order polynomial will introduce more error than an affine transformation on most orthographic maps, and an affine transformation is preferred. High-order polynomials allow more flexibility in warping the surface to fit the control points. Unfortunately, this warping may significantly deform the non-control-point coordinates, and add large errors when the transformation is applied to all data in a layer (Figure 4-27). Thus, high-order polynomials and others should be used with caution.

Finally, independent tests of the transformations make the best comparisons among transformations. A completely independent set of widely distributed test points are ideal, but these rarely exist. The extra points either haven't been collected, or suitable locations do not exist. The best way to test the accuracy of the transformation typically uses a "bootstrap" approach that treats each point as an independent test point. One point is withheld, the transformation estimated, and the error at the withheld point calculated. The point is replaced in the estimation set, and the next point withheld, fitting the same type of transformation. The equations will be slightly different. The error at this second withheld point is then calculated. This process is repeated for each control point, and a mean error calculated.

Control Point Sources: Surveying

Traditional ground surveys based on optical surface measurements are a common, although decreasingly used method for determining control point locations. Federal, state, county, and local governments all maintain a set of accurately surveyed locations (Figure 4-28), and these points may be used as control points or as starting points for additional surveys. Many of these known

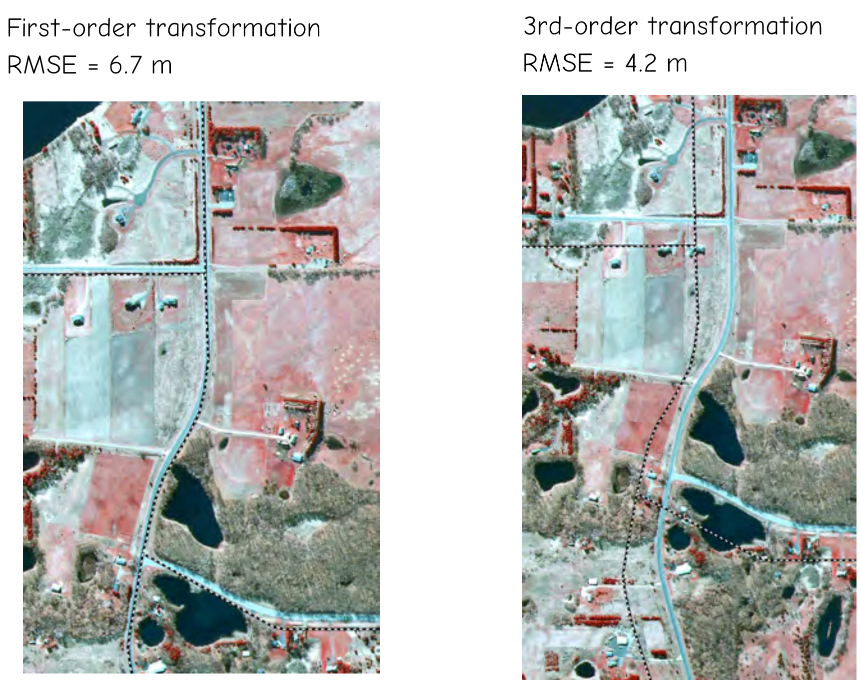


Figure 4-27: An illustration that RMSE should not be used to compare different order transformations, nor should it be used as the sole criterion for selecting the best transformation. Above are portions of a transformed image that was registered to a road network. This area is interstitial to 18 widely distributed control points. Because the 3rd-order polynomial is quite flexible in fitting the points and reducing RMSE, it distorts areas between the control points. This is shown by the poor match between image and vector roads (above right). Although it has a higher RMSE, the first-order transformation on the left is better overall.

points have been established using traditional surveying techniques. Indeed, the development of this "control network" infrastructure is one of the first and most important responsibilities of government. These survey points form the basis for distance, location, and area measurements used to define property, political, and municipal boundaries. As a result, this control network underlies most commerce, transportation, and land ownership and management. Coordinates, general location, and descriptions are documented for these control networks, and may be obtained from a number of government sources. In the United States, these sources include county surveyors, state surveyors, departments of transportation, and the National Geodetic Survey (NGS).

The ground survey network is often quite sparse and insufficient for registering many large-scale maps or images. Many may not be suitable for use as control points in a coordinate transformation of spatial data. First, our control points must be visible on the map, data layer, or image that we wish to register; and second, we must have precise ground coordinates in our target map projection. The first requirement, visibility on the source map or photograph, is often not met for survey-defined control. Therefore, we must often obtain additional control points.

Figure 4-28: Previous surveys are a common source of control points.

GNSS Control Points

The global positioning system (GPS), GLONASS, and Galileo are Global Navigation Satellite Systems (GNSS) that allow us to establish control points. GNSS, discussed in detail in Chapter 5, can help us obtain the coordinates of control points that are visible on a map or image. GNSS are particularly useful because we may quickly survey widely spaced points. GNSS positional accuracy depends on the technology and methods employed; it typically ranges from subcentimeter (tenths of inches) to a few meters (tens of feet). Most points recently added to the NGS and other government-maintained networks were measured using GNSS technologies.

Control Points from Existing Digital Data and Maps

Registered digital image data are common sources of ground control points, particularly when natural resources or municipal databases are to be developed for managing large areas. Digital images often provide a richly detailed depiction of surface features (Figure 4-29). Digital image data may be obtained that are registered to a known coordinate system. Typically, the coordinates of a corner pixel are provided, and the lines and columns for the image run parallel to the easting (E) and northing (N) direction of the coordinate system. Because the pixel dimensions are known, the calculation of a pixel coordinate involves multiplying the row and column number by the pixel size, and applying the corner offset, either by addition or subtraction. In this manner, the image row/column may be converted to an E, N coordinate pair, and control point coordinates determined.

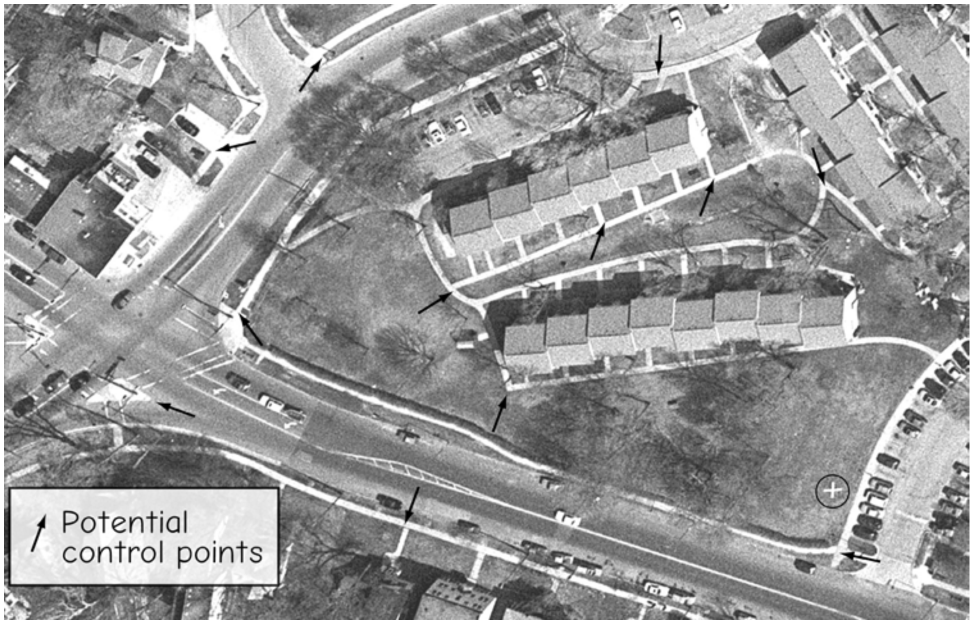


Figure 4-29: Potential control points, indicated here by arrows, may be extracted from digital images that have been geometrically corrected to provide an accurate, coordinate-projected rendering.

Existing maps are another common source of control points. Point locations are plotted and coordinates often printed on maps; for example the corner location coordinates are printed on USGS quadrangle maps. Road intersections and other well-defined locations are often represented on maps. If enough recognizable features can be identified, then control points may be obtained from the maps. Control points derived in this manner typically come only from cartometric maps, those maps produced with the intent of giving an accurate, map-projected representation of features on the Earth's surface.

Existing digital data may also provide control points. A short description of these digital data sources are provided here, and expanded descriptions of these and other digital data are provided in Chapter 7. For example, the USGS has produced Digital Raster Graphic (DRG) files that are scanned images of the 1:24,000-scale quadrangle maps. These DRGs come referenced to a standard coordinate system, so it is a simple and straightforward task to extract the coordinates of road intersections or other well-defined features that have been plotted on the USGS quadrangle maps. Vector data of roads are often widely available and, if of sufficient accuracy, may be used as a source of control points at road intersections and other distinct locations.

Raster Geometry and Resampling

Data often must be *resampled* when converting between coordinate systems, or changing the cell size of a raster data set (Figure 4-30). Resampling involves reassigning the cell values when changing raster coordinates or geometry. Resampling is required when changing cell sizes because the new cell centers will not align exactly with old cell centers. Changing coordinate systems may change the direction of the x and y axes, and GIS systems often require that the cell edges align with the coordinate system axes. Hence, the new cells often do not correspond to the same locations or extents as the old cells.

Common resampling approaches include the *nearest neighbor* (taking the output layer value from the nearest input layer cell center), *bilinear interpolation* (distance-based averaging of the four nearest cells), and *cubic convolution* (a weighted average of the sixteen nearest cells, Figure 4-30).

An example of a bilinear interpolation is shown in Figure 4-31. This algorithm uses a distance-weighted average of the four nearest cells in the input to calculate the value for the output. The new output location is represented by the black post. Initially, the height, or Z_{out} value, of the output location is unknown. Z_{out} is calculated based on the distances between the output locations and the input locations. The distance in the X direction is denoted in Figure 4-31 by d_1, and the distance in the y direction by d_2. The values in the input are shown as gray posts and are labeled as Z_1 through Z_4. Intermediate heights Z_b and Z_u are shown. These represent the average of the input values when taken in pairs in the x direction. These pairs are Z_1 and Z_2, to yield Z_u, and Z_3 and Z_4, to yield Z_b. Z_u and Z_b are then averaged to calculate Z_{out}, using the distance d_2 between the input and output locations to weight values at each input location. The cubic convolution resampling calculation is similar, except that more cells are used, and the weighting is not an average based on linear distance.

Map Projection vs. Transformation

Map transformations should not be confused with map projections. A map transformation typically employs a statistically fit linear equation to convert coordinates from one Cartesian coordinate system to another. A map projection, described in Chapter 3, differs from a transformation in that it is an analytical, formula-based conversion between coordinate systems, usu-

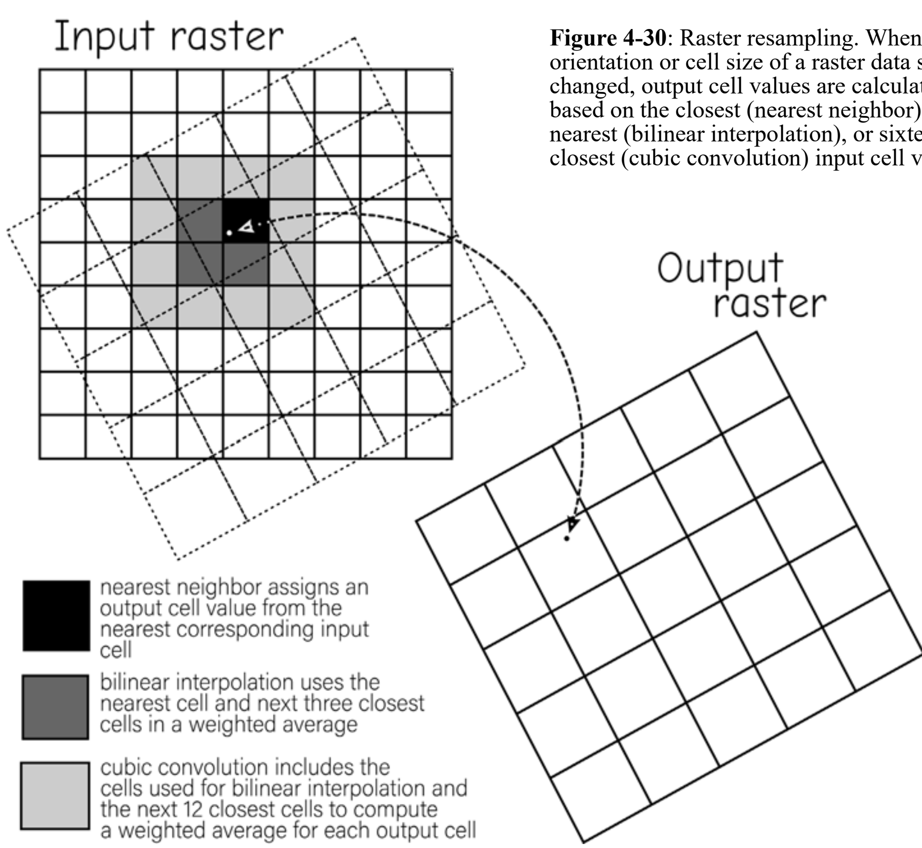
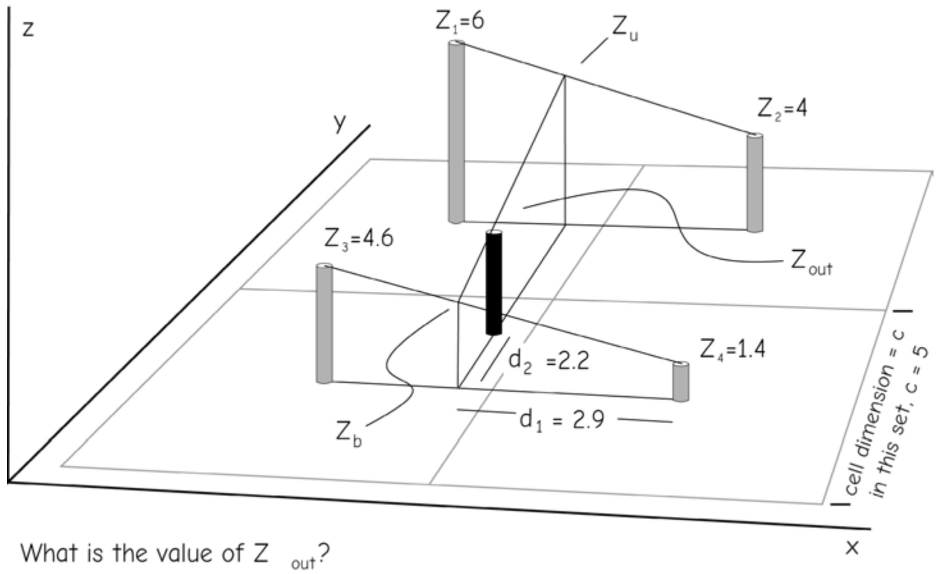


Figure 4-30: Raster resampling. When the orientation or cell size of a raster data set is changed, output cell values are calculated based on the closest (nearest neighbor), four nearest (bilinear interpolation), or sixteen closest (cubic convolution) input cell values.



What is the value of Z_{out}?

$$Z_b = Z_4 + \frac{(Z_3 - Z_4)*d_1}{c} \qquad Z_b = 1.4 + \frac{(4.6 - 1.4)*2.9}{5} = 3.26$$

$$Z_u = Z_2 + \frac{(Z_1 - Z_2)*d_1}{c} \qquad Z_u = 4 + \frac{(6 - 4)*2.9}{5} = 5.16$$

$$Z_{out} = Z_b + \frac{(Z_u - Z_b)*d_2}{c} \qquad Z_{out} = 3.26 + \frac{(5.16 - 3.26)*2.2}{5} = 4.1$$

Figure 4-31: The bilinear interpolation method uses a distance-weighted average to assign the output value, Z_{out}, based on input values, Z_1 through Z_4.

ally from a curved, latitude/longitude coordinate system to a Cartesian coordinate system. No statistical fitting process is used with a map projection.

Map transformations should rarely be used in place of map projection equations when converting geographic data between map projections. Consider an example when data are delivered to an organization in Universal Transverse Mercator (UTM) coordinates and are to be converted to State Plane coordinates prior to integration into a GIS database. Two paths may be chosen. The first involves projection from UTM to geographic coordinates (latitude and longitude), and then from these geographic coordinates to the appropriate State Plane coordinates. This is the correct, most accurate approach.

An alternate and often less accurate approach involves using a transformation to convert between different map projections. In this case, a set of control points would be identified and the coordinates determined in both UTM and State Plane coordinate systems. The transformation coefficients would be estimated and these equations applied to all data in the UTM data layer. This new output data layer would be in State Plane coordinates. This transformation process should be avoided, as a transformation may introduce additional positional error.

Transforming between projections is used quite often, inadvertently, when digitizing data from paper maps. For example, USGS 1:24,000-scale maps are cast on a polyconic projection. If these maps are digitized, it would be preferable to register them to the appropriate polyconic projection, and then reproject these data to the desired end projection. This is often not done, because the error in ignoring the projection over the size of the mapped area is typically less than the positional error associated with digitizing. Experience and specific calculations have shown that the spatial errors in using a transformation instead of a projection are small at these map scales under typical digitizing conditions.

This second approach, using a transformation when a projection is called for, should not be used until it has been tested as appropriate for each new set of conditions. Each map projection distorts the surface geometry. These distortions are complex and nonlinear. Affine or polynomial transformations are unlikely to remove this nonlinear distortion. Exceptions to this rule occur when the area being transformed is small, particularly when the projection distortion is small relative to the random uncertainties, transformation errors, or errors in the spatial data. However, there are no guidelines on what constitutes a sufficiently “small” area. In our example above, USGS 1:24,000 scale maps are often digitized directly into a UTM coordinate system with no obvious ill effects, because the errors in map production and digitizing are often much larger than those in the projection distortion for the map area. However, you should not infer this practice is appropriate under all conditions, particularly when working with smaller-scale maps.

Output: Hardcopy Maps, Digital Data, and Metadata

We create spatial data to use, share, and archive. Maps are often produced during data creation and distribution, as intermediate documents while editing, for analysis, or as finished products to communicate some aspect of our data. To be widely useful, we must also generate information, or "metadata," about the spatial data we've created, and we may have to convert our data to standard forms. This section describes some characteristics of data output. We start with a brief treatment of cartography and map design, by which we produce hardcopy and digital maps. We then provide a description of metadata, and some observations on data conversion and data transfer standards.

Cartography and Map Design

Cartography is the art and techniques of making maps. It encompasses both mapmaking tools and how these tools may be combined to communicate spatial information. Cartography is a discipline of much depth and breadth, and there are many books, journal articles, conferences, and societies devoted to the science and art of cartography. Our aim in the next few pages is to provide a brief overview of cartography with a particular focus on map design. This is both to acquaint new students with the most basic concepts of cartography, and help them apply these concepts in the consumption and production of spatial information. Readers interested in a more complete treatment should consult the references listed at the end of this chapter.

A primary purpose of cartography is to communicate spatial information. This requires identification of the

-intended audience,

-information to communicate,

-area of interest,

-physical and resource limitations;

in short, the whom, what, where, and how we may present our information.

These considerations drive the major cartographic design decisions we make each time we produce a map. We must consider the:

-scale, size, shape, and other general map properties,

-data to plot,

-symbol shapes, sizes, or patterns,

-labeling, including font type and size,

-legend properties, size, and borders, and

-the placement of all these elements on a map.

Map scale, size, and shape depend primarily on the intended map use. Wall maps for viewing at a distance of a meter or more may have few, large, boldly colored features. In contrast, commonly produced street maps for navigation in metropolitan areas are detailed, to be viewed at short ranges, and have a rich set of additional tables, lists, or other features.

Map scale is often determined in part by the size of the primary objects we wish to display, and in part by the most appropriate media sizes, such as the page or screen size possible for a document. As noted earlier, the map scale is the ratio of lengths on a map to true lengths. If we wish to display an area that spans 25 km (25,000 m) on a screen that spans 25 cm (0.25 m), the map scale will be near 0.25 to 25,000, or 1:100,000. This decision on size, area, and scale then drives further map design. For example, scale limits the features we may display, and the size, number, and labeling of features. At a 1:100,000 scale we may not be able to show all cities, burgs, and towns, as there may be too many to fit at a readable size.

Maps typically have a primary theme or purpose that is determined by the intended audience. Is the map for a general population, or for a target audience with specific expectations for map features and design? General purpose maps typically have a wide range of features represented, including transportation networks, towns, elevation, or other common features (Figure 4-32a). Special purpose maps, such as road maps, focus on a more limited set of features, in this instance road locations and names, town names, and large geographic features (Figure 4-32b).

Once the features to include on a map are defined, we must choose the symbols used to draw them. Symbology depends in part on the type of feature. For example, we have a different set of options when representing continuous features such as elevation or pollution concentration than when representing discrete features. We also must choose among symbols for each of the types of discrete features; for example, the set of symbols for points are generally different from those for line or area features.

Symbol size is an important attribute of map symbology, often specified in a unit called a *point*. One point is approximately equal to 0.467 mm, or about 1/55th of an inch. A specific point number is most often used to specify the size of symbols, for example, the dimensions of small squares to represent houses on a map, or the characteristics of a specific pattern used to fill areas on a map. A line width may also be specified in points. Setting a line width of two points means we want that particular line plotted with a width of 0.93 mm. It is unfortunate that "point" is both the name of the distance unit and a general property of a geographic feature, as in "a tree is a point feature." This forces us to talk about the "point size" of symbols to represent points, lines, or area fills or patterns, but if we are careful, we may communicate these specifications clearly.

The best size, pattern, shape, and color used to symbolize each feature depends on the viewing distance; the number, density, and type of features; and the purpose of the map. Generally, we use larger, bolder, or

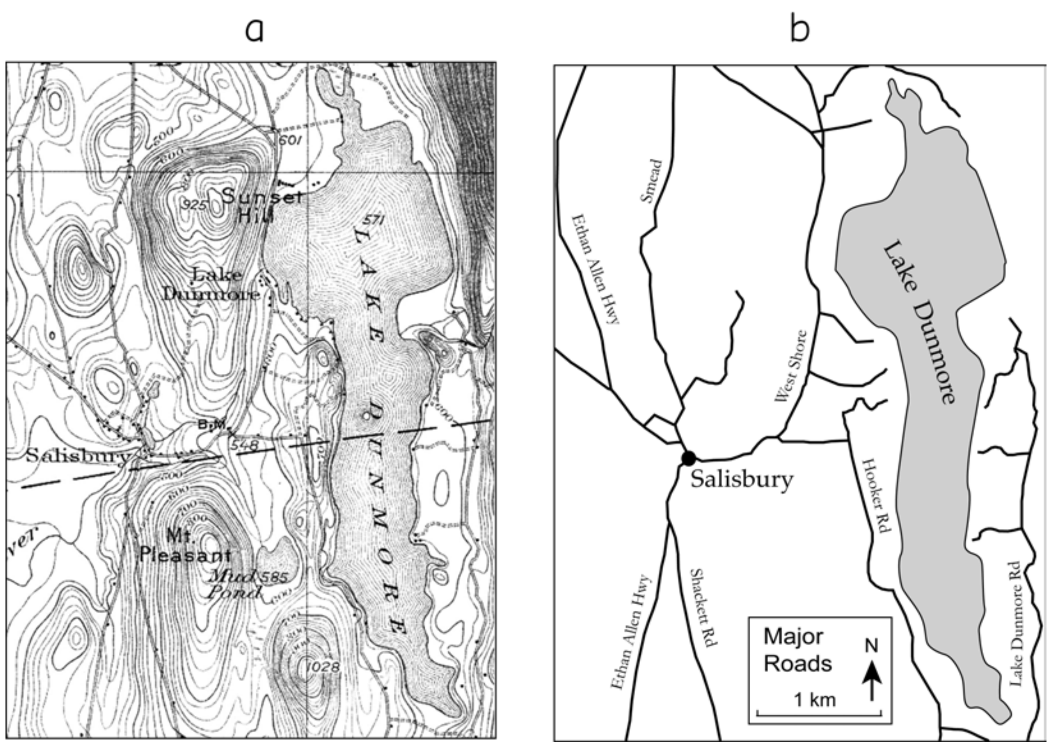


Figure 4-32: Example of a) a detailed, general-purpose map, here a portion of a United States Geological Survey map, and b) a specialized map focusing on a specific set of selected features, here showing roads. The features chosen for depiction on the map depend on the intended map use.

thicker symbols for maps to be viewed from longer distances, while we reduce this limit when producing maps for viewing at 50 cm (20 in). Most people with normal vision under good lighting may resolve lines down to near 0.2 points at close distances, provided the lines show good contrast with the background. Although size limits depend largely on background color and contrast, point features are typically not resolvable at sizes smaller than about 0.5 points, and distinguishing between shapes is difficult for point features smaller than approximately 2 points in their largest dimension.

The pattern and color of symbols must also be chosen, generally from a set provided by the software (Figure 4-33). Symbols generally distinguish among feature type by characteristics, and although most symbols are not associated with a feature type, some are, such as, plane outlines for

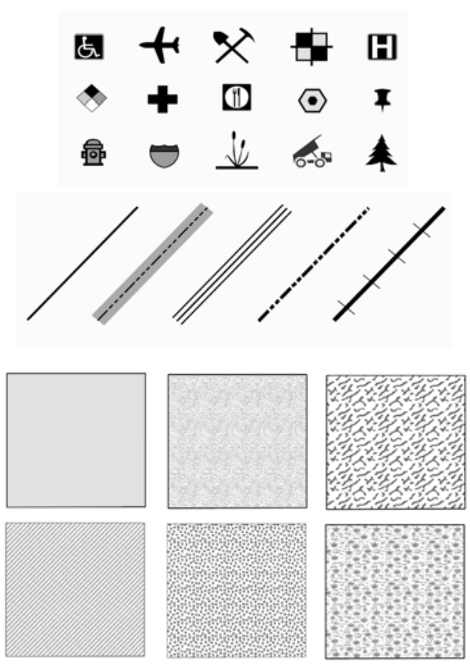


Figure 4-33: Examples of point (top), line (mid), and area (bottom) symbols used to distinguish among features of different types. Most GIS software provides a set of standard symbols for point, line, area, and continuous surface features.

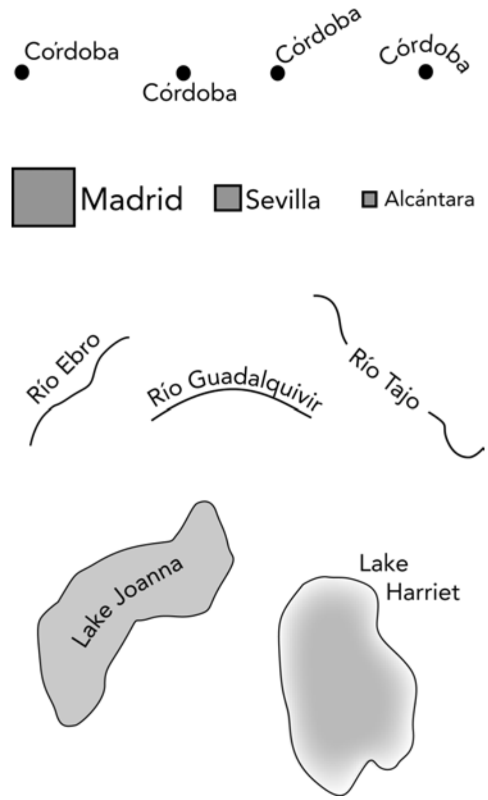


Figure 4-34: Common labeling options, including straight, angled, wrapped text, and graduated labels for points (top two sets), and angled, wrapped, fronting, and embedded labels for line and polygon features (bottom two sets).

airports, numbered shields for highways, or a hatched line for a railroad.

We also must often choose whether and how to label features. Most GIS software provides a range of tools for creating and placing labels, and in all cases we must choose the label font type and size, location relative to the feature, and orientation. Primary considerations when labeling point features are label placement relative to the point location, label size, and label orientation (Figure 4-34). We may also use graduated labels, that is, resize them according to some variable associated with the point feature. For example, it is common to have larger features and label fonts for larger cities (Figure 4-34). Labels may be bent, angled, or wrapped around features to improve clarity and more efficiently use space in a map.

Label placement is very much an art, and there is often much individual editing required when placing and sizing labels for finished maps. Most software provides for automatic label placement, usually specified relative to feature location. For example, one may specify labels above and to the right of all points, or line labels placed over line features, or polygon labels placed near the polygon centroid. However, these automatic placements may not be satisfactory because labels may overlap, labels may fall in cluttered areas of the map, or features associated with labels may be ambiguous. Some software provides options for automatic label placement, including removal or movement of overlapping labels. These often reduce manual editing, but sometimes increase it.

Figure 4-35 shows a portion of a map of southern Finland. This region presents several mapping problems, including the high density of cities near the upper right, an irregular coastline, and dense clustering of islands along the coast. Most labels are placed above and to the right of their corresponding city; however, some are moved or angled for clarity. Cities near the coast show both, to avoid labels crossing the water/land boundary where practical. Semitransparent background shading is added for Parainen and Hanko, cities placed in the island matrix. This example demonstrates the individual editing often required when placing labels.

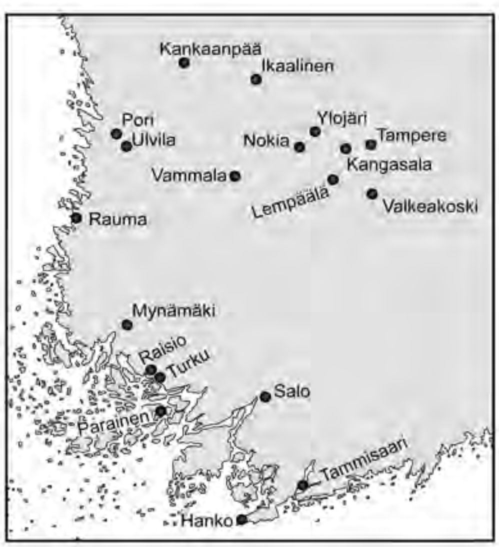


Figure 4-35: Example label placement for cities in southern Finland.

Most maps should have legends. The legend identifies map features succinctly and describes the symbols used to depict those features. Legends often include or are grouped with additional map information such as scale bars, north arrows, and descriptive text. The cartographer must choose the size and shape of the descriptive symbol, and the font type, size, and orientation for each symbol in the legend. The primary goal is to have a clear, concise, and complete legend.

The kind of symbols appropriate for map legends depends on the types of features depicted. Different choices are available for point, line, and polygon features, or for continuously variable features stored as rasters. Most software provides a range of legend elements and symbols that may be used. Typically, these tools allow a wide range of symbolizations, and a compact way of describing the symbolization in a legend (Figure 4-36).

The specific layout of legend features must be defined; for example, the point feature symbol size may be graduated based on some attribute for the points. Successively larger features may be assigned for successively larger cities. This must be noted in the legend, and the symbols nested, shown sequentially, or otherwise depicted (Figure 4-36, top left).

The legend should be exhaustive. Examples of each different symbol type that appears on the map should appear in the legend. This means each point, line, or area symbol is drawn in the legend with some descriptive label. Labels may be next to, wrapped around, or embedded within the features, and sometimes descriptive numbers are added, for example, a range of continuous variables (Figure 4-36, upper left). Scale bars, north arrows, and descriptive text boxes are typically included in the legend.

Map composition or layout is another primary task. Composition consists of determining the map elements, their size, and their placement. Typical map elements, shown in Figure 4-3 and Figure 4-4, include one or more main data panes or areas, a legend, a title, a scale bar and north arrow, a grid or graticule, and perhaps descriptive text. These each must be sized and placed on the map.

These map elements should be positioned and sized in accordance with their importance. The map's most important data pane should be largest, and it is often centered or otherwise given visual dominance. Other elements are typically smaller and located around the periphery or embedded within the main data pane. These other elements include map insets, which are smaller data panes that show larger or smaller scale views of a region in the primary data pane. Good map compositions usually group related elements and use empty space effectively. Data panes are often grouped and legend elements placed near each other, and grouping is often indicated with enclosing boxes.

Neophyte cartographers should avoid two tendencies in map composition, both depicted in Figure 4-37. First, it is generally easy to create a map with automatic label and legend generation and placement. The map shown at the top of Figure 4-37 is typical of this automatic composition, and includes poorly placed legend elements and too small, poorly placed labels. Labels crowd each other, are ambiguous, and cross water/land or other feature boundaries, and fonts are poorly chosen. You should note that automatic map symbol selection and placement is nearly always suboptimal, and the novice cartographer should scrutinize these choices and manually improve them.

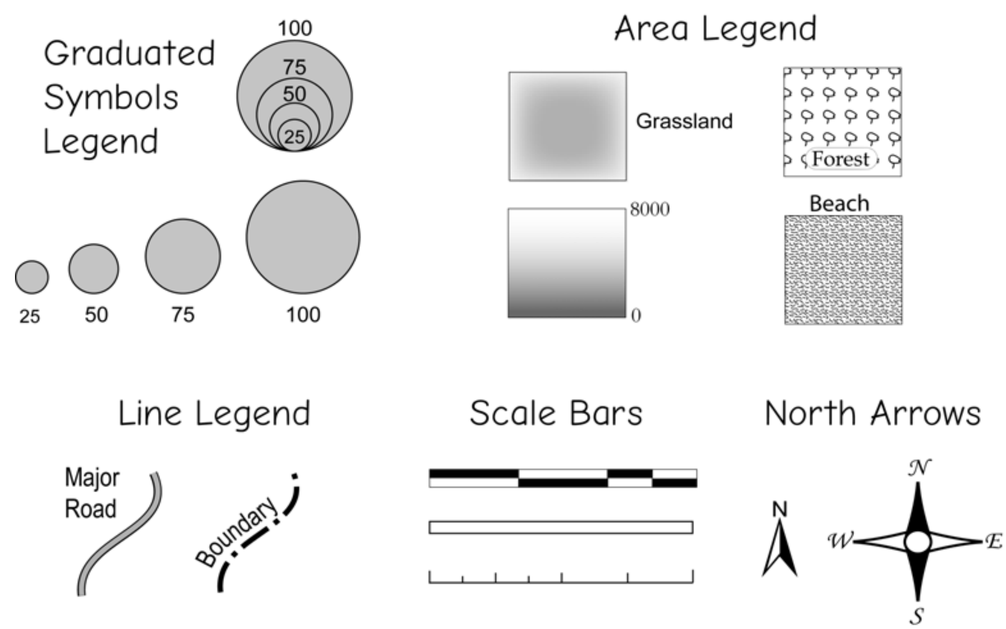


Figure 4-36: Examples of legend elements and representation of symbols. Some symbols may be grouped in a compact way to communicate the values associated with each symbol, e.g., sequential or nested graduated circles to represent city population size, area pattern or color fills to distinguish among different polygon features, line and point symbols, and informative elements such as scale bars and north arrows.

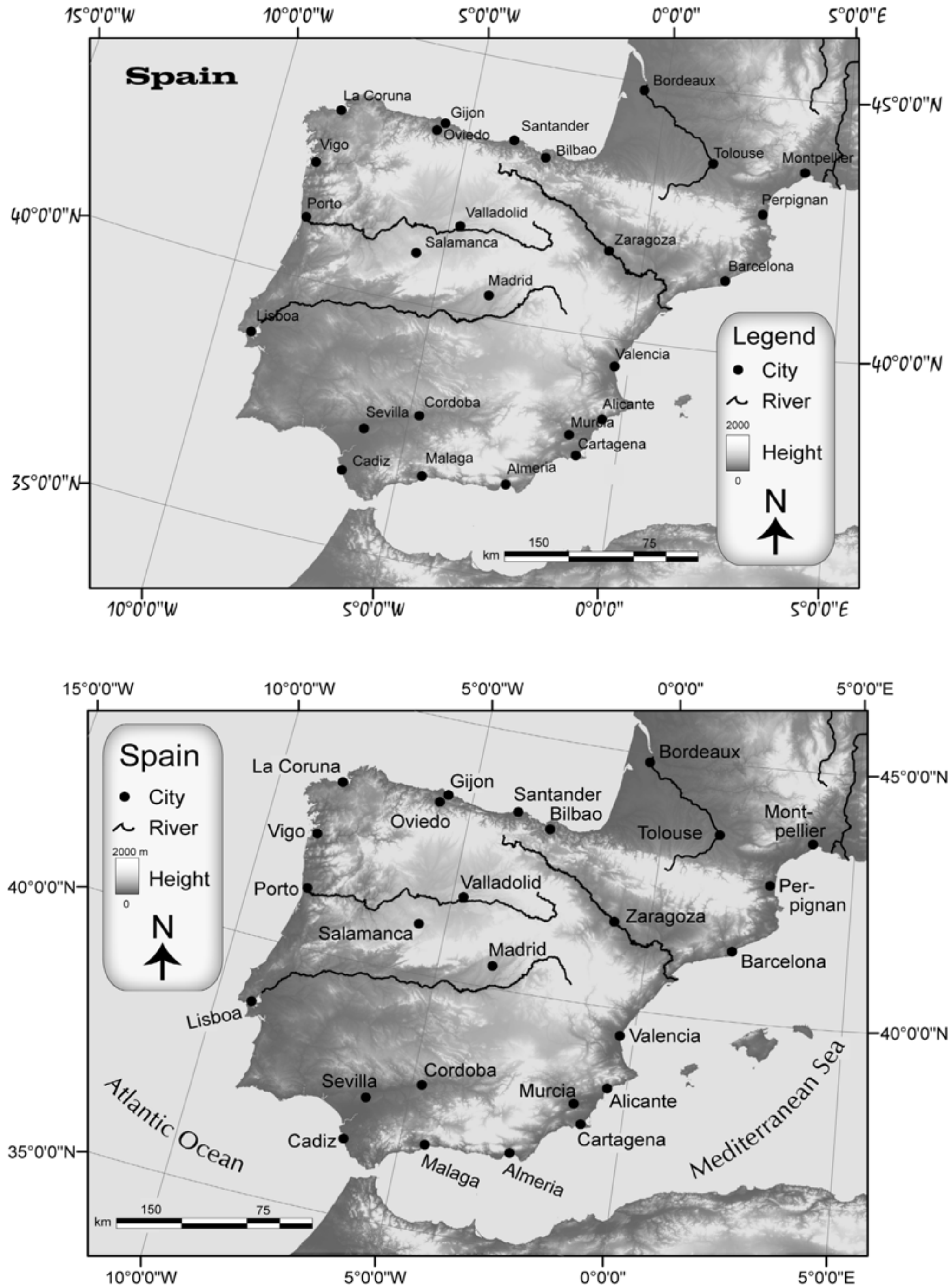


Figure 4-37: An example of poor map design (top). This top panel shows a number of mistakes common for the neophyte cartographer, including small labels (cities) and mismatched fonts (graticule labels, title), poor labeling (city labels overlapping, ambiguously placed, and crossing distinctly shaded areas), unlabeled features (oceans and seas), poorly placed scale bar and legend, and unbalanced open space on the left side of the map. These problems are not present in the improved map design, shown in the lower panel.

The second common error is poor use of empty space, those parts of the map without map elements. There are two opposite tendencies: either to leave too much or unbalanced empty space or to clutter the map in an attempt to fill all empty space. Note that the map shown at the top of Figure 4-37 leaves large empty spaces on the left (western) edge, with the Atlantic Ocean devoid of features. The cartographer may address this in several ways: by changing the size, shape, or extent of the area mapped; by adding new features, such as data panes as insets, additional text boxes, or other elements; or by moving the legend or other map elements to that space. The map shown at the bottom of Figure 4-37, while not perfect, fixes these design flaws, in part by moving the legend and scale bar, and in part by adding labels for the Atlantic Ocean and Mediterranean Sea. The empty space is more balanced in that it appears around the major map elements in approximately equal proportions.

As noted earlier, this is only a brief introduction to cartography, a subject covered by many good books, some listed at the end of this chapter. Perhaps the best compendium of examples is the Map Book Series, by ESRI, published annually since 1984. Examples are available at the time of this writing at www.esri.com/mapmuseum. You should leaf through several volumes in this series, with an eye toward critical map design. Each volume contains many beautiful and informative maps, and provides techniques worth emulating.

Digital Data Output

We often must transfer digital data we create to another user, or use data developed by others. Given the number of different GIS software, operating systems, and computer types, transferring data is not always a straightforward process. Digital data output typically includes two components, the data themselves in some standard, defined format, and *metadata*, or data about the digital data. We will describe data formats and metadata in turn.

Digital data are the data in some electronic form. As described at the end of the first chapter, there are many file formats, or ways of encoding the spatial and attribute data in digital files. Digital data output often consists of recording or converting data into one of these file formats. These data are typically converted with a utility, tool, or option available in the data development software (Figure 4-38). The most useful of these utilities supports a broad range of input and output options, each fully described in the program documentation.

All formats strive for complete data transfer without loss. They must transmit the spatial and attribute data, the metadata, and all other information necessary to effectively use the spatial data. There are many digital data output formats, although many are legacy formats that are used with decreasing frequency.

A common contemporary format is the *Geographic Markup Language (GML)*. This is an extension of XML for geographic features; XML is in turn the lingua franca for human/machine readable documents. As with most XML, there are two parts for any

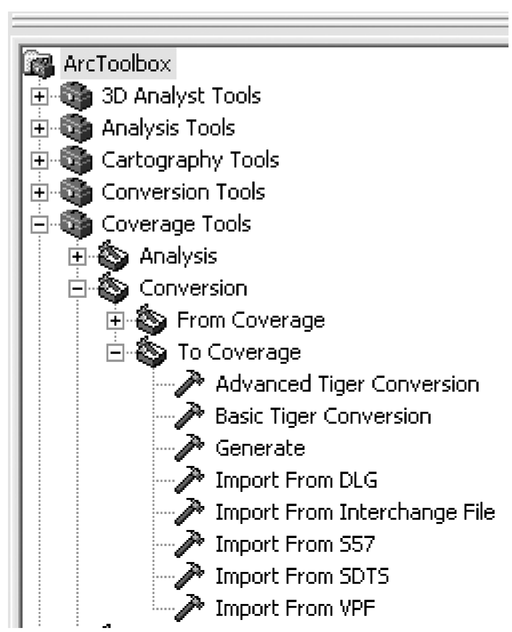


Figure 4-38: An example of a conversion utility, here from the ESRI ArcGIS software. Date may be converted from one of several formats to an ESRI-specific digital data.

GML dataset: a schema that describes the document, and the document containing the geographic data. GML is a standard, but there can be many extensions, so a community of users can extend the standard with additional features, and document the extension in a standard way.

There are many legacy digital data transfer formats that were widely used before GML. GML replaced a withdrawn standard, the *Spatial Data Transfer Standard* (SDTS), with translators available to or from this older format. There are several U.S. Geological Survey formats for the transfer of digital elevation models or digital vector data, or software-specific formats, such as an ASCII format (GEN/UNGEN) that was developed by ESRI. These were useful for a limited set of transfers, but shortcomings in each of these transfer formats led to the development of subsequent standards. These formats are not common, but sometimes arise in converting older data sets.

Metadata: Data Documentation

Metadata are information about spatial data. Metadata describe the content, source, lineage, methods, developer, coordinate system, extent, structure, spatial accuracy, attributes, and responsible organization for spatial data.

Metadata are required for the effective use of spatial data. Metadata allow the efficient transfer of information about data, and inform new users about the geographic extent, coordinate system, quality, and other data characteristics. Metadata aid organizations in evaluating data to determine if they are suitable for an intended use, e.g., to review accuracy, coverage, or information needs. Metadata may also aid in data updates by guiding the choice of appropriate collection methods and formats for new data.

Most governments have or are in the process of establishing standard methods for reporting metadata. In the United States, the Federal Geographic Data Committee (FGDC) has defined a Content Standard for Digital Geospatial Metadata (CSDGM) to specify the content and format for metadata. The CSDGM ensures that spatial data are clearly described so that they may be used effectively within an organization. The use of the CSDGM also ensures that data may be described to other organizations in a standard manner, and that spatial data may be more easily evaluated by and transferred to other organizations.

The CSDGM consists of a standard set of elements that are presented in a specified order. The standard is exhaustive in the information it provides, and is flexible in that it may be extended to include new elements for new categories of information in the future. There are over 330 different elements in the CSDGM. Some of these elements contain information about the spatial data, and some elements describe or provide linkages to other elements. Elements have standardized long and short names and are provided in a standard order with a hierarchical numbering system. For example, the westernmost bounding coordinate of a data set is element 1.5.1.1, defined as follows:

1.5.1.1 West Bounding Coordinate – westernmost coordinate of the limit of coverage expressed in longitude.

Type: real

Domain: -180.0 < = West Bounding Coordinate < 180.0

Short Name: westbc

The numbering system is hierarchical. Here, 1 indicates it is basic identification information, 1.5 indicates identification information about the spatial domain, 1.5.1 is for bounding coordinates, and 1.5.1.1 is the western most bounding coordinate.

There are 10 basic types of information in the CSDGM:

1) identification, describing the data set,

2) data quality,

3) spatial data organization,

4) spatial reference coordinate system,

5) entity and attribute,

6) distribution and options for obtaining the data set,

7) currency of metadata and responsible party,

8) citation,

9) time period information, used with other sections to provide temporal information, and

10) contact organization or person.

The CSDGM is a content standard and does not specify the format of the metadata. As long as the elements are included, properly numbered, and identified with correct values describing the data set, the metadata are considered to conform with the CSDGM. Indentation and spacing are not specified. However, because metadata may be quite complex, there are a number of conventions that are emerging in the presentation of metadata. These conventions seek to ensure that metadata are presented in a clear, logical way to humans, and are also easily ingested by computer software. There is a Standard Generalized Markup Language (SGML) for the exchange of metadata. An example of a portion of the metadata for a 1:100,000 scale digital line graph data set is shown in Figure 4-39.

Metadata are most often created using specialized software tools. Although meta-

```
4. Spatial_Reference_Information:
   4.1 Horizontal_Coordinate_System_Definition:
      4.1.2 Planar:
         4.1.2.2 Grid_Coordinate_System:
            4.1.2.2.1 Grid_Coordinate_System_Name:
                                 Universal Transverse Mercator
            4.1.2.2.2 Universal_Transverse_Mercator:
               4.1.2.2.2.1 UTM_Zone_Number: 10-19
         4.1.2.4 Planar_Coordinate_Information:
            4.1.2.4.1 Planar_Coordinate_Encoding_Method:
                                                coordinate pair
            4.1.2.4.2 Coordinate_Representation:
               4.1.2.4.2.1 Abscissa_Resolution: 2.54
               4.1.2.4.2.2 Ordinate_Resolution: 2.54
            4.1.2.4.4 Planar_Distance_Units: meters
      4.1.4 Geodetic_Model:
         4.1.4.1 Horizontal_Datum_Name: North American Datum 1927
         4.1.4.2 Ellipsoid_Name: Clark 1866
         4.1.4.3 Semi-major_Axis: 6378206.4
         4.1.4.4 Denominator_of_Flattening_Ratio: 294.98
   4.2 Vertical_Coordinate_System_Definition:
      4.2.1 Altitude_System_Definition:
         4.2.1.1 Altitude_Datum_Name:
                        National Geodetic Vertical Datum of 1929
         4.2.1.2 Altitude_Resolution: 1
         4.2.1.3 Altitude_Distance_Units: feet or meters
         4.2.1.4 Altitude_Encoding_Method: attribute values
      4.2.2 Depth_System_Definition:
         4.2.2.1 Depth_Datum_Name: Mean lower low water
         4.2.2.2 Depth_Resolution: 1
         4.2.2.3 Depth_Distance_Units: meters or feet
         4.2.2.4 Depth_Encoding_Method: attribute values
```

Figure 4-39: Example of a small portion of the FGDC recommended metadata for a 1:100,000 scale derived digital data set.

data may be produced using a text editor, the numbering system, names, and other conventions are laborious to type. There are often complex linkages between metadata elements, and some elements are repeated or redundant. Software tools may ease the task of metadata entry by reducing redundant entries, ensuring correct linkages, and checking elements for contradictory information or errors. For example, the metadata entry tool may check to make sure the westernmost boundary is west of the easternmost boundary. Metadata are most easily and effectively produced when their development is integrated into the workflow of data production.

Although not all organizations in the United States adhere to the CSDGM metadata standard, most organizations record and organize a description and other important information about their data, and many organizations consider a data set incomplete if it lacks metadata. All U.S. government units are required to adhere to the CSDGM when documenting and distributing spatial data.

Many other national governments are developing metadata standards. One example is the spatial metadata standard developed by the Australia and New Zealand Land Information Council (ANZLIC), known as the ANZLIC Metadata Guidelines. ANZLIC is a group of government, business, and academic representatives working to develop spatial data standards. The ANZLIC metadata guidelines define the core elements of metadata, and describe how to write, store, and disseminate these core elements. Data entry tools, examples, and spatial data directory have been developed to assist in the use of ANZLIC spatial metadata guidelines.

There is a parallel effort to develop and maintain international standards for metadata. The standards are known as the ISO 19115 International Standards for Metadata. According to the International Standards Organization, the ISO 19115 "defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data."

There is a need to reconcile international and national metadata standards, because they may differ. National standards may require information not contained in international standards, or vice versa. Governments typically create *metadata profiles* that are consistent with the international standard. These profiles establish the correspondence between elements in the different standards, and identify elements of the international profile that are not in the national profile.

Summary

Spatial data entry is a common activity for many GIS users. Although data may be derived from several sources, maps are a common source, and care must be taken to choose appropriate map types and to interpret the maps correctly when converting them to spatial data in a GIS.

Maps are used for spatial data entry due to several unique characteristics. These include our long history of hardcopy map production, so centuries of spatial information are stored there. In addition, maps are inexpensive, widely available, and easy to convert to digital forms, although the process is often time consuming, and may be costly. Maps are usually converted to digital data through a manual digitization process, whereby a human analyst traces and records the location of important features. Maps may also be digitized via a scanning device.

The quality of data derived from a map depends on the type and size of the map, how the map was produced, the map scale, and the methods used for digitizing. Large-scale maps generally provide more accurate positional data than comparable small-scale maps. Large-scale maps often have less map generalization, and small horizontal errors in plotting, printing, and digitizing are magnified less during conversion of large-scale maps.

Snapping, smoothing, vertex thinning, and other tools may be used to improve the quality and utility of digitized data. These methods are used to ensure positional data are captured efficiently and at the proper level of detail.

Map and other data often need to be converted to a target coordinate system via a map transformation. Transformations are different from map projections (discussed in Chapter 3), in that a transformation uses an empirical, least squares process to convert coordinates from one Cartesian system to another. Transformations are often used when registering digitized data to a known coordinate system. Map transformations should not be used when a map projection is called for.

Cartography is an important aspect of GIS, because we often communicate spatial information through maps. Map design depends on both the target audience and purpose, setting and modes of map viewing, and available resources. Proper map design considers the scale, symbols, labels, legend, and placement to effectively communicate the desired information.

Metadata are the "data about data." They describe the content, origin, form, coordinate system, spatial and attribute data characteristics, and other relevant information about spatial data. Metadata facilitate the proper use, maintenance, and transfer of spatial data. Metadata standards have been developed, both nationally and internationally, with profiles used to cross-reference elements between metadata standards. Metadata are a key component of spatial data, and many organizations do not consider data complete until metadata have been created.

Suggested Reading

Aronoff, S. (1989). *Geographic Information Systems, A Management Perspective*. Ottawa: WDL Publications.

Bolstad, P., Gessler, P., Lillesand, T.M. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, 4:399–412.

Burrough, P.A., Frank, A.U. (1996). *Geographical Objects with Indeterminate Boundaries*. London: Taylor & Francis.

Chrisman, N.R. (1984). The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21:79-87.

Chrisman, N.R. (1987). Efficient digitizing through the combination of appropriate hardware and software for error detection and editing. *International Journal of Geographical Information Systems*, 1:265–277.

DeMers, M. (2000). *Fundamentals of Geographic Information Systems* (2nd ed.). New York: Wiley.

Douglas, D.H., Peuker, T.K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, 10:112–122.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler C. (2002). The National Elevation Dataset. *Photogrammetric Engineering and Remote Sensing*, 68:5–32.

Holroyd, F., Bell, S.B.M. (1992). Raster GIS: Models of raster encoding. *Computers and Geosciences*, 18:419–426.

Joao, E.M. (1998). *Causes and Consequences of Map Generalization.* London: Taylor & Francis.

Laurini, R., Thompson, D. (1992). *Fundamentals of Spatial Information Systems*. London: Academic Press.

Maquire, D.J., Goodchild, M.F., Rhind, D. (Eds.). (1991). *Geographical Information Systems: Principles and Applications*. Harlow: Longman Scientific.

McBratney, A.B., Santos, M.L.M., Minasny, B. (2003). On digital soil mapping. *Geoderma,* 117:3–52.

Muehrcke, P.C., Muehrcke, J.P. (1992). *Map Use: Reading, Analysis, and Interpretation* (3rd ed.). Madison: J.P. Publications.

Nagy, G., Wagle, S.G. (1979). Approximation of polygonal maps by cellular maps. *Communications of the Association of Computational Machinery*, 22:518–525.

Peuker, T.K., Chrisman, N. (1975). Cartographic data structures. *The American Cartographer*, 2:55–69.

Peuquet, D.J. (1984). A conceptual framework and comparison of spatial data models. *Cartographica*, 21:66–113.

Peuquet, D.J. (1981). An examination of techniques for reformatting digital cartographic data. Part II: the raster to vector process. *Cartographica*, 18:21–33.

Shaeffer, C.A., Samet, H., Nelson, R.C. (1990). QUILT: a geographic information system based on quadtrees. *International Journal of Geographical Information Systems*, 4:103–132.

Shea, K.S., McMaster, R.B. (1989). Cartographic generalization in a digital environment: when and how to generalize. *Proceedings AutoCarto 9*, 56–67.

Warner, W., Carson, W. (1991). Errors associated with a standard digitizing tablet. *ITC Journal*, 2:82–85.

Weibel, R. (1997). Generalization of spatial data: principles and selected algorithms. In van Kreveld, M., Nievergelt, J., Roos, T., Widmayer, P., (Eds.). *Algorithmic Foundations of Geographic Information Systems*. Berlin: Springer-Verlag.

Wolf, P.R., Ghilani, C. (2002). *Elementary Surveying, an Introduction to Geomatics* (10th ed.). New Jersey: Prentice-Hall.

Zeiler, M. (1999). *Modeling Our World: The ESRI Guide to Geodatabase Design*. Redlands: ESRI Press.