4 Data Sources and Data Entry

Building a GIS Database

Introduction

Spatial data entry and editing are early and 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. Coordinate entry is often painstakingly slow, as when points are manually entered. Even with automated techniques and the most recent digital data entry methods, spatial data entry and editing take significant time for most organizations.

There are many spatial data sources, and they may be categorized as either hardcopy or digital forms. 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.

These hardcopy data are an important source of geographic information for a number of reasons. First, maps are a valuable record of historical knowledge. Nearly all geographic information produced prior to the 1960s was recorded in hardcopy form. Mapping has at least a 4000-year history. Advances in optics, metallurgy, and industry during the 18th and 19th centuries allowed the mass production of precise surveying devices, and by the mid 20th century, much of the world had been plotted on cartometric quality maps. Cartometric maps are those that faithfully represent the relative position of objects and may be suitable as a source of spatial data. This two-hundred year burst in mapping activity created a substantial archive of coordinate data, much of which merits conversion to digital forms.

Historical and current photographs are also a valuable source of geographic data. Although photographs do not typically provide an orthographic (flat, undistorted) view, they are a rich source of geographic information, and standard techniques may be used to remove major systematic distortions. Surveyor’s notes and coordinate lists may also provide positional information in a hardcopy format that may be entered into a GIS.

Digital forms of spatial data are those provided in a computer-compatible format. These include text files, lists of coordinates, digital images, and coordinate and attribute data in electronic file formats. Digitized data are a very common source of spatial information. These data are often from hardcopy maps that have already been converted to digital formats. Files and export formats may be used to transfer them to a local GIS system. The Global Positioning System (GPS), described in Chapter 5, is a direct measurement system which may be used to record coordinates in the field and report them directly into digital formats. Most modern surveying instruments also may be used to take direct measurements, reduce these measurements to coordinates using integrated computers, and output digital coordinate or attribute data in specific GIS formats.
Finally, a number of digital image sources are available, e.g., satellite or airborne images which are collected in a digital raster format, or hardcopy aerial photographs that have been scanned to produce digital images.

Our objective in this chapter is to introduce the forms, methods, and equipment typically used for spatial data entry. We will also cover basic editing methods and data documentation.

**Hardcopy Maps**

Hardcopy maps and tables were the most common storage medium for spatial data until the widespread adoption of GIS in the 1980s. Prior to this time, nearly all spatial data were collected with the aim of recording the numerical coordinates on paper and/or plotting them on hardcopy maps. Maps were and still are a relatively stable, permanent, familiar, and useful way to summarize spatial data, and because hardcopy maps are a source of so much digital data, most GIS users should be familiar with basic map properties.

Most maps contain several components (Figure 4-1). 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. (Figure 4-1).
Maps often depict coordinate lines (Figure 4-2). When the lines represent constant latitude and longitude, the set of coordinate lines is called the graticule (Figure 4-2a). These lines may appear curved, depending on the map scale, the map projection, and the location of the area on the Earth’s surface. Maps may also depict a grid consisting of lines of constant coordinates. Grid lines are typically drawn in both the x and y directions, and appear straight on most maps (Figure 4-2b).

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, as 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.

Many types of maps are produced, and the types are often referred to by the way features are depicted on the map. Feature maps are among the simplest, in that they map points, lines, or areas and provide nominal information (Figure 4-3, upper left). A road may be plotted with a symbol defining the type of road or a point may be plotted indicating the location of a city center, but the width of the road or number of city dwellers are not provided in the shading or other symbology on the map. Feature maps are perhaps the most common map form, and examples include many road maps, and standard map series such as the 7.5 minute topographic maps produced by the U.S. Geological Survey.

Choropleth maps depict quantitative information for areas. A mapped variable such as population density may be represented in the map (Figure 4-3, top right). Polygons define area boundaries, such as counties, states, census tracts, or other standard administrative units. Each polygon is given a color, shading, or pattern corresponding to values for a mapped variable, e.g., in Figure 4-3, top right, the darkest polygons have a population density greater than 1000 persons per square mile.

Dot-density maps are another map form commonly used to show quantitative data.
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 polygon. Note that the dots are typically placed randomly within the polygon area. Each dot on the map in the lower left of Figure 4-3 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 in Figure 4-3.

Isopleth maps, also known as contour maps, display lines of equal value (Figure 4-3, 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, e.g., a 10°C isopleth defines the position on the landscape that is at that temperature. Lines typically do not cross, in that it cannot be two different temperatures at the same location. However, isopleth maps are commonly used to 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. Note that the isopleths are typically interpolated surfaces and are not measured on the ground.

Figure 4-3: Common hardcopy map types depicting New England, in the northeastern United States.
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. Consider the dot-density map described above. Population is depicted by points, but the points are plotted with random offsets or using some method that doesn’t reflect the exact location of the population within each polygon. In truth the population may be distributed across the polygon in isolated houses and small villages. Dot density maps use a point symbol to represent a value that is aggregated from the entire polygon. If we digitized the point locations of each dot when entering data into a GIS we would record unwarranted positional information in our data. The map should be interpreted correctly, in that the number of dots within a polygon should be counted, this number multiplied by the population per dot, and the population value associated with the entire polygon.

**Map Scale**

Most hardcopy maps have a fixed map scale. The scale is typically reported as a distance conversion, e.g., one inch to a mile, or as a unitless ratio, such as 1:24,000, indicating a unit distance on the map is equal to 24,000 units of that same distance on the Earth’s surface. Maps may also report scale as a bar or line of known distance, labeled on the map.

The notion of large vs. small scale is often confused because scale is often reported as a ratio. 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. For example, a 1:100,000-scale map that is 20 inches on a side covers more ground than a 1:24,000-scale map that is 20 inches on a side. However, it is the size of the ratio or fraction, 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 4-4). It is also helpful to remember that large scale maps show more detail. 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. Thus, these errors which occur during map production are magnified more on a small-scale map than a large-scale map.

Table 4-1 illustrates the effects of map scale on data quality. Errors of one millimeter (0.039 inches) on a 1:24,000-scale map correspond to 24 meters (79 feet) on the surface of the Earth. This same one millimeter error on a 1:1,000,000-scale map correspond to 1000 meters (3281 feet) on the Earth surface. Thus, small errors or intentional offsets in map plotting or printing may cause significant positional errors when scaled to dis-

<table>
<thead>
<tr>
<th>Map Scale</th>
<th>Error (m)</th>
<th>Error (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:24,000</td>
<td>24</td>
<td>79</td>
</tr>
<tr>
<td>1:50,000</td>
<td>50</td>
<td>164</td>
</tr>
<tr>
<td>1:62,500</td>
<td>63</td>
<td>205</td>
</tr>
<tr>
<td>1:100,000</td>
<td>100</td>
<td>328</td>
</tr>
<tr>
<td>1:250,000</td>
<td>250</td>
<td>820</td>
</tr>
<tr>
<td>1:1,000,000</td>
<td>1,000</td>
<td>3,281</td>
</tr>
</tbody>
</table>
1:250,000-scale (smaller scale)

1:100,000-scale (mid-scale)

1:24,000-scale (larger scale)

Figure 4-4: Coverage, relative distance, and detail change from smaller-scale (top) to larger scale (bottom) maps. Note the changes in the depiction of Eureka (arrow), the city at the right center of the top two panels.
Map 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.

These data sources and mapmaking methods will unavoidably set limits on the size and shape of features that may be represented. Consider a lake mapping project based on image data with a 250 meter cell size (Figure 4-5). 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.

A finer resolution source, e.g., with 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 abstraction. 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 or printing a map. These generalizations, depicted in Figure 4-6, may be classed as:

*Simplified*: boundary or shape details are lost or “rounded off”,

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

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

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

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

Generalization is present at some level in every map, and should be recognized and evaluated for each map that is used as a source for data in a GIS (Figure 4-7). Large-scale maps typically cover smaller areas and show more detail. This usually results in less map generalization on large-scale maps, as shown by fewer omissions, less simplification, and fewer fused or aggregated features in larger-scale maps relative to smaller-scale maps. If generalization results in omission or degradation of data beyond acceptable levels, then the analyst or organization 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.

**Map Media**

Most maps have been printed on paper, introducing geometric errors because paper is not a stable medium. Paper is ubiquitous, inexpensive, and easily printed. However paper is most often composed of a mat of wood fibers, and these fibers shrink and swell with changes in humidity. Because many fibers are oriented in a common direction and because fiber shrinkage or expansion is greater longitudinally than laterally, the paper does not change shape uniformly. Under controlled environmental conditions this is often not a significant problem, because deformation is usually quite small. Shrinkage and swelling may at times be significant, particularly under extreme environ-
mental conditions or when the maps are constrained in some manner. For example, when maps are manually converted to digital coordinates they may be taped to the rigid surface of a digitizing board or tablet. Changes in humidity may cause the maps to expand and “pucker”, or rise off the digitizing surface, introducing horizontal uncertainty during digitization. Creases, folding, and wrinkling can also lead to non-uniform deformation of paper maps.

Maps are sometimes printed on dimensionally stable materials to avoid the deformations that occur when using paper. Plastic or other hydrocarbon-based media are sometimes used. These materials are highly resistant to expansion or contraction over broad ranges of temperature and humidity. They typically do not fold easily, although when crushed the materials will retain the deformation, as with paper. The materials are more likely to split or splinter than paper. However stable-base media are not commonly used because of higher costs relative to paper. These media are more expensive to produce per unit area due to lower production volumes. Printing on plastic or other dimensionally stable media is also more expensive because specialized inks and equipment are often required.

Figure 4-7: Examples of map generalization. Portions are shown for three maps for an area in central Minnesota. A large scale (1:24,000), intermediate scale (1:62,500), and small scale (1:250,000) map are shown. **Note that the maps are not shown at true scale to facilitate comparison.** 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.
Map Boundaries and Spatial Data

One final characteristic of maps impacts their use as a source of spatial data: maps have edges, and discontinuities often occur at these edges. Large-scale, high-quality maps generally cover small areas. This is because of the trade-off between scale and area coverage, and because of limits on the practical size of a map. Cartometric maps larger than a meter in any dimension have proven to be impractical for most organizations. Maps above this size can be produced by relatively few printers, they are difficult to store without folding, and it is often difficult to find a flat surface that is large enough to display or unroll them. Thus, human ergonomics set a practical limit on the physical size of a map.

The fixed maximum map dimension when coupled with a fixed map scale defines the area coverage of the map. Larger scale maps generally cover smaller areas. A 1:100,000-scale map that is 18 inches (47 centimeters) on a side spans approximately 28 miles (47 kilometers). A 1:24,000-scale map that is 18 inches on a side represents 9 miles (15 kilometers) on the Earth surface. Because spatial data in a GIS often span several large-scale maps, these map boundaries may occur in the area of the spatial database. Problems may arise at these boundaries when maps are entered into a spatial database, or when maps of different scales are combined.

Differences in the time of data collection for adjacent map sheets may lead to inconsistencies across the border. Landscape change through time is one major source of differences across map boundaries. For example, the U.S. Geological Survey has produced 1:24,000-scale map sheets for all of the lower 48 United States of America. The original mapping took place over several decades, and there were inevitable time lags between mapping some adjacent areas. As much as two decades passed between mapping or updating adjacent map sheets. Thus, many created features, such as roads, canals, buildings, or municipal boundaries are discontinuous or inconsistent across map sheets.

Different interpreters may also cause differences across map sheet boundaries. Large-area mapping projects typically employ several interpreters, each working on different map sheets for 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, discussed later in this chapter, is the process of converting 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 map sheet.

Spatial data stored in a GIS are not bound by the same constraints that limit the physical dimensions of hardcopy maps. Digital storage allows seamless digital maps of large areas. However, the inconsistencies that exist on hardcopy maps may be transferred to the digital data. Inconsistencies at map sheet edges need to be identified and resolved when maps are converted to digital formats.
Digitizing: Coordinate Capture

Digitizing is the process by which coordinates from a map, image, or other sources are converted into a digital format in a GIS. Points, lines, and areas on maps or images represent real-world entities or phenomena, and these must be recorded in digital forms before they can be used in a GIS. The 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 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. Digitizing software allows the operator to trace the points, lines, or polygons that are identified on the scanned map (Figure 4-8). Digitizing software allows the human operator to specify the type of feature to be recorded, the extent

Figure 4-8: 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.
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 affect an action. On-screen digitizing may be used for recording information from scanned aerial photographs, digital photographs, satellite, or other images.

**Hardcopy Map Digitization**

*Hardcopy digitizing* is human-guided coordinate capture from a map printed on paper, plastic, or other “hardcopy” material. An operator securely attaches a map to a digitizing surface and traces lines or points with an electrically sensitized puck (Figure 4-9). The puck typically has cross-hairs and multiple input buttons. When a button is pressed, a signal is sent to the digitizing device to record a coordinate location. Points are captured individually. Line locations are recorded by tracing over the line, capturing coordinate locations along the line at frequent intervals so that the line shape is faithfully represented. Areas are identified by digitizing the coordinates for all bounding lines.

The most common hardcopy digitizing devices are digitizing tables or digitizing tablets (Figure 4-10). The digitizing table typically has a hard, flat surface, although portable digitizing mats are available which are flexible enough to roll up and easily transport. Digitizer designs have employed several types of electrical and mechanical input devices, however most common designs 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. The location of the puck at the time the button is pressed is determined and sent to the computer to be recorded. Often there are buttons to erase the last digitized point or perform other editing functions. Digitizing tables may be quite accurate, with a resolution of between 0.25 and 0.025 millimeters (0.01 and 0.001 inches). If a puck is held stationary and points captured repeatedly, they will differ by less than this resolution.

During hardcopy digitization the map is securely fixed to the digitizing surface so that it will not move during digitizing. Maps may be taped to the surface, usually attached at each corner and each edge. Typically, one corner is taped and the map smoothed by hand; a slight downward pressure is applied to the map while moving the hand from the taped edge to the opposite corner. This opposite corner may then be taped, and the map smoothed from the middle outwards to the remaining opposing corners. Opposing edges are then taped in a similar manner. This taping sequence ensures a secure map surface, important because even small shifts of the map during digitizing can result in large errors that are difficult to remove. As an example, consider the error introduced with a shift of 2.5 millimeters (0.1 inches) while digitizing a 1:100,000-scale map. This shift would result in an error equal to 250
meters (800 feet) measured on the Earth surface. The Earth surface coordinate error would be less for an equivalent map shift when digitizing from a larger scale map, however it may still be quite large even when digitizing from a 1:24,000-scale map or larger, underscoring the need to firmly fix the map to the digitizing surface.

Digitizing tablets may include a mechanism for securing the map. Some tablets are built with a transparent plastic mat attached along an edge of the digitizing surface. The plastic mat may be lifted off of the surface, the map placed on the tablet, and the mat placed back down into position. The mat then holds the map securely to the surface. Other digitizing tablets are built with a dense pattern of small perforations in the table surface. A pump creates a partial vacuum just below the tablet surface. This pressure causes a suction at each perforation, pulling the map down onto the digitizing table, and ensuring the map does not move during digitizing.

**Characteristics of Manual Digitizing**

While manual digitizing can be slow, labor intensive, tedious, and inconsistent among human operators, manual digitizing, from either a scanned map on a computer screen, or using a digitizing table, is among the most common methods for hardcopy data entry. There are many reasons for this. Manual digitizing provides sufficiently accurate data for many, if not most, applications. Manual digitizing with precision digitizing equipment may record data to at least the accuracy of most maps, so the equip-
ment, if properly used, does not add substantial error. Manual digitizing also requires lower initial capital outlays than most alternative digitizing methods, particularly for larger maps. Not all organizations can afford the high cost of precise, large-format map scanners; or they may not digitize enough maps to justify the cost of purchasing such a scanner. Another limitation is the condition of the source material. Humans are usually better than machines at interpreting the information contained in faded, multicolor, or poor quality maps. Finally, manual digitizing is often best because short training periods are required, data quality may be frequently evaluated, and digitizing tablets are commonly available. For these reasons manual digitization is likely to remain an important data entry method for some time to come.

There are a number of characteristics of manual digitization that may negatively affect the positional quality of spatial data. As described earlier, map scale impacts the spatial accuracy of digitized data. Data collected from small-scale maps typically contain larger positional errors than data collected from large-scale maps.

Equipment characteristics also affect data accuracy. There is an upper limit on the precision of each digitizing tablet, and tablet precision reflects the digitizer resolution. Precision may be considered the minimum distance below which points cannot be effectively digitized as separate locations. The precision is often reported as repeatability: how close points are clustered when the digitizing puck is not moved. Although these points should be placed at the same location, many are not. There will be some variation in the position reported by the electronic or mechanical position sensors, and this affects the digitizing accuracy.

Both device precision and map scales should be considered when selecting a digitizing tablet. 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 (the “operator”) may also affect the geometric quality of manually digitized data. Operators vary in their visual acuity, steadiness of hand, attention to detail, and ability to concentrate. Hence, some operators will more accurately capture the coordinate information contained in maps. The abilities of any single operator will also vary through time, due to fatigue or difficulty maintaining focus on a repetitive task. Frequent breaks from digitizing, comparisons among operators, and quality and consistency checks should be integrated into any manual digitization process to ensure accurate and consistent data collection.

The combined errors from both operators and equipment have been well-characterized and may be quite small. One test using a high-precision digitizing table revealed digitizing errors averaging approximately 0.067 millimeters (Figure 4-11). Errors followed a random normal distribution, and varied significantly among operators. These average errors translated to

![Figure 4-11: 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)]
approximately 1.6 meter error when scaled from the 1:24,000 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.

On-screen digitizing offers advantages over both hardcopy digitizing and scan digitizing. Manual map digitization is often limited by the visual acuity and pointing ability of the operator. The pointing precision of the operator and digitizing systems translates to a fixed ground distance when manually digitizing a hardcopy map. For example, consider an operator that can reliably digitize a location to the nearest 0.4 millimeters (0.01 inch) on a 1:20,000-scale map. Also assume the best digitizer available is being used, and we know the observed error is larger than the error in the map. The 0.4 millimeter precision translates to approximately 8 meters of error on the Earth surface. The precision cannot be appreciably improved using manual digitization alone, because a majority of the imprecision is due to operator abilities.

In contrast, once the map is scanned, the image may be displayed on the screen at any map scale. The operator may zoom to a 1:5,000-scale or greater on-screen, and digitizing improved. While other factors remain that limit the accuracy of the derived spatial data (for example map plotting or production errors, or scanner accuracy), on-screen digitizing may be used to limit operator-induced positional error when digitizing.

On-screen digitizing also removes or reduces the need for a digitizing table. Large digitizing tables are an additional piece of equipment and require significant space. Digitizing tablets are specialized for a single use. Operator expenses are typically higher than hardware costs; if manual digitizing is an infrequent activity, the cost of digitizing hardware per unit map digitized may be significant. High quality scanning equipment is quite expensive, however maps may be sent to a third-party for scanning at relatively low cost.

**The Digitizing Process**

Manual digitizing involves placing a map on a digitizing surface or displaying a map on screen, 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 in this manner the locations and shapes of all required map features defined. Features that are viewed as points are represented by digitizing a single location. Lines are represented by digitizing an ordered set of points, and polygons by digitizing a connected set of lines. Lines have a starting point, often called a starting node, a set of vertices defining the line shape, and an ending node (Figure 4-12). Hence, lines may be viewed as a series of straight line segments connecting vertices and nodes.

Digitizing may be in point mode, where the operator must depress a button or otherwise signal to the computer to sample each

Figure 4-12: Nodes define the starting and ending points of lines. Vertices define line shape.
point, or in *stream mode*, where points are automatically sampled at a fixed time or distance frequency, e.g., once each second. Stream mode is not appropriate when digitizing point features, because it is usually not possible to find and locate points at a uniform rate. Stream mode may be advantageous when large numbers of lines are digitized, because points may be sampled more quickly and there may be less operator fatigue. The stream sampling rate must be specified with care to avoid over- or undersampled lines. Too rapid a collection frequency results in redundant points not needed to accurately represent line or polygon shape (Figure 4-13). Too slow a collection frequency in stream mode digitizing may result in the loss of important spatial detail. In addition, when using time-triggered stream digitizing the operator must remember to continuously move the digitizing puck; if the operator rests the digitizing puck for a period longer than the sampling interval there will be multiple points clustered together. These will redundantly represent a portion of the line and may result in overlapping segments. Pausing for an extended period of time often creates a “rats nest” of lines that must later be removed. Point mode digitizing allows the operator to specify the location of each point, vertex, and node, and hence precisely control the sampling frequency.

*Minimum distance* digitizing is a variant of streaming mode that avoids some of the problems inherent with simple streaming. Minimum distance digitizing is similar to stream mode digitizing in that points are collected automatically as the operator traces along a line or polygon boundary. Minimum distance differs from streaming mode in that 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 rats 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 cause problems when utilizing the data. These errors may 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 4-14 shows some common digitizing errors.

**Undershoots and overshoots** are common errors that occur when digitizing. Undershoots are nodes that do not quite reach line or another node, and overshoots are lines that cross over existing nodes or...
lines (Figure 4-14). Undershoots cause unconnected networks and unclosed polygons in the examples cited above (Figure 4-14). 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 relies on a snap tolerance or snap distance. This distance may be interpreted as a minimum distance, within which nodes or vertices are considered to occupy the same location (Figure 4-15). Node snapping prevents a new node from being placed within the snap distance of an already existing node; instead, the node is joined or “snapped” to the existing node (Figure 4-15). 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. Line snapping may also be specified. Line snapping inserts a node at a line crossing and clips the end when a small overshoot is digitized. Line snapping forces a node to connect 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. The snap process places a new node at the intersection point, and connects the digitized line to the existing line at the intersection point. 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 when node and line snapping are in force.

The snap distance must be carefully selected for snapping to be effective. If the snap distance is too short, then snapping has little impact. Consider a system where the operator may digitize with 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 very few nodes will be within the snap tolerance, and snapping has little effect. Another problem comes from setting the snap tolerance too large. If

![Figure 4-14: Common digitizing errors.](image-url)
the snap tolerance in our previous example is set to 100 meters, and we want the data accurate to the nearest 5 meters, then we may lose significant spatial information that is contained in the hardcopy map. Lines less than 100 meters apart cannot be digitized as separate objects. Many features may not be represented in the digital data layer. The snap distance should be smaller than the desired positional accuracy, and such that significant detail contained in the digitized map is recorded. The snap distance should not be below the capabilities of the system and operator 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 connected set of polynomial functions with constraints between functions (Figure 4-16). 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 4-16). A constraint may be placed that these functions connect smoothly, usually by requiring the first and second derivatives of the functions 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 interpolated on the line between digitized vertices 8 and 9, leading to the “smooth” curve shown in Figure 4-16.
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 of an issue 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 4-17). The Lang method exemplifies this approach. A spanning line connects two

---

Figure 4-16: Spline interpolation to smooth digitized lines.

Figure 4-17: The Lang algorithm, a common line-thinning method. Vertices are removed, or thinned, when they are within a weed distance to a spanning line. Thinned points are shown as open circles (adapted from Weibel, 1997).
non-adjacent vertices in a line. A pre-determined number of vertices is spanned initially. The initial spanning number has been set to 4 in the Figure 4-17, 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 an endpoint that is the 4th point down the line (Figure 4-17a). Any intermediate points that are closer than the weed distance are marked for removal. In Figure 4-17a no points are within the weed distance, therefore none are marked. The endpoint is then moved to the next closest remaining point (Figure 4-17b), and all intermediate points tested for removal. Again, any points closer than the weed distance are marked for removal. Note that in Figure 4-17b 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 drawn to connect 4 points (Figure 4-17c, d). The process may be repeated for successive sets of points in a line segment until all vertices have been evaluated (Figure 4-17e 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 are 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).

Global methods, such as the Douglas-Peuker algorithm, begin by using all vertex points in a test (Figure 4-18). The Douglas-Peuker method splits a line recursively until lines can get no smaller. Initially, the first and last points in a line are spanned (Figure 4-18a), and if all intermediate points are

![Figure 4-18: The Douglas-Peuker line-thinning method recursively splits line segments until all intervening points are within the weed distance, or there are no intermediate points (adapted from Weibel, 1997).](image)
within the weed distance, then they are all deleted. If there is at least one point further than the weed distance, the line is split into two lines (Figure 4-18b). The split occurs at the point that is farthest away from the initial spanning line. The algorithm is then applied to the two new line segments (Figure 4-18c). Spanning lines are drawn from the first to last points for each new line segment. Again, the rule is applied: if all points for a new line segment are within the weed distance, delete them, or else split this new line into two smaller lines. This process is repeated for each succeeding smaller line segment until each line segment can get no smaller (Figure 4-18d to f).

**Scan Digitizing**

Optical scanning is another method to convert hardcopy documents into digital formats. Scanners have light emitting and sensing elements. Most scanners pass a sensing element over the 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. Reflectance values are converted to numbers.

A threshold is often applied to determine if the sensed point is part of a feature to be recorded. For example, a map may consist of dark lines on a white background. A threshold might be set such that if less than 10% of the light striking the map is returned to the sensor, the sensed point is considered part of a line. If 10% or more of the energy is reflected back to the sensor, the point is considered part of the white space between lines. The scanner then produces 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.

Most scanners are either bed or drum type. Bed scanners provide a flat surface on which the map is placed (Figure 4-19). A mat or hinged cover is then placed on top of the map, flattening and securing the map to the bed. On some bed scanners an optical train is passed over the map, emitting light and sensing the light reflected back from the map. Sensing arrays are typically used to measure the reflectance so that one to several rows of cells may be scanned simultaneously. A motor then moves the optical train to the adjacent lines and the process is repeated. Positional accuracy depends on an optical device in one direction (along the sensing array) and on a mechanical device in the other direction (as the optical head travels down the scanning bed). Positional accuracy is generally greater in the direction of the sensing array. Significant distortion may be introduced by the motor or other drive mechanism, due to variations in the scanning speed or orientation of the optical train. These errors must be considered when selecting a flatbed scanning device.

Drum scanners differ from flatbed scanners in that they employ a rotating cylinder. A map is fixed onto the surface of this cylinder, and the cylinder set to rotate at a uniform velocity. The angular velocity of a rotating cylinder is easier to control than the straight-line motion of a bed scanner, so many of the early high-precision scanners used drums. Many drum scanners are similar to bed scanners in that they use optical detection of reflected light to sense map elements. Alternative designs are also available, using light sensitive lines or arrays on a drum. Sensors are pointed at the rotating drum and moved down the drum in a set of

---

**Figure 4-19**: A map scanner (courtesy Calcomp).
steps. Output is sensed and stored in a raster data set as with flatbed scanners.

Scanners work best when very clean map materials are available. Even the most expensive scanners may report a significant number of spurious lines or points when old, marked, folded, or wrinkled maps are used. These spurious features must be subsequently removed via manual editing, thus negating the speed advantage of scanning over manual digitizing. Scanning also works best when maps are available as map separates, with one thematic feature type on each map. Editing takes less time when maps do not contain writing or other annotation. Strongly contrasting colors are preferred, e.g. black lines on a white background, rather than dark grey on light grey. Finally, scanning is most advantageous when a large number of cartographic elements is found on the maps.

Scan digitization usually requires some form of skeletonizing, particularly if the data are to be converted to a vector data format. Scanned lines are often wider than a single pixel (Figure 4-20). 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, missing or extra points or labels are all errors that must be corrected. Spatial data may 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 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.

Error identification is the first step in editing. Errors may be identified by printing a map of the digitized data and verifying that each point, line, and polygonal feature is present and correctly located. Plots are often printed both at a similar scale and a significantly larger scale than the original source materials. The large-scale plots are often paneled with some overlap among panels. Plots at scale are helpful for identifying
missing features, and large-scale plots aid in identifying undershoots, overshoots, and small omissions or additions. Annotations are made on these plots as they are checked systematically for each feature.

Software may aid in identifying potential errors. Line features typically begin and end with a node, and 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, e.g., a cul-de-sac in a street network, while others will be the result of under- or overshoots. Dangling nodes that are plotted with unique symbols may be quickly evaluated, and if appropriate, corrected.

Attribute consistency may also be used to identify errors. This technique identifies 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 pre-defined precedence hierarchy.

Complete GIS software packages provide a comprehensive set of editing tools (Figure 4-21). 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...)

**Figure 4-21**: GIS software provide 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).
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 erasing 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 selected by dragging interactively on the screen to match point locations. All lines and points except the anchor points are interactively adjusted. One common application involves adjusting linework representing cultural features, such as a road network, when higher geometric-accuracy photo or satellite image data are available. The linework is overlain on an image backdrop and subsequently adjusted.

These edits should be made with due attention to the magnitude of positional change introduced 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 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. These features rarely have identical locations on each map, and often occur in different locations when digitized into their respective data layers (Figure 4-22). For example, water boundaries on soil survey maps rarely correspond to water boundaries found on USGS topographic quads. Features may be different 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 may compound the problem due to differences among layers in digitizing methods or operators.

There are several ways to remove this “common feature” inconsistency. One involves re-drafting the data from conflicting sources onto one base map. Inconsistencies are removed at the drafting stage. For example, vegetation and roads data may show stand boundaries at road edges that are

![Figure 4-22: Common features may be spatially inconsistent in different spatial data layers.](image-url)
inconsistent with the road locations. Both of these data layers may be drafted onto the same base, and the common boundaries fixed by a single line. This line is digitized once, and used to specify the location of both the road and vegetation boundary when digitizing. Re-drafting, although labor intensive and time consuming, forces a resolution of inconsistent boundary locations. Re-drafting also allows the combination of several maps into a single data layer.

A second, often preferable method involves establishing a "master" boundary which 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 quad 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 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.

Coordinate Surveying

While maps are a common source of digital data, most of the maps were themselves developed, at least in part, from coordinate surveys. Spatial data layers may be produced directly from field surveys, or from field surveys combined with measurements on aerial photographs. These survey measurements have historically been performed with optical or optical-electronic devices such as transits, theodolites, and electronic distance meters (Figure 4-23).

Surveying is particularly common when a highly valued data layer is to be developed or when very precise coordinates are required. Property lines are a good example of data that are often required to be of very high positional accuracy. Real estate in upscale markets may be valued at hundreds to thousands of dollars per square foot, and buildings, pools, or other improvements substantially more expensive. Zoning ordinances often specify setbacks, minimum distances between improvements and property boundaries. These factors raise the importance of precisely locating property lines, and justify precise and expensive coordinate surveys (Figure 4-24). Other commonly surveyed features include power lines, fiber-optic cables, sewers, pipelines, and other high valued and expensive utilities.
Plane surveying is horizontal surveying based on a flat surface. The assumption of a flat surface provides a significant computational advantage; the mathematics used to calculate positions in plane surveys are substantially less complicated than those required for geodetic surveys. The flat surface in a plane survey is usually defined by a map projection, with a known point serving as the starting location for the survey. Because map projected coordinates are used and an orthographic surface is assumed, the Earth’s surface is considered to be flat, with plumb lines perpendicular to the surface at all points in the survey. This means a plumb bob or weight suspended from a string is assumed to hang in a vertical direction and intersect the plane surface at a 90-degree angle. This is a valid assumption when the errors inherent with ignoring the Earth’s curvature are small compared to the accuracy requirements of the survey or to the errors inherent in the survey measurements themselves. The distance error due to assuming a flat rather than curved surface over 10 kilometers (6 miles) is 0.72 centimeters (0.28 inches). Therefore, plane surveys are typically restricted to distances under a few tens of kilometers. This restriction has affected relatively few surveys, and a substantial majority of the lines and points surveyed to date have been measured using plane surveying methods. Plane surveying is sufficient for most subdivisions, public works, construction projects, and property surveys.

Historically, plane surveys have been conducted with optical instruments similar to those described for geodetic surveys. These instruments typically have angle gauges in the horizontal and vertical planes and an optical sight, usually with some degree of telescopic magnification. The instruments go by a number of different names, including, in increasing order of sophistication and capabilities, a Dumpy level, a transit, a theodolite, and a total station.

![Figure 4-24](image-url): Surveying has been used to establish the coordinates for most property lines. Field measurements of distance and direction are used to establish the set of vertices that define property boundary lines. This is the only way to collect these data, as the features are not visible on any other source.
Distance and angle measurements are the primary field activities in plane surveying. Most surveys have been conducted as **traverses**, a series of connected lines that have a marked beginning and ending point. Traverses typically start at a known control point, or start at a point that has been referenced to a known control point. As described in the preceding sections, the control points are often part of a geodetic control network, or part of a subnetwork established by a municipal surveyor. A distance and angle may is measured from the control point to the first station, or surveyed location. Coordinate geometry may be used to calculate the station coordinates. Subsequent distance and angle measurements may be taken, and in turn used to calculate the coordinates of subsequent stations. A traverse may be open, with a different beginning and ending point, or closed with the traverse eventually connecting back to the starting location. Most of the millions of miles of property lines in North America have been established via plane surveys of open and closed traverses.

A common output from survey measurements, known as coordinate geometry or COGO, consists of a starting point (a station) and a list of directions (bearings) and distances to subsequent stations. The COGO defines a connected set of points from the starting station to each subsequent station. An example COGO description follows:

"The starting point is a 1-inch iron rod that is approximately 102.4 feet north 43.1 feet west of the northeast quarter of the southeast quarter section of section 16 of Township 24 North, Range 16 East, of the 2nd Principal Meridian. Starting from the said point, thence 102.7 feet on a bearing north 72.3 degrees east, to a 1-inch iron pipe; thence 429.6 feet on a bearing south, 64.3 degrees east to a 2-inch iron pipe....."

Basic trigonometric functions may be used to calculate the coordinates for each survey station. These stations are located at the vertices that define lines or areas of interest. In the past these distance and bearing data were manually plotted onto paper maps. Most survey data are now transferred directly to spatial data formats from the surveying instrument or associated software.

Field measurements may be directly entered and coordinate locations derived in the GIS software, or the coordinate calculations may be performed in the surveying instrument first. Many current surveying instruments contain an integrated computer and provide for digital data collection and storage. Coordinates may be tagged with attribute data in the field, at the measurement location. These data are then downloaded directly from a coordinate measuring device to a computer. Specialized surveying programs may be used for error checking and other processing. Many of these surveying packages then output data in formats designed for import into various GIS software systems.

Survey methods vary, but plane surveying used in developing coordinate geometry input are often based on measuring distances and angles for a set of points. These points are usually connected in a traverse, a combined set of distance and bearing angle measurements between traverse stations.

COGO calculations are illustrated on the left of Figure 4-25. Starting from a known coordinate, \(x_0, y_0\), we measure a distance \(L\) and an angle \(\theta\). We may then calculate the distances in the \(x\) and \(y\) directions to another set of coordinates, \(x_1\) and \(y_1\). The coordinates of \(x_1\) and \(y_1\) are obtained by addition of the appropriate trigonometric functions. COGO calculations may then be repeated, using the \(x_1\) and \(y_1\) coordinates as the new starting location for calculating the position of the next traverse station.

The right side of Figure 4-25 shows a sequence of measurements for a traverse. Starting at \(x_s, y_s\), the distance \(A\) and bearing angle, here 45°, are measured to station \(x_m, y_m\). The bearing and distance are then measured to the next station, with coordinates \(x_n, y_n\). Distances and angles are measured for all subsequent stations. Starting with the known coordinates at the starting station, \(x_s\),
coordinates for all other stations are calculated using COGO formulas.

Past survey records of bearings and distances may serve as a source of COGO input. Most of these measurements were recorded for property boundary locations, and are stored in notebooks, on deeds, and in plat books. These paper records must be converted to an electronic format prior to conversion to the coordinate locations. As described above, trigonometric functions may then be used to calculate the position of each station.

The direct entry of COGO data, when available and where practical, will usually lead to more accurate digital databases than the digitization of cartometric maps. Spatial errors, approximations, and positional uncertainty are introduced during the transition from survey measurement to hardcopy maps. Limits on plotting precision, printing alignment and distortion, deformation in the paper or other media, map generalization, and other factors compound to add uncertainty to the plotted locations on maps. Manual or automated digitizing may introduce additional error. However the survey data, converted to the coordinate geometry, contain none of these errors.

Another form of direct coordinate data entry exists, and is known as the Global Positioning System (GPS). Satellite-based measurements are used to determine positions and these may become part of the spatial data in a GIS. The GPS system is described in Chapter 5.
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. 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 coordinate system to a standard map coordinate system (Figure 4-26). The input coordinate system is usually based on the digitizer or scanner-assigned values. A hardcopy map may be taped to a digitizing table and coordinates recorded as a puck is moved across the map surface. These coordinates are usually recorded in pixel, inch or centimeter units relative to an origin located near the lower left corner of the digitizing table. The absolute values of the coordinates depend on where the map happened to be placed on the table prior to digitizing, but the relative position of digitized points does not change as long as the map is not deformed or moved. 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.

Source coordinate system

Target coordinate system

Figure 4-26: Control points in a coordinate transformation. Control points are used to guide the transformation of a source or input set of coordinates to a target or output set of coordinates. There are five control points in this example. Corresponding positions are shown in both coordinate systems.
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. 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 the digitizing tablet or photo point. Control points differ from other digitized points in that we know the map projection coordinates for these points as well as the digitizer coordinates.

These two sets of coordinates for each control point, one for the map projection and one for the digitizer system, are used to estimate a coordinate transformation. Control points 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 may 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 applied to convert all digitized data to map projection coordinates.

Control points should meet or exceed several criteria. Control points should be from a source that provides the highest feasible coordinate accuracy. Control point accuracy should be at least as good as the desired overall positional accuracy required for the spatial data. Control points should be as evenly distributed as possible throughout the data area. A sufficient number of control points should be collected. The minimum number of points depends on the mathematical form of the transformation, but additional control points above the minimum number are usually collected; this usually improves the quality and accuracy of the statistically-fit transformation functions.

The x, y (horizontal), and sometimes z (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. Sub-centimeter 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.

Control Point Sources: Surveying

Traditional ground surveys based on optical surface measurements are a common, although decreasingly used method for determining control point locations. Modern surveys use complex instruments such as transits and theodolites to precisely measure the relative location of points. If the survey starts from a known point, then the coordinate location of any survey station may be determined via simple trigonometric func-
Federal, state, county, and local governments all maintain a set of accurately surveyed locations (Figure 4-27), and these points may be used as control points or as starting points for additional surveys. Most of these known 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, and hence 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 and 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. Even when there is a sufficient number of ground-surveyed points in an area, many may not be suitable for use as control points in a coordinate transformation of spatial data. The control points may not be visible on the maps or images to be registered. For example, a surveyed point may fall along the edge of a road. If the control point is at a mapped road intersection, we may use the easting and northing coordinates of the road intersection as a control point during map registration. However if the surveyed point is along the edge of a road that is not near any mapped feature such as a road intersection, building, or water tower, then it may not be used as a control point. Our control points must have two characteristics to be useful: first, the point 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. Hence, we must often obtain additional control points.

One option is to perform additional surveys that measure the coordinates of visible features on the source materials. Precise surveys are used to establish the coordinate locations of a well-distributed, sufficient set of points throughout the area covered by the source map. While used sparingly because field surveys are often quite expensive, new surveys are chosen when the highest accuracies are required. Costs were prohibitive with traditional optical surveying methods, however, new positioning technologies are allowing more frequent, custom collection of control points.

**GPS Control Points**

The global positioning system (GPS) is a relatively new technology that allows us to establish control points. GPS, discussed in detail in Chapter 5, may help us obtain the coordinates of control points that are visible on a map or image. GPS is particularly useful for determining control point coordinates.
because we may quickly survey widely-spaced points. GPS is often preferred because it is typically faster and often less expensive than traditional surveying methods. GPS positional accuracy depends on the technology and methods employed; it typically ranges from sub-centimeter (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 GPS technologies.

**Control Points from Existing Maps and Digital Data**

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 (Figure 4-28). 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.

Registered digital image data are a common source of ground control points, particularly when natural resource 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 the multiplication of row and column number by the pixel size, and the application of 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.

Finally, 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. The coordinates of any identifiable feature may be obtained from the digital data layer. For example, the USGS has produced Digital Raster Graphics (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. Digital line graph data (DLGs) are also available that contain vector data, and control points may be identified at road intersections and other identifiable features.

Control points are used in coordinate transformation, irrespective of source. Typically a number of control points are identified for a study area. The x and y coordinates for control points are obtained from the digitized map, and the map projection coordinates, E and N, are determined from survey, GPS, or other sources (Figure 4-30). These coordinate pairs are then used with a transformation to convert data layers into a desirable map coordinate system.

There are several different types of coordinate transformations.
Figure 4-29: Potential control points, indicated here by arrows, may be extracted from digital reference images. Road intersections or other permanent, well-defined features are identified and coordinates determined from information provided with the digital image. Note the white cross, circled in the lower right corner. This is a photogrammetric panel, typically a plastic or painted wooden target placed prior to photo capture, and with precisely surveyed coordinates. These targets are used to create the systematically corrected digital image with a known coordinate system, a process described in Chapter 6. These digital images may then serve as a source of control for subsequent photographs.

Figure 4-30: An example of control points locations from a road data layer, and corresponding digitizer and map projection coordinates.
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:

\[ E = T_E + a_1 x + a_2 y \] (4.1)
\[ N = T_N + b_1 x + b_2 y \] (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 \( T_N \) are translation changes between the coordinate systems, and can be thought of as shifts 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 one coordinate system and the next. The affine is the most commonly applied coordinate transformation because it provides for these three main effects of translation, rotation, and scaling when converting from a digitizer to map coordinates, and because it requires relatively fewer control points and often introduces less error than higher-order polynomial transformations.

The affine system of equations has six parameters to be estimated, \( T_E, T_N, a_1, a_2, b_1, b_2 \). 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:

\[ 500,083.4 = T_E + a_1(103.0) + a_2(-100.1) \] (4.3)
\[ 4,903,683.5 = T_N + b_1(103.0) + b_2(-100.1) \] (4.4)

We cannot find a unique solution to this system of 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 this system. Each control point gives us two equations, one for the easting and one for the northing, so we need a minimum of three control points to estimate the parameters of an affine transformation. If we wish to use a statistical technique to estimate the transformation parameters, we need at least one additional control point, for 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:

\[ \text{RMSE} = \sqrt{\frac{e_1^2 + e_2^2 + \ldots + e_n^2}{n}} \] (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:

\[ e = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2} \] (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 4-31 shows examples of this lack of fit. Individual residuals may be observed at each control point location. The location of the control points predicted from the transformation equations do not equal the true locations of the control points, and this difference is the...
residual or positional error at that control point.

A statistical method for estimating transformation equations is preferred because it provides us with an indication of the quality of the transformation. Measurement and digitizing uncertainty introduce unavoidable spatial errors into the control point coordinate values. These uncertainties result in (hopefully) small differences in the transformed and true coordinates of our control points. The RMSE provides a summary of the difference between the “true” (measured) and predicted control point coordinates. It provides one index of transformation quality. Transformations are often fit iteratively, and control points with large errors inspected, corrected, or deleted until an acceptable RMSE is obtained (Figure 4-32). The RMSE will 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 will be an index of transformation accuracy, and a lower RMSE generally indicates a more accurate affine transformation.

Estimating the coordinate transformation parameters is often an iterative process for a number of reasons. First, the control point x and y coordinates may not be precisely digitized. Manual digitization requires the operator place a pointer at the control point and record the location. Poor eyesight, a shaky hand, fatigue, lack of attention, misidentification of the control location, or a blunder may result in erroneous x and y coordinate values. Control point locations may not be accurately or precisely represented on the map, for example when road intersections are used as control and wide road symbols or offsets are applied. There may also be uncertainties or errors in the E and N coordinates, all of which will introduce error into the transformation. Typically, control points are entered, the affine transformation parameters estimated, and the overall RMSE and individual point E and N errors evaluated (Figure 4-31, Figure 4-32). Suspect points are identified and the transformation re-estimated and errors evaluated. This process continues until a satisfactory transformation is fit. The transformation is then applied to all features to convert them from digitizer to map coordinates.

Other Coordinate Transformations

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

\[ E = T_E + cx - dy \]  \hspace{1cm} (4.7)

\[ N = T_N + dx + cy \]  \hspace{1cm} (4.8)

The coefficients \( T_E, T_N, c \) and \( d \) are estimated from control point data. As with the affine, the conformal transformation is also a first-order polynomial. The conformal transformation requires equal scale changes in the \( x \) and \( y \) directions. In the affine, scale
Figure 4-32: Iterative fitting of an affine transformation. The data from Figure 4-30 were used. Four control points were removed in the first two iterations. The final model fit in iteration three meets the \( \text{RMSE} \) criteria. It is better to examine control points with large residuals to determine if the cause for the error may be identified. If so, the control point coordinates may be modified, and the control points retained while fitting the transformation.
changes in the $x$ and $y$ directions can be different. Note the symmetry in the equations 4.6 and 4.7, 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:

\[ E = b_1 + b_2 x + b_4 y + b_5 x^2 + b_6 y^2 + b_7 xy \]  

(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 2nd-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.

The RMSE is typically lower for a 2nd 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. The RMSE is a useful tool when comparing among transformations that have the same model form, e.g., when comparing one affine to another affine as in Figure 4-32. The RMSE is not useful when comparing among different model forms, for example, when comparing and affine to a 2nd-order polynomial. Due to the nature of the statistical fitting process, the RMSE and related measures of fit will decrease as higher order polynomials are used, even if they introduce error. 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 non-linear errors when the transformation is applied to all data in a layer. Thus, higher-order polynomials and other “rubber-sheeting” methods should be used with caution, and the accuracy of the transformation tested with an evaluation that includes independent check points. These check points are withheld when estimating the transformation, and their transformed vs. measured coordinates compared.

### Raster Geometry and Resampling

Data often must be resampled when converting between coordinate systems or changing the cell size (Figure 4-33). 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 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.

Resampling involves re-assigning the cell values when changing raster coordinates or geometry. Cells must be resampled because the new and old raster coordinates represent different areas. Cell centers in the old coordinate system may not coincide with cell centers in the new coordinate system and so the average value represented by each cell must be re-computed.

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).
An example of a bilinear interpolation is shown in Figure 4-34. 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 \( Z_{\text{out}} \), value of the output location is unknown. \( Z_{\text{out}} \) is calculated based on the distances between the output locations and the input locations. The distance in the \( x \) direction is denoted in Figure 4-34 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 \( Z_{\text{out}} \), using the distance \( d_2 \) between the input and output locations to weight values at each input location.

### Map Projection vs. Transformation

Map transformations should not be confused with map projections. A map transformation typically employs a 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, usually from a curved, latitude/longitude coordinate system to a Cartesian coordinate system. No statistical fitting process is used with a map projection.

Transformations should rarely be used in place of projection equations when converting geographic data between map projections. Consider an example when data are delivered to an organization in Univer-

![Figure 4-33: 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.](image)
sal 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 involving projection from UTM to geographic coordinates (latitude and longitude), and then from these geographic coordinates to the appropriate State Plane coordinates.

The software may hide the intermediate geographic coordinates, but most projections among coordinate systems go through an intermediate set of geographic coordinates. This is because we know the forward and inverse conversions to and from latitude/longitude coordinates to the x and y grid coordinates for all analytical map projections, but we do not know the equations to convert from one projection x,y coordinate system to another projection x,y coordinate system. For example, there aren’t any general equations that will allow us to directly compute the a set of state plane coordinates directly from a known set of UTM coordinates. As described in Chapter 3, we must first calculate the geographic, latitudes and longitude coordinates from the UTM coordinates, then calculate the state plane coordinates from these geographic coordinates. This correct method of applying an inverse and then

![Figure 4-34: 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.](image)

\[
\begin{align*}
Z_b &= Z_4 + \frac{(Z_3 - Z_4) \cdot d_1}{c} \\
Z_u &= Z_2 + \frac{(Z_1 - Z_2) \cdot d_1}{c} \\
Z_{out} &= Z_b + \frac{(Z_u - Z_b) \cdot d_2}{c}
\end{align*}
\]

Where is the value of Z_{out}?

\[
\begin{align*}
Z_b &= 1.4 + \frac{(4.6 - 1.4) \cdot 2.9}{5} = 3.26 \\
Z_u &= 4 + \frac{(6 - 4) \cdot 2.9}{5} = 5.16 \\
Z_{out} &= 3.26 + \frac{(5.16 - 3.26) \cdot 2.2}{5} = 4.1
\end{align*}
\]
forward projection rather than a transformation may involve substantial computation because the projection equations may be quite complex, but this path doesn’t require the identification of control points and usually does not introduce additional error.

An alternative approach involves using a linear or polynomial 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 re-project 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 non-linear 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 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.

---

**Metadata: Data Documentation**

Metadata are information about spatial data. Metadata describes 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 extent, coordinate system, quality, and other data characteristics. Metadata aid organizations in evaluating data to determine if they are suitable for an intended use - are they accurate enough, do they cover the area of interest, do they provide the necessary information? Metadata may also aid in planning and executing data updates in that appropriate and compatible methods of new data collection may be more effectively chosen.

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 western-most bounding coordinate of a data set is element 1.5.1.1, defined as follows:

1.5.1.1 West Bounding Coordinate – western-most coordinate of the limit of coverage expressed in longitude.

Type: real

Domain: \(-180.0 \leq \text{West Bounding Coordinate} < 180.0\)

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 numbers for the other bounding coordinates, e.g. 1.5.1.2 is for the eastern-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 in conformance 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 4-35.

Metadata are most often created using specialized software tools. Although metadata 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 redundant information. 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 western-most boundary is west of the eastern-most 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 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 4-35: Example of a small portion of the FGDC recommended metadata for a 1:100,000 scale derived DLG data set.
data 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 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 cross-walk elements between 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. We have a long history of hardcopy map production, so centuries of spatial information are stored there. 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.

There are other common sources of digital spatial data, including COGO and GPS. COGO involves the input of coordinate geometry data, and may be used to build spatial data from surveyor’s measurements of distance and direction. GPS, discussed in-depth in the next chapter, is a system that allows rapid, accurate measurement of coordinates in the field.
Map and other data often need to be converted to a target coordinate system via a map transformation. A transformation is different from a map projection, discussed in Chapter 3, in that a transformation uses an empirical, least-squares process to convert coordinates from one Cartesian systems 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.

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 crosswalk 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


Study Questions

Why have so many digital spatial data been derived from hardcopy maps?

Which is a larger scale map, 1:20,000 or 1:1,000,000?

Can you describe three different types of generalization?

What are the most common map media? Why?

Is media deformation more problematic with large scale maps or small scale maps? Why?

Which map typically shows more detail - a large scale map or a small scale map? Can you give three reasons why?

Why is manual digitization so commonly used?

Can you describe the process of manual digitization from both hardcopy and digital images?

What features are typically digitized? What is the difference between a node and a vertex?

What is snapping in the context of digitizing? What are undershoots and overshoots, and why are they undesirable?

What is a spline, and how are they used during digitizing?

Why is line thinning sometimes necessary? Can you describe a line-thinning algorithm?

Can you contrast manual digitizing to the various forms of scan digitizing? What are the advantages and disadvantages of each?

What is the “common feature problem” when digitizing, and how might it be overcome?

What is COGO?

Can you define the basic trigonometric relationships, and use them to solve for coordinates given a list of distances and bearings?
Can you describe the general goal and process of map registration?

What are control points, and where do they come from?

Can you define an affine transformation, including the form of the equation? Why is it called a linear transformation?

What is the root mean square error (RMSE), and how does it relate to a coordinate transformation?

Is the average positional error likely be larger, smaller, or about equal to the RMSE? Why?

Why are higher order (polynomial) projections to be avoided under most circumstances?

Define and describe metadata. Why are metadata important?

Complete the table below for a traverse with the listed distances and bearings, given as azimuth degrees (drawing a rough sketch may help with the calculations). What is the distance and azimuth from P6 to P0?

**Starting point P0, X = 10,128.3, Y = 6,096.4**

<table>
<thead>
<tr>
<th>Point ID</th>
<th>Azimuth</th>
<th>Distance</th>
<th>Delta X</th>
<th>Delta Y</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>32.4</td>
<td>122</td>
<td>65.4</td>
<td>103.0</td>
<td>10,193.7</td>
<td>6,199.4</td>
</tr>
<tr>
<td>P2</td>
<td>91.7</td>
<td>207</td>
<td>206.9</td>
<td>-6.1</td>
<td>10,400.6</td>
<td>6,193.3</td>
</tr>
<tr>
<td>P3</td>
<td>123.3</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>212.5</td>
<td>193</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td>273.9</td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>355.5</td>
<td>145</td>
<td>-60.1</td>
<td>131.9</td>
<td>10,286.2</td>
<td>6,009.0</td>
</tr>
</tbody>
</table>