CHAPTER 3

Spatial Databases for Public Health

Spatial data sets are fundamental components of GIS. The success of health-related GIS projects depends critically on having access to accurate, timely, and compatible spatial data. For organizations embarking on GIS projects, spatial data can be viewed as both a cost and a resource. Developing spatial data sets is expensive; it is estimated that well over half the cost of GIS projects goes to database creation, updating, and improvement. Yet, database development is also an investment that creates long-term value for organizations and the people they serve. Spatial data sets are often useful for addressing a wide range of policy and planning issues. Their value extends well beyond the scope of the original projects for which they were created, and it increases as the data sets are used.

This chapter describes the major types of spatial databases for public health GIS. We begin by discussing the concept of foundation data and summarizing major types of foundation data sets like geodetic control, digital orthorectified imagery, and address-ranged street network data files. We then consider the diverse types of population and health data sets that can be incorporated in GIS by geocoding data on individual health events or by joining aggregated population and health data for areas to spatial databases for mapping and analysis. The final sections examine issues related to spatial data integration and sharing.

Foundation Spatial Data

In generating spatial databases for public health GIS, the key linkage among data layers is the spatial linkage. Layers are tied together by their common geographical location. If a house is located a quarter-mile east of a park and adjacent to a hospital, these features should appear in the same relative positions in a GIS that connects data layers of houses, parks, and hospitals. In a GIS, we cannot link the locations of features directly to their positions on the earth since we are working at a scale much smaller than the earth. Therefore, spatial data layers must be connected to a foundation that makes spatial integration and linkage possible. Foundation data provide a geographical frame of reference to which other data layers are tied.

Foundation spatial data are "the minimal directly observable or recordable data to which other data are spatially referenced" (National Academy of Sciences, 1995, p. 16). We use the term here to apply to the spatial data layer to which other data layers are linked in a public health GIS project. As in constructing a building, the foundation supports the other data layers and defines the ***footprint***, or geographic extent, of the GIS database. Many different types of spatial data can serve as foundation data, for example, digital imagery from aerial photographs or satellites, street centerline data, or property boundary data. These databases differ in their scale, resolution, degree of positional accuracy, and ease and cost of use. The choice of a foundation data set will be influenced by the scale of the analysis. A study of health problems at the neighborhood scale requires foundation data at an equivalent spatial scale.

This section explores the various types of foundation data and their characteristics. National mapping agencies guide the development of foundation data in different countries of the world. The discussion that follows focuses on geodetic control and foundation databases widely used in the United States. The International Cartographic Association website provides an interactive map to access contact information for national mapping agency members and also offers national reports of mapping activities (International Cartographic Association/Association Cartographique International, 2010). The African Geo Information Research Network (AGIRN), an initiative of the Human Sciences Research Council of South Africa and EIS-Africa, maintains a site with links to national mapping agencies in Africa (African Geo Information Research Network, 2011).

Geodetic Control

Geodetic control is a system for registering location information to a set of well-defined points on the earth's surface. It includes a set of ***survey monuments*** on the ground and a ***reference datum*** that gives geographic coordinates for those monuments based on our knowledge of the size and shape of the earth, as discussed in Chapter 2. The reference datum is a key feature of geodetic control. In North America, the currently accepted reference datum is the North American Datum, 1983 (NAD-83). This datum is linked to the World Geodetic System, 1984, a geodetic control system for geographical coordinate use worldwide. The reference datum for North America has changed in recent years. For decades, the reference datum was the North American Datum of 1927 (NAD-27), replaced by NAD-83 after its publication in 1986. Spatial databases that were created in the United States, Canada, and Mexico before the mid-1980s often use NAD-27.

In developing GIS databases, it is critically important that all data layers use the same reference datum. Longitude/latitude coordinates based on NAD-27 and NAD-83 can differ by up to 100 meters in the lower 48 states, leading to positional errors and inconsistencies (Keating, 1993). When linking different spatial data layers, analysts should check the reference datums associated with each data set and, if necessary, convert all data sets to a common datum. Most GIS include commands for converting among NAD-27, NAD-83, and other common reference datums.

In most GIS applications, geodetic control is not used directly as a foundation data layer. Geodetic control is transparent, never displayed or connected with attribute information. However, understanding geodetic control and reference datums is vital for developing GIS data sets and ensuring consistent, accurate data linkage. In addition, the growing use of GPS receivers for generating coordinates heightens the importance of geodetic control because GPS coordinates are directly tied to geodetic control. Furthermore, online systems like Google Earth® that are used for mapping incorporate imagery that is tied to specific reference datums.

Digital Orthorectified Imagery

Digital orthorectified imagery (DOI) comprises pictures of the earth's surface that show the locations of features like roads, coastlines, and buildings. The pictures are raster images generated from aerial photography or satellite data. Digital images are encoded records of spectral reflectance or emittance intensity for objects or areas. Sensors on satellites record energy reflected from the earth's surface for different wavelengths or "bands" of the electromagnetic spectrum. For each band, an individual pixel corresponding to a place on the earth's surface has a digital number representing the intensity of spectral reflectance.

Image files are generally very large and difficult to store. Compression reduces the size of the image file. ***Lossless compression***, as the name implies, results in a compressed image that can be reconstructed to produce an image identical to the original. Its main advantage is the ability to reconstruct the original image. Its main disadvantage is limited compression ratio. Wavelet compression is a ***lossy compression*** method, which means that some information is lost in order to achieve higher compression rates. The compressed image cannot be used to reconstruct the original image. A wavelet compression method commonly used with geographic imagery is ***MrSID (Multiresolution Seamless Image Database)*** (LizardTech, 2004). ***JPEG 2000*** is another wavelet compression technique used with geospatial imagery and many other types of images (Taubman & Marcellin, 2002).

Tied to geodetic control to permit matching with other spatial data layers, the images have the geometric properties of a map. The information necessary to make this tie may be stored in separate so-called world files for use with images in MrSID or JPEG 2000 format. The Open Geospatial Consortium has also created a metadata standard for georeferencing JPEG 2000 images with embedded Geography Markup Language (Open Geospatial Consortium, 2011b). Similarly, ***GeoTIFF*** image metadata allows georeferencing information to be embedded in a TIFF file so that the image displays properly when added to a GIS application. The GeoTIFF format is being adopted by a wide range of data providers including the U.S. Geological Survey, SPOT Image Corporation, and other agencies in the United States and other countries (Ruth, 2010).

DOI does not incorporate specific feature or attribute information: it simply provides an image of some part of the earth's surface. Identifying and recording features on the images requires image interpretation, field checking, or linkage with an attribute-based spatial data layer for the area. However, many significant landscape features are clearly visible on DOI.

An important kind of DOI for public health GIS is the ***digital orthophotoquarterquad*** (DOQQ). A DOQQ covers a "quarterquad," an area roughly 4 miles × 4 miles, at 1:12,000 scale. Produced by the U.S. Geological Survey in conjunction with other federal agencies, the DOQQs depict roads, houses, trees, and other detailed features (Figure 3.1). With their high resolution and high degree of positional accuracy, DOQQs form a useful foundation data layer for localized, large-scale public health assessments, such as mapping individual exposures to environmental contaminants. Other data layers can be matched to the DOQQs for detailed mapping and analysis.

The ***National Agriculture Imagery Program (NAIP)***, which began in 2003, is a source for high-resolution aerial photography imagery acquired during the

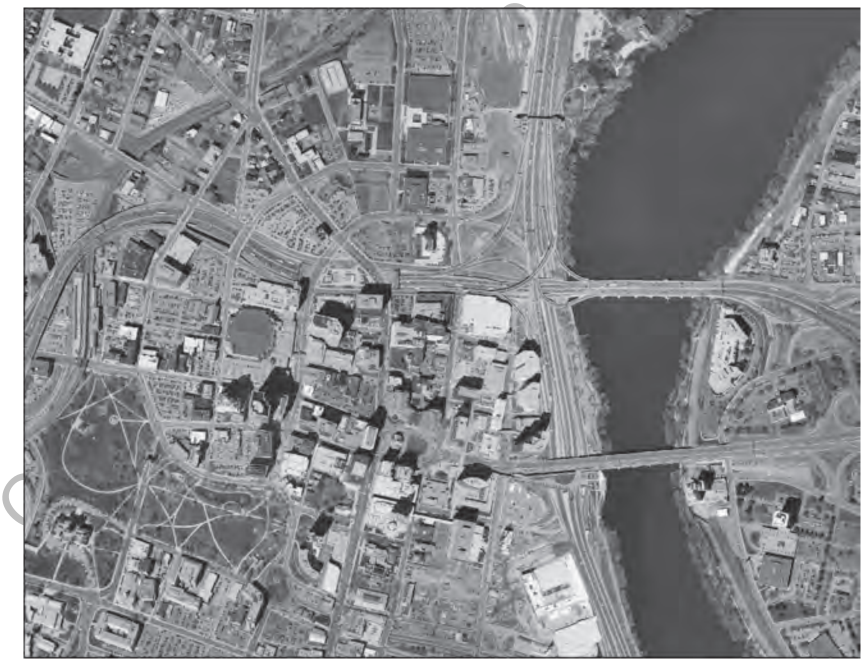


FIGURE 3.1. A portion of a digital orthophotoquad for the area around downtown Hartford. The dark area running north–south just east of the center of the view is the Connecticut River. The town boundary between Hartford to the west and East Hartford is the center of the river.

agricultural growing seasons for the continental United States (U.S. Department of Agriculture, 2010). NAIP imagery is acquired at one-meter ground sample distance. The imagery shows "leaf-on" conditions with no more than 10% cloud cover per quarterquad tile. Images correspond to the USGS quadrangles and are distributed in GeoTIFF format. The program makes imagery available to government agencies and the public within a year of acquisition. Many public agencies and private entities at every level are using these data for mapping, land classification, environmental monitoring, and a wide range of other activities including public health and safety (U.S. Department of Agriculture, 2008).

Smaller scale DOI includes ***satellite imagery*** from systems like SPOT and Thematic Mapper. Satellite images typically cover scales ranging from 1:50,000 to 1:100,000 at positional accuracies ranging from ± 25 meters to ±70 meters (Keating, 1993). Although scale and positional accuracy vary widely across satellite imagery, generally the images show major features such as roads, rivers, fields, and water bodies (Figure 3.2). As in other forms of imagery, features are not labeled or identified. However, methods for digital image interpretation that

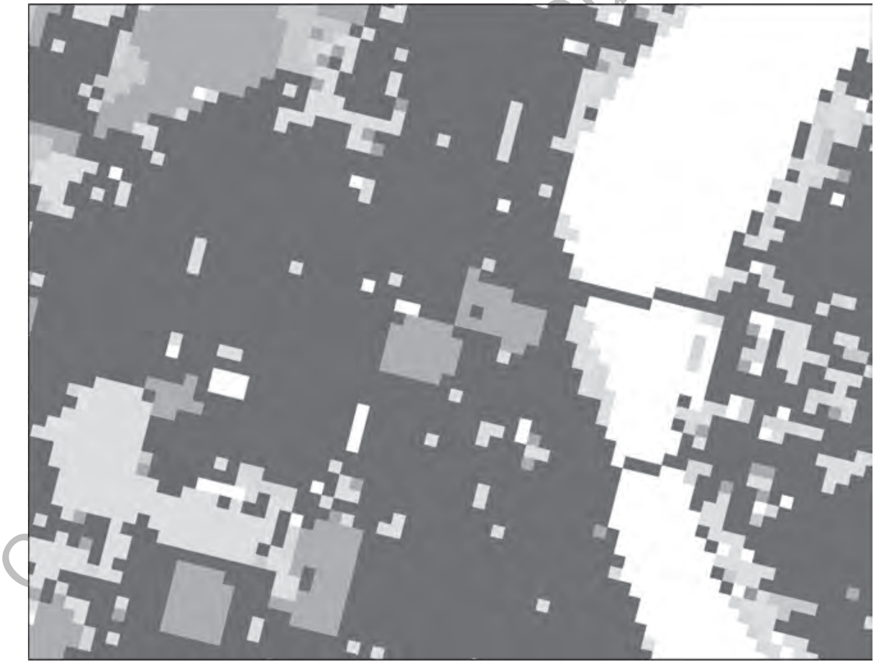


FIGURE 3.2. A portion of the land cover database for the area around downtown Hartford derived from Thematic Mapper Imagery. The areas shaded dark gray were classified as commercial/industrial/transportation. The areas shaded medium gray were classified as residential. The areas shaded light gray were classified as urban/recreational grasses. The areas shaded white were classified as open water. This figure shows roughly the same area as Figure 3.1.

distinguish land use/land cover features based on their distinct spectral characteristics are well developed and available in specialized computer software (Jensen, 2005). Some visible features in a satellite image vary seasonally because of changes in vegetation and precipitation. Cloud cover can also obscure features, complicating the interpretation of satellite images. In choosing a satellite image, the analyst should think through carefully the appropriate time of year for the image and the maximum allowable cloud cover. Detailed information is available for satellite imagery covering the United States to aid the analyst in selecting useful images (U.S. Geological Survey, 2010a).

Satellite images offer an important foundation data layer for regional-scale health analyses covering states or parts of states and for local analyses. The images have been widely used in displaying and analyzing land use, land cover, and natural resource patterns. In the public health field, the images have been utilized to analyze and predict outbreaks of vector-borne diseases such as Lyme disease (Glass et al., 1995; Ford et al., 2009).

Digital Line Graphs

Vector data also provide a foundation for regional-scale GIS development. ***Digital Line Graphs (DLGs)*** are vector databases that show transportation lines, water bodies, political boundaries, and elevation contour lines. Unlike imagery, DLGs include attribute information. Attribute codes describe the physical and cultural characteristics of points, lines, and areas on the DLG. DLGs are derived from the large- and intermediate-scale topographic maps created by the U.S. Geological Survey. They exist for all of the United States, excluding Alaska, at a scale of 1:100,000 (U.S. Geological Survey, 2010b). Large-scale DLGs, generated from the 7.5-minute topographic maps, have been created for many areas of the United States (Figure 3.3).

One concern in using DLGs is the accuracy and recency of attribute information. The sources of information for DLGs are topographic maps which may be years out of date. The Geological Survey has updated its topographic map series through a procedure known as “limited update,” focusing on features that are most likely to have changed such as roads and hydrography (Lemen, 1999). DOQQs from aerial photography are the basis for limited update revisions. The efficient, limited-update procedure has generated more timely information for topographic maps and DLGs, but time lags, naturally, exist. For GIS, these issues are especially relevant in communities experiencing rapid population and commercial development where feature and attribute information changes frequently.

As the national topographic mapping program of the United States has developed, data in DLG format are being incorporated into a new generation of topographic maps and spatial data products built in collaboration with local and state agencies. For example, DLG data have been used in the creation of the National Hydography Dataset. This and other data layers are available as part of The National Map and its developing Digital Map program (U.S. Geological Sur-

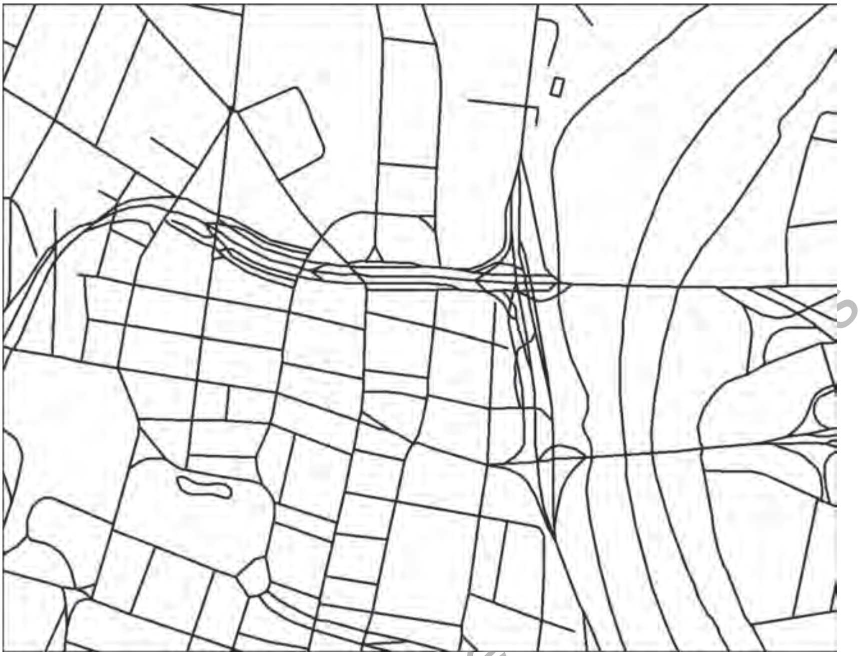


FIGURE 3.3. A portion of the 1:24,000 digital line graph database for the area around downtown Hartford, including roads, hydrographic features, and town boundaries. This figure shows roughly the same area as Figure 3.1.

vey, 2010b). These new spatial data products will be, like DLGs, used in health applications of GIS in the future.

TIGER/Line® Data

Another form of vector foundation data, compiled at 1:100,000 scale, is TIGER/Line data. The ***Topologically Integrated Geographic Encoding and Referencing (TIGER)*** data set was developed for the 1990 census (Marx, 1986). Since 1990, the TIGER/Line database has evolved into the ***MAF/TIGER® (Master Address File/Topologically Integrated Geographic Encoding and Referencing)*** database, which is the Census Bureau's set of digital files storing all of the geographic and attribute data necessary to conduct the census. The MAF portion contains a record for each potential housing unit. The TIGER portion contains all of the points and lines identifying the features used to form the areas for which the Census Bureau tabulates data. TIGER/Line data are an extract of selected geographic and cartographic information from the MAF/TIGER database (U.S. Census Bureau, 2009a). MAF/TIGER also included a redesign of the original TIGER/Line files database (U.S. Census Bureau, 2005). Although earlier versions of TIGER/Line data were distributed in Vector Product Format and required special utilities to convert them to formats that could be used in GIS software, TIGER/Line data have been converted to shapefile format in preparation for the 2010 census (Table 3.1). Various versions of TIGER/Line data and technical documentation can be downloaded from the census website. TIGER/Line data have been used widely in population, health, political, and transportation mapping.

TIGER/Line shapefiles may contain landmark point features, line features including ***street centerline*** data and other line features like political boundaries or rivers forming boundaries of census areas, or area features including states, counties, and census tracts. Depending on the data, shapefiles can be downloaded for the entire nation, a state, or a county within a state. Shapefiles containing line segments are distributed for individual counties (Figure 3.4). For TIGER/Line segments that are street centerlines, attributes for the left- and right-hand sides of the street segment are coded. These include a wide range of attributes for each side: street name, address range, ZIP Code, census and political unit identifiers, and congressional district identifiers.

A major benefit of the TIGER/Line files is that they provide a connection between street address ranges and locations on the ground. This makes it possible to locate or geocode address-based information such as hospital discharge records, birth certificates, and clinic locations. However, the TIGER/Line files do not record a precise location for each address, just an address range along a street segment; therefore, address locations can only be approximated by interpolation, as described later in this chapter. This may pose a few problems in urban and suburban areas where addresses are spread relatively evenly along street segments, but in rural areas TIGER/Line files should be used with caution for locating addresses if a high degree of positional accuracy is required.

Despite their wide coverage and applicability, the TIGER/Line files have had several important limitations. First, street and address coverage is incomplete and in some cases inaccurate in earlier versions of the data. Streets may be missing or misnamed. Address ranges may be missing, include incorrect values, or identify the wrong side of the street. These problems are especially relevant in rapidly growing communities where new residential development has taken

TABLE 3.1. Selected Shapefile Components

File extension	Content	Status
.shp	Geographic feature geometry	Mandatory
.shx	Index of feature geometry	Mandatory
.dbf	Attribute table with variables describing features	Mandatory
.sbn	Spatial index of features	Optional
.prj	Description of coordinate system and projection	Optional
.shp.xml	Metadata in XML format	Optional

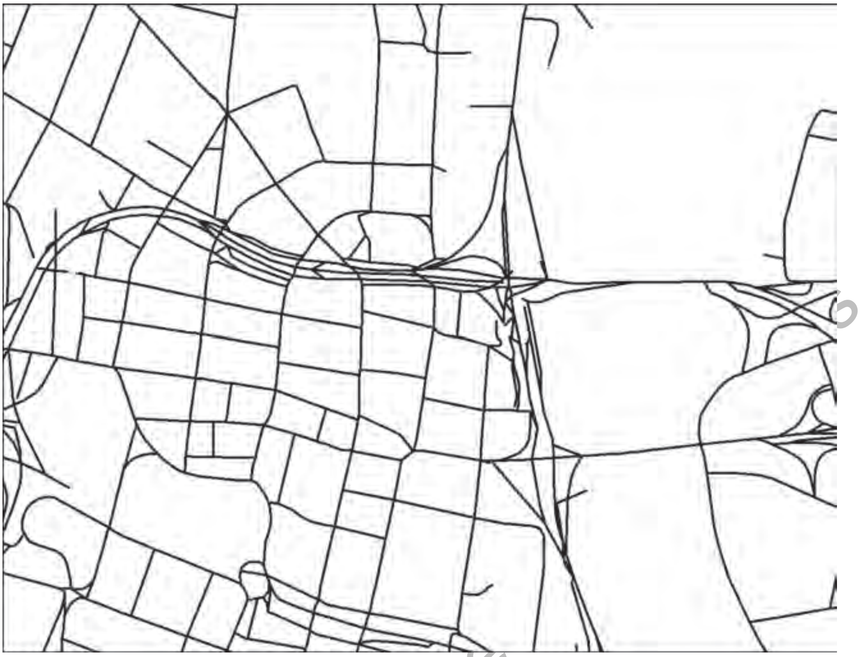


FIGURE 3.4. A portion of the TIGER/Line database showing road features in the area around downtown Hartford. This figure shows roughly the same area as Figure 3.1.

place. Many local governments enhanced and improved the TIGER/Line files for their local areas, and this information has been used to update TIGER in the years following its introduction (Sperling, 1995).

A second problem is that the positional accuracy of TIGER files as they were originally developed from multiple sources was unknown and varied from place to place. Positional accuracy of the original TIGER files was "no greater than the established National Map Accuracy Standards for 1:100,000-scale maps" (U.S. Census Bureau, 1992), and the files were logically consistent, but positionally inaccurate for large-scale mapping in some cases. These positional inaccuracies have implications for health studies. They may be of little consequence for a study of lead screening programs, for example, but extremely important in analyzing a problem like radon exposure.

Finally, TIGER data layers often did not match perfectly with layers generated from DLGs or DOQQs. Rubbersheeting techniques—techniques to adjust features to a foundation data layer—were often needed to combine TIGER data with data from other sources. Rubbersheeting is discussed in greater detail in the section on database integration later in this chapter.

In response to these problems, the Census Bureau has worked with various state and local agencies to improve the quality of address range information and

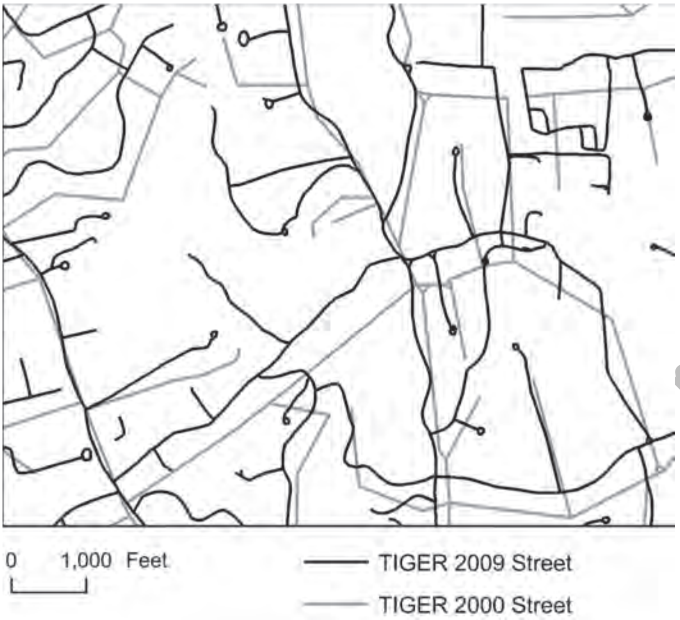


FIGURE 3.5. Overlaying street segments from the 2009 edition of the TIGER/Line database and street segments from the 2000 edition shows improvements in the positional accuracy of the data.

the positional accuracy of the TIGER/Line files for the 2010 census (Figure 3.5). The MAF/TIGER Accuracy Improvement Project, completed in 2008, resulted in the realignment of many, though not all, features based on data submitted by local and state agencies, imagery, and GPS data collected in the field (U.S. Census Bureau, 2009a). Even before these improvements, the TIGER/Line database has been one of the most important and widely used foundations for GIS-based health and socioeconomic analysis. Furthermore, the development of TIGER has prompted commercial firms to sell corrected and updated versions of the data. In fact, many GIS software packages come bundled with TIGER-based spatial data to facilitate mapping of census data. In developing a TIGER-based database for a GIS, it is well worth seeking out the most accurate and updated version. Analysts should also make sure that data used for geocoding health events is consistent with the data used for mapping census data by census tracts or other units so that health events will be correctly allocated to areas.

Cadastral Data

Another source of address-based spatial data that is generally more accurate than TIGER/Line for small geographic areas is cadastral information. ***Cadastral data*** are data associated with land ownership, and they are a matter of public record in the United States. Cadastral features are not visible on the ground, but are legally defined to specify ownership and administration of land parcels (Huxhold & Levinsohn, 1995). Digital cadastral data files contain property boundaries and a wide range of attribute data including land title, address, sale/resale information and building type, size, and characteristics (Figure 3.6). Property boundaries are stored in a vector format, with property attributes attached. Because the files describe land ownership, they often have a high degree of positional accuracy and represent large spatial scales—1:12,000 or larger. Address information is generally accurate and complete. Cadastral data also show street widths and thus better depict the built environment of a local area than do TIGER/Line files.

Despite these advantages, cadastral data have important limitations. Although most communities collect and maintain cadastral spatial data, conversion to digital form has been a relatively recent development. Some communities still rely on paper maps and written descriptions of property boundaries, some decades old. Furthermore, the quality and accuracy of cadastral data vary widely, depending on the recency and quality of the surveying or historical information on which it is based. Errors can creep into cadastral databases over time

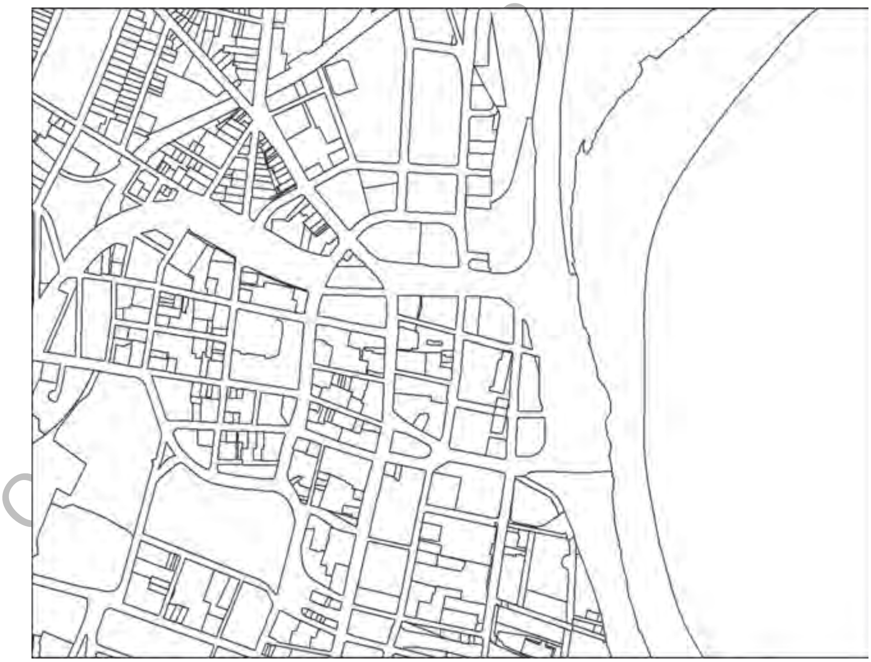


FIGURE 3.6. A portion of a cadastral database for the area around downtown Hartford. The figure shows roughly the same area as Figure 3.1. Property databases are generally maintained by local governments, so this database covers only Hartford and does not include properties in East Hartford.

depending on how well the registry is kept as boundaries are resurveyed, landscape change occurs, and properties are subdivided. In addition, communities use different formatting systems for digital cadastral information and capture different attributes, so conducting studies across community boundaries can be challenging. Efforts are currently underway to develop common standards for cadastral information. Cadastral data files can also be large, unwieldy, expensive to create, and unnecessarily detailed for some kinds of spatial analysis. Network models such as those discussed in Chapter 10, for example, require transportation routes to be represented as arcs, as in TIGER, rather than as double lines. Still, cadastral data offer an excellent foundation for address-matching and mapping in small areas.

Choosing a Foundation Database

The foundation data sets described in this section each offer a unique set of advantages and disadvantages for public health GIS. They differ in scale, resolution, positional accuracy, and display of features, as well as in their raster or vector structure. They are also evolving over time.

The choice among foundation data sets depends on the scale and scope of the project, the resources available for data creation, and the types and scales of other data sets to which the foundation data will be linked. Projects that are national or regional in scope are more likely to utilize intermediate scale foundation data such as satellite imagery, DLGs, and TIGER/Line data. In contrast, studies of single communities or neighborhoods can take advantage of the detail and positional accuracy of cadastral data and DOQQs.

Population Data

Foundation data create a platform for integrating spatial data layers that contain population and related health, social, and environmental information frequently used in health applications of GIS. The TIGER/Line files store spatial information for the geographical units the Census Bureau uses in tabulating and publishing the population data it collects. An understanding of census geography is important for any public health analyst who uses data compiled by the Census Bureau (U.S. Census Bureau, 2008a). Not all of the population data tabulated by the Census Bureau is published for every level of census geography (Peters & MacDonald, 2004).

The smallest unit is the census ***block***, and each block is bounded by a set of connected street segments or other linear features such as rivers, railroad tracks, or municipal boundaries (Figure 3.7). A ***block group*** is a cluster of blocks, typically containing from 600 to 3,000 people. Census ***tracts*** comprise groups of contiguous blocks (and block groups) and have populations ranging from 1,200 to 8,000. TIGER shapefiles are also available for state, county, and local political

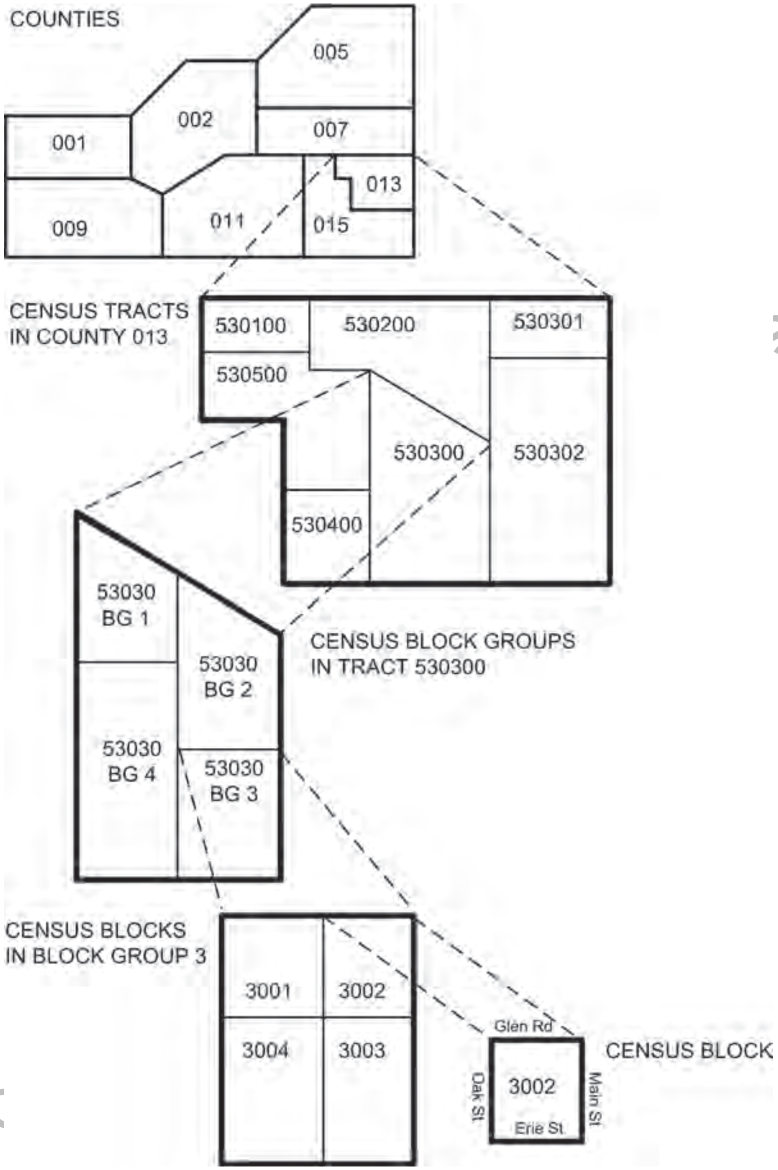
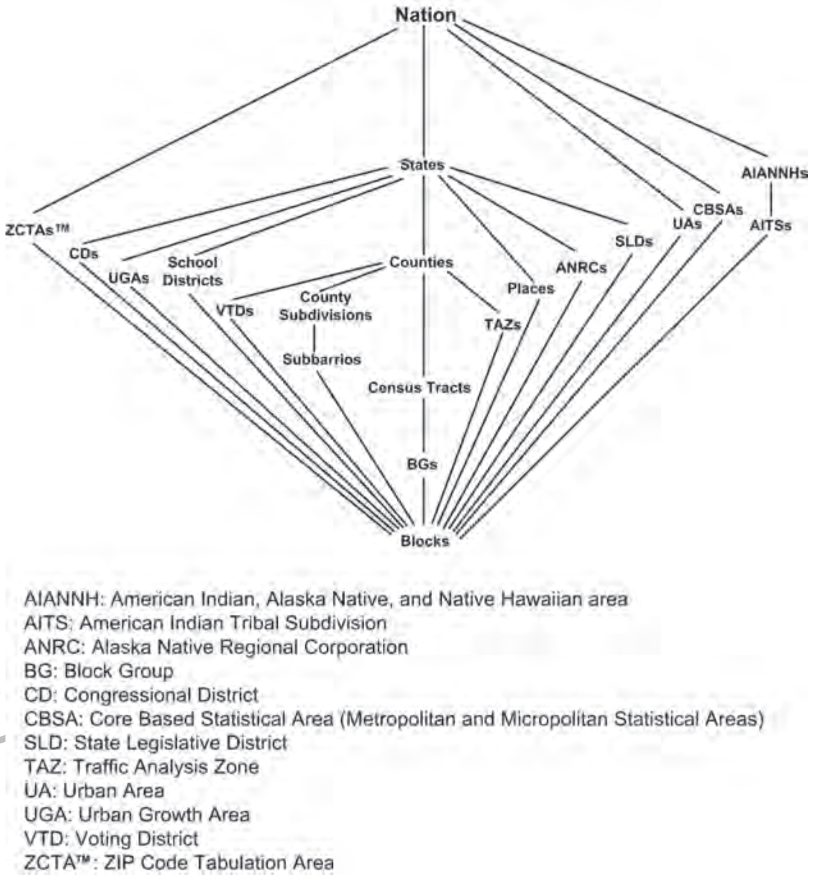


FIGURE 3.7. Geographic subdivisions for the U.S. Census. The smallest unit is the block. Each county is divided into census tracts, which are divided into block groups, and then into blocks. The first digit of the census block identifier corresponds to the block group.

boundaries, along with ZIP Code boundaries. As described in another section in this chapter, database tables describing the population, health, and socioeconomic attributes of these areas can be joined to databases describing the geography of the areas.

The hierarchy of census units in relation to other political and administrative units is complex, given the federal nature of the U.S. system of government (Figure 3.8). Census blocks, block group areas, and tracts nest perfectly within counties, but other local areas do not necessarily follow this pattern. In addition to data provided for counties, data are provided for places or minor civil divisions that are legally incorporated and bounded areas such as cities, towns, and villages. Census blocks nest perfectly within these units. Census block group and



AIANNH: American Indian, Alaska Native, and Native Hawaiian area
AITS: American Indian Tribal Subdivision
ANRC: Alaska Native Regional Corporation
BG: Block Group
CD: Congressional District
CBSA: Core Based Statistical Area (Metropolitan and Micropolitan Statistical Areas)
SLD: State Legislative District
TAZ: Traffic Analysis Zone
UA: Urban Area
UGA: Urban Growth Area
VTD: Voting District
ZCTA™: ZIP Code Tabulation Area

FIGURE 3.8. Hierarchical relationships of census and political or administrative areas in the United States for the 2010 census. Census block group areas and census tracts nest within counties, but their boundaries overlap the boundaries of many other political and administrative entities.

tract areas, however, do not always nest perfectly within places or minor civil divisions. Places may cut across county boundaries.

In some cases, census tracts may coincide with areas where many residents live in group quarters like prisons, military bases, or colleges and universities. Because these residents often differ from the general population in terms of age and sex and residential mobility, it is important in health applications of GIS to make explicit decisions about how to include group quarters populations. The population of a college town such as Mansfield, Connecticut, where the main campus of the University of Connecticut is located, is very different during the academic year than during the summer months (Figure 3.9).

In the United States, a complete enumeration of the population is conducted every 10 years, as mandated by the Constitution for the purposes of apportioning seats in Congress. Beginning with the 2010 census, the American Community Survey, a program initiated after the 2000 census for providing more up-to-date census data during the intercensal period, will be fully operational. The Census Bureau provides information on the dates of censuses conducted or scheduled in other countries from 1945 to 2014 (U.S. Census Bureau, 2008b) and links to statistical agencies in other countries responsible for population data (U.S. Census Bureau, 2010a).

Health Data

This section describes some of the major types of health information that can be incorporated in GIS for health planning, evaluation, and research. Our aim is to introduce these data sets and highlight geographical issues that affect data use and integration in GIS. Detailed discussion of the content of these data sources is available elsewhere (Halperin & Baker, 1992; Parrish & McDonnell, 2000; Huber, Boorkman, & Blackwell, 2008).

Registration System Data

VITAL STATISTICS

Local governments in the United States and other countries routinely collect information on all births and deaths that occur in their jurisdictions. These ***vital records*** are an important source of spatial data for public health GIS. Birth records document a wide range of conditions that affect newborn infants, including birthweight, gestational age, congenital malformations and obstetric procedures, along with the mother's demographic and social characteristics and her use of prenatal services (Friis & Sellers, 2009). Information about the infant and the birth process is generally accurate, but data for the mother, especially data based on recall of timing and events during pregnancy, can have errors and inconsistencies. Still, birth data offer a nearly complete summary of basic maternal and infant health indicators for the population.

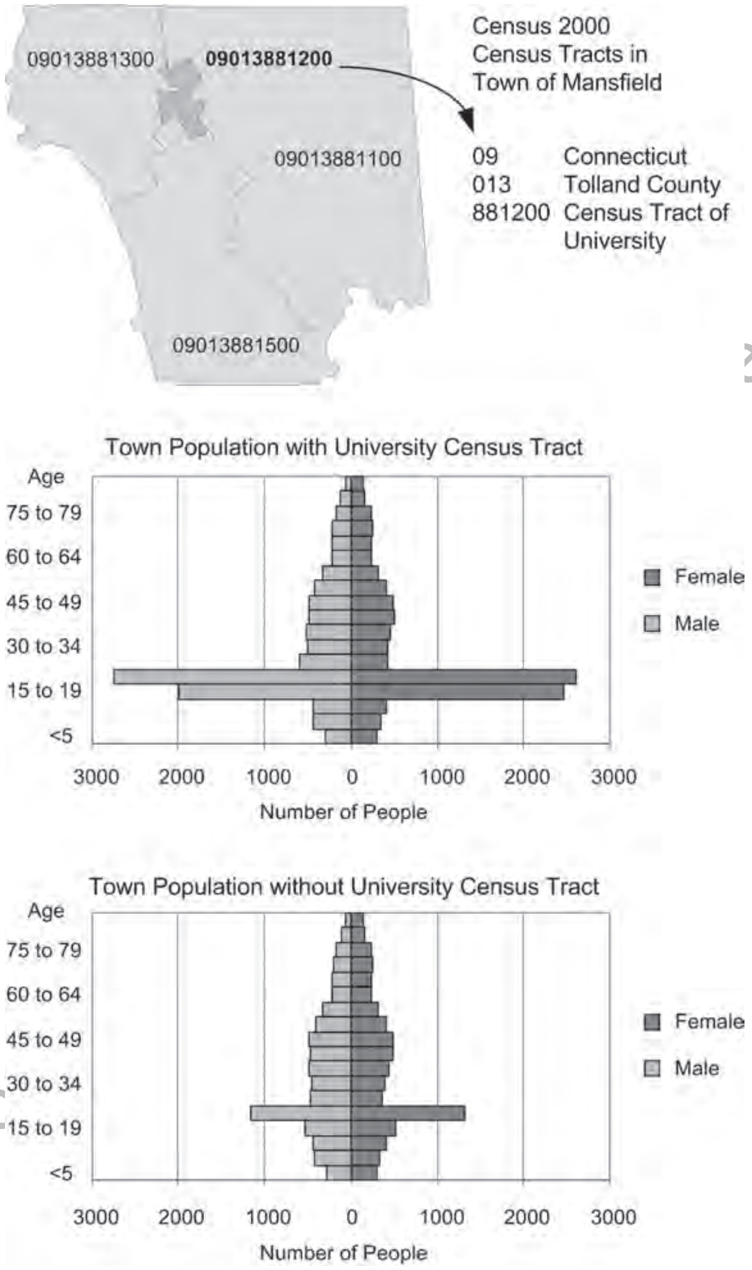


FIGURE 3.9. One census tract area from the 2000 census in the town of Mansfield has a large group quarters population because the University of Connecticut's main campus is located there. The size and age–sex structure of the town's population differs significantly depending on whether the tract where the university is located is included.

Birth records include the mother's residential address, a geographical identifier for GIS mapping and analysis. This information has been used to study environmental and neighborhood influences on maternal and infant health, for example, the effects of proximity to prenatal care services on prenatal care use and birth outcomes (McLafferty & Grady, 2004) or the clustering of birth defects in relation to hazardous waste sites (Rushton & Lolonis, 1996).

Health departments also collect and report data on deaths in ***mortality records***. Generated from death certificates, these data include demographic characteristics of the decedent and information about the cause of death, including the immediate cause and contributing factors (Friis & Sellers, 2009). Although demographic information is typically accurate, there are well-known problems with cause-of-death information stemming from errors and inconsistencies in diagnosis and difficulties in assigning causes when multiple causes are present (Garbe & Blount, 1992). Death certificates include two types of address-based geographic data: the place of death and the usual residence of the decedent. The place of death is often a hospital, nursing home, or other health care facility. This information can be used in analyzing health outcomes and service utilization by facility. In contrast, residential addresses provide a means for linking the residential environment to mortality outcomes.

Address-based vital statistics information presents several challenges to the GIS researcher. Addresses may be incorrectly coded, making it impossible to identify geographic locations. In a study of birth defects in Des Moines, Iowa, 8% of birth records could not be geocoded because of errors in the addresses and "P.O. Box" and "rural route" style addresses (Rushton & Lolonis, 1996), although geocoding using parcel or E911 data may increase the match rate (Mazumdar, Rushton, Smith, Zimmerman, & Donham, 2008). Also, because of privacy and confidentiality concerns, many health departments do not release address information (Istre, 1992). They only provide data in aggregate form, by ZIP Code, district, or census tract, making it impossible to analyze point locations. Finally, even if current residential address information is correct, it may not accurately represent the environment of the person before and during pregnancy or prior to death because the relevant exposure may have occurred someplace other than the residence (see Chapter 6). This is particularly problematic for mortality data, given that the conditions that lead to death can result from lifelong exposures and behaviors.

MORBIDITY DATA FROM SURVEILLANCE SYSTEMS AND DISEASE REGISTRIES

Looking beyond life's vital events, morbidity data are an essential source of information for public health GIS. ***Disease surveillance*** involves monitoring distributions and trends in morbidity and mortality data collected for a specified population and geographical area. There are many kinds of morbidity data, ranging from information gathered by government agencies and health care providers to information from survey research projects. These data differ greatly in content, coverage of the population, and geographic scale at which they are normally available.

Reportable disease data provide information on morbidity and mortality for certain "reportable" health conditions. Infectious disease has always been an important focus of public health surveillance in the United States (Centers for Disease Control and Prevention, 2008b). Authority to require notification of cases of disease resides in state legislatures, and there is considerable variation in state provisions. All 50 states require physicians to report cases of specified notifiable diseases to state or local health departments. Notifiable disease reports and vital records are the two health data sources available at the local level in all states.

The ***National Notifiable Diseases Surveillance System*** is operated by the Centers for Disease Control and Prevention (CDC) in collaboration with the Council of State and Territorial Epidemiologists (CSTE). Reporting by the states to the national system is voluntary. States generally also report internationally quarantinable diseases (cholera, plague, yellow fever) in compliance with World Health Organization (WHO) International Regulations. There are approximately 50 infectious diseases designated as notifiable at the national level (Council of State and Territorial Epidemiologists, 2009). The list of nationally reportable infectious diseases and other conditions changes periodically, and reporting practices may differ from state to state (Roush, Birkhead, Koo, Cobb, & Fleming, 1999).

In addition to notifiable disease reports by providers such as physicians, hospitals, and laboratories, key data sources for infectious disease reporting in the United States include sentinel systems, hospital surveillance, school surveillance, special surveys at the state and local level, vital records, and vector/host surveillance for zoonotic diseases. ***Sentinel health events*** are cases of illness that signal a need for immediate public health intervention or serve as a warning of hazardous conditions or poor quality medical care. A case of polio, for example, might signal a breakdown in the quality of immunization programs. A number of limitations of the current surveillance system have been described (Birkhead & Maylahn, 2000). The fragmentation of the system and voluntary reporting requirements affect the completeness of surveillance data. Generally the reporting system is thought to work well for diseases that are serious, have clear symptoms, and require medical attention. However, coverage is incomplete for conditions that can be asymptotic (tuberculosis), that do not necessarily compel medical treatment (animal bite, gastroenteritis), or that carry social stigma (HIV/AIDS) (Friis & Sellers, 2009).

Underreporting of infectious disease conditions may be explained by a number of factors, including provider lack of awareness of reporting requirements. The level of public concern also affects disease reporting. Infectious diseases that carry some social stigma may be concealed. For many infectious diseases, symptoms are either too mild to prompt a person to seek medical care or mimic flu-like symptoms associated with other common illnesses. For others, particularly emerging infectious diseases, the etiological definition may be incomplete or the case definition for surveillance purposes may be inadequate. "What is a case?" is not a trivial question. There may be differences of opinion about the criteria for defining a case of a disease. Sometimes, case identification requires laboratory confirmation. In addition, case definitions change with changes in scientific knowledge. Changes in case definitions over time have an impact on what is included in the surveillance database, as discussed in Chapters 7 and 8.

Active surveillance systems obtain data by searching and periodic contact with providers. ***Passive surveillance*** systems rely on reports by providers. Because of the costs associated with active surveillance, this type of system is often used strategically in limited areas or for limited time periods. Evaluation of active surveillance systems indicates two- to fivefold increases in reporting of specified diseases and other conditions not subject to active surveillance (Vogt, LaRue, Klaucke, & Jillson, 1983; Thacker et al., 1986). Surveillance method, therefore, has implications for completeness of the data, an important dimension of spatial database quality. Active surveillance systems offer a mechanism for completing and correcting information from the reportable disease record including address data used as geographical identifiers.

To protect privacy and confidentiality, federal agencies only release reportable disease statistics at the county level. Different policies exist in lower levels of government: some state or local health agencies will make information available for smaller geographic areas, or even by address, as long as the analyst agrees to maintain privacy and confidentiality. When address information exists, its accuracy can be problematic. Addresses may be missing or inaccurately coded. In an epidemiological study of reported rat bites in New York City, for example, almost 40% of bite reports had missing or incorrect address information and could not be geocoded (Childs et al., 1998).

Disease registries are centralized databases for the collection of information on specific diseases, the best examples being the cancer registries managed by state and local health authorities (Friis & Sellers, 2009). Disease registries use a reporting system similar to that for reportable diseases, with health providers reporting occurrences to the appropriate state or local registry. Some disease registries actively seek out case information, while others simply gather reports. Furthermore, some registries keep longitudinal information, following patients after diagnosis in order to track changes in health status and treatment regimes.

Cancer registries, the most extensive disease registries in the United States, offer a potentially valuable source of information for GIS analysis. Currently, all 50 states and a number of localities in the United States maintain cancer registries, some of which have existed for decades, funded through the CDC's National Program of Cancer Registries (NPCR) or the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program, or both (Centers for Disease Control and Prevention, 2011a). At the national level, the SEER program is an umbrella organization for a network of cancer registries that covers about 26% of the U.S. population. SEER includes active follow-up of living patients and is used to generate national estimates of overall cancer incidence and breakdowns by gender, race, age, and geographic location (National Cancer

Institute, 2011a). The North American Association of Cancer Registries promotes uniform data standards for cancer registration and the use of cancer surveillance data (North American Association of Central Cancer Registries, 2011). As with the other types of health data, registries include residential address information, and organizations like the North American Central Cancer Registries (NAACR) have developed valuable guidelines for geocoding health data (Goldberg, 2008).

The information in health databases and disease registries is protected by laws governing privacy and confidentiality. Some states will release addresses for research studies as long as appropriate measures are taken to ensure confidentiality; however, once again, policies differ among states. Other problems with address information arise from changes and errors in the coding and formatting of addresses.

Surveillance systems and disease registries have been sources of data for many GIS case studies but very few statewide surveillance systems or registries have been fully linked to GIS (Devasundaram, Rohn, Dwyer, & Israel, 1998; Cromley, 2000; South Carolina Vital Record and Statistics Integrated Information Systems Project Team, 2005). Implementation of a statewide or national surveillance system in GIS increases the likelihood that the case database will include cases identified using different case definitions and surveillance methods. To address this problem, case definition and surveillance method should be included as fields in a surveillance database.

Survey Data

To address a broader range of health issues than covered in standard vital statistics and morbidity data sets, public health researchers often turn to ***health survey data***. Surveys deal with a diverse array of health-related topics, topics that are beyond the scope of disease reporting systems and transcend biomedical concerns. Health surveys investigate health-related behaviors, psychosocial well-being, nutritional status, stress, and individual, family, and neighborhood circumstances that affect health. The major national surveys in the United States include the National Health and Nutrition Examination Survey (NHANES) and the National Health Interview Survey (NHIS). These surveys ask a detailed set of questions to a small, representative sample of the U.S. population. NHANES focuses on physiologic measures, measures of body weight and stature, and nutritional assessments. It has been conducted in several cycles since the early 1970s (Centers for Disease Control and Prevention, 2009a). NHIS, administered annually since 1957 with the U.S. Census Bureau serving as the data collection agent, collects information on health risk factors, chronic conditions, injuries, impairments and health service utilization, based on household interviews (Centers for Disease Control and Prevention, 2011b).

The purpose of these surveys is to develop a national picture of the health status of the population. Not every place is sampled. NHANES uses a four-stage sampling procedure. In stage 1, ***primary sampling units*** or PSUs are selected.

These are usually single counties but may be groups of contiguous counties. Samples are selected with probability proportional to size. In stage 2, the PSUs are divided into smaller areas called segments generally equivalent to city blocks, and segments are selected with probability proportional to size. In stage 3, within each segment, households are listed and selected by random sample. In geographic areas with high proportions of adolescents and elderly, minorities, and low-income whites, households are oversampled. In stage 4, individuals are chosen from a list of all persons in selected households. The NHIS uses a similar multistage design to sample individuals in all 50 states and the District of Columbia. Given the purpose and sample design of these surveys, they provide data primarily at the national level.

By the early 1980s, the need for more data on health risk behaviors at the state level led the Centers for Disease Control and Prevention to develop the ***Behavioral Risk Factor Surveillance System (BRFSS)*** (Centers for Disease Control and Prevention, 2009b). Initially, 29 states participated in the program and conducted telephone surveys of the adult population. By 1994, all states, the District of Columbia, and several territories were participating. In addition to a core set of questions, BRFSS includes a set of modules addressing specific health risk behaviors and provides the opportunity for individual states and territories to add state-specific questions.

Although some states from the outset developed telephone sampling designs that would make it possible to report results for selected areas below the state level, BRFSS provides primarily state-level data. In response to demand for more local-level data, the BRFSS SMART program offers data for selected metropolitan areas and small cities with 500 or more BRFSS respondents. Through the BRFSS Maps site, users can download shapefiles to which BRFSS data for individual survey years have been joined (Centers for Disease Control and Prevention, 2009b).

Other countries have developed and implemented similar health surveys. The Health Survey for England, for example, is a series of annual surveys conducted since 1991 (U.K. Department of Health, 2011). Studies comparing these surveys have highlighted differences in methodological approaches (Aromaa, Koponen, Tafforeau, Vermeire, & the HIS/HES Core Group, 2003). Despite these differences, the surveys provide data for international comparative research.

Surveys are also used to screen for problems like lead poisoning, PKU, and hypertension. ***Screening surveys*** are proactive public health activities that attempt to uncover health problems before symptoms appear, when the problems are difficult and expensive to treat. Screening surveys differ in the range and nature of population covered. Some cover the full population, as in screening of newborns for PKU, and thus can be used to estimate incidence rates and create maps of geographic variation in incidence. By contrast, many screening surveys only target high-risk populations and people likely not to have been screened as part of their regular health care. Estimates and maps prepared from such surveys only pertain to the screened population. Reported incidence will naturally be higher in areas where more people were screened, and GIS can be used to explore geographic variation in ***screening penetration***, the percent of risk population screened.

Health Care and Health Care Utilization Data

Medical care provision generates large quantities of information on patients and the treatment they receive, and secondary use of administrative data is made in many health studies. Most medical care providers and insurers maintain data on residential addresses of patients. The geographical organization of health care affects health care utilization, as discussed in Chapters 9 and 10. Nevertheless, data on the locations of medical care providers and the health problems of patients they treat are important sources of data for GIS applications.

HEALTH PROVIDER DATA

Health service information forms another valuable spatial data layer for public health GIS. Most health care providers—hospitals, physicians, clinics—offer their services from fixed locations and can be represented as point spatial data. A few health services, such as emergency medical services and mobile clinics, move from place to place and thus can be modeled as arc or network information. Beyond location, many other dimensions differentiate health services, including price, capacity, utilization, range of services provided, and the elusive quality of care.

Information about the locations and characteristics of health care providers is widely available. ***Gazetteers*** include geographical coordinates for major health facilities such as hospitals. These coordinates can be imported into GIS for mapping and display; however, one must be careful that the location coordinates use the same scale and projection system as the foundation data layer to which they will be linked. One shortcoming of gazetteers is that they do not include data on the characteristics of health service facilities. Such information must be brought in from other sources and linked to the facility sites.

Detailed information on health care providers comes from professional organizations like the American Hospitals Association (AHA) and the American Medical Association (AMA) and, increasingly, from commercial marketing database providers like InfoUSA. The AHA publishes an annual directory of hospitals that includes statistics on utilization, personnel, services, and finances for hospitals in the United States (American Hospital Association, 2009). Included in the directory is each facility's street address, which can be geocoded to a point location. Similar kinds of directories exist for nursing homes and mental health facilities.

For physicians, the AMA's Physician Masterfile offers analogous information and includes addresses that can be geocoded to identify point locations (American Medical Association, 2011). A directory of physicians based on data from the Masterfile is also available (American Medical Association, 2009). The Masterfile covers the vast majority of physicians, but certain important subgroups may be missing, for example, doctors who earned medical degrees outside the United States whose practices are often clustered in immigrant neighborhoods. As with other types of health data, the release of data on providers raises privacy and confidentiality concerns. Physicians have protested the sale of their data to businesses, including pharmaceutical companies (Saul, 2006). State laws prohibiting the sale of doctor-specific prescription drug data are being tested in the federal courts (Saul, 2008).

Data for other types of health care providers are often harder to come by. Health clinics, for example, are operated by federal, state, and local governments as well as voluntary organizations. Each type of agency maintains a list of its own clinics, but there may be no composite listing of facilities in an area. It may be necessary to piece together information from multiple sources or conduct fieldwork to uncover all health service locations. Despite these challenges, creating spatial data layers for health care providers is generally easier than preparing health and foundation data layers. Health services are limited in number, exist at discrete locations, and change relatively slowly over time, making them more manageable to deal with in a GIS context.

HEALTH CARE UTILIZATION DATA

Hospitals generate large quantities of spatial information on patients treated in their inpatient and outpatient facilities. These ***hospital discharge data*** provide an important base for examining hospital utilization and treatment patterns, though they are generally inadequate for population-based studies of morbidity because they are restricted to patients treated in hospitals. The large literature on small-area variations in the rates of medical and surgical procedures relies primarily on hospital discharge data (Wennberg & Gittelsohn, 1982), and the data sets are widely used in health policy analysis and planning. Included in the data sets are demographic information about the patient, primary and secondary diagnoses, diagnostic procedures, treatment procedures, length of stay, and insurance status. Hospital discharge data contain the patient's residential address, but that information is rarely released due to privacy considerations. Instead, hospital data can usually be obtained at the ZIP Code level, because ZIP Codes are part of the address and thus convenient geographical identifiers for the release of hospital information.

This section has described several important, widely available health data sets that can be incorporated in public health GIS. The data sets address a cross section of public health issues and offer a framework for diverse geographical investigations. Increasingly these information resources are available on electronic media, including Internet, and are readily accessible to users in a wide variety of settings (Lacroix & Backus, 2006). Many other health data sets exist. We have not even mentioned the vast proprietary databases held by health insurance companies or the specialized data sets in areas such as occupational, veterinary, and environmental health (Weise, 1997).

Spatial Resolution of Health Data

Regardless of which data sets are used, the spatial resolution of the data is crucial for GIS applications. Although all health data sets deal fundamentally with individuals and usually include address information, none routinely release those detailed geographical identifiers because of important privacy and confidentiality considerations. Thus, the analyst is typically faced with using health data that are aggregated to predefined geographical units, such as counties, ZIP Codes, or census tracts. This raises important substantive issues, as well as significant methodological issues as discussed in Chapter 5. Substantive issues concern the validity and usefulness of particular areal units for public health planning and analysis.

Most data from federal health agencies are available at the county level. Although counties are generally good geographical units for displaying health data at the national scale, they have important limitations (Croner, Pickle, Wolf, & White, 1992). Counties are administrative, political units that bear little relationship to areas defined according to socioeconomic, demographic, or environmental criteria. Counties often encompass diverse physical environments and heterogeneous populations. Moreover, the areas differ greatly in population size and areal extent. Counties large in area visually dominate the national map, despite the fact that they may have tiny populations. Small urban counties can hardly be seen on a national map, though they have huge populations. Thus, counties are not comparable to one another, and they have little basis in population and environmental factors relevant to public health. By comparison, census tracts, defined by the Census Bureau for tabulation purposes, are more similar than counties in population size and follow moderately well the fuzzy boundaries of social, economic, and ethnic areas.

ZIP Codes, commonly used for the tabulation of health data, have problems analogous to those for counties. Originally, ZIP Codes were devised by the U.S. Postal Service to facilitate mail delivery, each ZIP Code representing a collection of mail distribution points. The areas have little correlation with socially and environmentally defined areas. In cities, some ZIP Codes encompass neighborhoods with highly divergent economic and social characteristics. For instance, one ZIP Code in New York City includes census tracts whose 1990 median incomes ranged from $15,000 to $42,000, a threefold difference. A health statistic for such a ZIP Code would represent an “average” of statistics for two very different population groups. Another problem is that ZIP Codes occasionally cut across political and census boundaries and change over time, making it difficult to overlay and integrate ZIP Code data with other sociopolitical data (Kirby, 1996; Krieger et al., 2002). Despite these problems, ZIP Codes in the United States and postal codes in other countries are often used as convenient and practical reporting units for health data in small areas. Analysts should be aware of the strengths and limitations of using ZIP Codes for GIS-based health analysis.

Making Population and Health Data Mappable

In order to use population, health, and health care data sets in GIS, the data sets must first be captured and linked to a foundation spatial database. Data capture is a complex process that draws on an ever-increasing array of tools including scanning, digitizing, downloading from the Internet, and entering data directly from the field via the Global Positioning System. This section focuses on the two procedures typically used for capturing health information—address matching and joining.

Address Matching to Locate Health Events as Points

Health information is often georeferenced by street address. For example, we might have information on the residential addresses of people who died of breast cancer, or the addresses of hospitals, health clinics, schools, or workplaces. Using the process of ***address-match geocoding***, we can convert each address to a point on a map. The point is recorded as a pair of geographical coordinates that connect to the foundation database. At its simplest, address matching involves comparison of two data sets: one containing the addresses of health events and the other a foundation database with its own address information. An address (street name, number, and city, ZIP Code, or other zone) from the first database is compared against the full array of addresses in the second, and a "match" occurs when the two agree.

Address-match geocoding procedures differ depending on the type of foundation spatial database used in matching. Street centerline, address point, and cadastral or property databases have all been used in geocoding (Zandbergen, 2008). Address point databases are commonly developed as part of E911 systems in North America. E911 or Enhanced 911 is a telecommunications system that associates a calling party's telephone number with an address. Address point databases can also be created from parcel data. The point may be located at the centroid of the parcel or at a location where the driveway serving the property intersects with the road. When a match occurs, the health event is assigned the geographical coordinates of the corresponding property. Up-to-date property databases form an accurate platform for address matching because each address is associated with a unique property on the ground. The downside is that such databases are typically very large and cumbersome to work with and, like address point databases, compiled and maintained at the local level.

Although some GIS analyses have used more than one database to geocode addresses (Lovasi et al., 2007), street centerline databases, like TIGER/Line, are widely used as a foundation for address matching. Because street centerline databases do not include unique street addresses for specific structures, but only address ranges along street segments, address matching relies on interpolation. We match the particular address to a street segment (street name and address range), and we estimate the location of the address along the segment by interpolating within the corresponding address range. For example, the street address, 107 Oak Avenue, is assigned to the segment of Oak Avenue with address range 101–119 (Figure 3.10). By interpolation, the location of 107 Oak Street is estimated to be about one third of the way along the street segment. GIS users can specify an offset to take into account the setback of the structure from the street centerline.

This form of address matching does not place points at the exact locations of the structures, but rather at estimated locations along street segments. In urban and suburban areas, where properties are spaced fairly evenly along segments, spatial accuracy is generally quite good. In rural areas, the uneven spatial distribution of properties can cause significant spatial error from interpolation.

Address matching is an iterative procedure in which we first attempt to match all addresses and then correct those that fail to match. Typically one half to two thirds of addresses match in the first attempt. We then examine the unmatched addresses for obvious errors or inconsistencies. Often there are simple errors in spelling or abbreviation that can easily be corrected. After correcting obvious errors, it is typical to achieve a "match rate" of over 90% in most parts of the United States. Anything less calls for an assessment of the quality of both the address list and the spatial database used in matching.

Addresses can fail to match because of errors in the address list, errors in the street or property database, or inconsistencies between them. Errors in the address list are common, and they include misspellings and typographical errors

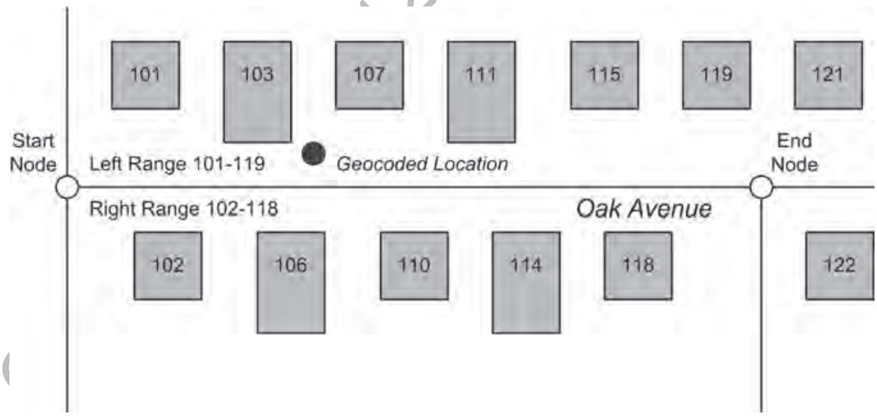


FIGURE 3.10. The TIGER/Line files include street centerline and address-range information that can be used in geocoding. This segment of Oak Avenue is represented by a start node and an end node, each with geographic coordinates. The address range for the left side of the segment contains odd-number addresses, while that for the right side contains even-number addresses. By interpolation, the location of 107 Oak Avenue is estimated to be approximately one third of the distance along the corresponding street segment of Oak Avenue. An offset was applied so that the geocoded point would not lie on the street centerline.

in the street number, street name, or zone (Table 3.2). These errors can easily be corrected by carefully inspecting and editing the address list. Furthermore, most GIS provide the option of automatically correcting the most common types of errors using simple rules and conventions. One should approach these automatic correction algorithms with caution, however, because they may falsely change the original address data and generate a false sense of accuracy.

Errors in the street or parcel database, including missing street segments and incorrect address range information, also create problems for address match-

TABLE 3.2. Sources of Error Affecting Address Match Outcomes

Record content	Street numbers	Street name	Street type	Zone (ZIP Code example)	Address match outcome for perfect match
Correct address	16	Main	St.	13501	Match at correct location
Correct street segment	Left 2–20 Right 1–19	Main	St.	13501	
Error in address record					
Incomplete address		Main	St.	13501	No match
Error in street number	166	Main	St.	13501	No match
Error in street name	16	Nain	St.	13501	No match
Error in street type	16	Main	Rd.	13501	No match
Error in zone	16	Main	St.	113501	No match
Address does not correspond to a real structure	16	Main	St.	13501	Match represents a structure that does not exist
Error in street segment record					
Missing range	Left Right	Main	St.	13501	No match
Error in range	Left 2–14 Right 1–19	Main	St.	13501	No match
Range applied to wrong side of street	Left 1–19 Right 2–20	Main	St.	13501	Match represents incorrect location
Error in street name	Left 2–20 Right 1–19	Nain	St.	13501	No match
Error in street type	Left 2–20 Right 1–19	Main	Rd.	13501	No match
Error in zone	Left 2–20 Right 1–19	Main	St.	113501	No match
Incomplete street network database					No match

ing. As the accuracy of spatial databases improves, it is less common than in the past to find true errors in such databases. Rather, most errors result from the time lag between new residential development and database update. Addresses fail to match because they are located in newly developed areas that have not been mapped or entered into a spatial database. Since these addresses must then be geocoded or digitized by hand, it is well worth the investment to use the most accurate and up-to-date street or parcel database.

Finally, addresses can fail to match because of inconsistencies between the address list and the foundation database. These include differences in street naming convention—for example, "6th Avenue" *versus* "Avenue of the Americas"—or in abbreviation—"St." *versus* "Str." Most GIS automatically correct obvious differences in abbreviation.

Although most analysts emphasize the "match rate," it is important to remember that a successful address match does not guarantee accuracy. Even if an address is successfully matched, it may not be assigned to the correct location. A field check of over 500 geocoded residential addresses to assess spatial accuracy uncovered a variety of errors (Cromley, Archambault, Aye, & McGee, 1997). The relative locations of 7% of the cases were incorrect. A few cases (less than 1%) had been geocoded to locations more than 500 feet away from the correct location. This type of error would be of particular concern in any study measuring distances from the geocoded location to another location because the true distance would be over- or underestimated. The remaining cases were estimated to be out of position by less than 500 feet. About half of these cases were on the wrong side of the street or on the wrong corner of an intersection. This type of error would be of particular concern in any study aggregating cases to an area like a census block or block group because census area boundaries often coincide with street centerlines, so cases on the wrong side of the street would be aggregated to incorrect spatial units. For 1% of the addresses, no residential structure could be found. Either the structure had been removed or the street number was incorrect but fell within a valid address range. The type of error—successfully geocoding an address that does not exist—may account for the higher match rate for street centerline geocoding versus parcel geocoding. Such errors can have significant impacts on spatial analyses based on geocoded data (Griffith, Millones, Vincent, Johnson, & Hunt, 2007).

These findings emphasize the importance of obtaining accurate address information and the need to look beyond the match rate in geocoding. Typically, the collection of addresses is decentralized. Addresses are entered at the source institution, for example, a hospital, doctor's office, or health clinic. From there, the institution transmits the information to a public health agency for mapping. Unless the addresses are used for billing or follow-up, the institution will have little stake in their accuracy and completeness. Errors emerge much later during address matching, and data editing and cleaning are performed by GIS personnel far removed from the source of data collection. Improving accuracy in geocoded address information requires not just better address-matching algorithms, but institutional arrangements that foster accuracy at the source.

The findings also emphasize the need for field checking of data, particularly when research findings are sensitive to the locations of cases in a few places. Researchers involved in GIS studies at a community scale can benefit from field trips to the study area before data collection and analysis to familiarize themselves with residential patterns and other landscape features of relevance to the particular study.

The widespread use of geocoded health data has led many health analysts to investigate methodological issues in geocoding (Rushton et al., 2008). A relatively neglected issue in the study of geocoding methods is the spatial distribution of errors. Many analysts have noted that geocoding rates and accuracy are lower in rural areas, but more research is needed to model other forms of error and their implications for spatial data analysis.

Joining Health Data to Geographical Areas

Many population and health databases only present information for geographical areas like counties, ZIP Codes, or census tracts. They include the area name and/or identifier and a set of variables that describe the health events, population, or other attributes of the area—for example, the census tract number and number of diagnosed cases of AIDS by tract. Capturing area data in a GIS involves ***joining***. We join the tabular data to a foundation spatial data set of area boundaries based on a common field like the census tract identifier. The data for each tract are attached to the corresponding tract in the foundation database.

Joining requires that each geographic area have a unique identifier, either a name or number. In the United States, state names are unique, as are ZIP Code numbers. However, many widely used areal units such as census tracts or blocks have identifier numbers that are unique only within larger units of geography. Census tract numbers are unique only within counties, and block numbers are unique only within tracts. A project that cuts across these larger units must create a new field that uniquely identifies each small area. In a tract-level study that encompasses many counties in several states, for example, the state number and county number must be included along with the tract number to define each tract (Figure 3.11).

For more than 30 years, the Census Bureau used the Federal Information Processing Standard (FIPS) codes for states, named populated places, primary county divisions, and other entities, issued by the National Institute of Standards and Technology. At the time of preparation for the 2010 census, the NIST standards were withdrawn and the Bureau was transitioning to a set of codes issued by the American National Standards Institute (ANSI). Many of the codes adopted by ANSI are unchanged, but users need to familiarize themselves with the new standards (U.S. Census Bureau, 2010b). The InterNational Committee for Information Technology Standards 446-2008 standard is tied to the Geographic Names Information System (GNIS) managed by the U.S. Geological Survey. Familiarity with the ANSI and Census Bureau identifiers for geographical units within GIS databases is important for accurately joining and manipulating geographic databases produced by the

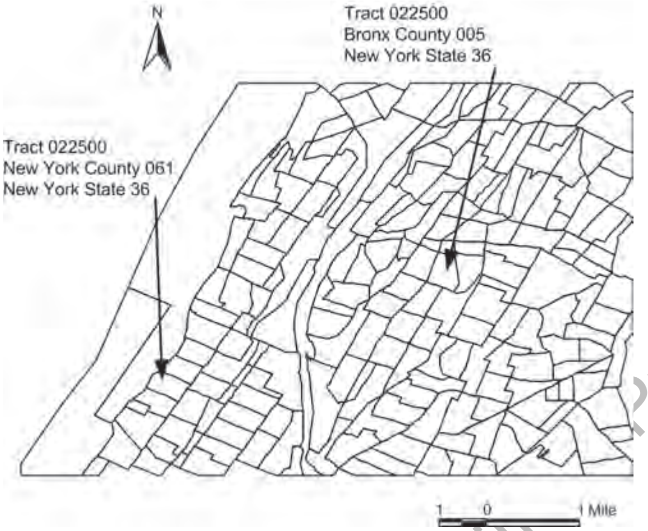


FIGURE 3.11. Census identifiers for tracts, block groups, and blocks are unique only in the context of the hierarchy of census units. Two tracts in New York City have the same tract identifier number, 022500. Use should be made of the full codes, 36061022500 for the tract in New York County and 36005022500 for the tract in Bronx County, to avoid confusion and errors in joining tables of data.

federal government (Table 3.3). State governments may have developed additional numeric identifiers for geographical units within their states.

Typically, joining links area-based health information to the corresponding geographic areas in a foundation spatial database. However, the procedure can also be used with address-based health data to find the area in which a health event is located. We match the address field in the health data set directly to the corresponding address field (street name and address range) in a foundation database table, like TIGER/Line, that contains area identifiers. Two tables, one for the health data set and the other for the foundation data set, are joined based on common address information. If this approach is used, the ability to map address-based health data as points is lost. As a consequence, the methods described in Chapters 4 through 10 for analyzing patterns of health data represented as points cannot be applied. Spatial information is lost when address-based health data are joined to areas rather than geocoded as points.

Database Integration

The power of a GIS lies in its ability to link, integrate, and manipulate the diverse types of spatial data described in this chapter. Integrating such data sets

TABLE 3.3. Comparison of ANSI, Census, and State Identifiers for an Area

	Hartford (Populated place)	Hartford (City)	Hartford (Town of)
ANSI/Census state code for Connecticut	09	09	09
ANSI/Census county code for Hartford County	003	003	003
ANSI GNIS feature identifier	213160	2378277	213442
Census code	37000	37000	37070
State of Connecticut town identifier	—	—	064

Note. These entries show that there are six different numerical codes for the same geographic area called Hartford. The ANSI and Census codes for the state of Connecticut and for Hartford County are the same. The ANSI standard using GNIS (Geographic Names Information System) codes has three different feature identifiers for Hartford. The census has two different codes for Hartford, one as an incorporated place and one as a minor civil division. Under a numerical coding system used by the state, the town of Hartford is 064. GIS users need to be aware of the coding systems that have been used to assign identifiers to geographic areas when joining and linking databases.

can be challenging, especially when the data sets differ in scale, resolution, and geographic extent. Most GIS packages include a series of cartographic and geographic procedures for linking and integrating spatial data sets (Table 3.4).

A common data integration problem arises when data layers that will be overlayed or linked in a GIS rely on different coordinate systems or different map projections. Typically this occurs when a health or an environmental data layer is being integrated with a designated foundation data layer. Common in all GIS are procedures for transforming coordinates so that they are consistent with those of the foundation data layer. ***Coordinate translation*** involves computing new coordinates as a mathematical function of the original set. Linear transformations, for example, can be used to move, stretch, or twist the coordinate axes (Figure 3.12). These simple linear transformations are often necessary when integrating spatial data from a digitizing tablet or scanner with existing geospatial foundation data sets like DOQQs or DLGs.

Sometimes geographical errors in overlaying data layers stem from positional inaccuracies that are unevenly or unpredictably distributed across the map. In this case, matching data layers requires nonlinear coordinate transformations that stretch or shrink different parts of a map until features align correctly with those on the foundation data layer. ***Rubbersheeting*** is the process of geometrically adjusting features to force a digital map to fit the designated foundation data layer (Antenucci et al., 1991). Rubbersheeting changes the relative locations of features, thus distorting the original map. Therefore the process should be used judiciously to make relatively small changes in map coordinates.

TABLE 3.4. Spatial Database Collection and Preprocessing Operations

Function class	Function
Data collection	Scanning Digitizing Address-match geocoding
Data conversion	Importing/exporting Edgematching Clipping Raster/vector conversion
Geometric transformation	Translation Rotation Map projection Rubbersheeting
Generalization	Line thinning Line smoothing

Of course, when the map is being linked to an up-to-date, planimetrically correct foundation data set, rubbersheeting can partially compensate for positional errors in the source map. Distorting an inaccurate source map may be a good thing. Rubbersheeting is often required in order to integrate data with low or unknown positional accuracy with more accurate foundation data layers, for example, in linking the TIGER/Line files, with their variable positional accuracy, to a DOQQ base.

Another kind of coordinate transformation is needed when data layers are based on different map projections. ***Map projection transformation*** is the change in coordinates from one map projection to another. Data that come from different sources often utilize different map projections, so that coordinates must be reprojected for the data to overlay properly. All GIS have built-in functions for converting among common map projections.

Another common problem in creating and linking spatial data sets involves changing the geographic extent of the data set. The analyst may want to focus on one portion of the mapped area—for example, one municipality within a county—or to join maps together to create a map layer that covers a larger geographic area. In GIS, one can extract a portion of a mapped data set by cutting out the portion from surrounding areas and saving it in a separate file (Figure 3.13). These ***clipping*** or ***windowing*** procedures are easy and efficient in GIS where they can be done by using the cursor to define a rectangle or irregular shape.

Edgematching is a procedure for joining maps together by matching common features along the shared map boundary. For example, a particular road

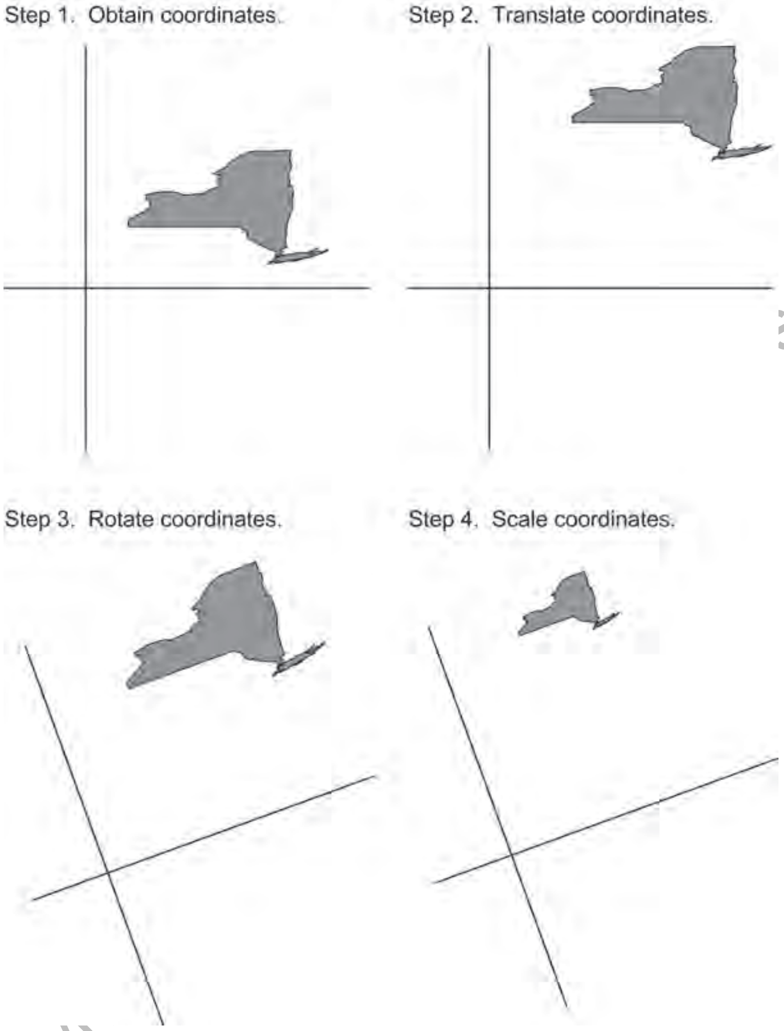

Step 1. Obtain coordinates.
Step 2. Translate coordinates.
Step 3. Rotate coordinates.
Step 4. Scale coordinates.

FIGURE 3.12. Coordinate translation of a spatial database of New York State.

appearing at the edge of one map is matched to its counterpart at the edge of the adjacent map (Figure 3.14). Most edgematching procedures also adjust features located in the area of overlap between the two adjacent maps to create a seamless map. For the procedure to work correctly, the maps being joined must have the same scale or resolution and contain comparable features. It is important to realize that edgematching creates a new topology for the database as, for example, two line segments are joined into one. The new topology enables new geographical analyses that address issues within the larger area.

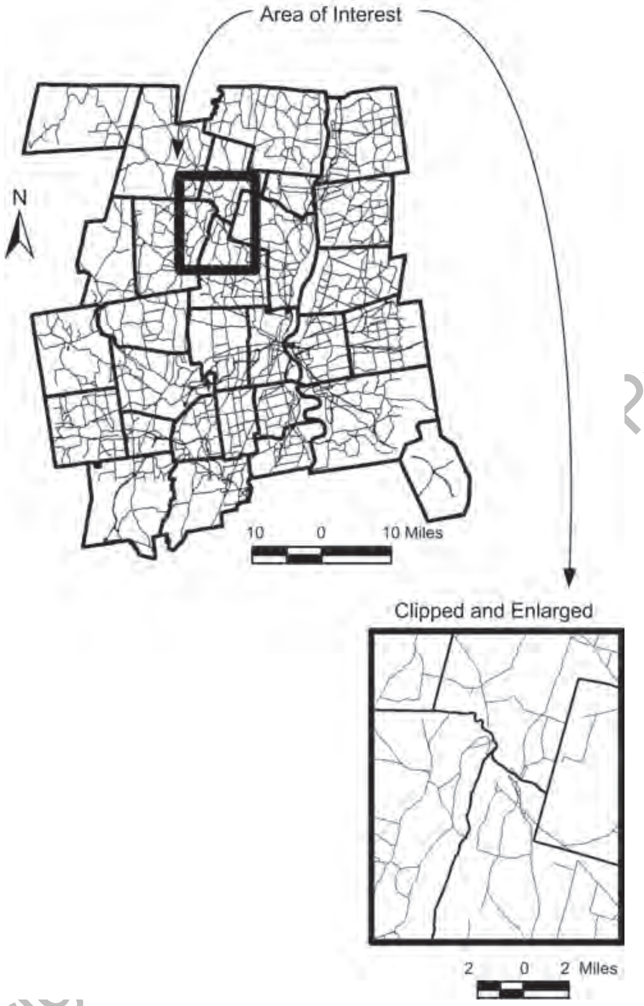


FIGURE 3.13. A window created around an area of interest can be used to “clip” the features of interest for viewing and analysis and for creating new databases containing just the clipped features.

Data Sharing

Assembling diverse spatial data sets and linking them with foundation spatial data is a time consuming, labor-intensive, and expensive process. The final product—an integrated ensemble of health, environmental, social, and foundation data—represents not only a major investment, but also a major resource, with value to other users analyzing issues in the same geographic area. ***Data sharing***, or the transfer of data between two or more organizations, offers many impor-

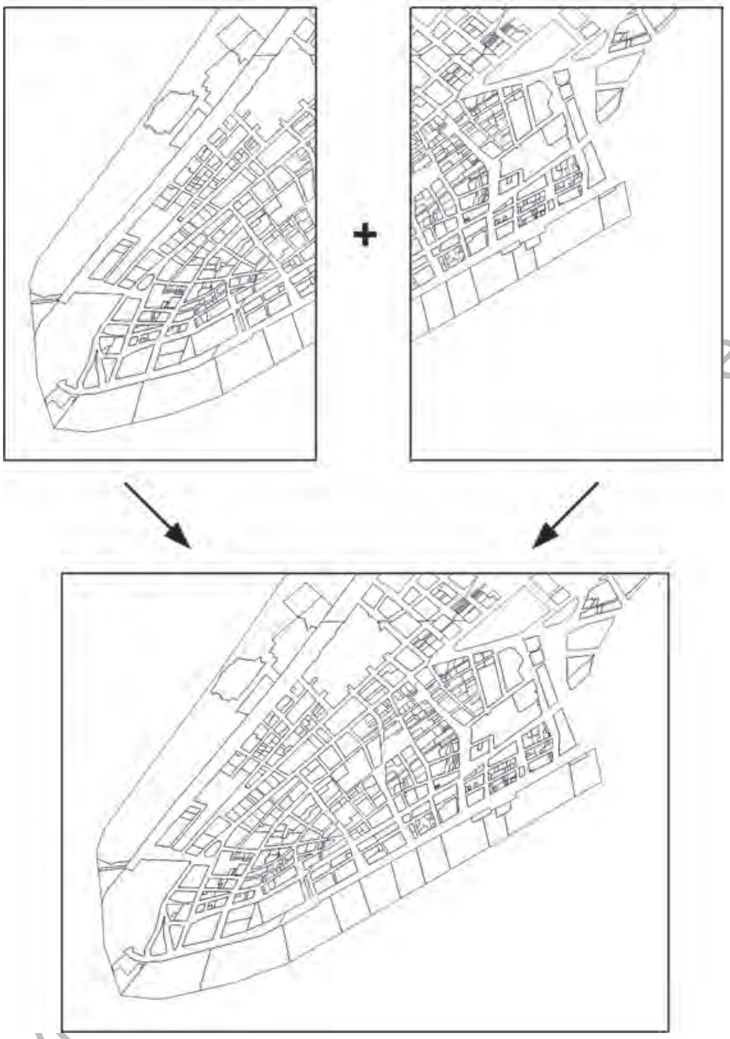


FIGURE 3.14. Two spatial databases of information for adjoining areas are joined by "matching" common features along the boundary to create a single seamless database.

tant benefits to the developers and users of geographic information (Onsrud & Rushton, 1995). The value of spatial data derives from its use, so enabling diverse groups to draw on the same data creates value by stimulating use. Data sharing is also a means for spreading the costs of database creation among multiple users and avoiding needless duplication of effort. Finally, there are often synergies in multiple use and analysis of a common spatial database. One group's insights spark another's, resulting in greater value overall.

Organizations at the regional, state, and national levels have increasingly recognized these benefits and taken steps to promote spatial data sharing. States have taken the lead by creating spatial data clearinghouses or unified state-level spatial databases. Most of these efforts involve extensive participation by local governments, which provide spatial data and draw upon it for local and regional planning purposes. At the federal level, the challenges of developing a ***national spatial data infrastructure*** in the United States over the last two decades have been acknowledged (Goodchild, Fu, & Rich, 2007; Craig, 2009).

Despite the many advantages of data sharing, technical and institutional barriers often get in the way. Sharing requires networked systems and agreements and common data formats that permit electronic exchange of information among users. Differences in hardware, software, and metadata standards impede spatial data sharing. As the volume of spatial data produced and used has grown, producers and users of data have needed to confront the legal implications relating to the dissemination and use of data. Issues of intellectual property rights, contract law, and liability affect data sharing in many countries (Cho, 2005).

More fundamentally, sharing requires cooperation among diverse institutions and branches of government and a shared sense of purpose. Differences in organizational needs, cultures, and interests make cooperation among organizations challenging at best (Obermeyer, 1995). Organizations often operate autonomously, emphasizing their particular needs and missions. In many cases, according to Craig (1995, p. 108), "agencies could share data, but they choose not to do so." Thus, data sharing is an inherently political process reflecting power, inertia, and access to resources. These political and institutional factors far outweigh the technical barriers to data sharing (Onsrud & Rushton, 1995).

An important barrier to sharing health data is the need to protect the privacy and confidentiality of health information. Many state and federal agencies and health care providers gather health data on individuals and are involved in data sharing. The development of health informatics, including electronic and personal health records, has raised additional privacy, confidentiality, and security concerns (O'Carroll, Yasnoff, Ward, Ripp, & Martin, 2003). A survey of public health professionals in Canada and the United Kingdom revealed that 71% identified privacy issues as an obstacle to public health practice (AbdelMalik, Boulos, & Jones, 2008). In the United States, the ***Health Insurance Portability and Accountability Act (HIPAA)*** of 1996 established standards for privacy of individually identifiable health information (U.S. Department of Health and Human Services, 2003). Geographic identifiers are included in the list of identifiers that must be removed to de-identify data so that an entity that passes the data to a third party can be held harmless under the law (Table 3.5). State departments of public health in the United States that use GIS are able to use individually identifiable health data in their work. Entities covered by HIPAA may also create limited data sets for use by other parties who enter into a data use agreement prior to the disclosure and use of the limited data. Methods for mapping individual geographic data to protect privacy are discussed in Chap-

TABLE 3.5. HIPAA Identifiers

Name

All geographic subdivisions smaller than a state, including:
- Street address and equivalent geocodes
- City and equivalent geocodes
- County and equivalent geocodes
- Precinct and equivalent geocodes
- ZIP Code and equivalent geocodes

Except for:

The initial three digits of a ZIP Code if, according to publicly available data from the Census Bureau;

The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people; and

The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people is changed to 000.

All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.

Telephone numbers

Fax numbers

Electronic mail numbers

Social Security numbers

Medical record numbers

Health plan beneficiary numbers

Account numbers

Certificate/license numbers

Vehicle identifiers and serial numbers, including license plate numbers

Device identifiers and serial numbers

Web Universal Resource Locators (URLs)

Internet Protocol (IP) address numbers

Biometric identifiers, including finger and voice prints

Full-face photographic images and any comparable images

Any other unique identifying number, characteristic, or code, except as permitted ter 7. There is growing awareness that advances in technology have made reidentifying health data easier and that alternatives to standard de-identification practices are needed.

Conclusion

This chapter has examined spatial data resources for public health GIS in the United States and geographical, technical, and institutional concerns in data integration. Investing in a GIS means investing in spatial data. Given the wide array of data sets available and the high costs of new database development, organizations need to assess carefully their spatial data needs and view development as a long-term investment rather than a short-term expense. As developers and users of spatial data, it is essential that public health organizations participate in the emerging efforts to create open, accessible, and integrated spatial data resources. Agencies need to plan how spatial data will be used internally, how to make it accessible to others, and how to promote spatial data sharing in partnership with other organizations.

Guilford Publications
72 Spring Street
New York, NY 10012
212-431-9800
800-365-7006
www.guilford.com