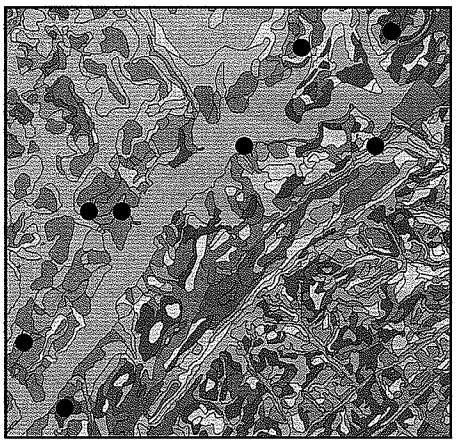

CHAPTER 26

GIS IN HUMAN HEALTH STUDIES

JOSEPH E. BUNNELL, ALEXANDER W. KARLSEN, AND ROBERT B. FINKELMAN*
United States Geological Survey

TIMOTHY M. SHIELDS
The Johns Hopkins University

CONTENTS

I. INTRODUCTION TO DATABASES AND GEOGRAPHIC INFORMATION SYSTEMS

Databases used in the field of medical geology are generally comprised of geospatial and/or temporal elements. Although these are not requirements for all medical geology research projects, much of the discussion in this chapter will be focused on databases incorporated into geographic information systems (GIS). GIS are computer-based (or manual) methods that allow a user to input, store, retrieve, manipulate, analyze, and output spatial data (Aronoff, 1989). There are four major systems of GIS: engineering mapping systems (computer-aided design/computer-assisted mapping; CAD/CAM), geographic base file systems, image processing systems, and generalized thematic mapping systems. Various software packages are available that perform one or more of these systems, and the relative ability to move data back and forth between them can be critical to the needs and success of a particular GIS. Relational databases are the most commonly used types of databases in GIS (Cromley & McLafferty, 2002). Relational database management models are convenient for linking formerly disparate databases together in a GIS. The databases to be joined must share one common attribute, usually an identifier such as coded patient number, sample site, or latitude/longitude. Other database management structures, such as hierarchical and network systems, are not as well suited to health GIS applications, although they may be useful for extremely large databases.

The capability to quickly and easily link large medical or public health databases with equally large geospatial databases represents an important technological breakthrough. Due to advances in computational power and speed, studies may be conducted today that could not have been done in reasonable time frames even just a few years ago. By linking disparate databases in a visually accessible manner (i.e., with maps), researchers are able to recognize relationships or discern patterns of

*The views expressed by the author are his own and do not represent the views of the United States Geological Survey or the United States.

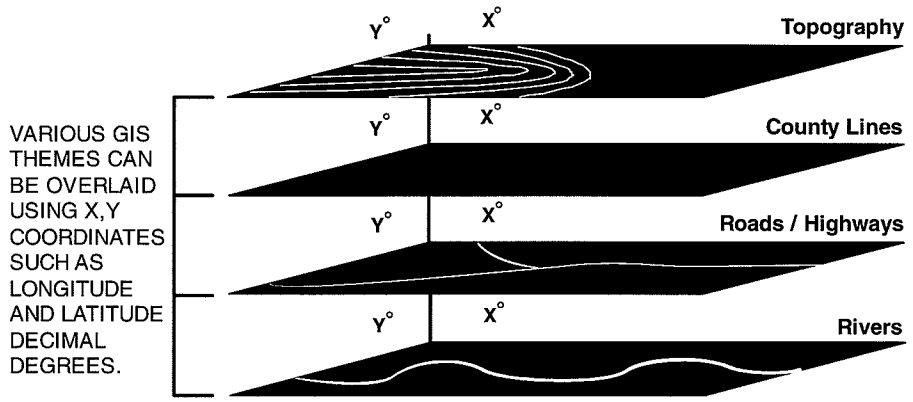

FIGURE 1 GIS conceptual diagram. Note that different data layers (covers or themes) are overlaid such that queries about individual point locations reveal numerous attributes from a variety of source databases. (Figure courtesy of Eric Morrissey, U.S. Geological Survey, Reston, VA.)

disease that can lead to an understanding of causality that was previously not apparent. The value in mapping disease occurrence is appreciated when doing so can illuminate the underlying cause of an outbreak, which may then enable mitigation measures to be taken to prevent further spread of a disease.

The earliest example of using such a spatioanalytical approach to solving an epidemiological riddle is generally credited to a physician named John Snow and the Broad Street Pump of London in 1854 (Cameron & Jones, 1983). Dr. Snow mapped a major outbreak of cholera, in a time before the germ theory was well accepted, and hypothesized that there was a causal association between the putative source of the contagion and locations of cases. He convinced city officials to remove the handle of the pump dispensing contaminated water—an intervention that promptly quelled the outbreak. Although modern tools are much more sophisticated than those at Dr. Snow's disposal, our goal remains the same in applying GIS to public health issues.

Currently, databases used in GIS applications are often developed by the user or are available on the Internet or from other sources. Databases are typically stored in electronic media formats such as a hard drive, floppy disk, and compact disc-read only memory (CD-ROM). GIS databases contain fields in columns and records in rows. A field, or item, is an element of a database record in which one piece of information is stored and represented as a column in a geodatabase table or spreadsheet (Kennedy, 2001). *Records* represent different entities with different values for the attributes represented by the fields. (Kennedy, 2001). *Attributes* are information about geographic features, and they are contained within GIS data layers, or themes (Figure 1). For example, a climate data layer (the feature) may contain the attributes of temperature, rainfall, and relative humidity for a specific geographic point location or region. Attribute data and spatial data comprise the two critical types of data in a GIS. A more in-depth look at how to assemble databases into a project is presented in Other Sources, part C.

Along with the individual databases, *metadata* are also included. Metadata, also called data dictionaries, are simply data about data. Metadata contain information such as the time/place the database was created, field and record identifier information (attributes), data development process (lineage), and individual(s) to contact regarding the data. If the data are displayed in a geographic environment, the metadata must also include additional information such as map scale and projection. Guidelines for what should be included in metadata are provided by the National Spatial Data Infrastructure, which is maintained by the Federal Geographic Data Committee (FGDC) (see Other Sources, part A). An interagency U.S. government organization, the FGDC sets guidelines for all aspects of spatial data, and works with offshore/international partners to develop the global spatial data infrastructure.

II. TYPES OF DATABASES AND THEIR FEATURES

There are enormous numbers of databases available in digital format, but the types of data likely to be used by

medical geologists fall into two broad categories: Earth science/geospatial databases and biomedical/health databases (see Other Sources, part A). What makes the field of medical geology innovative and unique is that it, by definition, brings together in a coherent manner databases from these two general areas in specific applications. This approach leads to fresh perspectives enabling recognition of connections between environmental factors and human health outcomes that may have previously gone unnoticed. Medical geological research can identify mechanistic connections that in turn can lead to new practices or policies. This may result in novel solutions to public health problems, which ultimately benefit large numbers of people. Two case studies are presented in Sections IV and V that illustrate the utility of such an approach.

Spatial data are represented in two models, vector and raster. In the vector models attribute data are attached or linked to one of three features: point, line, or polygon. Simply defined, a point is an x,y coordinate such as a mountain peak or soil sample location. The point feature does not have any length or area. A line is defined by the connection of two or more vertices (x,y coordinate pairs). A polyline is made up of numerous lines that represent the same feature, such as a road or river. The line or polyline feature has a length associated that is considered too narrow at the given scale to have an area. The polygon feature is defined by a series of lines that start and end at the same place, such as a state or country. Perimeter and area can be calculated for these features. In summary a line is a set of connected points whereas a polygon is a set of connected lines that have the same beginning and end points.

Features in a vector-based GIS can be linked or joined with attribute data from one or more databases provided a common identifier exists. The National Climatic Data Center (NCDC) generates climate data from ground-based observations. They supply a table of weather stations along with the stations' identifiers and x,y coordinates. This table can be imported into a GIS and point features can easily be made which correspond with weather station locations. Once these features exist in a GIS, they can be linked to other NCDC tables that contain station identifiers and hourly, daily, weekly, monthly, and/or annual climate data. Climatic conditions can be analyzed from the temporal frame of a single point in time or a complete historical compilation.

A key function of a vector-based GIS is topology. Topology is the spatial linkage between vector features. It stores the spatial relationship of the features with respect to each other. Topology enables the user to determine where a feature is in relation to other features, which parts of different features are shared, and how features are connected. Functions such as which sample points are located within a specific watershed or which counties does a river intersect can be performed (see Figure 1). Topology also reduces the amount of information that must be stored. If two polygons are adjacent, that is they share a common line as a border, the common line needs to be stored only once. It will be saved as the right side of one polygon and as the left side of the other.

Raster systems present data in a regular grid of squares or cells. These cells are often called "pixels," which is an abbreviation of the words picture elements. Each pixel is defined by a row and column number and in GIS these can be converted to x,y coordinates. Pixels contain a single attribute value relating to the feature they represent. In an image that represents soil, a unique value would be given for each type of soil.

The differences in the manner in which vector and raster systems present and store geospatial data often lead people to contrast the two in order to determine which is better. Each model has its advantages and disadvantages. Vector systems, with points, lines, and polygons produce maps that are more like those drawn by hand and therefore are more aesthetically pleasing. Raster systems tend to represent continuous surfaces such as elevation or vegetation well. At the same time they demand large storage capacity because they require a value for every pixel in the image. The answer to the question of which system is better is dependent upon the application. Both systems are effective, but the nature of the application or task generally determines which one should be utilized.

III. SOFTWARE, COMPUTATIONAL TECHNOLOGY, AND TECHNICAL ISSUES

Developments in software are facilitating the integration of these two forms of data. Vector-based systems now incorporate some raster-based system functionality and vice versa. The crossover of functionality has helped the user immensely. No longer do entire layers of data need to be converted from one data structure to the other. The integration of the two forms has made data management faster and less expensive while providing better quality of data generated.

Getting data to be used in GIS is often a challenge. Manual efforts such as digitizing (using a flat digitizing

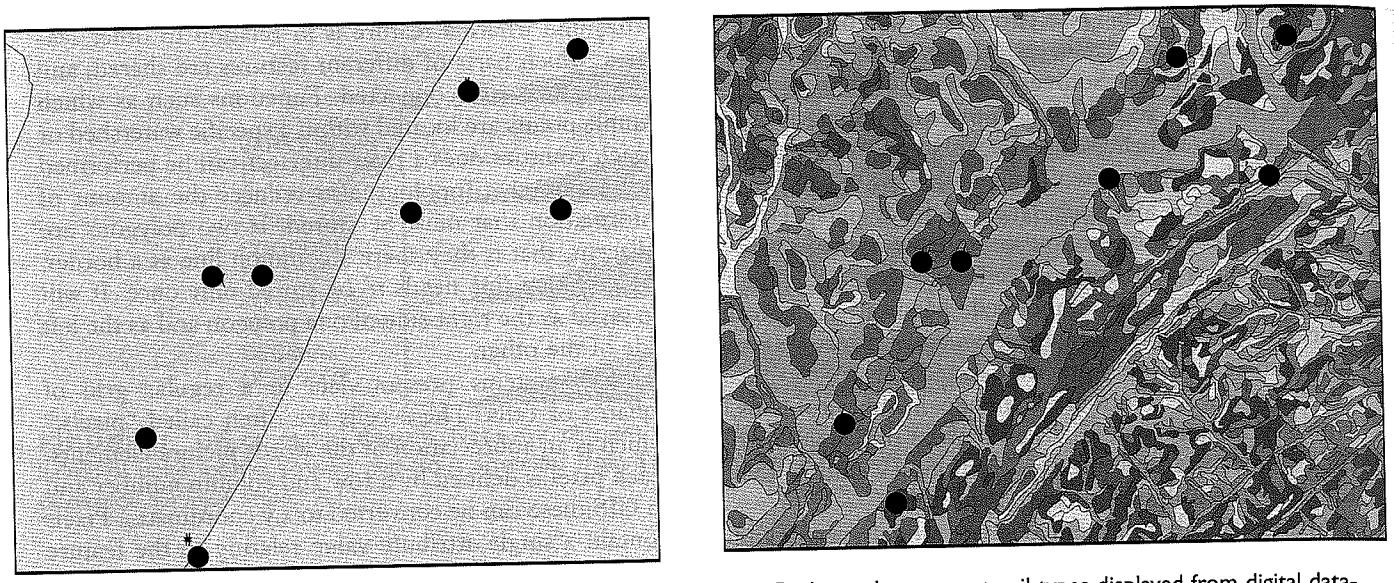

FIGURE 2 Example of the effect of using databases of different scale. Both panels represent soil types displayed from digital databases for the exact same field sites (locations of individual transects are indicated with black circles, and latitudes and longitudes have been determined with a hand-held global positioning system [GPS] device). Left panel shows level of resolution available with STATSGO data (from the U.S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE), 1:250,000. Right panel reveals much greater detail available with SSURGO data (from the U.S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE), 1:24,000. (Data from Bunnell et al., 2003.)

table to draw or copy a map into an electronic format), geocoding (using an electronic basemap to match an address to), or scanning (using a scanner to convert a paper map to electronic format) are time intensive. Using existing digital sources is often preferable, but these sources also have their share of issues pertaining to quality. These are outlined below.

Scale specifies the level at which real-world features have been reduced to be represented. It is usually stated as a ratio or fraction such as 1 : 1000 or 1/1000, where one unit on the map represents one thousand units in the area represented. An often confusing term for maps is small- or large-scale. Small-scale maps actually represent large areas but the ratio or fraction is a small number. Conversely, large-scale maps represent small areas. The map of a neighborhood would be considered large-scale whereas a map of Asia would be small-scale.

Resolution refers to the amount of detail in the features of the map. The scale of the map determines its resolution. It is the level at which features can be distinguished. On a small-scale map local features, such as ponds and small lakes, will not be represented. The terms fine (high) resolution and large-scale are synonymous; they contrast with coarse (low) resolution, and small-scale.

In GIS, issues of scale and resolution can determine which functions are appropriate as well as the level at which results can be stated (see Figure 2). Ecological fallacy occurs when statements or predictions are made at one level based upon observations made at another. A soil survey conducted in 1 of the 24 counties of Maryland would be insufficient to predict soil characteristics throughout the state. Likewise, a small-scale digital elevation model would be inappropriate to use as part of a local water drainage study.

Accuracy is an important concept when working with GIS. Spatial accuracy, or how well the mapped features are located, must be stated and understood. If the map you are using states a spatial accuracy of ±1000 feet, this will not be accurate enough to select a point to dig a well. Likewise, temporal accuracy must be stated. Population data from 1980 will not be accurate enough for current demographic studies. These are just a couple of ways error can be introduced in GIS.

Projections were developed to represent the curved surface of the Earth on a flat map. Projections are utilized to preserve local angles and shapes or relative size of areas. Every projection distorts the map in some manner and should be carefully chosen with respect to the study. Most GIS can convert data from one projec-

tion to another, which enables data layers from different sources to be compiled.

Metadata is information about data. All data sets utilized in GIS should have metadata accompanying them. This information should include the origin and characteristics of the data set, the purpose of the data set, and any problems the data set may have. This information is critical for the proper utilization of GIS data.

Remote sensing technology provides satellite imagery and high-resolution, high-altitude aerial photography. Such image data are also becoming more common in the context of GIS. Raster data can exist either simply or with multiple values, for example, representing spectral bands. Geodatabase features with the same type of geometry make up simple or topological feature classes (Zeiler, 1999). Object classes retain descriptive information related to geographic features, but they are not elements found on a map (Zeiler, 1999). Depths of wells could make up an object class, for example, in a medical geology GIS examining proximity of drinking water wells to sources of arsenic in Bangladesh (see also Chapters 11 and 27, this chapter). Due to advances in microcomputer technology, large raster data sets can now be manipulated and spatial data rigorously analyzed statistically with relative ease, thus making incorporation into epidemiological frameworks feasible (Robinson, 2000).

Medical geologists can now move beyond simply noting spatial coincidences of environmental features and disease patterns. By taking advantage of increasing computational speed and capabilities, sophisticated spatial statistics can be used in conjunction with GIS to reduce bias and correct for such potentially confounding effects as non-constant variance and autocorrelation (Haining, 1998). Moreover, spatial statistical models can rigorously test for clustering versus random distributions, and can incorporate a fourth, temporal dimension to better assess correlation and offer clues into disease etiology (Kulldorff, 1998).

IV. CASE STUDY 1: LYME DISEASE

The first is a study designed to identify environmental determinants of tick abundance in the Mid-Atlantic region of the United States (Bunnell et al., 2003). Lyme disease is the most commonly reported vector-borne disease in the United States and it is still rapidly growing with over ten thousand new cases annually (Centers for Disease Control and Prevention, 2001). In this part of the country, the black-legged, or deer, tick (*Ixodes scapularis*) transmits the microbial agents, that cause Lyme disease, ehrlichiosis and babesiosis, and other human and veterinary ailments. Tick abundance is likely a more reliable measure of the effect of landscape features on Lyme disease risk than human case data for several reasons, most notably because the location of Lyme disease cases will be, at best, the patients' home address, while many if not most of the cases are actually acquired elsewhere. Furthermore, by using tick distribution patterns, one avoids potentially misleading interpretations resulting from over- or underreporting due to the challenges in accurately diagnosing Lyme disease in humans. By better understanding the effects of environmental parameters on tick distribution, public health intervention strategies will likely be improved. Note that the word "vector" has a specialized meaning in the context of GIS (a mathematical definition, as above); in the biomedical community, a vector is an insect or other arthropod that actively transmits a pathogen from an infected reservoir host animal to another individual. Chapter 27 offers a thorough discussion of GIS technology applied to vector-borne diseases.

Environmental factors (covariates) that were previously known or suspected to correlate to tick abundance patterns included elevation, land cover, forest distribution, watersheds, and soil type (Glass et al., 1995; Ostfeld et al., 1996; Kitron & Kazmierczak, 1997; Jensen et al., 2000). Digital elevation model (DEM) databases were obtained in a raster format (U.S. Geological Survey, Reston, VA). Land cover attributes were obtained from the remotely sensed multiresolution landscape characterization (MRLC) Landsat thematic mapper (TM) source data, managed by the Earth Resources Observation Systems (EROS), an entity within the U.S. Geological Survey (USGS). Initially, soils data were obtained from the state soil geographic database (STATSGO) maintained by the Natural Resources Conservation Service (U.S. Department of Agriculture, National Soil Survey Center, Lincoln, NE).

From a five-state region, 320 field sites were randomly selected and ticks were collected along transects at each site. Latitude and longitude were recorded at the beginning and end of each transect. At first, locations of the transects were noted on a paper map, 7.5-minute topographical quadrangles were digitized by hand, and the sites were matched up. Digitizing a paper map requires a digitizing table and specialized hardware and

computer software. The digitizing process was not required when latitudes and longitudes were recorded directly to a global positioning system (GPS) device, and then entered into a desktop personal computer. Newer GPS devices can further streamline the process by coupling directly with a laptop or desktop computer. This obviates the need for data entry by hand. Each field site, or transect, became a unique identifier, and the latitude and longitude were the link to the environmental data downloaded. DEM data were used as is so that elevation in meters was obtained at each transect. Land cover data were of limited usefulness, as they only indicated whether a given field site was located in forest, low-intensity residential, open water, etc. Because tick population densities are positively associated with forest edges (ecotones), and because ticks are not found in open water, the GIS was used to calculate the distance from the midpoint of each transect to the nearest specific type of forest and body of water. The databases were thus manipulated from their native state, which resulted in new databases of particular utility to this specific research project.

Newly developed spatial statistics that incorporated spatial autocorrelation were applied to the data, and multiple regression analysis was performed (Das et al., 2002). These techniques revealed significant associations between tick abundance and certain environmental covariates that included soil type. This latter finding was particularly intriguing, and since the inception of the project, some soil data at much higher resolution was made available digitally. The soil survey geographic database (SSURGO) was released county by county as it became available (U.S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE). Only 75 of the 320 field sites happened to be located in counties with SSURGO data available at the time of analysis, and a second analysis was conducted on that subset of field sites. Because of the much finer resolution (1:24,000 vs. 1:250,000 for SSURGO and STATSGO, respectively), interpretations were made with greater precision (Figure 2).

The enhanced resolution available with SSURGO data, combined with the more extensive set of attributes of this database, revealed some surprising results. For example, with STATSGO data, well-drained soils were found to be positively associated with tick abundance, in keeping with previous reports in the literature. However, upon analysis using SSURGO data, it was found that poorly drained soils, too, could be positively associated with tick abundance. This seeming contradiction was apparently resolved by considering precipitation factors and water-holding capacity of the soil. This example demonstrates the power of a GIS approach to examining environmental influences on factors controlling human disease risk. Observation of previously obscured patterns has enabled a generation of hypotheses now being tested to explain factors responsible for spatioanalytical trends in biological terms. Important advances in our understanding of basic Lyme disease ecology are likely to follow from this application of newly acquired and improved computational technology.

V. CASE STUDY 2: FLUOROSIS IN CHINA

Elucidating the causes of fluorosis in the People's Republic of China offers another example of how GIS can be used to address the relationship between human health problems and geologic materials. Fluorosis, an abnormal condition of bones and teeth caused by exposure to excessive amounts of fluorine, affects millions of people throughout China. There are three principal pathways of exposure: drinking high-fluorine water, drinking tea made from tea leaves rich in fluorine, and exposure to fumes from residential combustion of high-fluorine coal or briquettes made with fluorine-rich clays as a binder (Zhang & Cao 1996; Ando et al., 1998) (see also Chapter 12, this chapter).

Until recently, only general information existed on the epidemiology of fluorosis in China. For example, it was known that Kazakhs in the Xinjiang Autonomous Region in northwestern China were exposed to high levels of fluorine due to their preference for "brick tea" made from tea leaves rich in fluorine (Ben et al., 2000).

To determine where fluorosis was likely to be caused by exposure to fluorine-rich coal or coal briquettes, two GIS layers were required: the distribution of fluorosis and the distribution of coal deposits in China. No digital versions of either layer could be located. A map of the distribution and prevalence rates of dental fluorosis by county in China was located (Jianan, 1989) and a map of the coal deposits of China was obtained (Ruiling et al., 1996). Both paper maps were digitized into electronic format at a computer workstation. Once in electronic form, the individual features of the maps (i.e., areas with the same prevalence rate of dental fluorosis) were assigned unique identifiers (attributes, in this case different colors). Parameters that control the way the map is displayed (projection) were adjusted so that the digital maps, representing the same geographic

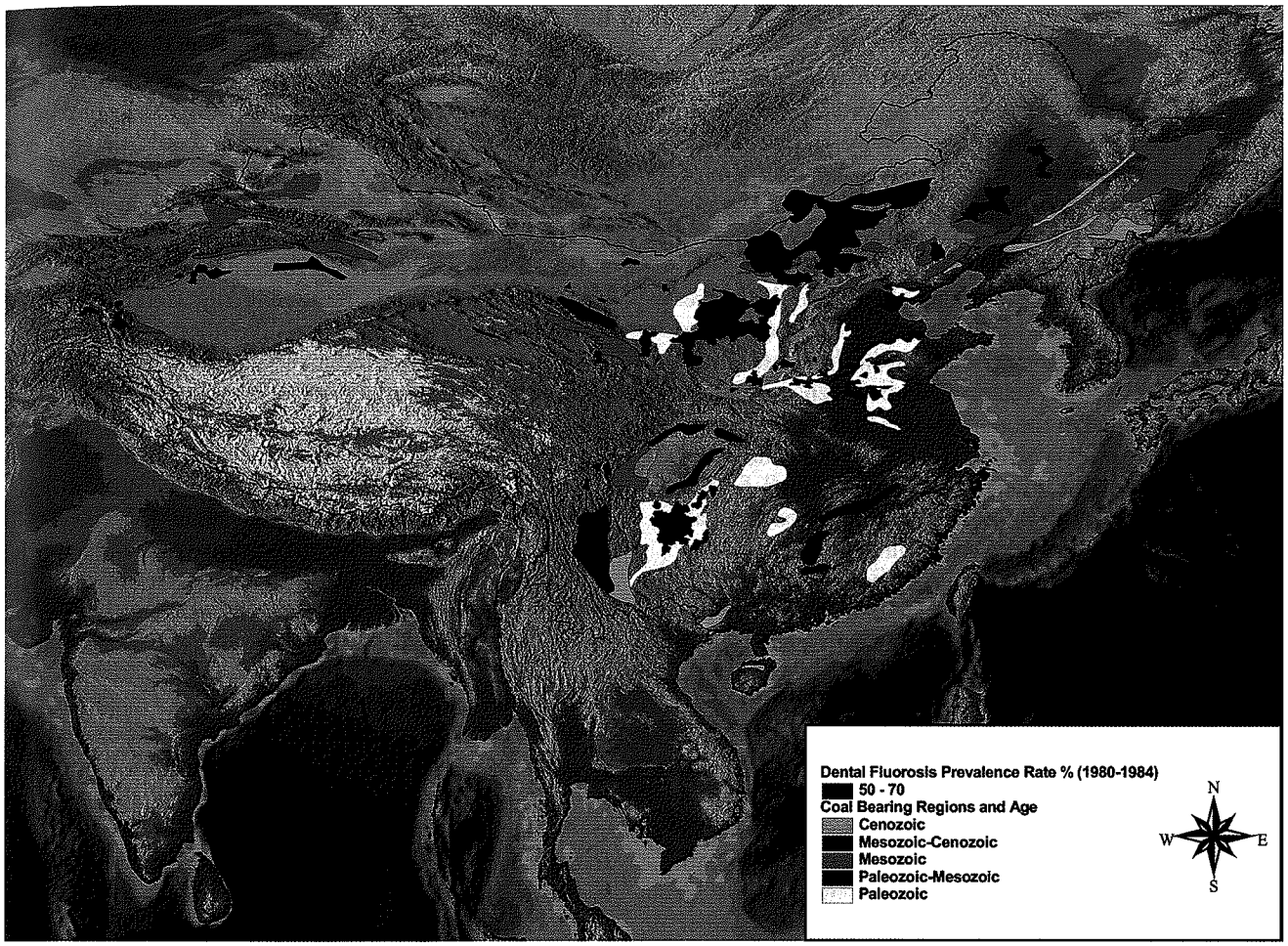

FIGURE 3 Relationship of high prevalence rates of dental fluorosis and coal deposits in the People's Republic of China. (From Karlsen et al., 2001.)

areas, would perfectly overlap each other (Karlsen et al., 2001).

In a GIS environment, the digital map of prevalence rates for dental fluorosis was overlain with the digital map showing coal distribution (Figure 3). The combined maps confirm the association of fluorosis and coal in Guizhou Province, where more than 10 million people are known to suffer from fluorosis (Zheng & Huang, 1989). Figure 3 indicates that the high incidence of fluorosis in north central China (Shaanxi Province, which is the largest coal-producing province in China) may not be related to coal use.

In Guizhou Province the fluorosis is caused primarily by combining moderately high fluorine coals (average of about 200 ppm) with clays having very high fluorine contents (average of about 800 ppm) to form briquettes (Belkin et al., 1999). The high-fluorine clays are the residual products from intensive leaching of the limestone substrate that formed the beautiful karst landforms for which the region is noted. In Shaanxi Province the substrate is primarily loess, a silicate-rich, wind-deposited sediment that is unlikely to have high-fluorine contents. Therefore, unless the coals in Shaanxi Province have exceptionally high fluorine contents, it is unlikely that the high incidence of dental fluorosis in that region is due to residential coal use.

Surprisingly, in the Xinjiang Autonomous Region the incidence of fluorosis also parallels the coal deposits, perhaps this indicates that the distribution of fluorosis may be controlled by sedimentary rocks that favor the growth of the trees from which the fluorine-bearing tea leaves are obtained.

VI. OTHER CASE STUDIES

A number of other examples of GIS applications to medical geology problems may be found. These include the examples below.

African trypanosomiasis—Environmental factors that influence temporal and spatial distributions of African trypanosomiasis (sleeping sickness) were analyzed in a GIS that incorporated temporal Fourier analysis and discriminating analytical techniques such as Mahalanobis distance metric to aid in data interpretation (Rogers, 2000). A relationship between the climate covariate included in the analysis was found to exist with vegetation patterns, which in turn influenced suitability of grazing for cattle, the main reservoir hosts for the tsetse fly (*Glossina* spp.) vector of the trypanosome parasites. Sequential statistical modeling of the tsetse fly populations and trypanosome disease transmission, when linked to biological modeling based on known differences due to the different species of tsetse, helped explain differences in the patterns of disease observed in different regions in Africa.

Hurricane Mitch—In 1998, one of the most powerful and deadly hurricanes in recorded history struck Central America. At least 6500 people died and over 11,000 went missing in Honduras alone as a result of Mitch's fury. Many cases of human disease and death were caused by flooding, even in areas not directly hit by the hurricane itself. Some of these floods, in turn, triggered lethal outbreaks of waterborne infectious diseases such as cholera, leptospirosis, Dengue fever, and malaria. GIS was used to predict high-risk areas for flood potential based on themes including river network configurations, elevation, and slope. This tool may have helped keep the casualty count low following this major disaster. Lessons learned from this experience may be helpful in using GIS to plan and execute preparedness and relief efforts before and after future catastrophic events (PAHO, 2000).

Cadmium in The Netherlands—From 1892 until 1973, a zinc works in the Kempen area of The Netherlands discharged zinc and cadmium (Cd) into the environment in an uncontrolled fashion and seriously contaminated the soil with up to 8 ppm Cd (Stein et al., 1995). Cleanup efforts undertaken in the 1980s made use of a GIS and geostatistics to contour the cadmium distribution and improve sampling efficiency. Data from more than 1700 soil samples were used as point data, and semi-variograms were created to compare stratified and ordinary kriging methods to interpolate the Cd concentrations. Neither mapping technique was uniformly superior; depending on the application (e.g., proximity to urban centers), one or the other map proved more useful. Had the GIS been used interactively, Stein et al. (1995) concluded that the number of soil samples necessary for testing could have been reduced approximately tenfold.

Malaria—A GIS analysis of malaria conducted in the Chiapas region, Mexico, and Peten, Guatemala, provides an example of how this analytical tool can be useful in active and iterative generation of testable research hypotheses. A priori hypotheses pertaining to environmental factors that influence malaria incidence by their impacts on the *Anopheles* spp. mosquito vectors invoked altitude, temperature, rainfall, land use, and vegetation type. In the course of developing the GIS, researchers observed that high-risk areas were often in close proximity to agricultural lands (SHA, 2000). Further analysis led to the generation of a novel hypothesis that relates malaria risk to deforestation, a potential linkage that is presently being investigated by several groups.

VII. CONCLUSIONS

In this chapter, the conceptual framework for GIS databases has been described, as have strategies and tools for conducting medical geology research projects. We have explained how recently developed GIS technology has truly revolutionized the study of human disease systems, which makes possible the simultaneous analysis of numerous interrelated factors that may exert unapparent and synergistic effects. Previously, such complex systems could only be addressed looking at one (or just a few) variable(s) at a time. With the case studies provided, the reader has seen examples of how to approach rigorous investigation causes of human disease patterns that have strongly suspected environmental influences. Finally, we have supplied an extensive, if admittedly selective, set of database resources that should at least provide a starting point for researchers wishing to conduct GIS studies of their own.

SEE ALSO THE FOLLOWING CHAPTERS

Chapter 11 (Arsenic in Groundwater and the Environment) · Chapter 12 (Fluoride in Natural Waters) ·

Chapter 27 (Investigating Vector-Borne and Zoonotic Diseases with Remote Sensing and GIS)

FURTHER READING

Ando, M., Tadano, M., Asanuma, S., Matsushima, S., Wanatabe, T., Kondo, T., Sakuai, S., Ji, R., Liang, C., and Cao, S. (1998). Health Effects of Indoor Fluoride Pollution From Coal Burning in China, *Environ. Health Perspect.*, 106, 239–244.

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

Belkin, H. E., Finkelman, R. B., and Zheng, B. S. (1999). Geochemistry of Fluoride-Rich Coal Related to Endemic Fluorosis in Guizhou Province, China, Pan-Asia Pacific Conference on Fluoride and Arsenic Research, Abstract 45, p. 47.

Ben, K., Hua, L., and Hongchao, H. (2000). The Current State of Epidemic Tea-Induced Fluorosis and Its Control Countermeasures in Urumqi County, Xinjiang. In *Metal Ions in Biology and Medicine*, Vol. 6, (J. A. Centeno, P. Collery, G. Vernet, R. B. Finkelman, H. Gibb, and J.-C. Etienne, Eds.), John Libby Eurotext, Paris, pp. 303–305.

Bunnell, J. E., Price, S. D., Lele, S. R., Das, A., Shields, T. M., and Glass, G. E. (2003). Geographic Information Systems and Spatial Analysis of *Ixodes scapularis* (Acari: Ixodidae) in the Middle Atlantic Region of the U. S. A., *J. Med. Entomol.*, 40, 570–576.

Cameron, D., and Jones, I. G. (1983). John Snow, the Broad Street Pump and Modern Epidemiology, *Int. J. Epidemiol.*, 12, 393–396.

Centers for Disease Control and Prevention (2001). Lyme Disease–United States, 1999, *MMWR*, 50(10), 181–185.

Cromley, E. K., and McLafferty, S. L. (2002). *GIS and Public Health*, The Guilford Press, New York, p. 340.

Das, A., Lele, S. R., Glass, G. E., Shields, T. M., and Patz, J. A. (2002). Modeling a Discrete Spatial Response Using Generalized Linear Mixed Models: Application to Lyme Disease Vectors, *Int. J. Geog. Inform. Sci.*, 16, 151–166.

Glass, G. E., Schwartz, B. S., Morgan, III, J. M., Johnson, D. T., Noy, P. M., and Israel, E. (1995). Environmental Risk Factors for Lyme Disease Identified with Geographic Information Systems, *Am. J. Public Health*, 85, 944–948.

Haining, R. (1998). Spatial Statistics and the Analysis of Health Data. In *GIS and Health*, *GIS Data VI* (A. C. Gatrell and M. Löytönen, Eds.), Taylor & Francis, London, pp. 29–47.

Hock, R. (2001). *The Extreme Searcher's Guide to Web Search Engines*, 2nd edition, CyberAge Books, Information Today, Inc., Medford, NJ, p. 241.

Jensen, P. M., Hansen, H., and Frandsen, F. (2000). Spatial Risk Assessment for Lyme Borreliosis in Denmark, *Scand. J. Infect. Dis.*, 32, 545–550.

Jianan, T. (Ed.) (1989). *The Atlas of Endemic Diseases and Their Environments in the People's Republic of China*, Science Press, Beijing.

Karlsen, A. W., Schultz, A. C., Warwick, P. D., Podwysocki, S. M., and Lovern, V. S. (2001). Coal Geology, Land Use, and Human Health in the People's Republic of China, U. S. Geological Survey Open File Report 01–318 (CD-ROM).

Kennedy, H. (Ed.) (2001). *Dictionary of GIS Terminology*, ESRI Press, Redlands, CA.

Kitron, U., and Kazmierczak, J. J. (1997). Spatial Analysis of the Distribution of Lyme Disease in Wisconsin, *Am. J. Epidemiol.*, 145, 558–566.

Kulldorff, M. (1998). Statistical Methods for Spatial Epidemiology: Tests for Randomness. In *GIS and Health*, *GIS Data VI*, (A. C. Gatrell and M. Löytönen, Eds.), Taylor & Francis, London, pp. 49–62.

Ostfeld, R. S., Hazler, K. R., and Cepeda O. M. (1996). Temporal and spatial dynamics of *Ixodes scapularis* (Acari: Ixodidae) in a rural landscape, *J. Med. Entomol.*, 33, 90–95.

PAHO (Panamerican Health Organization) (2000). Geographic Information Systems in Health, Special Program for Health Analysis, PAHO, Washington DC.

Robinson, T. P. (2000). Spatial Statistics and Geographical Information Systems in Epidemiology and Public Health. In *Remote Sensing and Geographical Information Systems in Epidemiology*, *Advances in Parasitology* 47 (S. I. Hay, S. E. Randolph, and D. J. Rogers, Eds.), Academic Press, San Diego, CA, pp. 81–128.

Rogers, D. J. (2000). Satellites, Space, Time and the African Trypanosomiases. In *Remote Sensing and Geographical Information Systems in Epidemiology*, *Advances in Parasitology* 47 (S. I. Hay, S. E. Randolph, and D. J. Rogers, Eds.), Academic Press, San Diego, CA, pp. 129–171.

Ruiling, L., Tianyu, H., and Jianping, W. (Compilers) (1996). Coalfield Prediction Map of China. Surveying and Mapping Institute of Jilin Province, Publishing House of Surveying and Mapping, 9 map sheets, scale 1:2,500,000.

SHA (Special Program for Health Analysis) (2000). Incidence of Malaria and Land Use in Chiapas, Mexico and Peten, Guatemala, PAHO, Scientific paper No. 104.

Sherman, C., and Price, G. (2001). *The Invisible Web*, CyberAge Books, Information Today, Inc., Medford, NJ.

Stein, A., Staritsky, I., Bouma, J., and van Groenigen, J. W. (1995). Interactive GIS for Environmental Risk Assessment, *Int. J. Geogr. Inf. Syst.*, 9(5), 509–525.

Zeiler, M. (1999). *Modeling Our World*, ESRI Press, Redlands, CA.

Zhang, Y., and Cao, S. R. (1996). Coal Burning Induced Endemic Fluorosis in China, *Fluoride*, 29(4), 207–211.

Zheng, B., and Huang, R. (1989). *Human Fluorosis and Environmental Geochemistry in Southwest China, Developments in Geoscience, Contributions to 28th International Geologic Congress*, Washington DC Science Press, Beijing, China, pp. 171–176.

SUGGESTED READING

Bernhardsen, T. (1999). *Geographic Information Systems: An Introduction*, John Wiley & Sons, New York.

Briggs, D. J., and Elliott P. (1995). The use of geographical information systems in studies on environment and health, *World Health Stat. Q.*, 48, 85–94.

Burrough, P. A., and McDonnell, R. (1998). *Principles of Geographic Information Systems*, Oxford University Press, Oxford, England.

Clarke, K. C. (1998). *Getting Started With Geographic Information Systems*, third edition, Prentice Hall, Upper Saddle River, NJ.

DeMers, M. N. (2000). *Fundamentals of Geographic Information Systems*, second edition, John Wiley & Sons, New York.

Glass, G. E. (2000). Spatial Aspects of Epidemiology: The Interface with Medical Geography, *Epidemiol. Rev.*, 22(1), 136–139.

Green, K. (1992). Spatial Imagery and GIS: Integrated Data for Natural Resource Management, *J. Forestry*, Nov., 32–36.

Lang, L. (2000). *GIS for Health Organizations*, ESRI Press, Redlands, CA, p. 100 plus CD-ROM.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (Eds.) (1999). *Geographical Information Systems*, second edition, John Wiley & Sons, New York, p. 1101 (2 volumes).

Meade, M. S, and Earickson, R. J. (2000). *Medical Geography*, second edition, The Guilford Press, New York.

Melnick, A. L. (2002). *Introduction to Geographic Information Systems in Public Health*, Aspen Publishers, Gaithersburg, MD.

Moore, G. S. (2002). *Living With the Earth: Concepts in Environmental Health Science*, second edition, Lewis Publishers, Boca Raton, FL.

de Savigny, D., and Wijeyaratne, P. (Eds.) (1995). *GIS for Health and Environment*, International Development Research Centre, Ottawa, Canada.

Vine, M. F., Degnan, D., and Hanchette, C. (1997). Geographic Information Systems: Their Use in Environmental Epidemiological Research, *Environ. Health Perspect.*, 105, 598–605.

OTHER SOURCES

A. Sources Of Earth Science/ Geospatial Information

FGDC database clearinghouse sources (all databases listed are accessible through the Internet at http://www.fgdc.gov). Information presented is current at the time of publication (accessed March 2002) and was prepared with sources readily accessible to Internet users in the United States. Additional sources are available internationally in various languages and formats.

- Alaska State Geospatial Data Clearinghouse (ASGDC)
- America Central Clearinghouse Nicaragua
- Australia–WALIS Interrogator–Spatial Data
- Australia–IndexGeo Pty Ltd–Eco Companion Catalogue
- Bureau of Land Management Spatial Data Clearinghouse
- Canada–Ecological Monitoring and Assessment Network Data Set Library (hosted by Environment Canada)
- Canada–National Forest Health Database–Archive of Insects and Diseases Found in Canadian Forests
- Canada–Newfoundland and Labrador Community Base Maps (1:2500 and 1:5000)
- Canada–Newfoundland and Labrador Geodetic Survey Data
- Canada–Purple Pages Business Directory
- Canada–RADARSAT Inventory held by CCRS
- Caribbean Environment Programme
- Columbia Environmental Research Centers Metadata Node
- Connecticut–Geospatial Data Clearinghouse
- Costa Rica Biological Resource Maps (KU)
- Earth Data Analysis Center, UNM–Prototype Hydrology WMS
- El Salvador, CNR Instituto Geografico Nacional
- Forest, aquatic, and rangeland Ecosystems in the Western United States
- Geography Network
- Geological Survey of Alabama Geospatial Data Clearinghouse Node
- Global Change Master Directory
- Illinois Natural Resources Geospatial Data Clearinghouse
- Kansas Ecological Reserves Clearinghouse
- Michigan GIS
- Minnesota Land Management Information Center
- Montana State Library
- NOAA Environmental Satellite, Data, and Information Services (SAT) Node
- NOAA NCDC Library Historical Data Sets (FDL) Node
- NOAA National Climatic Data Center (NCDC) Node

- National Biological Information Infrastructure Metadata Clearinghouse
- Natural Resources Conservation Service
- Nevada Dataworks Spatial Data Warehouse
- New Mexico Resource Geographic Information System
- Peru–Instituto Geografico Nacional
- Republica Dominicana–Nodo Nacional de Datos Geoespaciales
- Space Imaging 5-m Digital Ortho Quads
- Texas Natural Resources Information System
- Transboundary (U.S.–Mexico) Metadata Clearinghouse
- U.S. Geological Survey Advanced Very High Resolution Radiometer
- U.S. Geological Survey CORONA Satellite Photographs
- U.S. Geological Survey DOQ
- U.S. Geological Survey Digital Elevation Model 15-minute
- U.S. Geological Survey Digital Raster Graphics
- U.S. Geological Survey Geoscience data
- U.S. Geological Survey MAPS
- U.S. Geological Survey National Aerial Photography Program
- UK Node–British Geological Survey
- Uruguay–Clearinghouse Nacional de Datos Geograficos (Espanol)

For selected online earth science/geospatial journals see Appendix 2. For selected biomedical/health information see Appendix 2.

B. Libraries (for Further Research)

U.S. Geological Survey Library
950 National Center 12201 Sunrise Valley Drive
Reston, VA 20192
e-mail: library@usgs.gov
U.S. Library of Congress
101 Independece Avenue,
S. E. Washington DC 20540
e-mail: lcweb@loc.gov
U.S. National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bethesda, MD 20894
e-mail: NIHInfo@OD.NIH.GOV

C. Where Does One Start a Search for Relevant Databases?

The Federal Geographic Data Committee (FGDC) Web site http://www.fgdc.gov provides access to over 38 simultaneously searchable data clearinghouses in the United States and internationally. These include databases related to Earth sciences, geography, landform information, ecosystem health, biological resources, and satellite imagery (see Other Sources, part A). The focus of this chapter has been on electronic sources, but don't forget to check your library's reference section, trade journals, or other specialized periodicals and books. To find databases on the Internet, you might use a search engine. Be aware that different search engines work in different ways, and that what may be overlooked by one search engine might be found by another one. There are also metasearch engines that use several search engines simultaneously (Hock, 2001). Of course, another very efficient way to find out what databases are used by experts in a given field is to simply ask them. Contact information for university professors is often listed on their institution's Web site, which can be found on any search engine. Even if the principal investigator is hard to reach in person, his postdoctoral fellows, graduate students, and technicians may be willing to help.

An efficient strategy when starting out on medical geology research projects is to seek out relevant database clearinghouses. Using clearinghouses also offers some protection from rapidly changing unique or uniform resource locators (URLs), which are often referred to as Web site addresses. One example, is the U.S. Geological Survey (USGS), which is a major clearinghouse for Earth science data. The URL for some particular databases contained therein may change, but the URL for a clearinghouse such as USGS generally remains stable over time. The primary clearinghouse organization will maintain proper internal links and keep access to all of their individual databases current.

The geographic and temporal range of the data needed must be ascertained at the outset of any medical geology GIS. It is better to err on the side of obtaining more information then deleting unnecessary elements, because it can be difficult to add data later if it is decided to examine additional parameters. But the initial cost of the data and the cost in resources to store data must also be taken into account. Because different databases will likely contain data archived in a variety of formats, it is advisable to store the initial downloaded data as is and make copies of it before any subsequent manipulation. The text-only ASCII file format is a "common denominator" useful for merging data from different sources into a single data set.

The use of Internet-derived databases can be made frustrating and difficult by two realities of this medium. One reality is that URLs change quickly, and so the Web site address that worked in the past may not take you to the same page today. This potential pitfall can be avoided by using the "gatekeeper" URL to a database clearinghouse as mentioned above, rather than by using direct URLs to individual databases. For example, one is advised to use a main clearinghouse Web site rather than a more specific URL for some individ-

ual database, such as for water table depths in India. The other major problem is the so-called Invisible Web (Sherman & Price, 2001). There are a great many databases accessible via the Internet with no easy way to find them or to find out about them. Many databases can only be accessed after registering and entering a password. Search engines will miss these and other relevant sites, and they will often come up with totally irrelevant sites. Investigators must be mindful, too, of the reliability of database sources accessible via the Internet. *If associated metadata are not available, that database should not be used.*

Once a database of interest has been identified and the legitimacy of the organization that maintains it is verified, you are ready to download. Appendix 2 of this volume lists a number of earth science/geospatial and biomedical/human health databases as examples. Make sure you have the minimum requirements and sufficient memory space on your computer before proceeding. As always when downloading any software or data to a personal computer, remember to have some tool in place for screening computer viruses. It is critical to ensure that once downloaded, the data were not corrupted in the process. You must examine the source data carefully and confirm that they match the data in the form that has been downloaded. Problems can arise, for example, if the source data are tab delimited and your default download is space delimited. As soon as you have downloaded the source data, you should make a backup copy before doing anything with the data. It is generally convenient to keep such files on a compact disc (CD). Now you are ready to open up your data with your spreadsheet software package and import it to your GIS application or to a statistical analysis package. In a GIS environment, you can easily query the data. That is, by clicking with a mouse on a location visibly displayed on a map, you can extract attributes of that point.

You will need to join data from different databases for use in a GIS project. Get to know the raw data well as you must always maintain quality assurance/quality control (QA/QC). It is easy to mix up or somehow corrupt data when manipulating it. For instance, if you sort the data for some reason, make sure you keep a copy of the original unsorted data. Also be careful not to sort only one field, but rather keep your unique identifiers tied to the data in the proper order. If your ultimate aim is to do some statistical analysis of the data, you should work closely with a statistician right from the start. The statistician will help you determine the appropriate data you need to answer the questions you are asking.