

# 2

# SPATIAL DATA MODELS, VECTOR DATA, AND DATA QUALITY

## Introduction

In chapter 1, you learned what public domain means, its effect on various spatial data sources that are increasingly available, and its impact on GIS practitioners. In the accompanying exercise, you downloaded public domain data to tackle global issues surrounding population, climate, and the environment. In this chapter, you will examine some of the main vector datasets available in the public domain and some of the issues associated with their use. The accompanying exercise is based on US Census demographic and street data to determine the best location for an Internet café.

As you saw in chapter 1, solving a problem using GIS and spatial analysis begins with the acquisition and investigation of spatial data. Most analyses involve some base spatial data—roads, hydrology, administrative boundaries, property outlines, and so

on—that will be either central to your analysis or simply reference layers for data you generate yourself. As you work with GIS and spatial data to solve problems and make decisions, it is important to consider the intended use of the data, which influences the format you use. However, before learning more about the plethora of datasets available today, you should consider some issues that affect *all* data—no matter what format, no matter what source.

## Spatial data models

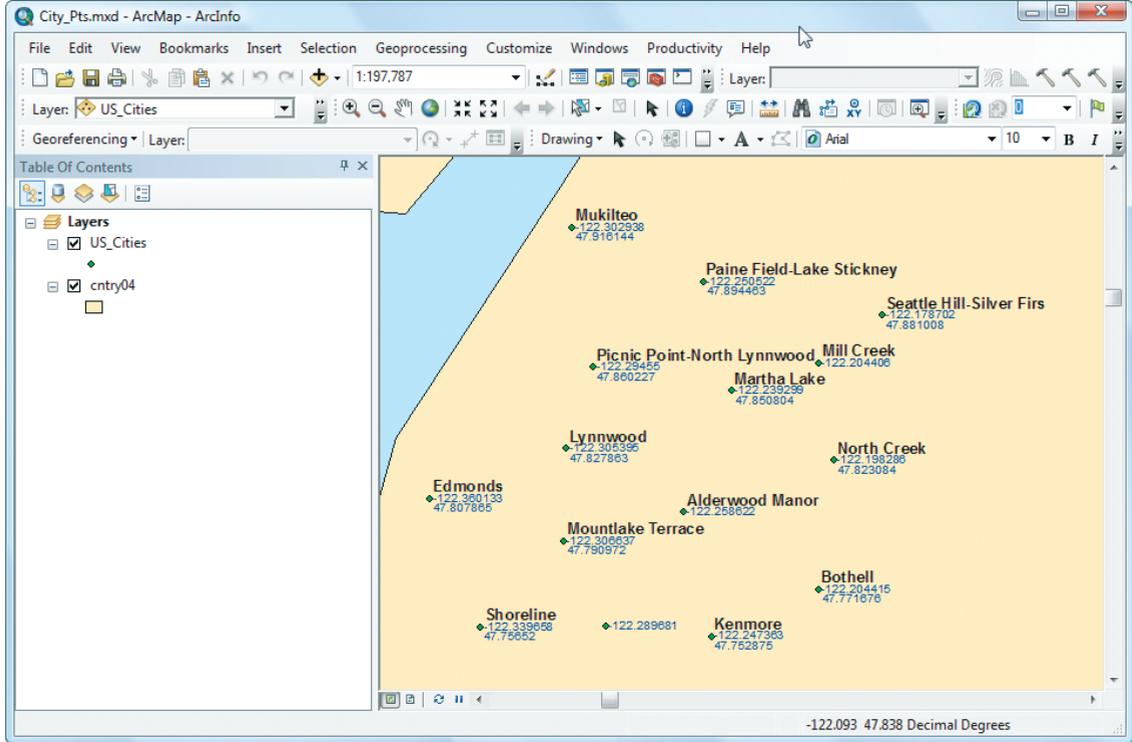
Within a digital environment, graphically representing and interrogating real-world entities, such as school campus zones or the average path of hurricanes as they head west over the Atlantic Ocean, involves storing them within the framework of a spatial data model. The data model's coordinates define the geographic location and extent of each entity. It is this ability to manage location information, in either two- or three-dimensional space, that sets a GIS apart from other information systems. Another unique characteristic of a GIS is that the data model maintains not just the coordinates representing the location of each entity but also the relationships between entities. This is critical in examining issues such as:

- What is the relationship between birth rates and life expectancies between countries?
- How does the underlying geology affect the conductivity of streams?
- How do contour plowing techniques affect soil moisture and erosion on specific soil types and slopes?

Real-world entities like these are generally represented in a GIS using one of two data models—vector or raster.

### Vector data model

Vector data models use discrete elements to represent objects as points, lines, or areas. This allows all positions, lengths, and dimensions to be defined precisely. As well as storing the x,y coordinate information describing the location of each entity, each vector type also stores related attribute information to identify, for example, what type of feature it is and how it should be displayed. In the vector model, collections of similar features (a feature layer) are stored as vector points, vector lines, or vector areas. A single



Vector point data: Cities, Washington State, United States. Each city has a unique latitude-longitude coordinate value, shown in blue.

feature layer never includes all three data types because individual entities are either points or lines or areas. Water wells, mountain peaks, and power poles are examples of entities typically represented as points. Railways, bird migration routes, and rivers are represented as lines, while census enumeration areas, wetlands, and geologic units are often represented as areas.

Map scale refers to the ratio of the distance on a map to the corresponding distance on the ground and is an important component in spatial data. For example, a scale of 1:10,000 indicates that one unit on the map (or in a GIS database) represents 10,000 units on the ground. One centimeter, or inch, or meter, on the map represents 10,000 centimeters, inches, or meters on the ground. As a paper or digital map at 1:1 scale would be too large and unwieldy, the earth is shown at a reduced scale in a map or GIS database and described by a ratio or representative fraction. All spatial data are collected at a specific scale. Spatial data at a large scale, such as 1:10,000—or written as a fraction, 1/10,000—show a great amount of detail but cover a small area. Conversely, spatial data at a small scale, such as 1:1,000,000, show a small amount of detail but cover a large area.

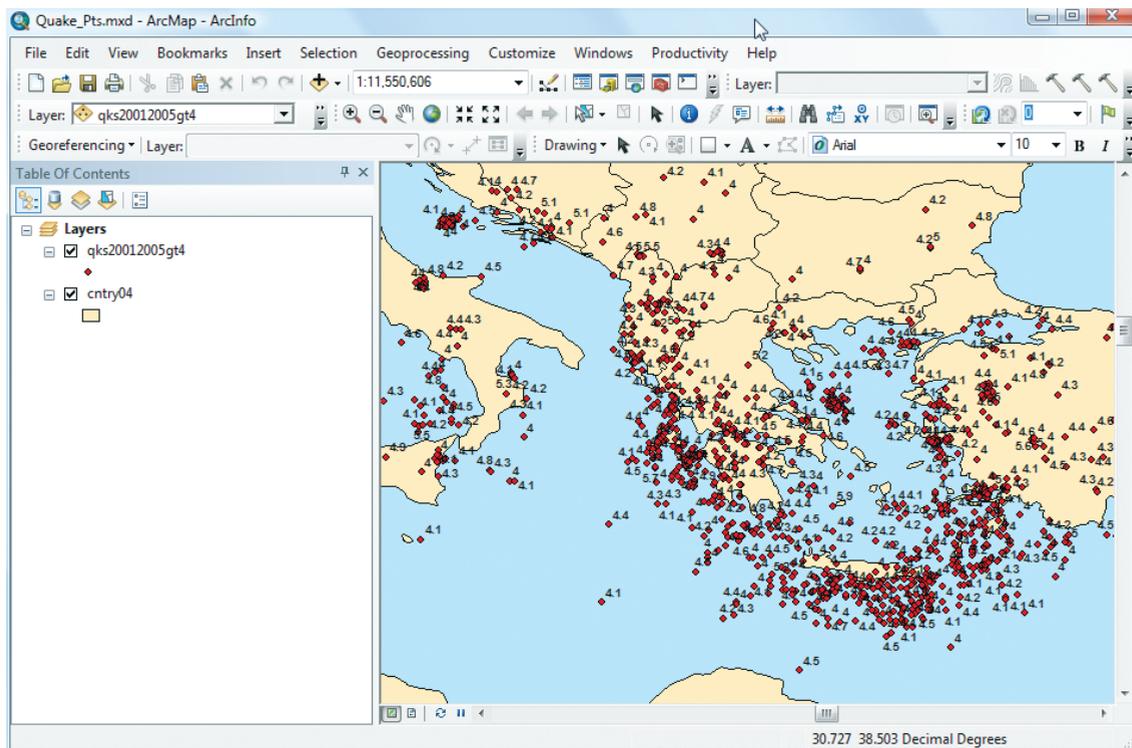
## ORIGIN OF THE TERM

The term *large scale* was coined as the fractions representing large scale, such as 1/10,000, are larger numbers than those representing small scales, such as 1/1,000,000.

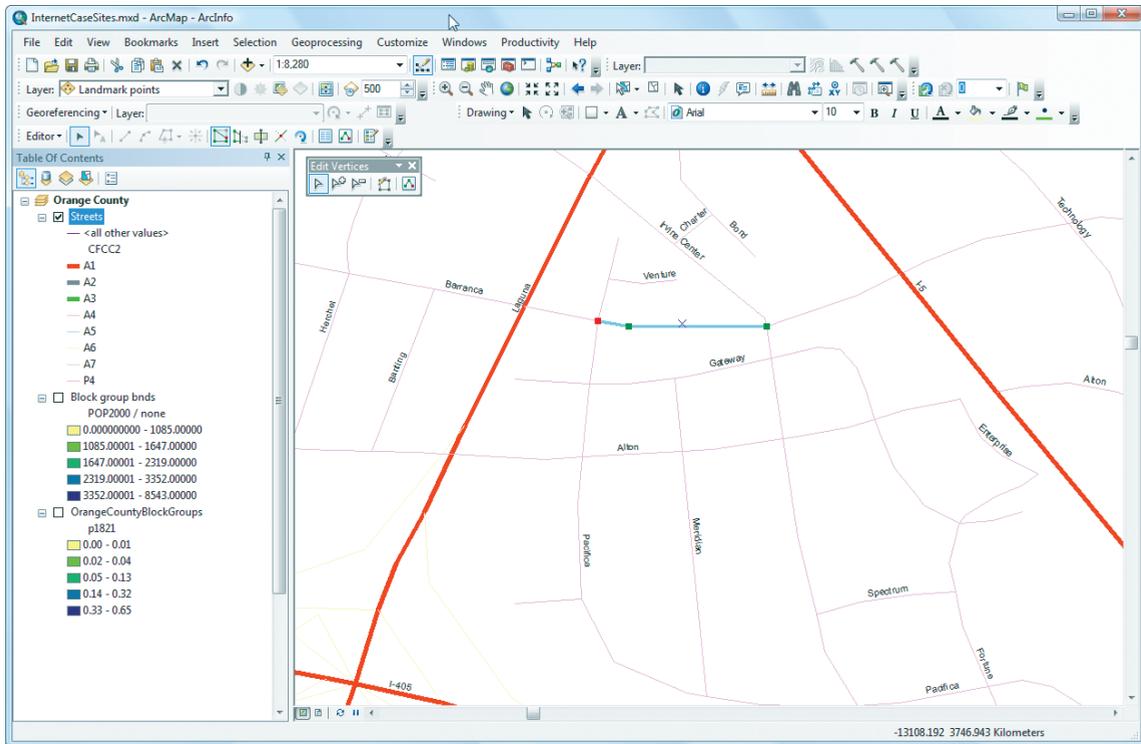
Scale has a significant influence on deciding which vector type best represents a particular entity. Cities may be better represented as points at a small scale but as areas at a larger scale, just as points best represent tornadoes at a small scale but at a larger scale, lines provide a better indication of the tornado's path.

## Vector points

Vector points represent all geographical entities that are referenced by a single x,y coordinate pair.



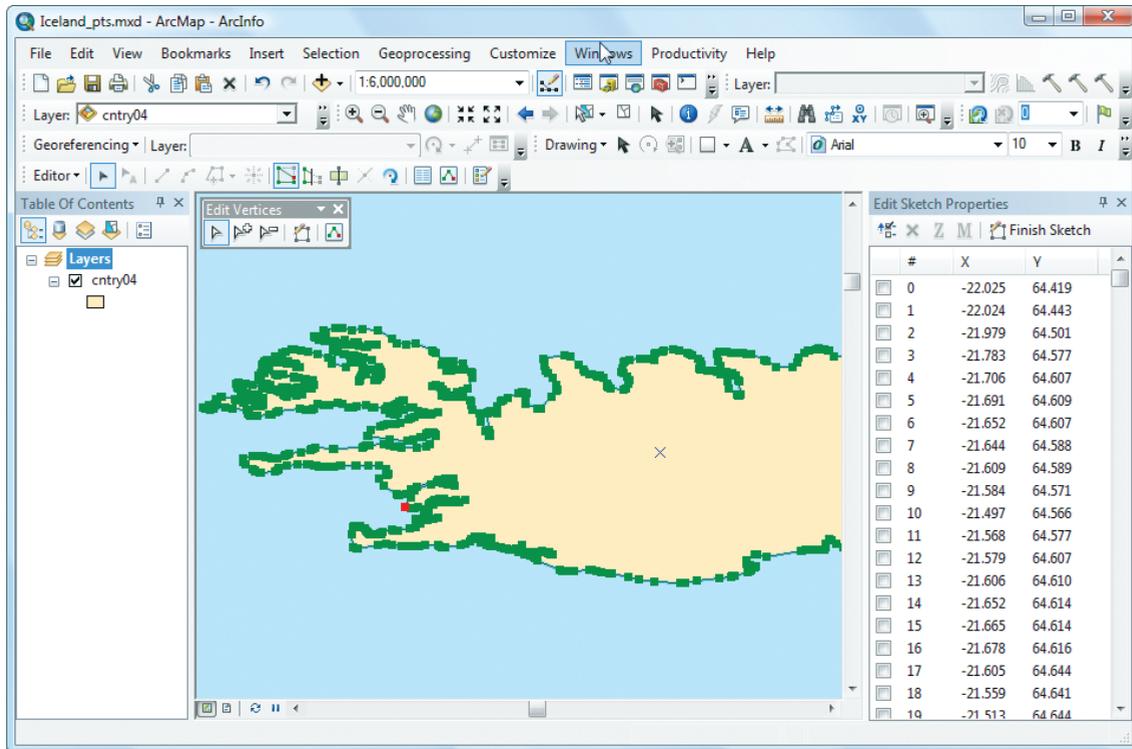
Vector point data: Feature annotation. Annotations indicating the magnitude for each earthquake have been created for each epicenter.



Vector line data: The vertices that comprise a single street segment in Orange County, California.

## Vector lines

Vector lines representing linear geographical features are composed of a number of straight line segments, with each line segment defined by a number of vertices. The simplest linear feature contains a start and an end node (two x,y coordinate pairs). In ArcGIS, lines are referred to as *polylines* if they are stored as shapefiles and simply *lines* if they are stored in a geodatabase.



Areas such as the country boundary illustrated here, start and end at the same node.

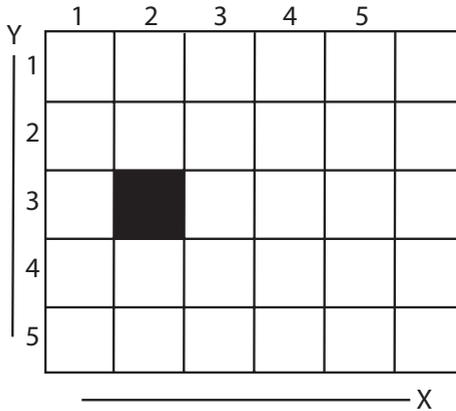
## Vector areas

Vector areas represent features that have a perimeter, or boundary, and an extent. The boundary is composed of line segments and x,y coordinate pairs, where the start and end node share the same geographic location. For topological consistency, to support near neighbor, point-in-polygon searches and so on, the line segments must not overlap or intersect.

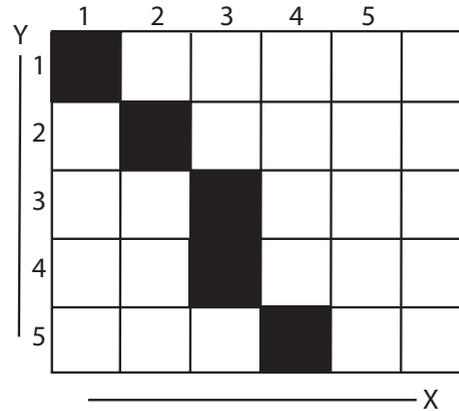
## Raster data model

In the raster data model, objects are represented by cells or pixels. A simple raster data structure consists of a 2D array of grid cells. Each grid cell is referenced by a row and a column number and contains a number representing the type (value) of the attribute being mapped. For example, in a raster dataset representing snow cover, a pixel with a value of 0 may represent no snow and a value of 1 represents areas of snow. As each cell can only contain one number or attribute, different attributes must be represented in separate 2D arrays (or overlays). Most raster data analysis involves combining 2D arrays to create new layers and cell values (Burrough and McDonnell 1998).

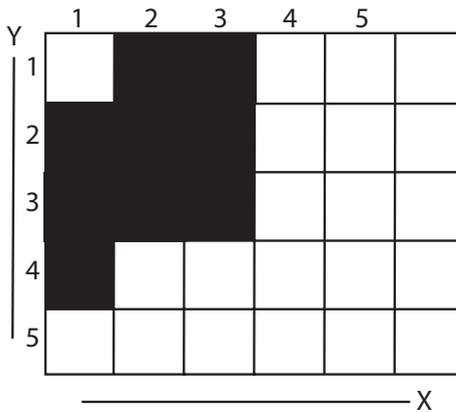
Raster data models are either *discrete* (thematic) or *continuous*, depending on what they represent. A forestry land use raster with a value of 1 that indicates deciduous trees, a 2 representing coniferous trees, and a 0 representing no trees is an example of a single-layer thematic raster, as the cells are encoded with a value from a discrete range of possible values. An elevation grid, or digital elevation model (DEM), is an example of a single-layer continuous raster where each cell is encoded as a floating point decimal number and each cell has an infinite number of potential values. Satellite imagery, such as the multi-band imagery from the Landsat Thematic Mapper satellite, is an example of a multi-layer continuous raster format; cells could have an infinite combination of reflectance values in different bands of the electromagnetic spectrum.



Raster points are represented as a single grid cell.



Raster lines are represented by a number of neighboring cells stretching out in a given direction.



Raster areas are represented by a group of neighboring cells.

Although hybrid data models such as the object data model and triangulated irregular networks (TINs) do exist, most public domain spatial data are available in either vector or raster format. Certain features such as water wells, trails, and municipal land use zones are typically represented as points, lines, and areas (respectively) in the vector model, and other characteristics of the earth such as elevation or land cover are represented in the raster data model. However, it is possible to convert data from one data model format to another. For example, point vector data representing soil pH could be converted to a raster grid for the purpose of estimating, or interpolating, what the pH would be in areas not specifically sampled. The same data could be visualized in 3D, with mountains representing areas of high pH, valleys representing low pH soils, and cliffs representing significant changes over short distances, which merit further investigation.