NORTH AMERICAN ASSOCIATION OF CENTRAL CANCER REGISTRIES

Executive Director:

Holly L. Howe, PhD
2121 West White Oaks Drive
Suite C
Springfield, Illinois 62704-6495
(217) 698-0800 Ext. 2
(217) 698-0188 Fax
hhowe@naaccr.org

http://www.naaccr.org

The NAACCR Board of Directors voted on November 13, 2002, to accept this report of the GIS Task Force, "Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices." GIS Task Force members compiled state-of-the-art information to assist all cancer registries in their decisions about GIS tools, practices, and current issues. The Handbook is a current reflection of how GIS can be applied to cancer registry operations, practices, and even research using cancer registry data. For all cancer registries that are considering a GIS initiative for their programs, this Handbook will be an invaluable resource.

The report includes seven specific recommendations for NAACCR. These recommendations pertain to continuing NAACCR involvement in the application of GIS techniques to cancer registry activities. With the understanding that awareness of the evolution of GIS technology and tools will be helpful and will provide guidance to cancer registries that choose to incorporate GIS activities into their program, the NAACCR Board of Directors accepted all of the recommendations.

Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data:

A Handbook of Basic Practices

October 2002

Table of Contents

Acknowledgments

Much of the material in this handbook was generated at a North American Association of Central Cancer Registries (NAACCR) GIS Workgroup meeting held in Princeton, New Jersey, on September 26-27, 2001. The following individuals contributed to the development of this document:

Meeting Participants:

Toshi Abe, M.S.W., C.T.R., Chairperson, New Jersey State Cancer Registry
Robert Borchers, Wisconsin Department of Health and Family Services
Francis P. Boscoe, Ph.D., New York State Cancer Registry
Dianne Enright, North Carolina Division of Public Health
Betsy A. Kohler, M.P.H., C.T.R., New Jersey State Cancer Registry
Mary Mroszczyk, C.T.R., [telephone link], Massachusetts Cancer Registry
David O'Brien, Ph.D., [telephone link], Alaska Cancer Registry
Linda Pickle, Ph.D., National Cancer Institute
Thomas B. Richards, M.D., Centers for Disease Control and Prevention
Ric Skinner, GIS Coordinator, Baystate Medical Center
Lyna Wiggins, Ph.D., Facilitator and Report Editor, Rutgers University
James Wilson, North Carolina Division of Public Health
Meeting notes by Edith H. Konopka, Ph.D., State of New Jersey, Office of GIS

Other GIS Workgroup Members:

Tim E. Aldrich, Ph.D., M.P.H., University of South Carolina
Andy Amir-Fazli, New Mexico Tumor Registry
Cheryl Bowcock, Texas Cancer Registry
Scott Horel, Texas Department of Health
Christopher Johnson, M.P.H., Cancer Data Registry of Idaho
Rich Ann Roche, Texas Department of Health
Amy Stoll, Arizona Cancer Registry
Marc-Erick Theriault, M.Sc., CancerCare Ontario
Donna Turner, Ph.D., CancerCare Manitoba
Ray Vezina, Vermont Cancer Registry

Comments and Reviews:

Carolyn (Virginia) Lee, M.D., Agency for Toxic Substances and Disease Registry (reviewer)
Gerard Rushton, Ph.D., University of Iowa (reviewer)
Robert Semenciw, Cancer Bureau, Statistics Canada (comments)

The workshop was supported by Cooperative Agreement Number U75/CCU515998 from the Centers for Disease Control and Prevention. The report is based on the meeting proceedings and its contents are solely the responsibility of the authors and do not necessarily represent the official views of Centers for Disease Control and Prevention. The report was funded in part by the National Cancer Institute, National Institutes of Health, under Contract No. N02-PC-05030.

Executive Summary

The North American Association of Central Cancer Registries (NAACCR) is pleased to present this handbook of basic practices on the use of Geographic Information Systems (GIS) with cancer registry data. The idea for this Handbook arose from NAACCR's first Strategic Planning Meeting held in Monterey, California, on July 26-27, 2000. Participants at that meeting decided that NAACCR should produce a document that outlines the appropriate uses of GIS by central cancer registries. The NAACCR Information and Technology Committee formed the GIS Workgroup to address this objective; this document is one product of the Workgroup's efforts. It is hoped that this handbook answers the basic questions that cancer registry GIS specialists have as they approach the task of performing spatial analyses with their data.

The field of GIS is vast and it is the intent of this Handbook to address the most basic issues. The main sections of this document present an overview of GIS, some of the problems of accurately geocoding the patient address at diagnosis, issues of data confidentiality, an introduction to spatial analysis, and the basics of cartography. The Handbook also provides many references to additional printed resources and Web sites. In addition, the Workgroup presents seven recommendations to NAACCR and several areas that require additional research to effectively implement the ideas presented in this document.

The Workgroup sincerely wishes to thank the NAACCR Board of Directors for their support, and especially Dr. Lyna Wiggins of Rutgers University, Workshop Facilitator and Handbook Editor, who helped to consolidate ideas over several meetings this past year.

Suggested citation: Wiggins L (Ed). Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices. Springfield (IL): North American Association of Central Cancer Registries, October 2002, 68 pp.

Section I: Introduction to GIS

Purpose of the Handbook

Geographic Information Systems (GIS) technology can enrich the collection, analysis, and presentation of health data. Health practitioners and researchers are now more aware of GIS through publications and conferences, and many are starting to use this technology in their daily work. The GIS Workgroup of the North American Association of Central Cancer Registries (NAACCR) concluded that it would be timely to consolidate current knowledge of GIS practice in cancer research into a single handbook.

The handbook begins with a brief introduction of GIS and descriptions of some GIS success stories in cancer studies. The second section discusses the importance of address geocoding for the spatial analysis of cancer data. The third section discusses the important issue of confidentiality of data. The fourth and fifth sections provide introductions to spatial analysis and cartography. The sixth section addresses Internet accessibility issues for the disabled, and the final section presents seven recommendations to NAACCR and areas for further GIS research.

Brief Introduction to GIS

What questions do maps help answer? Why conduct spatial analyses? What do they add to traditional data analyses? The following are some of the questions that GIS can help to answer in a richer way than traditional statistical analyses:

- What are the cancer rates in place X?
- How do these rates compare to rates in other places adjacent or close to place X?
- How do these rates compare to standard rates in the state, region, or nation?
- How do these rates compare to expected rates?
- Are there spatial patterns and trends that should be studied?
- Are there temporal trends that can be identified?
- Are there spatial/temporal interactions that could be important?
- Are the spatial patterns similar across maps of different regions?
- How can spatial analyses help to evaluate equity in access to care?
- How can spatial analyses help to predict future trends in rates?

The National Research Council defines GIS as *"a structural approach to collecting, archiving, analyzing, manipulating, and displaying data having one or more spatial components, using a combination of personnel, equipment, computer software, and organizational procedures."* Notice that this definition gives equal importance to the personnel and organizational issues as it does to the software and hardware. Successful GIS implementation strategies are widely discussed in the literature (Huxhold and Levinsohn, 1995; Obermeyer and Pinto, 1994). Some of the factors of successful implementation of GIS include upper-level management support, the presence of a champion of the technology, the completion of a user needs and requirements analysis by trained GIS professionals, and the early development of a data acquisition and maintenance plan.

GIS originated in university research laboratories where the application focus was in evaluating site selection and suitability using a number of environmental and landscape factors. The technology is approximately 35 years old (Foresman, 1998). This early focus led to a model of layers of spatial information that could be combined and analyzed with the aid of computer software. Computer-generated maps took the place of hand-drawn maps on acetate sheets that were manually overlaid. The concept of layers is still central to GIS (see Figure 1).

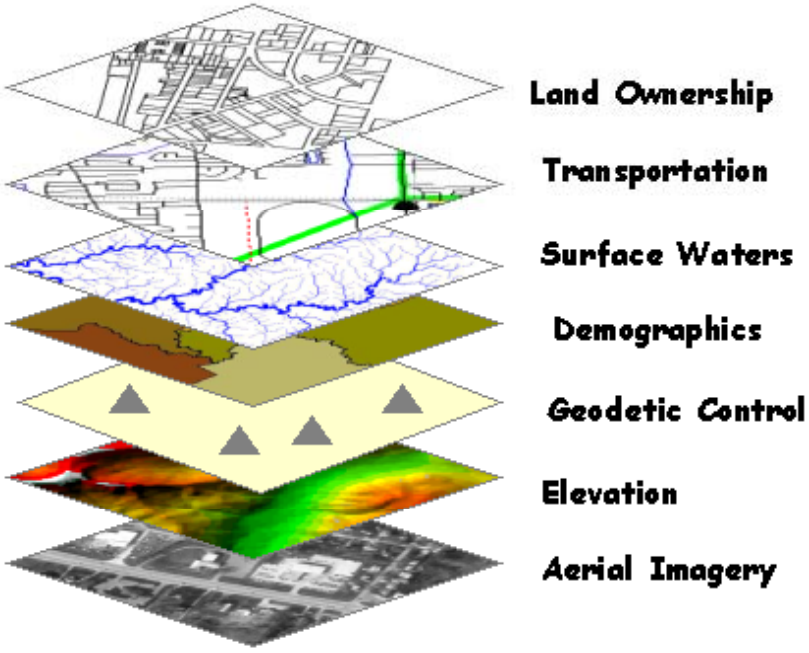

Figure 1. Layers in a GIS.

Layers belong to two major data types: (1) raster data (where the layer is divided into cells of uniform size, forming a matrix of digital information, as in satellite images, aerial photographs, and elevation grid data); and (2) vector data (where the layer is composed of points, lines, or polygons, as in maps of census tract or ZIP code boundaries). An example of points, lines, and polygons and their related attribute database is presented in Figure 2.

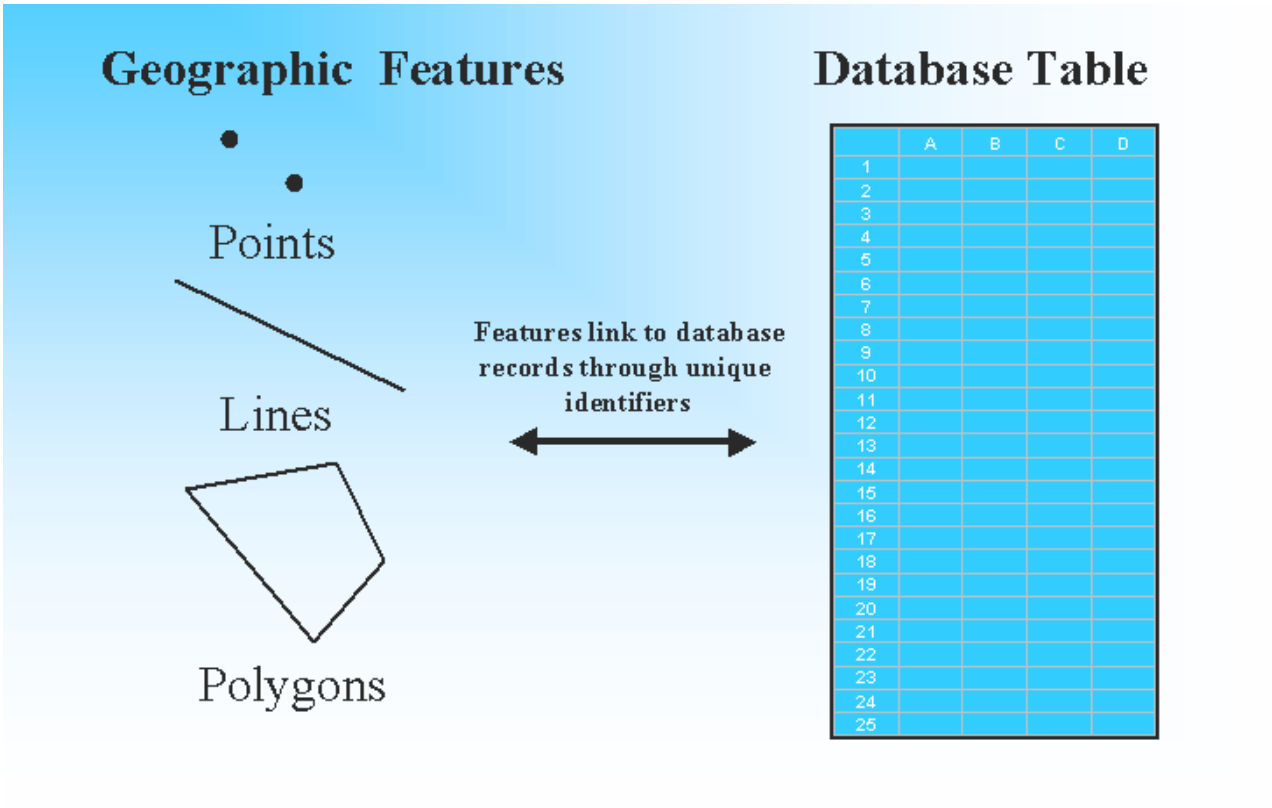

Figure 2. Types of geographic features in vector GIS.

For health applications, typical layers might include:

- **Digital orthophotography:** A base layer of digital photography that has been corrected so that buildings, etc. appear as in their geometrically corrected, true map position. Errors from taking the aerial photography—wing tilt, distortion, etc.—are eliminated through the use of an underlying surface elevation model. This base layer is available and in the public domain for many of the states. For most of the public domain imagery, the nominal scale is 1:12,000, and they are often referred to as Digital Ortho Quarter Quads (DOQQs). Examples of many types of remote images, in addition to DOQQs, can be found on the Terraserver Web Site (http://www.terraserver.com/). State GIS clearinghouses also are good sources of information on locally available imagery.

- **Streets:** A line layer of street centerlines with information on street names and address ranges.

- **Administrative boundaries:** Polygon layers for ZIP code boundaries, census blocks, census tracts, health care areas, minor civil divisions (MCDs), counties, states.

- **Environmental layers:** Point layers for drinking water wells, water or air quality measurement points; polygon layers for wetlands and contaminated sites.

- **Facility layers:** Point layers for the locations of environmentally regulated facilities (e.g., hospitals, health care offices, mammography screening facilities, physician offices).

- **Event locations:** Point layers showing the locations of particular public health events (e.g., residences of patients with cancer, houses with elevated radon levels).

When layers are collected and translated into the same coordinate system and projection, GIS helps integrate the information in the various layers through their common geography. Analytic questions can then be asked that combine the information across the various layers. Another core technical concept of GIS is that the spatial data (e.g., the x,y point location of a hospital) can be joined to attribute data (e.g., the name of the hospital, its street address, patient capacity, whether or not it has an approved cancer program, etc.) corresponding to that spatial feature. The attribute data often reside in a separate data table that is joined, when needed, to its corresponding geographic feature. The spatial and attribute data are joined through a system of unique identifiers. For example, each census tract polygon has a unique census tract number that allows attribute data from the census to be joined to it. This allows investigators to map any of the fields of census data (such as sex, age, race, income, and education) by census tract. The linkage of attribute data to spatial features gives GIS its analytical power.

GIS data can be acquired from both public and commercial sources. For example, layers for streets, environmental factors, and most administrative boundaries typically are available at little or no cost from public agencies. It is rare for cancer researchers to need to create original GIS data layers because most of the core GIS layers are created and maintained by other researchers. Other data, detailed demographic data suitable for market research for example, are available commercially and may be quite expensive. For example, one commercial vendor of demographic data classifies socioeconomic status (SES) based on retail data by census tract (or ZIP code) into 62 clusters that have been given such creative names as *hard scrabble* or *shotguns and pickups* (see Claritas, Inc., Prizm Lifestyle Segmentation, San Diego, CA, http://www.Claritas.com). Credit card and point-of-sale information also are used to construct these data sets. Within public health, the roles of these lifestyle data are controversial. One of the problems with the use of these data are the lack of metadata about the procedures used to generate them (e.g., the data are developed using proprietary methods). Another problem is that the marketing terminology is perceived as politically incorrect (i.e., could not be used in a report to the public).

To find geospatial data efficiently, avoid duplicate geospatial data collection, and share geospatial data more effectively, a network of National Spatial Data Clearinghouses has been established (to access the current Clearinghouses, go to http://www.fgdc.gov/clearinghouse/clearinghouse.html). The search process for geospatial data is aided by the use of a metadata standard. Metadata are data about the data, and are useful to both data users in judging the suitability of certain data for use in a specific application as well as to users searching for desired data. Most GIS producers follow the Geospatial Metadata Standard (see http://www.fgdc.gov/metadata/contstan.html), which allows one to search the Clearinghouses using keyword, place-based, and temporal queries. Some GIS software products now include easy-to-use metadata creation tools to help GIS data producers create and keep their metadata up to date. Along with other GIS data producers, cancer registry GIS users need to create and maintain metadata for their geospatial data sets. The importance of developing and maintaining metadata cannot be overemphasized.

An emphasis on geospatial metadata is important because beginning in Fiscal Year 2003, agencies will have to meet the Federal Geographic Data Committee (FGDC) standards (or fund remediation

to standardize systems). Every Federal agency, as well as non-Federal entities that produce similar data through contracts and collaborative activities, must comply. The requirements cover any geospatial data collected either directly or indirectly (e.g., through grants, partnerships, or contracts with other entities). Office of Management and Budget approval of project funding will be contingent on the compliance of geospatial activities.

Other GIS data are acquired directly in the field. For example, global positioning systems (GPS) are taken into the field to collect the spatial data for the location of public drinking water wells. Depending on the quality of the GPS equipment and the post-processing methods used, the location measured in latitude and longitude (as well as the elevation) can be determined quite precisely, from a few centimeters to several meters in accuracy. Using GPS in the field provides much greater positional accuracy than marking a location on a paper map. Water quality measurements for the wells also are collected through field monitoring, and these non-spatial attribute data are then joined to the point features. This process allows investigators to map and analyze the spatial patterns of water quality. Another example of point data that might be collected in the field through GPS is the location of companies with hazardous chemicals that are potential carcinogens. Analysts might be particularly interested in companies located in or near residential communities. Non-spatial attribute data for this example might include information about the various chemicals and the amounts of these chemicals stored on the site.

Still other GIS data are derived from other spatial or attribute data. Many administrative databases include street addresses of clients, employees, and so on. The address fields in these databases can be matched against the attribute information included in street files. When the address records match, an approximate geographic location (point) can be determined. Because this is an important data source for cancer registries, it will be discussed later in greater detail.

GIS data are increasingly available on the Internet. There are various types of Web sites that include GIS-based information. Some sites provide static maps. Others allow various levels of interaction. One health example is a Web page containing a state map with county boundaries that are linked to tables of data. The user can select a county with the mouse cursor, and a Web page is then displayed with data and/or statistics for that county. See the Kentucky Cancer Registry's Web Site, at http://web.kcr.uky.edu/cir/incidence.html, for a demonstration of this application. Other sites allow users to complete some basic spatial query functions over the Web. Many of the server-side solutions to GIS data on the Web also allow users to zoom, pan, query, and even complete buffer and selection operations.

Although GIS provides a powerful tool for the display and analyses of geospatial data, there are recognized limitations of the technology. There is a growing critical literature on GIS and society (see the research agenda of the University Consortium for Geographic Information Science at http://www.ucgis.org). As with any statistical analysis tool, there also is the risk of inappropriate use of GIS. Some of these issues are addressed in Section IV: Spatial Analysis and Section V: Cartography of this document. In a chronic disease/cancer context, of particular concern are issues of the correct cartographic display of rates.

GIS Success Stories Using Cancer Registry Data

GIS is already widely used in health applications. There are many ways in which the geographic analyses of cancer incidence, mortality, and treatment can inform cancer control activities. A popular perception is that geographic investigations are most often associated with efforts to ascertain environmental carcinogens. There have been some successes of this type, but these comprise but a fraction of the many ways in which GIS has proven to be a useful epidemiological tool, as the following examples demonstrate:

- The first *Atlas of Cancer Mortality for United States Counties,* published in 1975, generated a host of hypotheses that led to substantial findings regarding cancer etiology (Mason et al., 1975). Among these was the observation that elevated rates for respiratory tumors appeared to be clustered in port cities. Subsequent case-control studies established that these elevations stemmed at least in part from asbestos exposure among shipyard workers during World War II.

- During the 1970s and 1980s, cervical cancer mortality rates declined across the country, but the rate of decline was slower in Appalachia, particularly in West Virginia. These patterns escaped scrutiny as long as the information was confined to tabular reports. Once mapped, an unmistakable pattern led West Virginia to approve Medicaid funding of Pap (Papanicolaou) smears. Cervical cancer mortality rates subsequently declined in this region.

- Researchers at the Nova Scotia Cancer Registry used GIS in an investigation of variations in delivery of palliative radiation. Residents were less likely to receive palliative radiation the further they lived from the nearer of the two cancer centers in the province. GIS-derived estimates of driving time proved to be a better predictor than straight-line distance measures.

- The Missouri Department of Health used maps to educate the public about a suspected cluster of brain cancer. Fueled by extensive media coverage, many residents were convinced that brain cancer was uniquely occurring in their community. A series of prevalence and incidence maps were effective in conveying the message that brain cancer, although rare, was nevertheless a ubiquitous phenomenon in all Missouri communities. Media coverage essentially ceased following the release of these maps. An accompanying statistical report concluded that evidence for a statistical elevation in brain cancer in the one community was equivocal.

Section II: Patient Address Data

Where does cancer strike? Is there a geography of cancer that can be mapped in essentially the same way that a geography of urbanization or native plant distributions can be mapped? Does where one lives or works increase the cancer risks?

Analysts at central cancer registries are familiar with the requirement of collecting the address of the cancer patient at the time of diagnosis. Researchers want to relate cancer patient data to other data. Much of the relevant other data also are geographic (e.g., attributes of census tracts and their SES). It is useful to know where a patient lived to identify areas of unusually high or low cancer rates and for integration with other spatial data sets (e.g., toxic release inventory, areas of pesticide application, etc.), as the mechanisms of carcinogenic exposure are explored. The idea of geospatial lifelines is important to public health. This is an area of research interest within Geographic Information Science (GIScience) (Rushton G, Elmes G, McMaster R., 2000).

A Broad Perspective on Patient Address Data

From medical record to cancer map, there are several major control points along the data stream at which accurate, useful patient address data can be preserved, improved, or lost. Consider the following questions:

Who had the address in the cancer abstract? Was it:
- The patient who noted his or her main residence? (preferred)
- The patient who reported a temporary residence?
- The address noted by the payer of health services?

When was the address? Was it the patient's residence at the time of:
- Diagnosis? (preferred)
- Treatment?
- Data entry at the central registry?
- Follow up?

What type of address was it? Was it a:
- Street mail delivery address? (preferred)
- Rural mail delivery route stop?
- Post office box?

NAACCR, the Surveillance, Epidemiology, and End Results (SEER) Program, and the National Program for Cancer Registries currently require cancer registries to assign census tract-level geocodes. The specification of this level of geographic detail reflects what is needed for the purposes of reporting at a national level while still protecting confidentiality. Some of the advantages of mapping cancer cases to a point location (latitude, longitude) rather than to a polygon (ZIP code, census tract, etc.) include:

- Many knowledgeable analysts recommend using the smallest possible spatial unit of analysis. This is most often the street address. Addresses geocoded to latitude and longitude allow GIS analyses within a state for cluster identification and analysis. Aggregation can then be performed as needed to protect confidentiality before reporting to external organizations. The smallest geographic unit of analysis for reporting may vary depending on the state. However, the street address may be needed to assign the case to the correct census tract or ZIP code polygon.

- Software can induce errors if there are defaults for county and or ZIP code, city, and/or state. Researchers often wish to perform time trend analyses, but census tracts and ZIP code boundaries (or enumeration areas and postal codes in Canada) change over time.

- To complete longitudinal studies, a point location needs to be correctly assigned to the correct polygon in any desired time period. For example, the boundaries of census tracts change over time. The commercial firm Geolytics (http://www.censuscd.com/) has taken the census files from the1970-2000 Censuses and created a file with population changes for the 40-year period that can be used for trend analyses. They fitted the area from previous censuses to the 2000 tract boundaries. If a patient's address is geocoded to a specific point, then he or she can be assigned to the correct census tract for any specified time point.

- Point locations allow for a more disaggregate level of analysis. For example, if proximity to a point source of pollution is important in an analysis, the distance can be much more accurately approximated from a set of address-matched points instead of assuming that everyone lives at the centroid point of their census tract or ZIP code. Assigning a patient to a block group, census tract, county, or municipality (or any other polygon feature) is not valid if the patient's location is based solely on a ZIP code centroid. The reason for this is that the ZIP code centroid may geographically be located in a different census tract, etc., than where the patient actually resides as determined by the street address.

- There are many spatial analysis tools that help examine the spatial pattern and concentrations of points. Analysts of crime events have developed many useful geostatistical techniques to map "hot spots" of crime events. Several software packages designed for crime analyses provide easy access to many of these tools (see information about the CrimeStat package in the Appendix).

The Basics of Address Geocoding

Address geocoding relies on at least two data sets: (1) a georeferenced street file with attributes (address ranges, street name, street type, etc.); and (2) a database that contains street address information.

Georeferenced Street Files

Many counties and cities have active GIS programs, and the street file is one of the layers often given high priority for the creation and maintenance of GIS implementation. Street files created for the city and county GIS programs are often accurate in their spatial position and are generally up-to-date and (relatively) complete. Metadata, or data about the data, are becoming increasingly more available from GIS staff able to describe the spatial positional accuracy of their data in some detail.

To be useful for address geocoding, the street files also must have attributes that include the street name, address ranges, and so on.

In the absence of spatially accurate and complete street files from local governments, national data sets are available. The best known of these street files are the Topologically Integrated Geographic Encoding and Referencing (TIGER) files from the U.S. Census Bureau. These street files are used to assign households to the correct census block for the census counts. Although originally designed for this purpose, these street files rapidly became valuable for other GIS analyses. For a historic perspective on TIGER files, their development, and their limitations, see Marx (1986), Broome and Meixler (1990), and Marx (1990). The records in a TIGER file correspond to the line between two intersections in a street network (like the block of Main Street that falls between Church Street and School Street). A street address that is matched to a specific city block and an address range (including left or right side of street) can then be matched to a census block and tract.

The most recent TIGER files are the Redistricting Census 2000 TIGER/Line files, which consist of line segments that represent physical features and legal and statistical boundaries. The geographic coverage is the county or statistical equivalent. The TIGER files do not include the demographic statistics from the Census 2000. Seventeen separate record types are used, including the basic data record, the shape coordinate points (feature shape records), and geographic entity codes. The 17 record types contain data that must be linked in specific ways to represent specific spatial objects (e.g., a street). The records are in ASCII text format, and to create maps with the files, one typically uses a GIS package. Detailed technical documentation on the TIGER files can be found at http://www.census.gov/geo/www/tiger/rd_2ktiger/tgrrd2k.pdf. After processing through a GIS package (in this case Environmental Systems Research Institute's [ESRI] ArcView 3.2), a typical record for a block of one street appears in Figure 3. This record is for a block of S Brook Drive in Middlesex County, New Jersey. The ZIP code to the left and right is 08850. The address range on the left (given the from and to direction of the chain) is from 40 to 62, and the address range on the right is from 49 to 73.

Shape	PolyLine
Tlid	60268875
Fnode	8128
Tnode	7981
Length	0.16183
Fedirp	S
Fename	Brook
Fetype	Dr
Fedirs	
Cfcc	A41
Fraddl	40
Toaddl	62
Fraddr	49
Toaddr	73
Zipl	08850
Zipr	08850
Census1	0
Census2	0
Cfcc1	A
Cfcc2	A4
Source	A

Figure 3. Example of a GIS-processed record for TIGER 2000.

Figure 4 is a graphic of a few blocks of a city as represented in a TIGER file, with their related attribute data. Note that the streets do not meet at right angles, as they do in real life. This is a function of the scale of the TIGER base (1:100,000) and the digitizing process used to create the files.

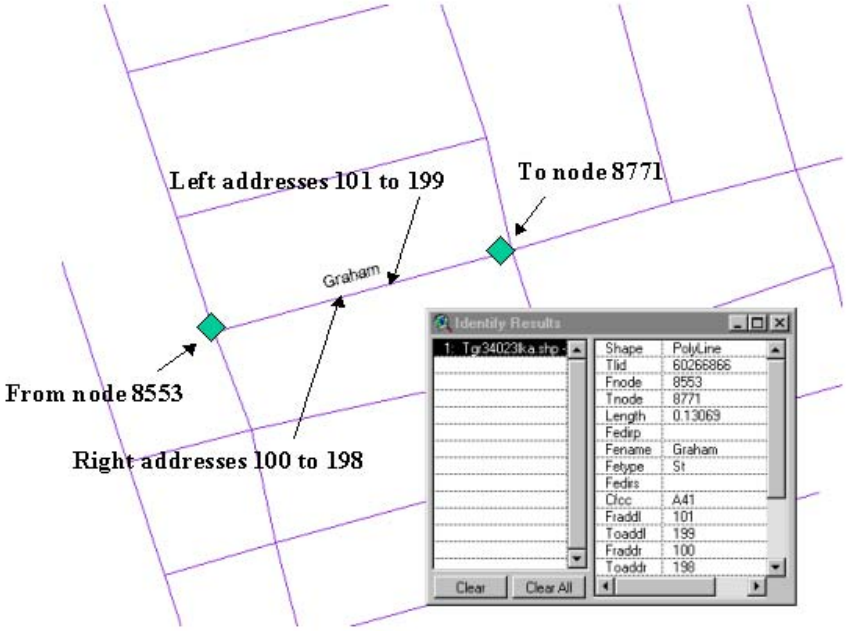

Figure 4. TIGER example.

Because of the scale of the TIGER files, the street centerlines often do not run down the center of streets in digital orthophotography. Figure 5 shows the street files against a DOQQ background. The DOQQs are at a scale of 1:12,000.

Figure 5. Street centerline files against a DOQQ.

Attribute Databases

The second required data set is an attribute database that contains the street address of the case in one or more fields. An example of a portion of a database with the street address in one field is shown in Figure 6. Notice that some of the addresses will be difficult to match against the street files because of inconsistent format and spelling.

Sic	Name	Street	City	State	Zip
0752	T BLUMIG KENNELS INC	645 OLD STAGE RD	EAST BRUNSWICK	NJ	08816
0752	CALLING ALL PAWS -CORP-	1400 MAIN ST	SAYREVILLE	NJ	08872
0782	CURRY LANDSCAPING INC	64 WILCOX RD	NEW BRUNSWICK	NJ	08901
0782	DOUGLAS BALLENTINE	ONE STARODUB DR	MILLTOWN	NJ	08850
0782	LAWN MATE INC	447 SMITH ST	PERTH AMBOY	NJ	08861
0782	TAKE FIVE LANDSCAPING & LA	76 MYRTLE AVE	METUCHEN	NJ	08840
0782	S L BENCZE & ASSOCIATES INC	51 CRANBURY NECK RD	CRANBURY	NJ	08512
0782	STANIS LAWN SERVICE & LAN	455 SPOTSWOOD GRAVEL HILL	JAMESBURG	NJ	08831
0782	FOUR SEASONS LANDSCAPIN	4 AVENUE D	JAMESBURG	NJ	08831
0782	A & J MOWER SERVICE INC	27 MONMOUTH RD	SPOTSWOOD	NJ	08884
0782	REASONABLE RATE LANDSCA	27 PARKER ST	EAST BRUNSWICK	NJ	08816
0782	NOWICKI LANDSCAPING INC	14 KRUMB ST	SAYREVILLE	NJ	08872
0782	THE TRUGREEN CO L P	117 CORPORATE BOULEVARD	SOUTH PLAINFIELD	NJ	07080
0782	ERIC LEWCZAK	178 10 EYCK ST	SOUTH PLAINFIELD	NJ	07080
0782	JOSEPH KREMER	263 FRIENDSHIP RD	CRANBURY	NJ	08512
0782	RICHARD WILLIAMS	9 YORKTOWN CT	SPOTSWOOD	NJ	08884
0782	DREAMSCAPE LANDSCAPING	78 BUTTONWOOD DR	EAST BRUNSWICK	NJ	08816
0782	RONALD DOMANSKI	28 MILLER RD	CRANBURY	NJ	08512
0782	OLIVERS LAWNCARE & BACKH	87 OLD TRENTON RD	CRANBURY	NJ	08512
0782	J Q LANDSCAPING INC	7 HASTINGS PL	CARTERET	NJ	07008
0782	ADDARIO LANDSCAPING L L C	154 MARTIN DR	SOUTH PLAINFIELD	NJ	07080
0782	GRO RITE LANDSCAPING INC	26 TICE AVE	SOUTH RIVER	NJ	08882

Figure 6. Database with single street address field.

An example of a portion of a database with the street address in more than one field is shown in Figure 7.

The current NAACCR data exchange record layout includes the following address fields (NAACCR data item numbers appear in brackets):

- [2330] Street address at diagnosis
- [2335] Street address at diagnosis--supplemental
- [70] City of residence at diagnosis
- [80] State of residence at diagnosis
- [90] County of residence at diagnosis
- [100] ZIP Postal Code at diagnosis
- [110] Census Tract
- [362] Census Tract Block Group
- [364] Census Tract Certainty
- [120] Census Tract Coding System

Sic	Name	N3	Prefix	N5	Type	Suffix	Unit	Street	City	State	Zip
0752	T B	645		OLD STATE	RD			645 OLD STAGE RD	EAST BRUNSWICK	NJ	08816
0752	CAL	1400		MAIN	ST			1400 MAIN ST	SAYREVILLE	NJ	08872
0782	CUR	64		WILCOX	RD			64 WILCOX RD	NEW BRUNSWICK	NJ	08901
0782	DOU	1		STARODUB	DR			ONE STARODUB DR	MILLTOWN	NJ	08850
0782	LAW	447		SMITH	ST			447 SMITH ST	PERTH AMBOY	NJ	08861
0782	TAK	76		MYRTLE	AVE			76 MYRTLE AVE	METUCHEN	NJ	08840
0782	S L	51		CRANBURY NECK	RD			51 CRANBURY NECK RD	CRANBURY	NJ	08512
0782	STA	455		SPOTSWOOD GRAVEL				455 SPOTSWOOD GRAVEL H	JAMESBURG	NJ	08831
0782	FOU	4		AVENUE D				4 AVENUE D	JAMESBURG	NJ	08831
0782	A &	27		MONMOUTH	RD			27 MONMOUTH RD	SPOTSWOOD	NJ	08884
0782	REA	27		PARKER	ST			27 PARKER ST	EAST BRUNSWICK	NJ	08816
0782	NOW	14		KRUMB	ST			14 KRUMB ST	SAYREVILLE	NJ	08872
0782	THE	117		CORPORATE	BLVD			117 CORPORATE BOULEVAF	SOUTH PLAINFIELD	NJ	07080
0782	ERI	178		EYCK	ST		10	178 10 EYCK ST	SOUTH PLAINFIELD	NJ	07080
0782	JOS	263		FRIENDSHIP	RD			263 FRIENDSHIP RD	CRANBURY	NJ	08512
0782	RIC	8		YORKTOWN	CT			9 YORKTOWN CT	SPOTSWOOD	NJ	08884
0782	DRE	78		BUTTONWOOD	DR			78 BUTTONWOOD DR	EAST BRUNSWICK	NJ	08816
0782	RON	28		MILLER	RD			28 MILLER RD	CRANBURY	NJ	08512
0782	OLI	87		OLD TRENTON	RD			87 OLD TRENTON RD	CRANBURY	NJ	08512
0782	J Q	7		HASTINGS	PL			7 HASTINGS PL	CARTERET	NJ	07008
0782	ADD	154		MARTIN	DR			154 MARTIN DR	SOUTH PLAINFIELD	NJ	07080
0782	GRO	26		TICE	AVE			26 TICE AVE	SOUTH RIVER	NJ	08882

Figure 7. Database with multiple street address fields.

Other database formats are sometimes encountered. One format of some concern is the case where some records have the street address in Field 1, while other records have the street address in Field 2. One solution (for software packages that allow this option) is to process the records in several passes. The first pass standardizes the address information in Field 1, and places the standardized information into Field X. The second pass standardizes the address information in Field 2, and places the standardized information into Field X.

Address Geocoding Process

If an investigator has the street files, an attribute database containing address fields, and GIS software or other software with address-matching functionality, address geocoding can be performed. If the address field for a record in the attribute database matches the attributes of a line in the street file, the software assigns that record to the matching line in the street file. To exactly match, each component of the address (prefix, number, name, type, suffix) would need to match. Because the street files include the address ranges of each side of a street, the point also can be assigned to one side of the street. To determine an approximate location for the point along the line (between the two intersection end points), the software approximates the location by interpolating between the address ranges on that side of the street. This obviously introduces additional positional error, because it assumes that the street numbers are uniformly distributed between the endpoints.

There are many identical addresses within a state or county (or city). For example, there are many "*Main St*" addresses across a state. Other polygons such as ZIP codes or MCDs are therefore needed as limiting boundaries. This additional geography helps narrow the street address down to a smaller geographic area where the address is likely to be unique.

The software alternatives for address geocoding vary in their ability to add flexibility in the matching process. For example, the software may provide spelling flexibility tools to allow for

minor misspellings. Other tools may allow flexibility to soften the requirement that all of the different components of the address need to match perfectly. This is a case where careful consumer research on the various products can be of benefit. Some of the software alternatives are listed in the Appendix.

If an investigator does not have GIS software with address-matching functionality, or if he/she does not wish to complete the address geocoding himself/herself, there are commercial firms that provide this service. Contact information for a selection of these firms is provided in the Appendix. Both the pricing and quality of service is known to vary widely, so this is a case of "buyer beware." Some studies of the performance of these services are included in the GIS literature (Johnson 1998). Because different results are obtained from the use of different software or geocoding services, it is important to include this information as part of the metadata of the geospatial data set (this last point is discussed in more detail in later sections).

Nancy Krieger (Krieger et al., 2001) reports the results of a study in Massachusetts that compared address processing by four geocoding vendors. The same test data were provided to all of the vendors to check on differences in algorithms. There were wide variations in match rates between the four vendors. The study also evaluated the customer service provided by the vendors.

Online geocoding services can track submission sources. This could potentially breach patient confidentiality if addresses are traced back to a cancer registry. So far, use of online services has not been forbidden. In the case of cancer cluster data, patients usually do not mind, but it is an issue for other kinds of studies. The security of such sites should be investigated before they are used. In a recent application to the Centers for Disease Control and Prevention's (CDC) Institutional Review Board (IRB), a cancer registry proposed using a commercial geocoding vendor. No details were provided as to whether this was an online service or whether there was a confidentiality agreement with the commercial geocoding service. CDC's IRB reviewer raised questions about whether use of the commercial geocoding was a violation of confidentiality/privacy. The cancer registry subsequently revised its study protocol so that all geocoding was performed by in-house staff (i.e., the commercial geocoder would not be used). Different reviewers may have different opinions regarding this issue.

NAACCR's GIS Workgroup advises that cancer registries exercise caution in the use of free, online commercial Web-based geocoding services, because the free services may not involve any confidentiality/privacy agreement. The commercial Web-based geocoding services also may be able to link the identity of the cancer registry (via cookies) with the list of anonymous addresses and derive the locations of a set of cancer patients.

The Workgroup also advises that vendors be evaluated closely, and that cancer registries compare the experiences of others before making a vendor decision. The quality, timeliness, and accuracy of data may vary for a given vendor across various geographic areas. If acquiring software for in-house use, cancer registries should complete a test of any package under consideration. Studies have shown that vendor-quoted rates of matching and quality may not be accurate (Krieger et al., 2001). It may be advisable for NAACCR to promote the development of model language for cancer registry contracts with online geocoding services so that there is a greater level of protection to the cancer registry. The sense of some of the members of the Workgroup is that it would be best for all

geocoding to be performed in-house. The cancer registry could still contract out the work, but the contractor would need to conduct the geocoding within the cancer registry (so the cancer case address information does not leave the cancer registry).

Address Geocoding in Canada

Canada also publishes standards regarding street addresses. They can be found on the Internet at http://www.canadapost.ca/tools/pg/manual/b03-e.asp#top. In Canada, the address fields include a Canadian postal code. Canadian cancer analysts have used the postal code to map cases to geographic areas. The TIGER files seem to correspond closely to the Postal Code Conversion File (PCCF) in Canada, although it is more precise (exact street address as opposed to postal code only). Once a postal code has been assigned to a census enumeration area, the latitudes and longitudes of each centroid are used to geocode the cases.

The PCCF has several problems. The first problem is that postal codes may cross enumeration area boundaries. When this happens, the analyst has the following options:

- Assign the postal code using the single link indicator, which assigns the postal code to the enumeration area with the largest percentage of population in the postal code. This is available for the 1986, 1991, and 1996 Censuses.

- Assign the postal code using the Health PCCF+ software from Russell Wilkins at Statistics Canada (Wilkins, 2001). This software randomly assigns cases to all enumeration areas within the postal code using the proportions from the 1991 Census population.

The second problem is that this procedure is quite precise in urban areas, but not necessarily precise in rural (especially northern) areas. SES data are not available for all Canadian enumeration areas.

The third problem is temporal. Using postal codes limits how far back in time an analyst can investigate. This affects longitudinal studies.

Geocoding Process: Problems With Street Files

One hundred percent of the address records never exactly match a line in the street files within a specific limiting boundary. Matching problems can occur due to errors in either of the two data sources—the reference street files or the address field(s)—in the attribute database. First, consider the street files. Collecting and maintaining street files and their attribute data for an entire country is a daunting task. Some of the likely errors in the national-level files include:

- Missing streets or portions of streets (due to errors or new growth).

- Inclusion of streets that are not there (due to errors or elimination of streets).

- Missing data for attributes (some combination of street name, type, address ranges, etc. are missing).

- Errors in attributes (e.g., names are incorrectly spelled, address ranges are incorrect or incomplete).

- Rural route addresses are typically not included in the street files.

- Changes in ZIP code boundaries.

- Different streets within a single municipality or ZIP code boundary with the same name. Although this name duplication should be corrected through various 911 programs, this has not yet happened in many cases. For example, Boston contains several different Washington Streets because the city absorbed several preexisting adjacent cities and towns that each contained its own Washington Street. The analyst must rely on their ZIP codes or localities (neighborhoods) to tell them apart. This is very commonplace wherever such annexing occurred.

With the growth of on-board car navigation systems, positional accuracy is more of an issue in everyday life. The TIGER files were originally digitized from 1:100,000 scale maps from the U.S. Geological Survey. Under National Map Accuracy Standards, the street linework for this scale is plus or minus 167 feet. Because the point location also is interpolated between the address ranges, there is potential for another few hundred feet of error. The technical documentation for the TIGER files covers these issues in some detail (http://www.census.gov/geo/www/tiger/rd_2ktiger/tgrrd2k.pdf).

Because of many reported errors in the TIGER street files, commercial vendors have entered the street file market. Information on some of these commercial vendors is provided in a list of resources in the Appendix. Prices for these commercial files vary widely. Some of the commercial vendors simply add value to the TIGER files by minimally correcting the attribute data. Others conduct extensive work in the field to upgrade and maintain their files. Some of the commercial vendors of street files have positionally corrected their files. Again, users of these files should carefully evaluate their need for positional accuracy before making purchase decisions. Because commercial street files give different results when address geocoding, it is important that the metadata include information on the source of the street files. Most street files from commercial vendors have copyright restrictions (one user, one computer). Copyright restrictions may make the use of commercial data sets difficult for public health agencies where the desired use is to share data sets at no added cost.

Longitudinal studies present additional GIS analyses issues. The base street files used to geocode also can change over time. The base street files may be improved in quality and accuracy, or entirely new sources of street files may be obtained. The assignment of latitude and longitude coordinates depends on the accuracy and validity of the base street file. A potential methodological issue is whether all cases should be geocoded at the time of a study using the most recent base street file or whether latitude and longitude coordinates should be assigned to cases on an incremental basis (using the most recently available street file for any given year). A related question is which base street file should be used for studies involving several different states (i.e., should there be a uniform street file recommended for use by all states in geocoding)?

Geocoding Process: Problems With the Attribute Database

To complete the address geocoding process, it is ideal to have clean address fields that exactly match the conventions used by the geocoding software. To create clean address fields in standard formats, it is best to have the data originally collected with normalized and quality-controlled address fields. Some of the general difficulties with the address fields in many administrative databases include:

- Post office box (billing or mailing addresses) is given rather than a street address.
- Rural route address is given (or other types of postal delivery routes, like Star Routes and Highway Contract Routes).
- A building or facility name is given instead of an address (e.g., just the nursing home or apartment tower name is given).
- The address is missing one or more components (building number is missing, street type or direction is missing or incorrect).
- Error(s) in the attribute field (e.g., spelling errors).
- Apartment numbers or street addresses included in the building field (unit designations belong in a separate field).

GIS analysts encounter many variations in the formats of addresses while geocoding from address fields included in administrative databases. Automated address matching works best when the attribute address is in the same format as the georeferenced addresses against which it is to be matched. The basic standardized format is: {house or building number}_{prefix}_{street name}_{street type}_{suffix}. Not all valid addresses can be expressed in this simple format, and different people may provide or record the same address in different ways. Some variations on the standardized address format are shown below. Experienced professionals can interpret such unorthodox formats easily, but they are difficult to geocode in an automated mode. Consider the following examples:

Foot of Elm Street
Twelve Brown St
15 Second Ave vs. 15 2nd Avenue
50R Elm St vs. 50 Rear Elm St. vs. Rear 50 Elm Street
23 Green St, Apt A vs. 23A Green St. vs. Apt A, 23 Green Street
Old Age Village Bldg D #16
34 Highway 96 vs. 34 Rte 96 vs. 34 Old Post Rd (local name)
Livingston and New Streets
Ocean View Apartments
10 E. West St vs. 10E West St
25 E Street
75-100 Water St
401 ½ School St

Off Old State Road
88 Maple Blvd West vs. 88 W. Maple Blvd

Some of the address format variations can be modified into standardized addresses by specialized software. Sometimes this standardization results in the correct geographic location, and at other times it results in placement errors. Forcing an address that has an unorthodox format into a more standardized format can produce an address that is easier to geocode. Standardization can lose valuable and distinguishing address details, however, and it may become impossible to differentiate separate buildings. The following addresses could all be standardized to simply "10 Main St":

10 Main St Bldg D #25
10 Main St Ext
10 Rear Main St
10 ½ Main St
10 Main St Lot 4

Other address geocoding difficulties can be alleviated by allowing for alias street names (e.g., Big Boulevard was recently renamed President Clinton Boulevard, but no one living there uses the new name yet, or Washington Boulevard also is frequently referred to by residents as Route 22). Other format variations can be corrected by manual or semi-automated data cleanup techniques. Some address variations are not possible to code, even with manual methods (e.g., Foot of Elm Street).

Geocoding Process: Problems With Geocoding Software

There also are some recognized problems with geocoding software. The software uses algorithms and/or assumptions when it encounters a problematic address. Not all commercial vendors design geocoding software with the same algorithms or assumptions; therefore, different results are obtained when different vendor products are used. Also, because the geocoding software is proprietary (trade secret), a commercial vendor may make changes in the geocoding algorithms without providing any notice to the cancer registry.

Geocoding Process: Dealing With Unmatched Addresses

GIS analysts use a general rule that 70-80 percent of attribute address records should match using stock TIGER files, following some cleanup of the address fields in the attribute database. What do analysts do with the records that still do not match after data cleanup efforts? For some very important analyses, manual placement of points for records may be made through field checks or use of secondary paper maps. This is obviously time consuming. The more common strategy is to assign unmatched records to the centroid of a ZIP code or census tract boundary (its center point). For records that cannot be geocoded to the polygon (e.g., ZIP code is missing or incorrect), many analysts then assign that record to the centroid of a larger geographic unit like an MCD. It should be noted that patients geocoded to something other than a street address will most likely not be assigned to the correct census tract. There are some obvious related data quality issues here, and a recommendation for recording metadata about match quality is provided in this section.

One related issue is that with the increase in availability of urban GIS data at the parcel level, it also is possible to match an address attribute database against parcel address attributes. Matching to a parcel centroid is generally more spatially accurate than matching to a street file.

Geocoding Process: Recommendations for Improving the Process

Some of the common problems associated with the process of address geocoding have been described in the previous section. The following paragraphs detail some additional recommendations on address geocoding from the members of the NAACCR GIS Workgroup.

To the fullest extent possible, central registries should assure that reporters of cancer information in the field are supplied with up-to-date tabular information and maps to facilitate the recording and checking of correct, valid patient addresses at the time of original case abstraction. At the central registry, three general steps should be followed to ensure the maximum value of case level data for geographical analyses: (1) address checking, (2) address correction, and (3) address geocoding.

Is the address received with the case record plausible? Could it be correct? If not, changes will be required before the data are mapped. Raw address data should be checked against current reference lists both by individual address element fields (e.g., locality and ZIP code separately) and in combinations (e.g., locality in conjunction with ZIP code).

Individually, major address element fields (e.g., state, locality, county, ZIP code) should be compared against current inventory lists. If any do not match, this should be recorded. Locality presents practical ambiguities. Although a correct United States Postal Standard (USPS) residential mailing address is required, a political jurisdiction (e.g., official name of an incorporated village or town that identifies a unit of local government) may have been recorded on the abstract. In some cases, the USPS-approved name of a locality associated with a ZIP code will be identical to the name of a unit of local government; however, this will not be the case in many instances. This problem may be further exacerbated by the extent to which mail is routinely delivered, without complaint from the USPS, despite incorrect locality designation on a mailing address. Unfortunately, automated processing of addresses can be less forgiving than an address-error tolerant mail carrier.

The initial checking and verification of an address is an important step. A Yes/No flag field can indicate if the address has been reviewed and verified, not just in terms of checking that it is some valid address, but that it was the patient's usual residence at the time of diagnosis. It is very difficult to correct a list of problem addresses without having the patient identifiers and temporal fields attached. If all an analyst has are the address fields, all he or she can do is make educated assumptions about what the correct addresses should be. If an analyst researches the patient instead of the address, the analyst will not have to rely on guessing. Not all states have the resources to do this, nor a caseload small enough to give individual records attention, but this is a sound process.

The initial checking of address validity, as received at the central cancer registry, should record (perhaps with the conventional binary "1" for yes and "0" for no) how each of the address elements compared with the master list. Likewise, the address element pairs county:locality, locality:ZIP code, and ZIP code:county may be fairly easily compared against valid combinations in independent lists. If, for example, locality and county are a valid pair but locality and ZIP code are not, the ZIP code

may be either incorrect or out of date (from time to time ZIP code delivery areas emerge, change, merge, or vanish).

Automated address geocoding needs all the clean and standardized data that can be obtained. Validity checks may provide a substantial weight of circumstantial evidence supporting changes (corrections) in individual elements. For example, pursuant to matching an incoming patient address against standard street segments and address ranges, a corrected ZIP code might be added to the original patient record by virtue of the valid combination of all the other address elements (state, locality, county, street name, and number) or information about the closing of a post office that resulted in a ZIP code change. When such alterations are made, two actions should be taken:

- The original data should be retained in the record (to preserve the option of re-correcting the data) and each correction should be documented to the extent of recording which fields were changed to produce an address valid against all references.

- The source of each replacement data file (e.g., November 25, 2001 ZIP code table from the USPS) should be specified.

When automated geocodes are produced from matched addresses, the present state-of-the-art of automated geocoding makes it advisable to record both the tools used and the geographical reference data used with the tools. When batches of addresses are processed, the tools record would describe the use of a geocoding extension included with a specific GIS or a stand-alone geocoding package. Software version numbers also should be recorded.

The reference data (e.g., the street files) are generally independent from the tools. These might include Census TIGER 2000 files, national files from commercial vendors, or digital street files produced by a state or provincial highway or transportation authority. The source, versions, and dates of these reference data sets also should be retained in a metadata record associated with the specific batches of processed addresses. Batches of geocoded addresses from contractors or other external (out-sourced) providers also should have appropriate documentation so far as the protection of contractual obligations, disclaimers, proprietary technologies, or related limitations allow.

If address geocoding is used to assign a record to a census tract or ZIP code polygon, it is important to include the date of the boundary definition used as part of the metadata for the data set. It may be desirable also to include a processing date, or batch processing ID code containing the date, to help trace the lineage of this derived data. Again, full documentation in the metadata is necessary. Another temporal issue includes changes in the way a given address is expressed. The same point on the earth may be described by different addresses at different points in time (e.g., an area's ZIP code/post office name may change or the street may be renamed or renumbered). When an address is changed in such ways, it is not automatically updated in hospital medical records. A renumbered street is especially difficult, because it could result in "*10 Main St*" in two different records representing two different places.

Because of the dynamism of administrative boundaries, the definition date for the administrative boundaries used should be included in the metadata for the data set. This dynamism underscores the usefulness of obtaining point data based on addresses.

Good data quality is an appropriate goal for cancer data analyses. To solve some of the quality problems with the address field(s) in attribute databases, an address standard is needed. The GIS Workgroup examined the FGDC Address Standard, now under review, and felt that this standard met most of their needs. The FGDC Address Standard incorporates appropriate components of the USPS standards. The text of the full standard is available on the Web (http://www.fgdc.gov/standards/status/sub2_4.html). Several of the important tables from the standard are included in the Appendix. One crucial note is the recommendation in the Address Standard that data quality is easier to ensure if addresses are parsed by their component parts into separate data fields. For example, a "pick list" can be provided to ensure that the street type field (ST, RD, etc.) adheres exactly with USPS standards. Data entry staff are left with no discretion on spelling. Multiple fields are easy to recombine into a single field through concatenation when desired.

The Workgroup believes that some new fields to the FGDC Address Standard are needed. Obviously, there is a cost to each additional field of data collected, but the group felt that the added cost would be outweighed by the extra quality of the information gained. These additional fields (with a rationale for each addition) are:

- **Add a field for institutional names.** Many cases are reported from health care facilities, veterans homes, correctional facilities, and other institutions. Including this as a specific additional field should clarify the locational information. One Workgroup member (Robert Borchers) reported that he once had 75 percent of the cases in one ZIP code reported as the veterans home with the name of the veterans home spelled dozens of different ways.

- **Add a field for unit/multi-building complexes.** Such details can be important. Different units within a building may experience different environmental exposures (radon in a basement apartment, outdoor air pollution in a penthouse). In very large buildings, two units may be far apart geographically. Complexes of separate buildings that have been given a single street address may cover very large geographic areas (e.g., mobile home parks, dormitory complexes, retirement villages). The street address assigned to such complexes may even be far removed from the buildings themselves (e.g., a mobile home park's address is 100 Main Rd, but the residences themselves are set back a long distance from the road).

- **Add a field for the quality of the geocode.** Because the quality of the match varies by record, this field would capture detailed information). The Workgroup senses that a hierarchical measure of quality would work well. Codes should be assigned to represent each class in the following hierarchical classification:
  - GPS with metadata included on accuracy of instrument
  - Parcel centroid
  - Address match, with quality measure as generated by software
  - Street intersection
  - Street name only, within a ZIP code or tract or other polygon
  - Mailing address ZIP code +4
  - Mailing address ZIP code +2
  - Mailing address ZIP code only
  - ZIP code of post office or Rural Route

- Centroid of postal address city (when residence ZIP code is unknown and there are multiple ZIP codes for the city)
- Centroid of MCD or jurisdictional city
- Assigned from paper map (of specified scale and accuracy)
- Assigned from digital map source (of specified scale and accuracy)
- Assigned to hospital or other facility
- Unable to assign coordinate.

The Workgroup also thinks that criteria and testing procedures should be established for state test data sets, and that process records should be kept county by county. The reason for this is that some counties have very poor street files, and there is a wide variation is quality across counties and states.

Section III: Confidentiality

Confidentiality is another important policy issue. Most cancer registries operate under various disclosure rules. Some of these rules are generally referred to as a Rule of 5 or Rule of 6. Simply expressed, no cells in a table will contain a count of less than five (or six) persons. For maps, a geographic unit (census tract, ZIP code, county, etc.) is treated like a cell in a table. Such rules generally state that areas with less than a certain number of cancer counts (such as five or six) do not have the exact number of cancer counts displayed in a table. The exact rules and procedures vary from state to state. There also are Federal regulations. For example, the National Center for Health Statistics (NCHS) has a regulation that for mortality, 20 deaths or less are not shown (tabular) except for large populations of at least 100,000 or unless aggregated by 3 years or more in time.

Analysts need to be aware that spatial data can breach confidentiality that tabular data would not. Maps also can be created that breach confidentiality inadvertently. For example, locals may be able to identify a person from data that would not be meaningful to the general public. Data linked to geography requires additional care. Joining data with other spatial data sets, for example, can lead to specific identifications of individuals. Point locations and phone numbers can be back-matched, and individuals may be identifiable.

Map scale also must be considered when evaluating confidentiality problems. One GIS Workgroup member mentioned a brain cancer study of Sugar Creek, Missouri (Simoes et al., 2000). On the face of it, the maps used in the publication may appear to have a confidentiality problem. However, with each dot covering 250 meters or so, this may not be a significant problem. The use of the map in the publication was to illustrate that brain cancer is everywhere, not just in a few localities. Data were aggregated over quite a few years. However, in very rural areas, individuals might still be identifiable.

Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act (HIPAA; 45 CFR § 164.514) implements privacy standards and covers health plans, health clearinghouses, and health providers who transmit health information electronically. The rule became effective April 14, 2001, and health care providers must comply with the requirements by April 14, 2003. There is a Web site that provides information for users within NCHS. For users outside of NCHS, the U.S. Department of Health and Human Services Web Site at http://aspe.hhs.gov/admnsimp/ provides information and other links.

The regulation restricts the use and disclosure of certain protected health information. For some disclosures, individual authorization is required. Use without individual disclosure is allowed for research and public health purposes subject to certain conditions. Disclosures that are made must be *"limited to the minimum information necessary to accomplish the purpose of the disclosure and must be disclosed consistent with the practices described in the entity's privacy notice (Public Health GIS News and Information, November 2001)."*

Of particular concern in GIS analyses are requirements relating to uses and disclosures of protected health information. Under Federal law, a street address is considered to be a personal identifier. Two

possible procedures are specified for determining that health information is de-identified. The first is that a *"knowledgeable person applying appropriate methods determines that the risk is very small and that the information could be used alone or in combination with other reasonably available information to identify an individual."* The second procedure for determining that health information is de-identified is to use a specific set of 18 identifiers. These range from names to medical record numbers, and include ZIP codes. The following text is from the HIPAA regulations:

(b) Implementation specifications: requirements for de-identification of protected health information. A covered entity may determine that health information is not individually identifiable health information only if:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names;

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of a ZIP code if, according to the current publicly available data from the Bureau of the Census:

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

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

Masking Methods

If aggregated, over space or time, researchers can use data where the Rule of 5 or Rule of 6 is violated. Data can still be used to create a map, even if they cannot be point mapped.

Cartographic means have been suggested to obscure exact locations. These include generating random variation from points that can then be constrained to a specific census tract, grid cell, etc. (Armstrong MP et al., 1999; also *"Statistical Disclosure and Disclosure Limitation,"* a course by Larry Cox that can be found at http://projects.isr.umich.edu/jpsm/materials/ 2002-0129.html). Some tools to generate random variation are available in ArcView scripts, for example. Reconstituting point data as a dot density map is not a preferred solution because this type of map may lead some viewers to give specific spatial connotations incorrectly to the randomized points. Maps using graduated circles to summarize the counts of points within a polygon provide a better solution.

The basic principle is that cancer registries should not publish maps with point locations of cases without masking techniques or without confidentiality agreements in place. Often, it can be deduced from mapped data that a certain group of individuals do not have cancer. This is usually not perceived as a problem, but could it be? This might be called a negative cluster or cold spot, and may deserve additional study.

Section IV: Spatial Analysis

Spatial analysis has been described as "*the crux of GIS, because it includes all of the transformations, manipulations, and methods that can be applied to geographic data to add value to them, to support decisions, and to reveal patterns and anomalies that are not immediately obvious – in other words, spatial analysis is the process by which we turn raw data into useful information"* (Longley et al., 2001, p. 278).

Spatial analysis covers a wide range of topics that are of particular interest to cancer registries, including the following:

- The importance of developing a partnership with an epidemiologist and/or statistician before developing a spatial analysis
- Selected references on spatial statistics and analyses
- Examples of software programs for spatial statistics
- Maps of expected values
- Maps of the locations of individual cancer cases (points) as part of a cancer cluster investigation
- Maps of cancer incidence rates by various geographic units (e.g., census tracts, single counties, or multi-county polygons)
- Evaluating whether observed spatial patterns in cancer incidence rates are random
- Using GIS technology and methods to enhance cancer surveillance efforts
- Applying GIS as part of efforts to analyze access to care
- Methods to minimize the ecologic fallacy problem.

A more detailed report on spatial statistics and analyses methods for cancer registries is currently being developed by the National Cancer Institute (NCI) through a Task Order with Lance Waller at the Emory University School of Public Health.

At this time, a national database does not exist that allows accurate, quantitative analyses of the maps prepared by cancer registries (e.g., as part of annual reports). For state cancer registries to learn what other registries are doing in the area of mapping and to avoid each cancer registry having to re-invent the wheel, the GIS Workgroup suggests that NAACCR or one of its partners periodically compile a collection of cancer registry annual reports (e.g., every 3-5 years), and then report highlights of innovative uses of spatial analyses and mapping. The Workgroup also suggests that maps be routinely included in annual reports. Mapping is an effective communication tool, and NAACCR should promote mapping as a useful descriptive and analytic tool.

The general impression of the Workgroup is that GIS is still in the early stages of development in cancer registries, and even for registries that have started to develop GIS capability, efforts have been focused more on geocoding cancer cases, rather than on spatial and statistical analyses of the geocoded data. Also, the general impression is that annual reports in many cancer registries do not include maps. If maps are included, they are more likely to be simple, thematic maps (e.g., a map of cancer incidence rates by single county or multi-county unit), rather than maps illustrating the application of spatial statistics or more complex spatial analyses (e.g., a smoothed map of cancer incidence). The *Canadian Cancer Incidence Atlas Volume 1* (published by Health Canada in 1996) appeared noteworthy as one of the first cancer registry atlases to apply spatial statistics. Appendix E of the *Canadian Cancer Incidence Atlas Volume 1* includes an evaluation of spatial autocorrelation using Moran's I. Additional details on the Canadian approach to methodological issues are provided in Semenciw, et al. (2000).

Epidemiology, Statistics, and Spatial Analysis

Epidemiologic principles and methods provide the foundation for spatial analyses. To avoid drawing erroneous conclusions from maps, users of GIS technology need to understand and apply these principles and methods in formulating study questions; testing hypotheses about cause-and-effect relationships; and critically evaluating how data quality, confounding factors, and bias may influence the interpretation of results.

Along these lines, GIS users in cancer registries might find it advantageous to develop close partnerships with epidemiologists in the cancer registry. That is, GIS users should design their studies as epidemiologic studies, rather than as GIS studies. Within an epidemiologic study, traditional questions are about person, time, and place. Within the epidemiologic conceptual framework, the question for a GIS user then becomes: how might GIS technology and methods be applied to enhance analyses of questions about place (i.e., questions starting with the word "where")?

The interpretation and analysis of a map for chronic diseases with long latency periods such as cancer often are much more challenging than the interpretation and analysis of a map for diseases with relatively short latency periods. For example, point sources or the cause of the outbreak may be readily identifiable for an infectious disease with a relatively short latency period. In contrast, *"while maps showing the spatial distribution of non-infectious disease are useful in generating hypotheses about disease causation, they are of more limited value in establishing the precise nature of a causal relationship. This is due to the influence of many confounding variables, such as genetics, behavioral characteristics, the difficulty in establishing a dose-response relationship, and the often long latency between the stimulus and the overt response"* (Lawson and Williams, 2001). Alternatively expressed, although maps of cancer incidence or mortality rates might be useful for hypothesis generation and to stimulate questions for future research studies, additional epidemiologic studies will be needed before coming to any definite conclusions about cause-effect relationships.

As part of the efforts to identify and select a GIS contractor for the Long Island Breast Cancer Study Project (http://www.healthgis-li.com/), NCI identified 10 key tasks or functions that a GIS user would likely need to perform in any epidemiologic study. These tasks are: (1) data set loading,

(2) Internet downloading, (3) address matching, (4) geographic area construction, (5) rate estimation, (6) spatial cluster detection, (7) buffering, (8) contour/choropleth map design, (9) spatial smoothing, and (10) map overlay (Broome, 1999).

GIS users in cancer registries may want to go beyond simple maps and measures of spatial correlation to actually conduct statistical analyses of their data. Consultation with a statistician familiar with spatial analyses is recommended prior to using these more advanced methods. For example, there are more than 100 different statistical tests for spatial clustering. Some of these are appropriate for focused tests (clustering of cases around a known exposure location), others for unfocused tests (clustering in general), and still others make restrictive assumptions about the data. The user needs to know which of these is best for their data. Statistical software packages now include procedures for analyzing geographic data; S-Plus (http://www.insightful.com/products/product.asp?PID=17) includes a spatial analysis module that will generate variograms, use Moran's I, and perform other similar functions. Both S-Plus (S+) and SAS/GIS (http://www.sas.com/products/gis/) provide functions to include spatial correlation in regression models. A very active area of statistical research is the application of hierarchical models to spatial and spatial-temporal data. For example, the *NCHS Atlas of U.S. Mortality* published by NCHS (Pickle et al., 1996) includes maps of expected age-specific rates and graphs of regional rates calculated by a hierarchical model. These complex methods are very useful for removing background noise from a map to highlight its underlying patterns, but extensive collaboration with a statistician familiar with these techniques is necessary. Further discussion of these advanced methods is beyond the scope of this document; for more details see, for example, the texts by Elliott et al. and by Lawson listed below.

Selected References on Spatial Statistics and Analyses

As part of the Long Island Breast Cancer Study Project, NCI conducted a review of the literature on spatial analytic methods. This bibliography is provided in Appendix A of Rushton et al. (2000). A number of textbooks and articles on spatial statistics and analytic methods may be useful to GIS users in cancer registries. Several examples (listed in alphabetic order) include:

Anselin L. Spatial Econometrics: Methods and Models (Studies in Operational Regional Science). Norwell, MA: Kluwer Academic Publishers; 1988 (out of print).

Best N, Elliott P, Richardson S. Web Site for Short Course on Spatial Epidemiology Held at the Imperial College, London, March 12-14, 2002. http://stats.ma.ic.ac.uk/~ngb30/.

Chou Y-H. Exploring Spatial Analysis in Geographic Information Systems. Albany, NY: OnWord Press; 1997 (provides an introductory-level description of basic spatial queries, point pattern analysis, network analysis, spatial modeling, surface analysis, and grid analysis).

Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial Epidemiology: Methods and Applications. New York, NY: Oxford University Press; 2000.

Fisher MM, Getis A. Recent Developments in Spatial Analysis: Spatial Statistics, Behavioral Modeling, and Computational Intelligence (Advances in Spatial Science). New York, NY: Springer-Verlag; 1997.

Kulldorff M. Geographic information systems (GIS) and community health: some statistical issues. J Public Health Manag Pract 1999;5(2):100-106.

Lawson AB. Statistical Methods in Spatial Epidemiology. New York, NY: John Wiley & Sons, Ltd.; 2001.

Lawson AB, Williams FLR. An Introductory Guide to Disease Mapping. New York, NY: John Wiley & Sons, Ltd.; 2000.

Lee J, Wong DWS. Statistical Analysis With ArcView. New York, NY: John Wiley & Sons, Ltd.; 2000.

Mitchell A. The ESRI Guide to GIS Analysis: Volume 1: Geographic Patterns and Relationships. Redlands, CA: ESRI Press; 1999.

Rushton G. CD-ROM and Web Site: Improving Public Health Through Geographical Information Systems: An Instructional Guide to Major Concepts and Their Implementation. http://www.uiowa.edu/~geog/health.

Software for Spatial Statistics

In developing a budget for GIS activities, cancer registries should note that specialized software for spatial statistics may be required at additional expense ($2,500 or more) beyond what is required for basic GIS software. Also, as the number of points included in an analysis increases (e.g., more than 1,000 records), upgraded computer capacity (memory) may be required to process a spatial statistical analysis.

Several examples of spatial statistical software (listed in alphabetic order) include:

Cluster Software, Version 3.1 (Note: An updated version of Cluster is scheduled for release in the near future)
http://www.atsdr.cdc.gov/HS/cluster.html

Crime Stat
http://www.icpsr.umich.edu/NACJD/crimestat.html

Distance Mapping and Analysis Program (DMAP) Software
http://www.uiowa.edu/~geog/health/

EI: A Program for Ecological Inference (by Gary King); and
EzI: An Easy Program for Ecological Inference
http://gking.harvard.edu/stats.shtml

ESRI ArcGIS Geostatistical Analyst Extension
http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/index.html

ESRI ArcGIS Spatial Analyst Extension
http://www.esri.com/software/arcgis/arcgisxtensions/spatialanalyst/index.html

Point Pattern Analysis
http://www.dpi.inpe.br/gilberto/csiss/papers/aldstadt.pdf

S+ SpatialStats
http://www.insightful.com/products/product.asp?PID=17

SAS/Stat
http://www.sas.com/rnd/app/da/stat.html

SaTScan Version 2.1 (Software for Calculating the Spatial, Temporal, and Space-Time Scan Statistics) (Note: An updated version of SaTScan is scheduled for release in the near future)
http://www3.cancer.gov/prevention/bb/satscan.html

TerraSeer Environmental Insight Software (BoundarySeer and ClusterSeer)
http://www.terraseer.com

WinBugs/GeoBugs
http://stats.ma.ic.ac.uk/~ngb30/

Maps of Expected Values

The process of converting raw GIS data into useful information for decisionmakers often requires more than a single map. That is, a single map frequently generates questions that require additional maps, tables, graphs, charts, or explanatory text.

A critical part of the process of map interpretation is the ability to compare observed values with a set of expected values. Lawson and Williams (2001, p. 3) write: *"disease mapping refers to the visual representation of the geographical distribution of actual data, but…before any… interpretation can be made or…hypotheses generated concerning the disease displayed in these maps, it is important to consider how many cases would have been expected to be found in the mapped area."* That is, suppose that the distribution of a disease for individual cases or groups has a spatial distribution. The population demographics (e.g., age, gender, ethnicity, etc.) also have a spatial distribution. To assess whether any particular pattern of disease has arisen by chance, the additional knowledge of expected values is needed to discover the pattern that could arise from the demographic characteristics of the underlying population.

Cancer Cluster Investigation

A cancer cluster is the occurrence of a greater than expected number of cases of a particular disease within a group of people, a geographic area, and a period of time. Cancer clusters may be suspected when people report that several family members, friends, neighbors, or coworkers have been diagnosed with cancer. Cancer registries are frequently asked to evaluate whether a cancer cluster exists.

Several software programs are available that may be helpful when a cancer registry is asked to evaluate a cancer cluster. These include: SaTScan, Cluster, CrimeStat, ClusterSeer, and PointPattern Analysis. The URLs for more information about these software programs is provided under the list of spatial statistical software and in the Appendix. The starting point for each of these software programs is a data set with latitude-longitude coordinates for the location of a cancer case.

Cancer cluster detection software provides the ability to detect cold spots (lower than expected number of cases) in addition to hot spots (higher than expected number of cases). High positive spatial autocorrelation can be generated by low values close together (cold spots) as well as by high values close together (hot spots) (Lee J, Wong DWS, 2001). As noted earlier, there are more than 100 statistical tests available for cluster assessment (Kulldorff, 2002). Some of these are appropriate for focused tests (clustering of cases around a known exposure location), others for unfocused tests (clustering in general), and still others make restrictive assumptions about the data. The user needs to consult with a statistician to determine the most appropriate method to use for a particular data set.

Explorations for local or global clustering are facilitated by studies of rare events. An application from South Carolina illustrates this process and the capabilities of GIS in cluster evaluations. In this study, the investigation concerned mesothelioma in a community with a history of many asbestos-exposed workers (see Figure 8). In the figure, the p-value for the risk ratio was mapped and scaled as a halo to provide the classic hot spot representation associated with a cancer cluster. This use of GIS graphic utilities also preserved patient confidentiality without revealing the locations of individual cases.

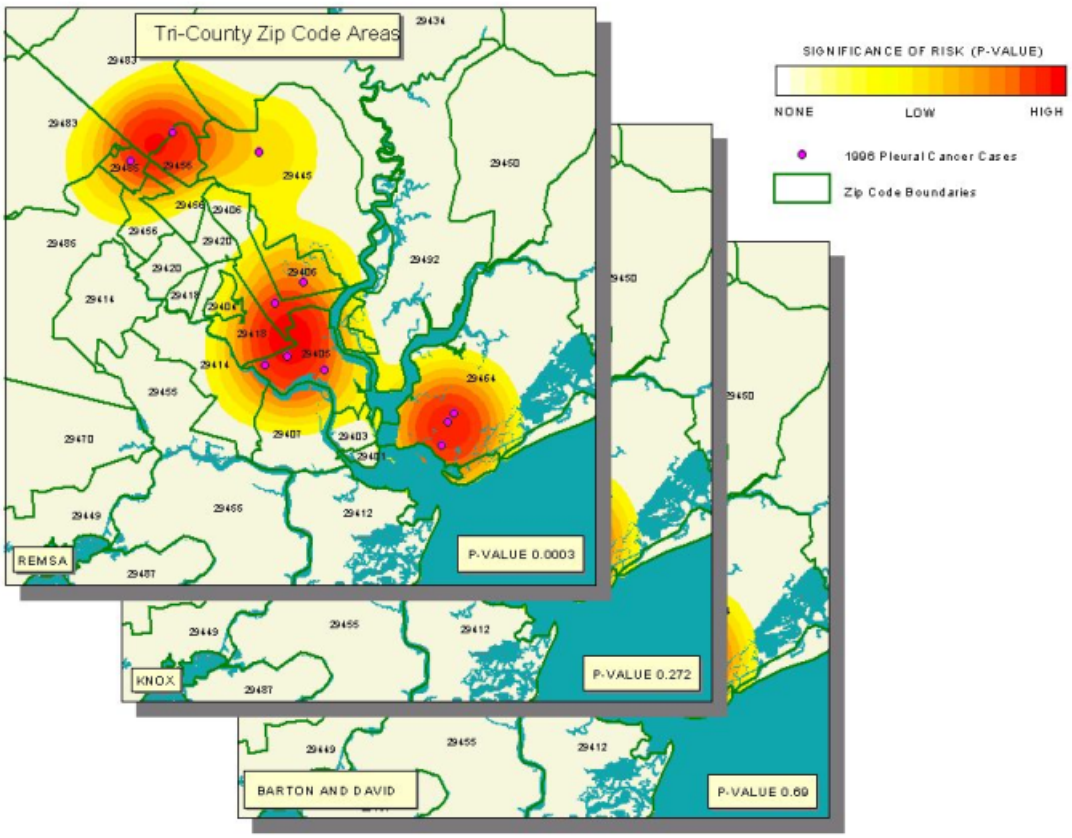

Figure 8. Cancer cluster example.

It is generally agreed that it is the responsibility of public health officials to respond to inquiries regarding potential cancer clusters and address concerns of the public.

Maps of Cancer Incidence Rates

GIS users in cancer registries are often asked to develop maps of cancer incidence rates. Along these lines, the following are several useful resources or starting points:

- The templates shown on page 7 of: Pickle LW, Mungiole M, Jones GK, White AA. Atlas of United States Mortality. Hyattsville, MD: National Center for Health Statistics. DHHS Publication No. (PHS) 97-1015; 1996.

- The Web-based approach to interactive maps and drill down tables and charts developed by NCI for its cancer mortality atlas (http://www.nci.nih.gov/atlasplus).

- The templates proposed by Carr et al. for micromap plots in states with many counties (Carr DB, Wallin JF, Carr DA, 2000; Carr DB, 2001).

- Monmonier's books on *"How To Lie With Maps"* (1996) and *"Mapping It Out: Expository Cartography for the Humanities and Social Sciences"* (1993).

An example of a cancer mortality rate map is shown in Figure 9.

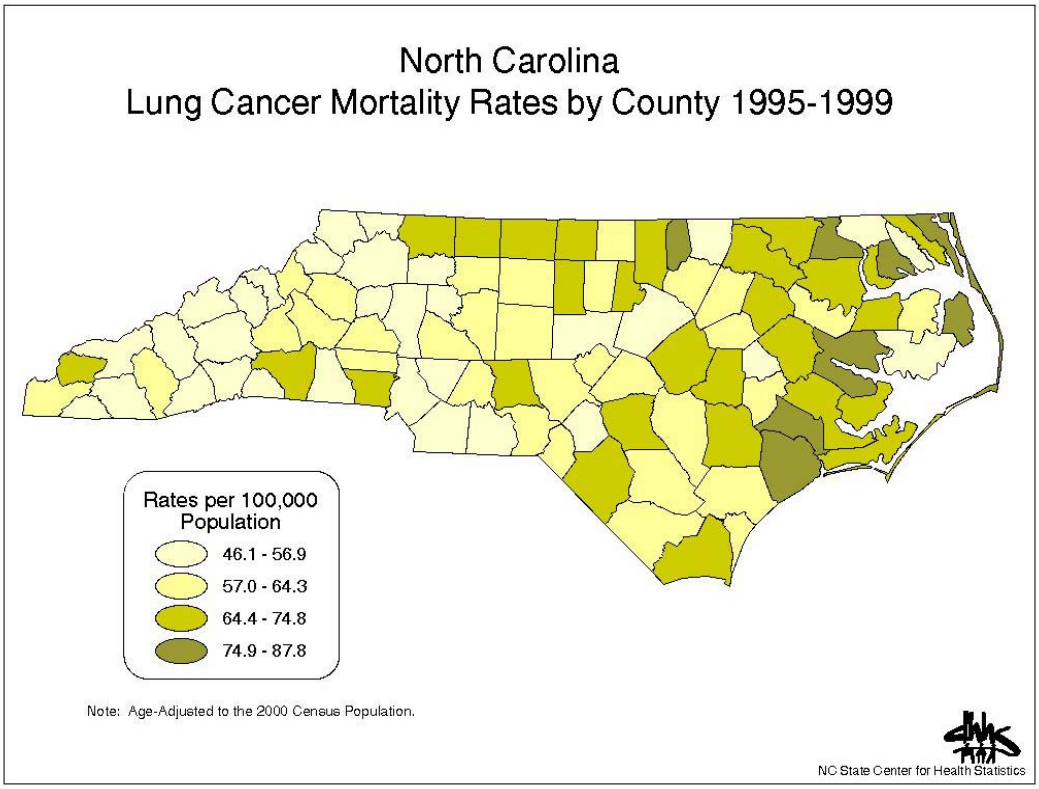

Figure 9. North Carolina lung cancer mortality map.

A number of methodological questions exist in mapping cancer incidence rates. The following questions are of special interest (and will therefore be briefly discussed): (1) What geographic units should be used for mapping cancer incidence rates? (2) What methods should be used for mapping spatial-temporal patterns? (3) What methods should be used to adjust rates? and (4) What methods should be used for representing the reliability of the rate estimates?

What Geographic Units Should Be Used for Mapping Cancer Incidence Rates?

A frequent question in developing GIS maps of cancer incidence rates is whether there might be an optimal geographic unit of analysis. Several examples of geographic units (polygons) that might be of interest include: (1) census block groups; (2) census tracts; single county boundaries; (3) multi-county units (e.g., local health jurisdictions or local health department districts); (4) state economic areas used by NCI in its U.S. cancer mortality atlas; (5) ZIP codes; (6) the health service areas used in the *NCHS Atlas of U.S. Mortality*; and (7) hospital service areas and hospital referral regions used in the *Dartmouth Atlas*. For additional details on the Dartmouth hospital service areas and referral regions, see the Dartmouth Web Site (http://www.dartmouthatlas.org) and the ESRI CD-ROM on *"GIS Solutions for Health and Human Services"* (includes a copy of the shape files used in the *Dartmouth Atlas*).

Examples of potential questions about geographic units of analysis might include: (1) Is there an optimal geographic unit (e.g., should census tract be used, or should ZIP code be used, and so forth)? (2) How should denominators be calculated for rates (e.g., if the denominator data from the census were collected using a different geographic unit than the numerator data from the cancer registry)? and (3) What methods should be used when geography changes over time (e.g., when the census tracts for the 1990 Census are not identical to the census tracts in the 2000 Census)?

With GIS technology and methods, a potential solution to the optimal geographic unit question (being explored by some researchers and states) is to use a grid to develop smoothed maps of cancer incidence rates that are independent of the boundaries of administrative/political units. Additional details on smoothed maps are provided on the University of Iowa Web Site (http://www.uiowa.edu/~geog/health/).

Aggregated data sets have a problem called the Modifiable Areal Unit Problem (MAUP). The MAUP actually consists of two closely related problems (Openshaw, 1983). The first problem arises when a set of areal units are aggregated into fewer and larger units. There are many different grouping strategies for a set of data. The geographical areas are termed modifiable, because the choice is arbitrary. Spatial-temporal data have the additional difficulty that the temporal component also is modifiable. Analysis results often change with the aggregation. The second MAUP problem relates to the fact that alternative combinations of areal units can change the results of analysis.

What Methods Should Be Used for Mapping Spatial-Temporal Patterns?

Spatial-temporal issues are important in many cancer analyses. For example, patients that spent many hours on the beach as a teenager may be diagnosed for melanoma in their 40s, a significant amount of time between exposure and symptoms. Including a temporal dimension in geospatial data is a recognized area of GIScience research. For example, this issue is included in 1 of the 10 research

priorities identified by the University Consortium for Geographic Information Science (see http://www.ucgis.org).

One of the first GIScience research efforts in the spatial-temporal area was Langran's *Time in Geographic Information Systems* (1992). A new advanced textbook by Peuquet (*Representations of Space and Time*) also focuses on these issues (Peuquet, 2002). Other recent work has examined geovisualization of activity-travel patterns in time (Kwan, 2000), geospatial lifelines for studying environmental health issues (Mark et al., 1999), and residential history models (Yang, 2001).

For cancer registries, a specific spatial-temporal issue is how to design cancer registry databases to record the address at diagnosis, plus information describing the geospatial lifeline for the patient. A geospatial lifeline is the record of all the locations that an individual has occupied during a specified time period. The theoretical framework for the geospatial lifeline is Hagerstrand's time geography. The specific area of interest for cancer research is in the spatial-temporal history of people's place of residence. Recent research by Yang is of particular interest. He studies hot spot detection techniques and discovers that when a snapshot view of longitudinal data is used, incorrect conclusions may be drawn. Some of the noted difficulties in this work are the large size and ill-structured nature of this type of data set. He also suggests alternative visualization technologies.

There are many health examples of the problems of attempting to infer exposure based on residence for diseases that have multi-year latency periods. An example reported by the Director of the Minnesota Cancer Surveillance System of a less than productive effort is the current activity to identify mesothelioma occurrence around vermiculite processing plants in metropolitan areas. Many of the people now in these neighborhoods were not living there during the likely exposure period, and most of the people who may have been exposed when they lived in the neighborhoods in the past have since moved away.

Cancer registries currently geocode the data collected about an individual to the address at the time of diagnosis. The years of residency at that address are not currently collected. The "usual residence at diagnosis" rules are based on Census 2000 rules, found at http://www.census.gov/population/www/censusdata/resid_rules.html. The address of the death certificate is matched later, if possible. In some cases, the date of the death certificate is the diagnosis date, and that address is used. There are often problems matching address at time of diagnosis to address of death certificate across state lines.

Some registries keep all of the historical addresses, starting with the address at diagnosis. Many agencies keep such data (e.g., Social Security), but access to these data is limited. Years at address might be available for individuals, but comparison data for others are not available. As noted previously, such data sets are large, complex, and ill structured.

Another issue identified by the GIS Workgroup is seasonal residency (Boscoe, 2002). Resort/vacation communities have no or low population counts due to the date of census data collection, but many have a large number of cancer cases. Seasonal residency also has been identified as an issue in the collection and analysis of crime data. A New Jersey law adds people to resort municipalities for the purposes of computing crime statistics (P.L. 1998, c.50 [S233]).

Another spatial-temporal issue is the change in administrative boundaries over time. For example, ZIP code boundaries may change over time; and although more stable than ZIP codes, changes also occur in census boundaries geography (e.g., census tracts in 2000 are not always the same as census tracts in 1990).

A recent example of the dynamism of ZIP code boundaries is the following. In 1998, the U.S. Postal Service changed the ZIP codes in 10 suburbs west of Boston. Five of the suburbs that had been covered by one ZIP code received at least one additional ZIP code (Boston Globe (May 2, 1998): Journal Code: BG Standard No: ISSN: 0743-1791). The noteworthy issue about this particular set of ZIP code changes was that the intent was to release ZIP code ranges for reuse. Most ZIP code changes involve the simple retirement of a code, the creation of entirely new codes, or the organizational change of a postal designation from a community or branch post office into a type of office higher in the postal hierarchy. What was interesting about this case was that many preexisting ZIP codes were replaced so that they can be reassigned in the future to other geographic areas. For example, the ZIP code 02172 belonged to Watertown; it was replaced with a new code, 02472, and the old code (02172) will be reused in the future within the city of Boston, a completely different geographic area. If an analyst did not update the old ZIP codes in their data, the single ZIP code 02172 would end up representing two completely different geographic areas.

Because of the dynamism of administrative boundaries, the definition date used should be included in the metadata for the data set. Even with this information included, a difficult spatial-temporal question arises: Which census tract should be used for data reported after the next boundary revision, but from an incidence datum before that? How quickly data comes in varies enormously (i.e., after the incidence date). There is a lag between diagnosis and geocoding; there is no known usual or customary lag period, it may be 20 years or more. Even if reporting is required much sooner, work on data may continue. Some research has been completed on adjusting incidence rates based on reporting delay (Clegg et al., 2000).

Because spatial-temporal issues are of core importance in the spatial analysis, the Workgroup suggests that cancer registries explore the feasibility of collecting longitudinal data.

What Methods Should Be Used To Adjust Cancer Incidence Rates for Maps?

Maps of cancer incidence rates need to be adjusted to control for potential confounders (e.g., age distribution). At least three methods have been proposed for calculating adjusted cancer incidence rates. Currently, guidance does not exist for GIS users in cancer registries as to the relative strengths and weaknesses of the various methods, and whether any specific method might be preferable depending on the purpose of the map.

The first method is direct age adjustment, in which the rate for a small area geographic unit of analysis is based on the specific numerator and denominator for that geographic unit. This approach has been used by NCI to develop national-level maps of cancer mortality (http://www.nci.nih.gov/atlasplus). A potential limitation of this approach is that the directly standardized rates can be statistically unstable for areas with small numerators and denominators.

The second method is a direct age adjustment method, in which the rate displayed in a small area geographic unit is calculated using the numerators and denominators from the immediately contiguous geographic units in addition to the information from the target geographic unit. This type of approach has been applied in the CDC heart disease mortality atlases (http://www.cdc.gov/nccdphp/cvd/womensatlas/spatialsmooth.htm). One result of this method is the introduction of more spatial correlation than was in the original data set, which can violate the assumption of independent rates required by some statistical procedures.

The third method is indirect age adjustment, in which the expected numbers of cases are calculated by multiplication of the number of individuals in each demographic group within a small geographic unit of analysis (e.g., a census block) by the corresponding, demographically stratified rate for a much larger geographic area (the entire state). Indirect age standardization has been used in a computer program developed at NCI for cancer cluster analysis (http://srab.cancer.gov/satscan).

In one comparison of methods, researchers determined that the direct adjustment method is more appropriate for mapped data because each geographic unit is adjusted to the same standard population, and thus, the data are comparable (Pickle and White, 1995). NCI and NCHS publish cancer statistics using the direct method. However, the indirect method has been recommended as a measure of the burden of cancer in the population (Rushton, 2003, in press).

What Methods Should Be Used for Representing the Reliability of Cancer Rate Estimates on Maps?

When a state cancer registry develops a cancer incidence map, the size of the denominator may be small in certain geographic units. In turn, the relatively small numbers in the denominators for these units may result in unstable and/or unreliable cancer rate estimates. Thus, an issue for GIS users developing cancer incidence maps is the appropriate method to identify and display unreliable rates on a map.

In its *Atlas of U.S. Mortality*, NCHS added hatched lines to convey rate variance information to readers without hampering their ability to identify underlying patterns on the maps. Other atlases have used alternative methods. For example, in its *Atlas of Cancer Mortality*, NCI used a gray color to indicate areas with sparse data (i.e., unstable rates). Other methods also may exist (e.g., mapping spatial patterns of residuals from regression equations). In an experiment that compared several methods of representing rate reliability on a map, the hatch overlay and separation of rate and reliability information into two separate maps worked well, but a bivariate color scheme did not (MacEachren, 1998). The commonly used method of blanking or graying out unreliable areas interfered with the reader's ability to identify trends and clusters on the map (Lewandowsky 1993).

Are Observed Spatial Patterns Random?

After a GIS user in a cancer registry has prepared a map of cancer incidence rates using a selected geographic unit of analysis (e.g., census tract or census block group), the next analytic question is whether the observed spatial patterns are random. Spatial statistical methods may be helpful in providing a quantitative answer to this type of question, especially in situations where clustering may not be borderline and/or not immediately obvious to the map reader.

GIS methods can be used to calculate the centers or centroids of the geographic units (polygons) of interest, and the rate value for each unit can be assigned to the centroid. Spatial statistical tests (e.g., SaTScan) can then be applied, using the latitude-longitude coordinates for the centroids.

Cancer Surveillance Efforts

GIS technology and spatial analysis tools potentially could be applied to help add value to cancer surveillance efforts. An example is provided by Kulldorff (2001) for thyroid cancer among men in New Mexico from 1973 to 1992. Kulldorff points out that most disease registries are updated at least annually. If a geographically localized health hazard suddenly occurs, the author suggests that it would be desirable to have a surveillance system in place that can detect a new geographical disease cluster as quickly as possible, regardless of its location and size. Although a surveillance system is desirable in such instances, false alarms may unduly trouble the public. The author proposes a system for regular, periodic disease surveillance. The system would detect any currently active geographical clusters of disease and test the statistical significance of such clusters, adjusting for possible geographical locations, sizes, and time intervals.

Analysis of Access to Care (Distance)

As part of cancer prevention and control efforts, cancer registries may want to evaluate patterns of health care. GIS potentially provides tools and methods to quantify the distance between the residence of the patient and the location where the diagnosis was first made and/or where certain types of treatment were provided. The GIS estimate potentially can range from a simple distance (i.e., straight line between two points) to a more complex measurement (e.g., the time that would be required by a patient to drive along a street).

To date, a relatively small number of cancer prevention and control studies have been published that have used GIS to measure travel. Some of these studies have come to conflicting results (e.g., whether the choice of therapy for breast cancer may vary as the distance increases that a patient needs to travel to receive radiation treatment). For example, a study in Iowa suggested that travel distance was a factor in choice of radiation treatment, but a study in New Mexico did not observe such an effect (Rushton G, West M, 1999; Athas WF et al., 2000).

Methods To Help Minimize the Ecological Fallacy Problem

The ecological inference problem refers to whether researchers can reliably infer individual-level behavior from aggregate (ecological) data. In the 1980s, epidemiology textbooks tended to be critical of studies with an ecologic correlational design, that is, where a GIS user attempted to draw inferences about cause-effect relationships in individuals based on correlations of aggregated, population-level data. Conclusions about cause-effect relationships may be considerably different if a study is based on analysis of individual-level data, as opposed to population-level correlations. In the 1990s, however, some epidemiologists began to revisit the methodologic problems in ecologic studies, and to explore alternatives such as multilevel analyses and designs to better distinguish biologic, contextual, and ecologic effects (Morgenstern H, 1998).

Gary King (Harvard University) also has proposed a solution to the ecological inference problem (King G, 1997), and has developed a software package to facilitate ecologic analysis (http://gking.harvard.edu/stats.shtml). King's methods and software have not yet been evaluated and used by any of the GIS Workgroup members for specific application in a cancer registry context. King's methods involve reformulating the data by generalizing the method of bounds both algebraically and with graphical methods as well as providing narrower bounds for the aggregate-level quantities of interest. King's methods also involve modeling the remaining uncertainty within the observation-level deterministic bounds.

Section V: Cartography

Introduction to Cartography

Cartography has been described as "the meeting place of science and art." The primary purpose of a map is to convey information and to illustrate a geographical concept or relationship (this is the science component). It also is desirable to produce attractive maps from an artistic perspective. There are many excellent cartography textbooks to help create correct, clear, and attractive maps (e.g., Robinson et al, 1995. *Elements of Cartography*).

Maps have a variety of roles. They are used as aids in thinking visually and spatially. They also are used to communicate analyses. Researchers use maps for exploration (to generate questions and hypotheses). Finally, maps are used to help confirm (or not confirm) hypotheses, communicate what analysts think they know, and synthesize and present findings.

Maps have limitations. Because there is a tremendous amount of information that could be portrayed, map designers need to select, generalize and simplify. There is a large body of academic literature on these topics (Buttenfield, 1991). If too much information is presented, it may hide the core theme of the map and obscure its understanding for the audience. If too little detail is included, it also can lose the audience. A poorly designed map fails to communicate effectively, and may in fact deliver the wrong message to the audience (Monmonier, 1993). The underlying spatial analysis may be complex and effective, yet the audience may not be convinced if the final output products fail to effectively portray the message of the results. Another useful point is that all of the information does not have to be incorporated into a single map. Using more than one map can be an effective approach.

It is important to understand both the audience and the viewing environment of maps. For the audience, it is important to understand the demographics of age, education, map-reading experience, and knowledge of the substantive area of the map contents. Color blindness and other disabilities must be considered as well. The viewing environment provides a context for map design. Will the map be in a newspaper, report for public information, scientific paper, or atlas? Will it be distributed on the Internet in either static or interactive formats? Will it be viewed on paper, or in a room with a computer projection system? Will the reader spend a large amount of time studying the map, or flip the page quickly?

One of the fundamental cartographic design principles is that the map designer must have a clear understanding of the major communication goal of the map. In other words, the cartographer must have a clear understanding of what the map reader should conclude after studying the map. Maps have different purposes, including purposes of convincing and persuading in the case of maps used for advertising (Monmonier, 1993).

Design Elements

Most good maps include a common set of design elements. Common elements are shown in Figure 10.

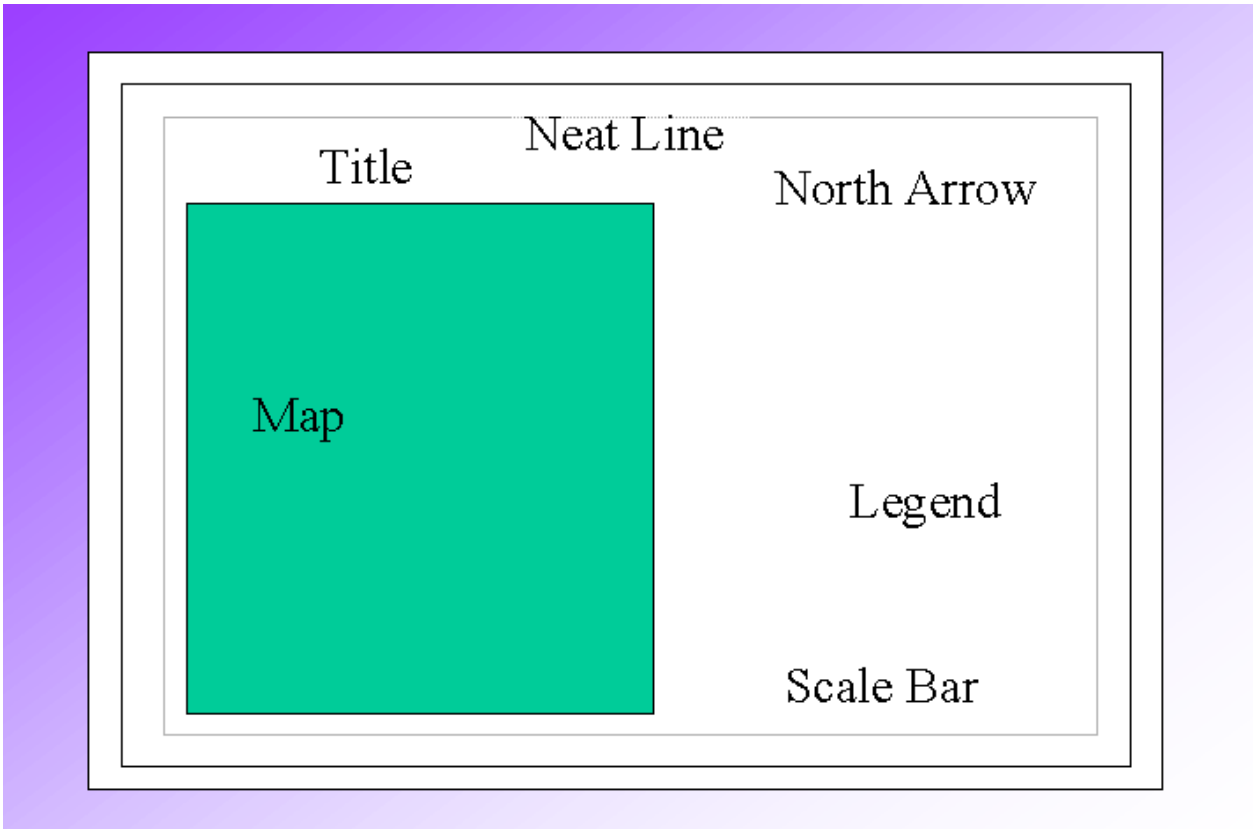

Figure 10. Common design elements in a map.

These common elements include:

- **Title:**
  - Should be matched to theme and audience.
  - Should be composed to be concise, but accurate.
  - Titles should be simple and clearly indicate the purpose of the map. In other words, what information the map designer is trying to convey to the audience.

- **Legend:**
  - Provides symbol interpretation.
  - Should be designed with ease of interpretation and clarity in mind.
  - Should portray any features on the map that may cause confusion or be an unknown symbol to the audience.
  - If the map designer is not absolutely sure that the audience will understand a symbol, include it in the legend.

- **Map body:**
  - The map itself should have the necessary amount of data and detail.
  - Too much detail can result in losing the intended message.

- **Scale:**
  - May be verbal, a representative fraction, or graphic.
  - Direction indicator.
  - Representing true north, magnetic north.
  - North arrows help to orient viewers who are unfamiliar with the area portrayed on the map. Choosing a north arrow is a matter of personal style and taste. Ensure that the graphic does not overwhelm the information.

- **Labels:**
  - Include place names, data values, etc., as needed, but not to the extent that the major communication goal of the map is obscured
  - Labels on the map should identify features on the map, but not every feature.

- **Source:**
  - Should provide clear reference links to data sources.
  - Other documentation.

Other map elements appear often or selectively, such as the following:

- **Projection:**
  - Needed if the theme of the map involves area, distance, direction, or shape.

- **Cartographer:**
  - Identify the individual and/or organization.

- **Date of production:**
  - May not be needed if the map is in a dated book, journal, or periodical.
  - Especially important for time-sensitive data, such as a weather map.

- **Neat line:**
  - Line around the map extent to indicate exactly where map begins and ends.

- **Locator maps:**
  - Small-scale map used to provide locational context for a larger scale map.

- **Inset maps:**
  - Large-scale map of a zoomed-in portion of the main map.

- **Index maps:**
  - Depicts the location of each of several map compositions that comprise coverage of an area.

After the map elements are complete, they need to be combined in a pleasing design. The map designer needs to decide on the relative importance of each element in communicating the map's objective to the audience. Many cartography textbooks recommend the use of a visual hierarchy of elements. This hierarchy helps to emphasize the more important elements. The typical hierarchy

suggestion is that the more important elements be larger, near the top of the map, and near the left of the map. Less important elements are then smaller, near the bottom of the map and to the right side of the map. Also, it is important to maintain a pleasing visual balance. Sketching the map out helps in the design process, and the flexibility of the GIS layout tools is a significant aid in evaluating design alternatives.

Another important consideration is the scale of the map. This should not be confused with the scale bar, but it is related. The map designer needs to be aware of the scale of a map and the space on the page. The small-scale map shows a large amount of area and enough information to make the map useful in a general way; however, a small-scale map can have too much information and may be difficult to read. The large-scale map shows a small area in great detail, but a large-scale map may not have enough information to make it useful (see Figure 11).

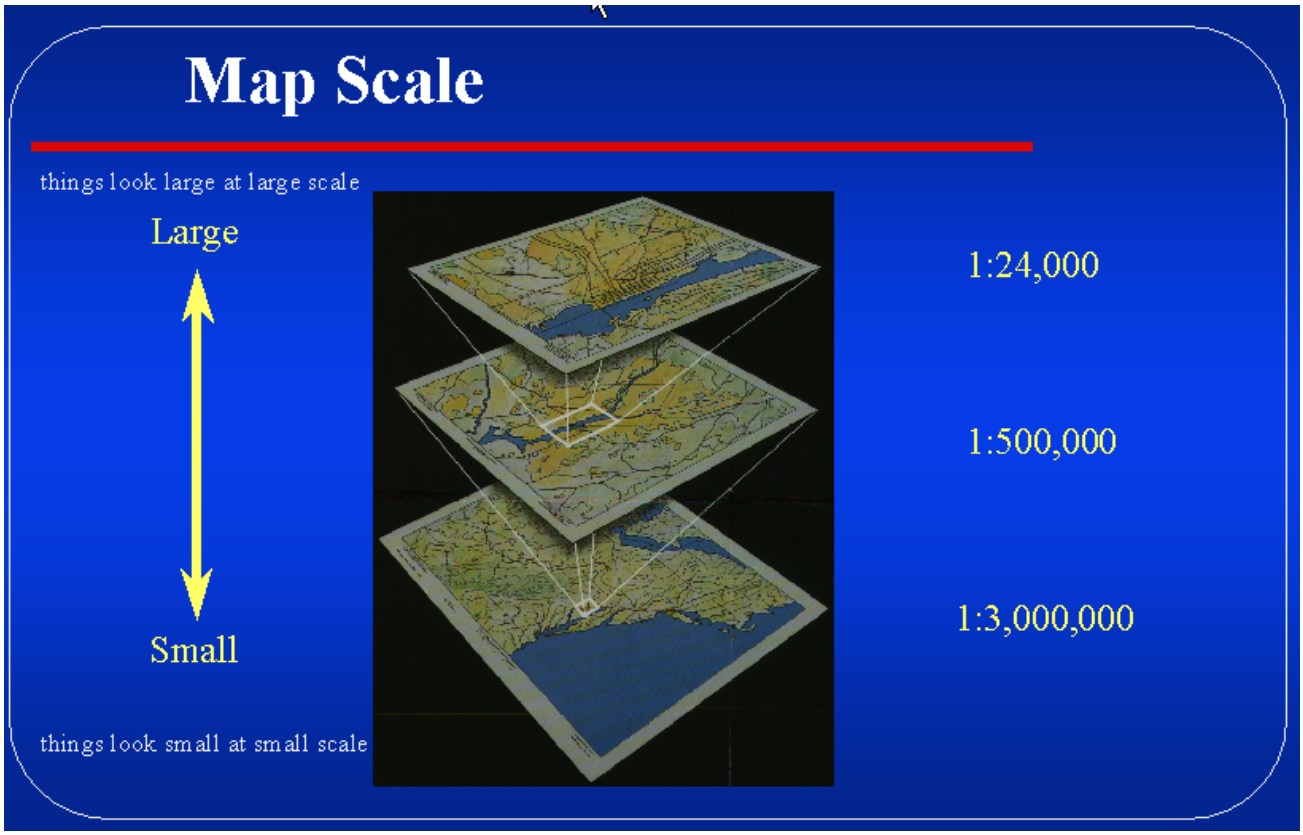

Figure 11. Small versus large scale in maps.

An additional important consideration is the projection used for the map. A detailed discussion of projections is beyond the scope of this document, but is covered in detail in cartography and GIS textbooks. Many organizations choose to standardize on particular projections. Information about projections and coordinate systems also is one of the most important fields of geospatial metadata. Without this specific information, it is difficult to share data across organizations.

Media of Graphic Communication

Cartographers make good use of the media of graphic communication. Cartographers consider the careful use of: (1) shape, (2) size, (3) value, (4) pattern, (5) hue or color, and (6) direction. Color choices are of particular concern in that most persons have learned expectations about color. Colors have connotations that can "make or break" a map. For example, individuals are accustomed to blue representing water and green representing vegetation, lowlands, and forests. Some professional groups and organizations adopt particular color standards for their maps. Color conventions are useful, although there is the potential for incorrect use unless the map designer is aware of their limitations. The convention for quantitative data is that either darker or warmer colors represent higher values. A map with blues for the highest values and yellows for the lowest values will be confusing because it conflicts with this convention—does the darker blue or the warmer yellow represent the high values? When there is a strong expectation about colors, such as the historical use of reds for high rates and blues for low rates in cancer mortality maps, the reader will be confused if these conventions are violated, even when the legend is clear (Carswell, 1995).

Colors attract the eye and affect the perception of the feature. When data are classified into groups (classes), colors need to be assigned that work well in distinguishing between the classes. A recent National Science Foundation (NSF) study has funded research by Cynthia Brewer that has produced a Web site that is particularly useful for making the color choices for sequential (light to dark or dark to light); diverging (dark to light of one color, then light to dark of another color); and qualitative color schemes. Diverging schemes are useful when one of the goals of the map is to show where rates are higher or lower than some middle value (e.g., the median or U.S. overall rate). The Brewer Web Site (http://www.colorbrewer.org) also helps inform the map designer of concerns about supporting the needs of the color blind, printing in black and white, displays on a laptop computer, and room projection from a computer projection system. An experiment comparing various color schemes for mortality rate maps showed that color enhanced the effectiveness of the maps compared to gray-scale maps, although no one color scheme was superior (Brewer, 1997).

Above all, remember that a map is a document that will be used for many purposes; it also can be considered graphic art. A map that appears confused and/or unbalanced can be difficult to interpret, to the point of being useless, or remembered for the wrong things. GIS systems make the creation of thematic maps almost too easy. If the map designer quickly chooses the defaults used by the program, he or she needs to stop and closely examine what is produced, because it might not be suitable for making the salient points.

Map Symbolization

Map designers need to be aware that maps are a graphic rather than a written form of communication. Written communication presents information in a cumulative and logical sequence. The graphic quality of maps means that a large amount of information is presented at one time. Therefore, the symbolic language of mapping is extremely important.

Some map symbols are chosen to replicate their real-world counterparts (e.g., the representation of a stream system on a map). Other symbols are much more abstract (e.g., the color light green means lower median income, darker green represents higher median income, darkest green means the

highest income). Map scale also influences how map designers classify a type and assign symbology. For example, a street network may be represented as a street centerline on a small-scale map, but as a street with width on an engineering drawing at a large scale. Similarly, cities may be represented as points on a small-scale map but as polygons with areas on a large-scale map.

The combination of symbols and colors can be effective in conveying two types of information on a single map. For example, an experiment by MacEachren (1994) showed that overlaying color shading by hatched lines was more effective in communicating where mapped rates were unreliable than using a bivariate color scheme. This method was used in the *NCHS Atlas of U.S. Mortality* (Pickle, 1996) to permit the reader to judge spatial trends in the data while at the same time being warned that individual rates were based on small numbers.

Map Types

Good cartography textbooks help map designers select the appropriate map type for their application. Examples of point symbol maps include:

- **Dot density map:**
  - Used to map the number of things in space.

- **Proportional symbol map:**
  - Point symbols that vary in size.
  - Size or area of a symbol is directly proportional to the data value it represents.
  - Used to represent the relative magnitude of data within a set of defined areal units, a set of counties, for instance.

- **Graduated symbol map:**
  - Use a discrete set of symbol sizes.
  - Each size represents a range of data values (an alternative design is one in which the symbol size is directly proportional to the underlying data value, but it can be difficult to distinguish among similar-sized symbols).

Line symbol maps are used to represent:

- **Boundaries:**
  - True geometric lines on the earth.
  - Lines that have extent, but not thickness.
  - Political boundaries.

- **Networks:**
  - Physical or information links among a set of places (e.g., roads, migration, commodity flows, communication links).

- **Flow:**
  - Quantity of movement among a set of points, in a network, or among areas.

Area symbol maps are used to represent data that in concept or in fact extend over an area. To correctly create area symbol maps, the map designer needs to be aware of the level of measure of their data. For nominal (qualitative) data, the symbol used shows the boundary of a nominal category. These thematic maps are often called unique value maps. Examples include maps of land use (see Figure 12).

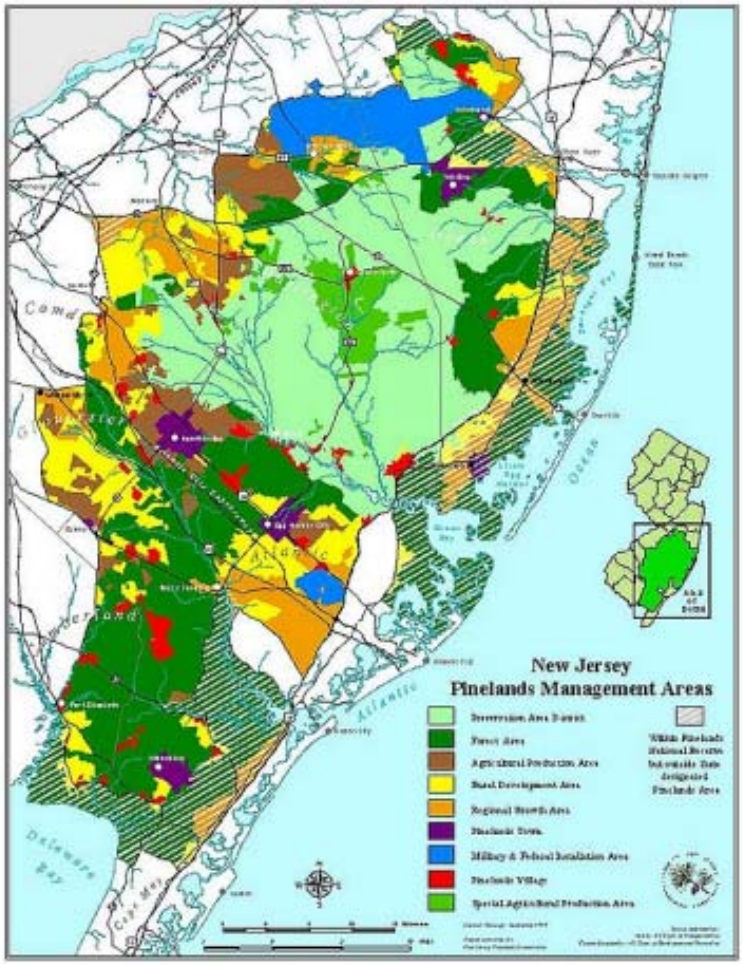

Figure 12. Unique value map: Land Use.

For ordinal, interval, or ratio-level data, the area symbol is used to show the magnitude of an attribute for a bounded area. Thematic maps for quantitative attributes need to have the data classified into groups before they can be mapped. An example of a graduated color map from the *Atlas of Cancer Mortality in the U.S., 1950-94*, is presented in Figure 13. The maps from this document are available online (http://www3.cancer.gov/atlasplus/).

Map designers need to integrate statistical data properly with choice of map type and also with choice of classification method. Most GIS packages provide alternative methods of classification, typically including equal interval, quantiles, standard deviation, natural breaks (typically using Jenk's algorithm), equal area, and user defined. The choice of numbers of classes and the classification method used impacts the usefulness and readability of the map. Research work in this

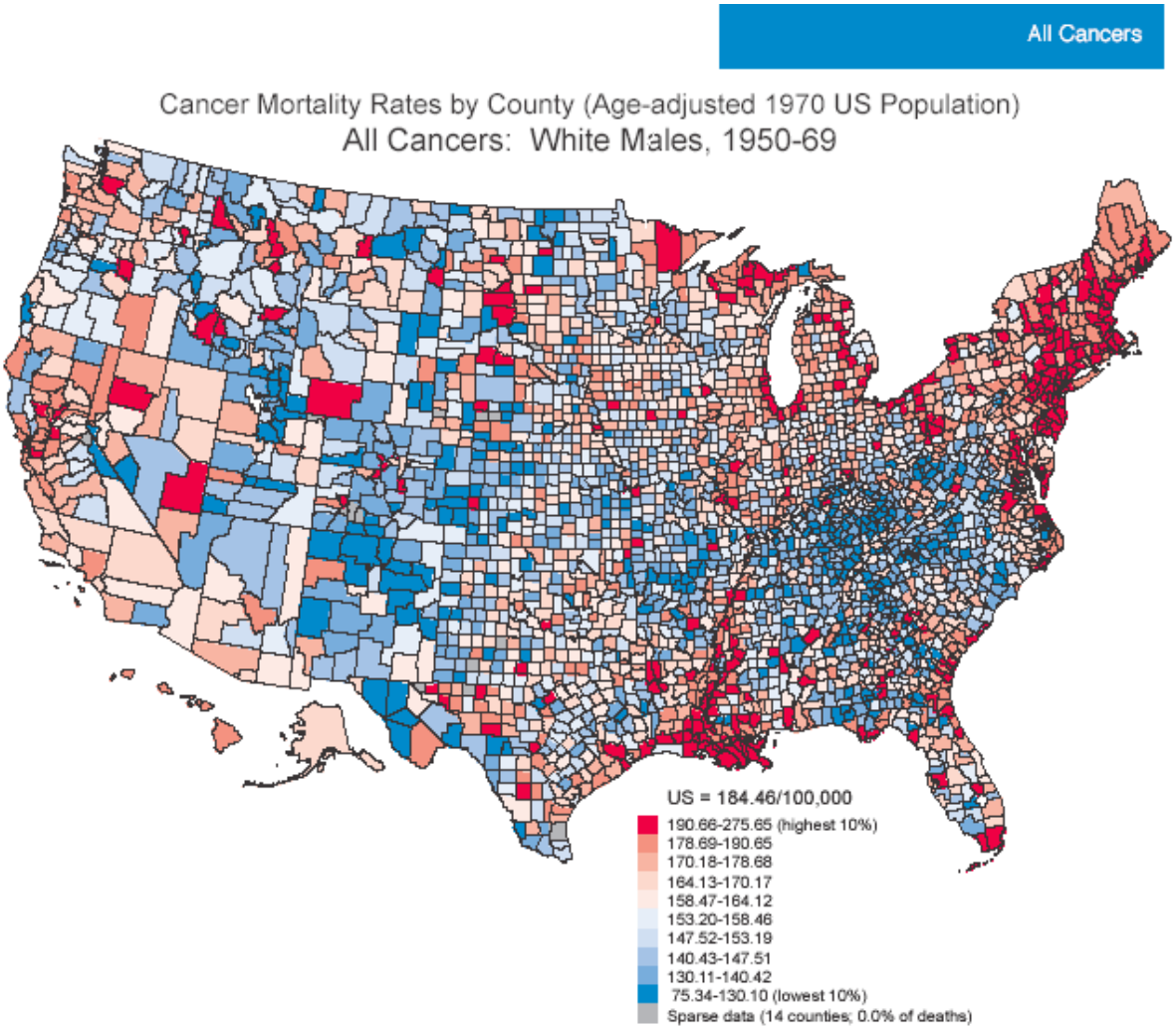

Figure 13. Example of a choropleth map: Cancer Atlas.

area is helpful in making good decisions. Equal interval classification is useful when the mapped quantity is in familiar units, such as the number of packs of cigarettes smoked per day. However, adjusted incidence or mortality rates are only meaningful in relation to other similarly adjusted rates, and so quantile classification is typically used. A recent experiment showed that the quantile method was the best of seven methods at conveying patterns of mapped rates (Brewer C, Pickle L, in press, *Annals of the Association of American Geographers*).

Maps using surface or volume symbols include maps used to represent phenomena that vary continuously over space such as rainfall, elevation, and temperature. They also can be used to represent phenomena that can usefully be conceptualized as a surface, such as population density.

Animated Maps

In Section IV: Spatial Analysis, the importance of spatial-temporal data analysis in cancer mapping was discussed. Map animations provide a visualization tool for geographic data over time. Map animations are now appearing on Web sites, along with static maps and interactive maps (maps that allow users to perform simple GIS functions from Web browsers without local GIS desktop software). Kraak and Brown (2001) provide guidance on cartography on the Web.

Examples of animated maps for communicable diseases in North Carolina can be found at http://www.schs.state.nc.us/SCHS/healthstats/healthatlas.html. The NCI Cancer Mortality Atlas Web Site (http://www.nci.nih.gov/atlasplus) now includes the ability to develop a series of maps over time (e.g., to show the decline in cervical cancer mortality). An example of a series of 5-year maps by state economic area is included in Figure 14.

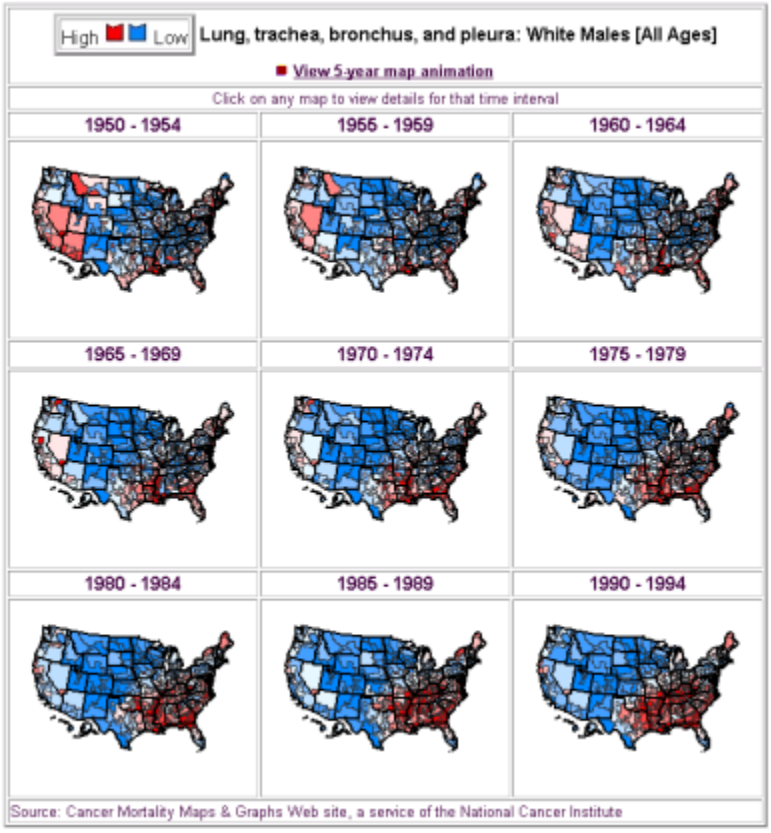

Figure 14. Example of map animation.

Section VI: Internet Access Issues for the Disabled

On August 7, 1998, President Clinton signed into law the Workforce Investment Act of 1998, which includes the Rehabilitation Act Amendments of 1998. Section 508(a)(2)(A) of the Rehabilitation Act (29 U.S.C. 792) requires that when Federal agencies (or people funded by a Federal agency) develop, procure, maintain, or use electronic and information technology, they shall ensure that the electronic and information technology allows Federal employees with disabilities to have access to and use of information and data. This access shall be comparable to the access to and use of information and data by Federal employees who do not have disabilities, unless an undue burden would be imposed on the agency. The Rehabilitation Act also applies to *"individuals with disabilities who are members of the public seeking information or services from a Federal department or agency to have access to and use of information and data that is comparable to the access to and use of the information and data by such members of the public who are not individuals with disabilities."* Technical standards for compliance were published on December 21, 2000, and took effect on June 21, 2001.

The criteria for Web-based technology and information are based on access guidelines developed by the Web Accessibility Initiative of the World Wide Web Consortium. Many of these provisions ensure access for people with vision impairments who rely on various assisting products to access computer-based information, such as screen readers that translate what is on a computer screen into automated audible output, and refreshable Braille displays. Certain conventions, such as verbal tags or identification of graphics and format devices, such as frames, are necessary so that these devices can read them for the user in a sensible way. The standards do not prohibit the use of Web site graphics or animation. Instead, the standards aim to ensure that such information also is available in an accessible format. Generally, this means use of text labels or descriptors for graphics and certain format elements (HTML code already provides an *Alt Text* tag for graphics that can serve as a verbal descriptor for graphics). This section also addresses the usability of multimedia presentations, image maps, style sheets, scripting languages, applets and plug-ins, and electronic forms.

Cancer Registry personnel should be aware of these access issues as they prepare information for publication on the Internet. This section describes organizations and activities of particular relevance. The World Wide Web Consortium (W3C) was created in October 1994. It is an international, vendor-neutral, industry consortium of more than 483 organizations. The W3C plays a central role in the development of Web technologies by developing common protocols that promote the evolution of the Web and ensure its interoperability. The W3C has four domains: (1) Architecture, (2) User Interface, (3) Technology and Society, and (4) the Web Accessibility Initiative (WAI). It works aggressively with government, industry, and community leaders to establish and attain Internet accessibility goals.

W3C is committed to removing accessibility barriers for all people with disabilities—including the deaf, blind, physically challenged, and cognitive or visually impaired. The WAI was launched during the Sixth International World Wide Web Conference in 1997. WAI's mission is to lead the Internet to its full potential, and includes promoting a high degree of usability for people with disabilities. Many Web sites are not accessible to large segments of the disabled community. The number of affected people is large, estimated at 54 million individuals in the United States, and

10-20 percent of the population in most countries. With the average age of the population increasing, changes in vision, dexterity, hearing, and memory problems are certainly important issues to consider in the utility of Internet sources. More information is available on the WAI homepage (http://www.w3.org/WAI/).

The WAI develops guidelines for accessibility. These guidelines play a critical role in making the Internet accessible, and explain how to use Web technologies to create accessible Web sites. Following the guidelines also will make information via the Internet (Web content) more available to all users, whatever their user agent or the constraints under which they may be operating. A document entitled *Web Content Accessibility Guidelines* is available. This document includes 16 guidelines, or general principles, of accessible design. Each guideline is followed by a more detailed statement of the principle and the rationale behind the guideline. The most important part of each guideline is the list of checkpoints explaining how the guideline applies in typical content development scenarios. Each checkpoint definition links to a section of the techniques document, where implementations and examples of the checkpoint are discussed. Each checkpoint is specific enough so that an individual reviewing a page or site may verify that the checkpoint has been satisfied.

Each checkpoint has a priority level assigned by the WAI based on the checkpoint's impact on accessibility:

- **Priority 1:** A Web content developer must satisfy this checkpoint. Otherwise, one or more groups will find it impossible to access information in the document. Satisfying this checkpoint is a basic requirement for some groups to be able to use Web documents.

- **Priority 2:** A Web content developer should satisfy this checkpoint. Otherwise, one or more groups will find it difficult to access information in the document. Satisfying this checkpoint will remove significant barriers to accessing Web documents.

- **Priority 3:** A Web content developer may address this checkpoint. Otherwise, one or more groups will find it somewhat difficult to access information in the document. Satisfying this checkpoint will improve access to Web documents.

The *Web Content Accessibility Guidelines* also defines three conformance levels:

- **Conformance Level A:** A document or process satisfies all Priority 1 checkpoints.

- **Conformance Level Double-A:** A document or process satisfies all Priority 1 and 2 checkpoints.

- **Conformance Level Triple-A:** A document or process satisfies all Priority 1, 2, and 3 checkpoints.

An example of a Guideline that is relevant to data presentation by cancer registries is Guideline #5. This Guideline states: *"Create tables that transform gracefully."* Checkpoints for this Guideline include:

5.1 For data tables, identify row and column headers [Priority 1].
5.2 For data tables that have two or more logical levels of row or column headers, use markup to associate data cells and header cells [Priority 1].
5.3 Do not use tables for layout unless the table makes sense when linearized. Otherwise, if the table does not make sense, provide an alternative equivalent [Priority 2].
5.4 If a table is used for layout, do not use any structural markup for the purpose of visual formatting [Priority 2].
5.5 Provide summaries for tables [Priority 3].
5.6 Provide abbreviations for header labels [Priority 3].

Many states are starting to establish state accessibility guidelines. An example is New York State. In 1998 in New York, a workgroup formed (the New York State Accessibility to Information Technology Workgroup) to make recommendations to the New York State Office for Technology on improving access to state agency information via the Internet. In September 1999, New York State released *Technology Policy 99-3: Universal Accessibility for NYS Web Sites*, with the requirement that all New York State agencies' Web sites provide universal accessibility to persons with disabilities. This clarified New York State policy and provided a framework to achieve access. Agencies were directed to use W3C's *Web Content Accessibility Guidelines* in the design, creation, and maintenance of all newly created official agency Web sites. Web content should conform with level "A" conformance, satisfying all Priority 1 checkpoints. Each site should have a contact mechanism so individuals who might have trouble accessing any portion of the site can report the problem. The tasks of review, prioritization, and modification of existing content and pages was to be completed within 1 year (by September 2000). On October 4, 2000, the deadline for compliance was extended to December 31, 2000.

The New York State Cancer Registry (NYSCR) has developed a plan to comply with a state accessibility policy. NYSCR was one of the first programs in the Department of Health to post data in an accessible format. SAS programming was used to generate accessible data tables.

Section VII: Recommendations to NAACCR

The current GIS Workgroup effort is an extremely important first step towards the eventual development of best practices, but much remains to be done (e.g., research to define a scientific basis for making recommendations about best practices within the context of cancer prevention and control). Similar to the way that levels of evidence are used in the *Guide to Community Preventive Services*, these Workgroup recommendations are based more on consensus than on scientific studies. The following recommendations are made to NAACCR to help continue and expand this work:

- **Recommendation 1: Use latitude-longitude coordinates.** The Workgroup recommends that the data format of geocoded points be latitude-longitude stated in decimal degrees. This format allows maximum portability between various GIS software packages, and thereby maximizes efficiency for data sharing. Storing point locations rather than just boundary identifiers like a census tract also enables more accurate spatial statistical analyses and permits different boundaries to be generated corresponding to any desired spatial scale or point in time.

- **Recommendation 2: Use North American Datum 1983 (NAD83).** The Workgroup recommends that the accepted NAD83 be adopted as the standard. The Earth has a complex shape that is only approximately spherical. Over the past hundreds of years, scientists have generated a number of ellipsoids that approximate the Earth's shape. The ellipsoid known as WGS84 (the World Geodetic System of 1984) is now widely accepted. In North America, the virtually identical NAD83 brings North American mapping into conformity with WGS84. NAD83 is a common standard accepted by many state GIS coordinating councils and by most major GIS data producers. Again, adopting this standard will allow maximum portability for data sharing.

- **Recommendation 3: Adopt the Federal Geographic Data Committee (FGDC) Geospatial Metadata Standard.** Because of the growing importance of geospatial metadata, the Workgroup recommends the adoption of the FGDC Geospatial Metadata Standard (http://www.fgdc.gov/ metadata/contstan.html). Many important characteristics of address data, including the completeness and accuracy of reporting, are already included as fields in the Geospatial Metadata Standard.

- **Recommendation 4: Develop a written policy on confidentiality and disclosure rules.** The Workgroup recommends that cancer registries develop a written policy addressing the confidentiality, security, and disclosure of data before beginning GIS analysis and reporting. The *NAACCR Inventory on Data Security and Confidentiality* should be used as a guide. All registry staff should be trained on this policy. The policy should cover such things as data security and permissions, discussion of people or locations in ways that could be overheard by non-registry employees, replacing confidential information on the screen when the user is away from the computer for an extended period of time, immediately retrieving data and other information sent to computer printers, and so on. For more specific information, visit the NAACCR Web Site, http://www.naaccr.org/training/confidentiality.html, to review NAACCR Policy Statement 99-01 on Confidentiality, and the NAACCR report on *Data Security and Confidentiality*. It is further recommended that cancer registries develop policy guidelines on disclosure rules to help them in

their reporting of data and to assist them in making decisions on data requests from outside groups. Every state has different rules on this issue (e.g., North Carolina requires a waiver from every patient before releasing point maps in cluster investigations), but many states have similar rules.

- **Recommendation 5: Participation in Federal policy and standard setting.** The Workgroup recommends that NAACCR further adopt relevant GIS standards and address standards. A long-term goal could be to develop a common set of standards and recommendations on GIS that could be endorsed by NAACCR, NCI, and CDC (i.e., form an ongoing workgroup or subcommittee that would have representatives from the major cancer prevention and control partners). NAACCR also should consider participating in the FGDC and other relevant Federal GIS committees.

- **Recommendation 6: Organizational infrastructure.** The Workgroup recommends that it continue its work as either a NAACCR Subcommittee or by extending the life of the existing GIS Workgroup. These experts would continue to provide a point-of-contact for continuing communication relevant to GIS issues, enable NAACCR to support a GIS user listserv and GIS user group, provide newsletter articles, interface with other groups, and provide information via the Internet. NAACCR also should perform periodic surveys of registries regarding GIS capability and use.

- **Recommendation 7: Education.** Cancer registries need trained staff in the GIS area and a commitment of time and resources to continuing education. Registries need to budget for GIS-competent staff, find ways for them to keep current in a rapidly changing field, and provide sufficient commitment for them to produce maps frequently and competently. NAACCR should promote the education of staff through workshops and online and classroom-based courses. A review of the quality of online courses, particularly in the public domain, would be valuable. Additional written material for education also could be completed. This material should include tutorials and guidelines on the issues of the spatial analysis of data, calculating and applying distance measures in studying patterns of health care, and data confidentiality.

NAACCR's GIS Workgroup also encourages the continued pursuit of new knowledge in the following areas of GIS science:

- **Research on estimating expected values and their reliability on maps.** Several different methods exist for estimating expected values on maps. The Workgroup encourages research on the relative strengths and weaknesses of methods for estimating these expected values and their reliability.

- **Research on cancer cluster detection.** Given that cancer registries are often asked questions about cancer clusters, the Workgroup encourages additional research on cancer cluster detection, comparative evaluation of software products, and the preparation of step-by-step tutorials on cancer cluster detection software for GIS users in cancer registries. Registries need to respond appropriately to inquiries about cancer clusters. Improved statistical tools, updated guidelines, and a proactive attitude will not eliminate, but may significantly reduce, the cost of responding to cluster inquiries.

- **Research on adjusted cancer incidence maps.** Given that cancer registries often want to develop adjusted cancer incidence maps, the Workgroup encourages additional research on the relative strengths and weaknesses of the various methods, and whether any specific method might be preferable depending on the purpose of the map.

- **Encourage research and education in cartography.** Cartography plays an important role in disease analysis. The education of staff members on basic cartographic principles and GIS fundamentals will be increasingly important as GIS diffuses through the cancer registries. Research and educational activities related to innovations in cartography should be promoted. For example, cancer registries (especially in rural states) would welcome research on the relative strengths and weaknesses of various methods for representing the reliability of rate estimates.

- **Other research projects.** Research support from the appropriate funding sources for GIS and cancer data analysis is needed to address other research topics, including:

  - How can GIS technology and methods be made simpler so central cancer registries can easily use GIS in their annual reports, research, and publications?
  - Looking towards the future, how can models and templates be developed that help central cancer registries add value to maps (e.g., start to make maps more along the lines of the templates/models included in Linda Pickle's *NCHS Atlas of U.S. Mortality*)?
  - Research-based guidelines for posting mapped information on the Web are an important current issue. Examples of research areas include accessibility, map design, formats, and Web standards. Section 508 (accessibility) governs state and Federal Web publications, and Web sites must at least have a verbal description of the existence of a map as well as some information about the map.
  - Research and development on address parsing and other data cleaning and standardizing software and procedures.
  - Spatial-temporal analysis of data.
  - Expanded research on the evaluation of data quality and locational accuracy for address geocoding.

References

General GIS References

Antenucci JC, Brown K, Croswell PL, Kevany MJ, Archer H. Geographic Information Systems: A Guide to the Technology. New York, NY: Van Nostrand Reinhold; 1991.

Bailey T, Gatrell A. Interactive Spatial Data Analysis. New York, NY: Wiley; 1995.

Clarke KC. Getting Started With Geographic Information Systems. Upper Saddle River, NJ: Prentice Hall; 1997.

Foresman TW. The History of Geographic Information Systems: Perspectives From Pioneers. Upper Saddle River, NJ: Prentice-Hall, Inc.; 1998.

Heywood I, Cornelius S, Carver S. An Introduction to Geographical Information Systems. Upper Saddle River, NJ: Prentice Hall; 1998.

Huxhold WE, Levinsohn AG. Managing Geographic Information System Projects. New York, NY: Oxford University Press; 1995.

Longley PA, Goodchild M, Maguire DJ, Rhind DW, Lobley J. Geographic Information Systems and Science. Chichester, Sussex: John Wiley; 2001.

Martin D. Geographic Information Systems: Socioeconomic Applications. New York, NY: Routledge; 1996.

Obermeyer NJ, Pinto JK. Managing Geographic Information Systems. New York, NY: The Guilford Press; 1994.

Star J, Estes J. Geographic Information Systems: An Introduction. Englewood Cliffs, NJ: Prentice Hall; 1990.

GIS in the Health Sciences

Albert DP, Gesler WM, Levergood B. Spatial Analysis, GIS, and Remote Sensing Applications in the Health Sciences. Michigan: Ann Arbor Press; 2000.

Cromley EK, McLafferty SL. GIS and Public Health. New York, NY: Guilford Press; 2002.

Gatrell AC, Loytonen M. GIS and Health. London: Taylor & Francis, Inc.; 1998.

Address Geocoding

Broome FR, Meixler DB. The TIGER database structure. CaGIS 1990;17(1):39-47.

Bureau of the Census. TIGER/Line Files, 1992: Technical Documentation. 1993.

Johnson SD. Address matching with Commercial spatial data. Bus Geographics 1998;6(3):24-36.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91(7):1114-1116.

Marx RW. The TIGER System: automating the geographic structure of the United States census. Gov Pub Rev 1986;13:181-201.

Marx RW. The TIGER System: yesterday, today and tomorrow. CaGIS 1990;17(1):89-97.

Wilkins R. PCCF+ Version 3F. User's Guide. Ottawa: Statistics Canada; January 2001.

Spatial Epidemiology

Andrew B. Statistical Methods in Spatial Epidemiology. New York, NY: John Wiley & Sons; 2001.

Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press; 2000.

Lawson AB. Statistical Methods in Spatial Epidemiology. New York, NY: John Wiley & Sons, Ltd; 2001. (Chapter 10, pp. 207-22, is on Ecological Analysis.)

Lawson AB, Williams FLR. An Introductory Guide to Disease Mapping. New York, NY: John Wiley and Sons, Ltd.; 2001.

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia, PA: Lippincott-Raven; 1998. (Chapter 23, pp. 459-80, by Morgenstern H, is on Ecologic Studies.)

Integrating GIS Into Medical Research

Ali M, Emsch M, Ashley C, Streatfield PK. 2001. Implementation of a medical Geographic Information System: concepts and uses. J Health Popul Nutr 2001;19(2):100-110.

Boulos MN, Roudsari AV, Carson ER. 2001. Health geomatics: an enabling suite of technologies in health and healthcare. J Biomed Inform 2001;34(3):195-219. http://www.idealibrary.com.

Higgs G, Gould M. Is there a role for GIS in the 'new NHS'? Health Place 2001;7(3):247-259.

Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. Geographic Information Systems and public health: mapping the future. Public Health Rep 1999;114:359-373.

Rushton G. Methods to evaluate geographic access to health services. J Public Health Manag Pract 1999;5(2):93-100.

Rushton G, Elmes G, McMaster R. Considerations for improving Geographic Information System research in public health. URISA J 2000;12(2):31-49.

Walter SD, Birnie SE. Mapping mortality and morbidity patterns: an international comparison. Int J Epidemiol 1991;20(3):678-689.

Confidentiality

American Statistical Association Committee on Privacy and Confidentiality. http://users.erols.com/dewolf/pchome.htm.

Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Stat Med 1999;18(5):497-525 (on the Web at: http://www3.interscience.wiley.com/cgi-bin/issuetoc?ID=45002089).

General Accounting Office Report on Record Linkage and Privacy: Issues in Federal Research and Statistical Information. GAO-01-126SP; April 1, 2001 (see especially pages 91-3; on the Web at: http://www.gao.gov/special.pubs/pubshort.htm, proceed to "Record Linkage").

McLaughlin CC. Confidentiality protection in publicly released central cancer registry data. J Reg Manag (in press).

Simoes EJ, Kinman E, Chang J. Sugar Creek Cancer Inquiry Report: Level Three Investigation. Division of Chronic Disease Prevention and Health Promotion. Missouri Department of Health; April 27, 2000. http://www.health.state.mo.us/publications/sugarcreek2000.pdf.

Spatial Analysis

Anselin L. Spatial Econometrics: Methods and Models (Studies in Operational Regional Science). Norwell, MA: Kluwer Academic Publishers; 1988 (out of print).

Athas WF, Adams-Cameron M, Hunt WC, Amir-Fazli AM, Key CR. Travel distance to radiation therapy and receipt of radiotherapy following breast-conserving therapy. J Natl Cancer Inst 2000;92:269-271.

Best N, Elliott P, Richardson S. Web site for Short Course on Spatial Epidemiology held at the Imperial College, London; March 12-14, 2002. http://stats.ma.ic.ac.uk/~ngb30/.

Blot WJ, Fraumeni JF Jr. Geographic patterns of lung cancer: industrial correlations. Am J Epidemiol 1976; 103:539-550.

Blot WJ, Harrington M, Toledo A, Hoover R, Heath CW Jr, Fraumeni JF Jr. Lung cancer after employment in shipyards during World War II. N Engl J Med 1978;299:620-624.

Blot WJ, Morris LE, Stroube R, Tagnon I, Fraumeni JF Jr. Lung and laryngeal cancers in relation to shipyard employment in coastal Virginia. J Natl Cancer Inst 1980;65:571-575.

Boscoe FP, McLaughlin CC. The effect of seasonal residence on cancer incidence rates. J Reg Manag, 2002;29:3-7.

Carr DB. Designing linked micromap plots for states with many counties. Stat Med 2001;20(9-10):1331-3139.

Carr DB, Wallin JF, Carr DA. Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. Stat Med 2000;19(17-18):2521-2538.

Chou Y. Exploring Spatial Analysis in Geographic Information Systems. Santa Fe, NM: Onword Press; 1997.

Clegg LX, Midthune DN, Feuer EJ, Fay MP, Hankey BF. Cancer Incidence Rates Adjusted for Reporting Delay. In: Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg LX, Edwards BK, editors. SEER Cancer Statistics Review, 1973-97, National Cancer Institute. Bethesda, MD: NIH Pub. No. 00-2789; 2000.

Elliott P, Wakefield J, Wakefield J. Disease clusters: should they be investigated, and, if so, when and how? J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):3-12.

Fisher MM, Getis A. Recent Developments in Spatial Analysis: Spatial Statistics, Behavioral Modeling, and Computational Intelligence (Advances in Spatial Science). New York, NY: Springer-Verlag; 1997.

Fotheringham S, Rogerson P, editors. Spatial Analysis and GIS. Philadelphia, PA: Taylor & Francis, Inc.; 1998.

Gelman A, Park DK, Ansolabehere S, Price PN. Models, assumptions and model checking in ecological regressions. J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):101-118.

Guthrie KA, Sheppard L. Overcoming biases and misconceptions in ecological studies. J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):141-154.

Hoover R, Mason TJ, McKay FW, Fraumeni JF Jr. Cancer by county: new resource for etiologic clues. Science 1975;189(4207):1005-1007.

King G. A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior From Aggregate Data. Princeton, NJ: Princeton University Press; 1997.

Kinman EJ, Chang J. Sugar Creek Cancer Inquiry Report: Level Three Investigation. Missouri Department of Health; 2000.

Kulldorff M. Geographic Information Systems (GIS) and community health: some statistical issues. J Public Health Manag Pract 1999;5(2):100-106.

Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):61-72.

Kulldorf M. Tests for spatial randomness adjusted for inhomogeneity: A general framework. Unpublished manuscript; 2002.

Kwan MP. Interactive geovisualization of activity-travel patterns using three-dimensional Geographical Information Systems: A methodological exploration with a large data set. Transportation Research C 2000;8:185-203.

Langran B. Time in Geographic Information Systems. London: Taylor & Francis; 1988.

Lawson AB, Williams LR. An Introductory Guide to Disease Mapping. New York, NY: John Wiley & Sons, Inc.; 2001.

Lee AM, Fraumeni JF Jr. Arsenic and respiratory cancer in man: an occupational study. J Natl Cancer Inst 1969;42(6):1045-1052.

Lee J, Wong DWS. Statistical Analysis With ArcView GIS. New York, NY: John Wiley & Sons, Ltd.; 2001.

Lewandowsky S, Herrmann DJ, Behrens JT, Li S-C, Pickle LW, Jobe JB. Perception of clusters in statistical maps. Appl Cog Psych 1993;7:533-551.

Longley P, Batty M, editors. Spatial Analysis: Modeling in a GIS Environment. New York, NY: John Wiley & Sons, Inc.; 1996.

Longley PA, Goodchild M, Maguire DJ, Rhind DW. Geographic Information Systems. New York, NY: John Wiley & Sons, Ltd.; 2001.

MacEachren AM, Brewer CA, Pickle LW. Visualizing georeferenced data: representing reliability of health statistics. Environ Planning A 1998;30:1547-1561.

Mark DM, Egenhofer M. Geospatial Lifelines. In: Guentehr O, Sellis T, Theodoulidis B (Eds). Integrating Spatial and Temporal Databases. Dagstuhl Seminar Report No. 228; 1998.

Mark DM, Egenhofer MJ, Bian L, Hornsby KE, Rogerson PA, Vena J. Spatial-Temporal GIS Analysis for Environmental Health: Solutions Using Geospatial Lifelines. GEOMED 99; 1999.

Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni JF Jr. Atlas of Cancer Mortality for U.S. Counties: 1950-1969. Washington, DC: U.S. Government Printing Office (DHEW pub. no. [NIH] 75-780); 1975.

Mitchell A. The ESRI Guide to GIS Analysis: Volume 1: Geographic Patterns and Relationships. Redlands, CA: ESRI Press; 1999.

Morgenstern H. Ecologic Studies. Second Edition. In: Rothman KJ, Greenland S, editors. Modern epidemiology. Second Edition. Philadelphia, PA: Lippencott-Raven; 1998, pp. 459-480.

Openshaw S. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography. No. 38. Norwich: Geo Books; 1983.

Peuquet DJ. Representations of Space and Time. New York, NY: Guilford; 2002.

Pickle LW. Mapping Mortality Data in the U.S. In: Elliott P, et al., editors. Spatial Epidemiology. London: Oxford Press; 2000, pp. 240-252.

Pickle LW, Mungiole M, Jones GK, White AA. Atlas of United States Mortality. Hyattsville, MD: National Center for Health Statistics. DHHS Publication No. (PHS) 97-1015; 1996.

Pickle LW, White AA. Effects of the choice of age-adjustment method on maps of death rates. Stat Med 1995;14:615-627.

Rogerson PA. Monitoring point patterns for the development of space-time clusters. J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):87-96.

Rushton, G. Public health, GIS and spatial analytic tools. Annu Rev Public Health 2003 (in press). Rushton G. CD-ROM and Web Site: Improving Public Health Through Geographical Information Systems: An Instructional Guide to Major Concepts and Their Implementation. http:www.uiowa.edu/~geog/health.

Rushton G, West M. Women with localized breast cancer selecting mastectomy treatment, Iowa, 1991-1996. Public Health Rep 1999;114:370-371.

Semenciw RM, Le ND, Marrett LD, Robson DL, Turner D, Walter SD. Methodological issues in the development of the Canadian Cancer Incidence Atlas. Stat Med 2000;19(17-18):2437-2449.

Sheehan TJ, Gershman ST, MacDougall LA, Danley RA, Mroszczyk M, Sorensen AM, Kulldorff M. Geographic assessment of breast cancer screening by towns, ZIP codes, and census Tracts. J Public Health Manage Pract 2000;6(6):48-57.

Steward J, John G. An ecological investigation of the incidence of cancer in Welsh children for the period 1985-1994 in relation to residence near the coastline. J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):29-43.

Tagnon I, Blot WF, Stroube RB, Day NE, Morris LE, Peace BB, Fraumeni JF Jr. Mesothelioma associated with the shipbuilding industry in coastal Virginia. Cancer Res 1980;40:3875-3879.

Wakefield J, Quinn M, Raab G. Disease clusters and ecological studies (Editorial). J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):1-2.

Wartenberg D. Investigating disease clusters: why, when and how? J Royal Stat Society: Series A (Statistics in Society) 2001;164(1):13-32.

Winn DM, Blot WJ, Shy CM, Pickle LW, Toledo A, Fraumeni JF Jr. Smokeless tobacco and oral cancer among women in the Southern United States. N Engl J Med 1981;304:745-749.

Yang, Z. Modeling and Reasoning with Geospatial Lifelines in Geographic Information Systems. Dissertation. University of New York at Buffalo; 2001.

Data Quality

Minnesota Land Management Information Center Positional Accuracy Handbook (hardcopy, 33 pp from State of Minnesota or online as pdf file at: http://www.mnplan.state.mn.us/press/accurate.html

Guptill S, Morrison JL, editors. Elements of Spatial Data Quality. 1995 (Published on behalf of the International Cartographic Association by Elsevier Science).

Cartography

Brewer CA, MacEachren AM, Pickle LW, Herrmann D. Mapping mortality: evaluating color schemes for choropleth maps. Ann Assoc Am Geographers 1997;87(3):411-438.

Brewer CA, Pickle LW. Comparison of methods for classifying epidemiological data on choropleth maps in series. Ann Assoc Am Geographers. (in press).

Brewer CA, Pickle LW. Evaluation of methods for classifying epidemiological data on choropleth maps in series. Ann Assoc Am Geographers (in press).

Buttenfield BP, McMaster RB, editors. Map Generalization: Making Rules for Knowledge Representation. London: Longman; 1991, pp. 172-186.

Carswell CM, Kinslow HS, Pickle LW, Herrmann D. Using Color To Represent Magnitude in Statistical Maps: The Case for Double-Ended Scales. In: Pickle LW, Herrmann D, editors. Cognitive Aspects of Statistical Mapping. Working Paper Series No. 18. Hyattsville, MD: National Center for Health Statistics; 1995.

DiBiase D, MacEachren AM, Krygier J, Reeves C. Animation and the role of map design in scientific visualization. CaGIS 1992;19(4):201-214.

Eicher CL, Brewer CA. Dasymetric mapping and areal interpolation: implementation and evaluation. CaGIS 2001;28(2):125-138.

Kraak M-J, Brown A, editors. Web Cartography: Developments and Prospects. London: Taylor and Francis; 2001.

MacEachren AM. How Maps Work: Representation, Visualization, and Design. New York, NY: The Guilford Press; 1995.

MacEachren AM. Some Truth With Maps. Washington, DC: Association of American Geographers; 1994, 129 pp.

MacEachren AM, Brewer CA, Pickle LW. Visualizing georeferenced data: representing reliability of health statistics. Environ Planning A 1998;30:1547-1561.

Monmonier M. Cartographies of Danger. Chicago, IL: The University of Chicago Press; 1997.

Monmonier M. How To Lie With Maps. Chicago, IL: The University of Chicago Press; 1996.

Monmonier M. Mapping It Out: Expository Cartography for the Humanities and Social Sciences. Chicago, IL: The University of Chicago Press; 1993.

National Center for Health Statistics. Working Paper Series #18 1995, Cognitive Aspects of Statistical Mapping.

Pickle LW, Herrmann DJ, editors. Cognitive Aspects of Statistical Mapping. National Center for Health Statistics Working Paper Series Report, No. 18. Hyattsville, MD: Centers for Disease Control and Prevention/National Center for Health Statistics; 1995.

Pickle LW, Herrmann DJ. Cognitive Research for the Design of Statistical Rate Maps. In: Sirken MG, editor. National Center for Health Statistics Working Paper Series No. 28, Survey Research at the Intersection of Statistics and Cognitive Psychology, Centers for Disease Control and Prevention/National Center for Health Statistics; 2000.

Robinson AH, Morrison JL, Muehrcke PC, Kimerling AJ, Guptill SC. Elements of Cartography. New York, NY: John Wiley & Sons, Inc; 1995.

Appendix: Resources

Geocoding Resources

Data Sources—Street Files

Street files are available from the Federal Government, commercial vendors, and state and local GIS sources.

National data sources include:

- **TIGER.** Free files are available from the census (http://www.census.gov/geo/www/ tiger/ index.html).

- **Geography Network.** Versions of the TIGER files in ESRI data formats also are available on the Geography Network (http://www.geographynetwork.com).

- **USPS (master address file).**

Also useful is the United States Postal Service ZIP+4 State Directory product. Formerly, this product was available by individual state in hardcopy format; however, it is now only available for the entire United States and is available on CD-ROM.

Commercial vendors include:

- **Geographic Data Technology (GDT).** GDT sells street network products (e.g., Dynamap), postal databases, and boundary files. See GDT's Web page at http://www.geographic.com.

- **Tele Atlas USA (formerly Etak).** Tele Atlas USA sells street map products (e.g., MultiNet™ USA street data). See their Web page at http://www.etak.com. Spatial Data Demonstrations are shown at http://www.na.teleatlas.com/products/pdemos.html. The Route Planning Demonstration includes the ability to identify cross streets.

- **NavTech.** This company sells the NAVTECH Map Database. See their Web page at http://www.navtech.com.

- **Semaphore Corporation Delivery Point Confirmation.** See http://www.semaphorecorp.com/ cgi/dpv.html. The U.S. Postal Service Delivery Point Validation (DPV) Database is essentially a "yes/no" table for checking the validity of any individual mailing address (more than 145 million locations). The DPV Database is distributed on an optional $79 CD-ROM and can be installed to enhance ZP4's ability to validate mailing addresses. Without DPV, ZP4 is only capable of confirming whether addresses fall within the "low-to-high"' address ranges specified in ZIP+4 records. For example, "100 to 198 (even) MAPLE AVE" is a typical house number range listed in the ZIP+4 database, and any even house number in that range will validate and become CASS-certified, whether or not the address actually exists. With DPV installed, ZP4 can confirm

whether any individual address really exists, regardless of what the ZIP+4 database range implies. For example, only three addresses might actually exist in a given range like "100 to 198 (even)." Using DPV, ZP4 will validate the three correct addresses and reject any other address in the range. In other words, with just ZP4 installed, the proper ZIP+4 can be determined for any address. With ZP4 and DPV installed, it can also be determined whether mail is actually delivered to the address. In addition to allowing ZP4 to return a "yes/no" indication for any specific mailing address, DPV also allows ZP4 to detect whether an address is a commercial mail receiving agency (such as a Mail Boxes, Etc. store).

Local street files (state, county, local):

- For local data sources, check state GIS Clearinghouses (http://www.fgdc.gov) and local GIS users groups. These organizations also may be helpful:
  - E911
  - State Department of Transportation
  - Utility companies (e.g., electric, gas, cable)
  - Local government GIS units (parcel data, street files).

Geocoding Software

- **GDT, Inc.** The Matchmaker® SDK Professional is a set of development tools using the GDT Address Coding Guide (ACG). The ACG is accessed through an application programming interface that allows users to build applications to geocode to the street level. The geocoding process attaches census as well as coordinate information to address files allowing users to derive valuable information from address files.

- **MapInfo.** MapMarker Plus is available for both the United States and Canada, and runs in interactive or batch mode. The software costs approximately $1,700. There also is a less expensive version, called MapMarker, which is available for $1,300. The difference is that MapMarker's address dictionary is based on TIGER 1998, but MapMarker Plus incorporates enhanced data from GDT, Inc., and thus is more precise in street-level geocoding. MapMarker data are updated every 6 months and MapMarker Plus data are updated every 3 months. See http://www.MapInfo.com.

- **Sagent Technology** (see their Web page at http://www.sagent.com). Tools include:
  - AddressBroker, a client-server solution for real-time and batch address standardization, geocoding, spatial analysis, and data enhancement
  - GeoStan, an embeddable address quality and geocoding application
  - Spatial+, a spatial analyses tool
  - Merge/Purge, which performs householding and finds duplicate records, and matches records across multiple databases using "fuzzy" matching.
  - The Centrus Desktop, a Windows-based desktop application that performs address cleansing and geocoding.

Geocoding Services

- **Tele Atlas.** Tele Atlas offers two geocoding services: EZ-Locate™ (an Internet service), and a Geocoding Service Bureau. See their Web page at http://www.etak.com.

- **GDT.** GDT provides geocoding services. See their Web page at http://www.geographic.com.

Spatial Analysis Software

Once addresses have been geocoded into valid geographic points they can be mapped or, for some applications, aggregated into small areas (e.g., census tracts) for visual inspection and pattern determination. Spatial pattern analysis is useful for generating ideas or hypotheses about disease events and related factors. However, simple visual inspection of points representing tumor events, sites, and stages may not be adequate for determining the existence of a true spatial pattern. Several statistical and mathematical techniques have evolved to assist the investigator in discerning whether an observed distribution of points or small areas are random or have some structure that may be related to a disease process. These techniques are found within many software packages that are free and readily available on the Internet.

Many of these software packages are accompanied by extensive documentation on not only the operation of the software and how to prepare the data, but also they describe the statistical methodology and the appropriate uses and interpretation of the tests employed. Bibliographies are also provided. Most of the spatial analytical techniques found in the software listed in this section are commonly used in spatial epidemiological studies of disease. The software applications are available in a wide variety of operating system formats. This section includes only software packages that are typically used by cancer registries, the entire spatial analysis software market is not represented here.

- **Cluster 3.1 (1993).** This program is currently DOS based, although there has been discussion about a pending Windows version. Cluster 3.1 is available as a free download from http://www.atsdr.cdc.gov/HS/cluster.html. This software contains a number of cluster evaluation routines. The data input is basic (text format), and some pre-processing by the user is necessary before employing the selected test.

- **CrimeStat 1.1 (1999).** CrimeStat 1.1 is a Windows-based software package that was developed for studying spatial patterns of crime incidents. Crime incidents are analyzed as points in much the same manner as public health events or the addresses of particular tumor sites. This software also interfaces graphically with many of the geographic information systems in use today. It is available for free download at http://www.icpsr.umich/NACJD/crimestat.html. Data input is either in .dbf or text format.

- **Point Pattern Analysis (PPA) (2001).** PPA is available in three operating system formats: (1) DOS, (2) UNIX, and (3) as a Web interface. The Web site for online use is located at http://xerxes.sph.umich.edu:2000. Like Cluster 3.1, the input data set needs to be in text format. The Web interface is actually one of several modules of epidemiological interest at this site. There are plans to make the DOS version available from an FTP site in the near future. See

http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cg for more information.

- **SaTScan (NCI freeware).** SaTScan (http://dcp.nci.nih.gov/bb/satscan.html) analyzes spatial, temporal and space-time point data using the spatial, temporal, or space-time scan statistic. The software was written by Martin Kulldorff, Katherine Rand, Greg Gherman, Gray Williams, and David DeFrancesco, and resides on the NCI, Division of Cancer Prevention, Biometry Branch Web Site. Specifically, it identifies clusters of similar rate places (high or low) that are unlikely due to chance alone. The program is being extended to detect elliptical clusters; it now only detects circular ones. For a short description of SaTScan, see http://www.cancer.gov/prevention/bb/software.html.

- **SpaceStat (Anselin).** This commercial software package is oriented to spatial econometrics. Information is available at http://www.spacestat.com/.

- **DMAP (Disease Mapping and Analysis Program).** For references on this program, see: Rushton G. Improving Public Health Through Geographical Information Systems: An Instructional Guide to Major Concepts and Their Implementation, CD-ROM Version 2.0. Iowa City, Iowa: Department of Geography, University of Iowa; December 1997 (on the Web at: http://www.uiowa.edu/~geog/health) under "GIS Lab, Demo: Spatial Analysis of Data," which states: "*DMAP is a Windows-compatible program that produces disease rates using variable spatial filters and tests for their statistical significance using Monte Carlo simulations. The program computes values that are inputs to GIS software that will produce disease rate maps and maps of statistical significance. Input data are either individual disease records and individual at risk records or are aggregates of the above. DMAP produces disease rate maps from two types of data: (1) Address-matched records of individuals with the disease (known as 'numerator' events) and individuals at risk for having the disease (known as 'denominator' events). In this case weights equal one; (2) Address-matched data for spatial aggregates of people with the disease ('numerator' events) and people at risk of having diseases ('denominator' events). An example of spatial aggregation might be census blocks. When spatial aggregates are used, the weights are greater than one.*"

- **EpiAnalyst.** The EpiAnalyst extension for ArcView® GIS is a productivity tool and resource kit for spatial-epidemiologic research. The extension contains spatial cluster analysis software from NCI, has a link to the CrimeStat spatial analysis software, and includes the Spatial Data Modeler extension. The EpiAnalyst also includes software from the University of Iowa that allows users to perform the computations that are required to make smoothed maps of disease rates and tests of statistical significance. The extension interfaces with the latest version of Epi Info 2000 from CDC.

- **Biomedware ClusterSeer.** Biomedware's ClusterSeer can evaluate whether disease clusters occur in the vicinity of a risk factor (such as a chemical plant or injection well), or it can locate clusters without a known focus or cause. It also offers retrospective surveillance methods to identify possible clusters that have not yet been noticed. Available methods include those proposed by Besag and Newell, Rogerson, Kulldorff, Bithell, Diggle, Levin and Kline, Moran, and Ripley.

- **Geostatistical Analyst (ESRI).** This is an extension to the ArcGIS 8.1 family of software. Functionality includes a variety of exploratory data tools (histogram, Normal QQ Plot, trend analysis, Voronoi map, and semivariogram/covariance cloud. Various surface interpolations are included: Inverse Distance Weighting, local polynomial interpolation, radial basis functions, kriging, and cokriging. Information is available on the ESRI Web Site at http://www.esri.com

- **S-Plus (spatial analysis module).** See http://www.insightful.com/products/spatial/default.html. This is a spatial statistics module for S+; see "*Feature List*" for functions. In general, this includes geostatistical analysis, point pattern, and lattice data analysis.

- **SAS/GIS.** SAS/GIS software provides an interactive GIS within the SAS System. See their Web Site at http://www.sas.com/products/gis/.

- **Orius Cancer Data Analysis and Mapping (SAS-based) Software.** The basic Orius calculates rates and related statistics once the counts are grouped into areas. It has only been linked directly to GIS software at the NAACR Web Site (CINA online). It is good at working with multiple lists, but requires a manual workaround to stratify by period in the analysis. The standard choice is annual rates or one period. References for this software are available at http://www.hc-sc.gc.ca/ hpb/lcdc/publicat/cdic/cdic213/cd213f_e.html.

- **TerraSeer Environmental Insight Software.**

GIS Software Resources and Hardware

Included in this section are only software packages typically used by cancer registries. The entire spatial analysis software market is not represented here.

GIS Viewers

- **ArcExplorer 2.0 and 3.1.** These are two free viewers from ESRI (see http://www.esri.com., then go to "Free Resources"). The 3.1 Java-based viewers include basic GIS functionality and allow users to connect to map services on the Geography Network.

- **MapInfo ProViewer.**

- **Epimap.**

GIS Software

- **ESRI.** http://www.esri.com.
  - ArcView 3.2 and extensions
  - ArcGIS 8.1 and extensions (Spatial Analyst, Geostatistical Analyst, U.S. Streets).

- **MapInfo.** http://www.mapinfo.com.

- **Maptitude.** Caliper Corporation, http://www.caliper.com.

- **Idrisi.** Clark University, http://www.clarklabs.org/.

Web Interactive Map Server Software

- **ArcIMS.** ESRI (http://www.esri.com).

- **WebMap.** Intergraph (http://www.ingr.com).

- **MapXtreme.** MapInfo (http://www.mapinfo.com).

- **MapGuide.** Autodesk (http://www.autodesk.com).

Recommended Hardware

It is recommended that investigators contact software vendors and double their recommendations for hardware (triple ESRI recommendations). Check for package hardware deals that software vendors advertise.

GIS Data Resources

- **National Spatial Data Infrastructure (NSDI) Clearinghouses.** Information about free geographical data sources and other GIS data are available through searching the metadata on the NSDI Clearinghouses (see http://www.fgdc.gov). Most state GIS Clearinghouses also are accessible from the NSDI site.

- **Landview.**

- **National Atlas Project, U.S. Geological Survey (USGS).** The National Atlas Project is a collaboration of many Federal agencies, led by USGS, to display and disseminate geospatial data via the Internet. The National Atlas of the United States® is intended to provide a comprehensive, map-like view into the enormous wealth of data collected by the Federal Government. The new National Atlas Web Site includes downloadable data and maps on mortality rates, weather, transportation infrastructure, soils, county boundaries, volcanoes, watersheds, crime patterns, and population distribution. It provides easy-to-use tools to display, manipulate, and query geospatial data so that customers can produce their own relevant information, includes links to current and real-time events and to other Federal producers of geospatial information, and furnishes a national framework of well-maintained and documented base cartographic data for use by others. Additional information can be found at http://www.nationalatlas.gov.

- **Geography Network.** The Geography Network provides a metadata search of many data and service products. Map services from the Geography Network can be accessed directly via ArcExplorer (a free GIS viewer from ESRI) and from any of the ArcGIS 8.1 software products.

- **Orthos** and other remote-sensed data available on:
  - Federal and State NSDI clearinghouse sites
  - http://gisdatadepot.com/
  - http://www.terraserver.com/home.asp.

Professional Conferences and Organizations

- GIS and Health Conference (organized by Ric Skinner, has been held several times, but is not necessarily ongoing).
- ESRI Health User Conference (the first Conference was held in the fall of 2001).
- GIS/Spatial Statistics tracks at the American Statistical Association Annual Meeting.
- NAACCR Education Committee.
- List servs as a source of help/information.
- Health users groups.
- State GIS committees/councils.
- GIS and health conferences from the International Health Geographic/ESRI Health Users Group/American Statistical Association track in Spatial Statistics.
- ESRI Users Conference.
- Urban and Regional Information Systems Association, American Congress on Surveying and Mapping, American Society for Photogrammetry and Remote Sensing, Association of American Geographers, and Geospatial Information and Technology Association.

Journals and Newsletters

- *Statistics and Medicine* has published articles and has had special issues on spatial topics.
- *Journal of Public Health Management and Practice* has spatial issues and scattered articles.
- *International Journal of Health Geographics* (a soon-to-be-launched online peer-reviewed journal that will enable publishing of color maps).
- *Health and Place* (relatively new and small).
- *American Journal of Epidemiology*.
- *Journal of the National Cancer Institute*.

- *ESRI Healthy GIS Newsletter*, produced 2-3 times per year.

- *Public Health GIS News and Information*, a newsletter produced by Chuck Croner (cmc2@cdc.gov).

Web Sites

The Web Site for NAACCR's GIS Workgroup is http://www.schs.state.nc.us/NAACCR-GIS/. This site has links to NAACCR's data standards document by David O'Brien and Ric Skinner and also to Kevin Liske's PowerPoint presentation on cartography. Other Web sites of interest include:

- **http://www.uiowa.edu/~geog/health/.** Excellent online course material by Professor Gerry Rushton, University of Iowa.

- **http://www.nci.nih.gov/atlasplus.** The new and improved NCI Cancer Mortality Atlas Web Site.

- **http://facfinder.census.gov/servlet/BasicFactsServlet.** Users can enter a street address and obtain the enclosing state, county, tract, block group, and block. This is convenient for investigators who only have a few addresses to check.

- **http://www.census.gov/geo/www.garm.html.** This site was last updated in May 2001. The *Geographic Areas Reference Manual* defines and discusses census geographic units, including a historical perspective on census evolution via decades of Decennial Censuses.

- **http://usps.gov/ncsc/ziplookup/lookupmenu.htm.** This site is useful for investigators who have a small batch of addresses to verify. All of the USPS standard abbreviations for street types are included.

- **http://cythera.ic.gc.ca/dsol/cancer/help_e.html.** Health Canada's cancer surveillance data online. Incidence and mortality maps are customizable by the user. The site also allows users to generate their own animations (time trends) on the Web.

- **http://www.personal.psu.edu/faculty/c/a/cab38/.** Interesting information on research in cartography is available on Cynthia Brewer's Web Site.

- **http://cancer.gov/atlasplus/.** The newly enhanced Cancer Mortality Maps and Graphs Web Site.

- **http://www.cdc.gov/nchs/gis.htm.** The NCHS GIS home page.

- **http://www.cdc.gov/nchs/products/pubs/pubd/other/atlas/atlas.htm.** The NCHS mortality atlas.

- **http://www.nationalatlas.gov.**

- **http://www.hhs.gov/ocr/hipaa/.** The U.S. Department of Health and Human Services confidentiality rules Web Site.

- **http://www.hhs.gov/ocr/hipaa/genoverview.html.** Confidentiality rules overview.

- **http://seer.cancer.gov/Publications/Data1973_1998/.** Instructions for printing a public use data agreement (confidentiality) for NCI SEER data.

- **http://www.geovista.psu.edu/research/healthvisualization/index.html.** Health data visualization research at Penn State, funded by NCHS, NCI, and NSF.

- **http://www.geovista.psu.edu/grants/dg-qg/intro.html.** Specific information on the NSF grant for digital government quality graphics.

- **http://www.insightful.com/products/spatial/default.html.** Spatial statistics module for S+.

- **http://www.cast.org/bobby.** Bobby is a Web site validation tool, developed by the Center for Applied Special Technology, that helps Web page authors identify and repair significant barriers to access by individuals with disabilities.

- **http://www.w3c.org.** World Wide Web Consortium.

- **http://www.w3c.org/.** WAIWeb Access Initiative.

- **http://www.w3.org/TR/WCAG10.** Web content accessibility guidelines.

- **http://www.healthgis-li.com/.** Long Island Breast Cancer Study Project GIS site; the only Congressionally mandated GIS.

- **http://www.geocomputation.org/2000/GC045/Gc045.htm.** GeoComputation 2000. Proceedings of the 5th International Conference on GeoComputation. University of Greenwich, United Kingdom; August 23-25, 2000.

- **http://www.geog.ubc.ca/courses/klink/g370_472.html.** Brian Klinkenberg, University of British Columbia, Department of Geography, GIS and Cartography On-Line Resources. Unit 41-Spatial Interpolation II (compiled with Nigel M. Waters, University of Calgary). Also see **http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u41.html**, **http://www-personal.umich.edu/~cgc/interpolation_3.htm**, and for course home **http://www-personal.umich.edu/~cgc/up_507_1.htm.** Chae Gun Chung (Ph.D. Student; Taubman College of Architecture and Urban Planning; Urban, Technological, and Environmental Planning Program, University of Michigan, Ann Arbor. UP507 Information Systems Course Project: Project II (Areal Interpolation).